Physics 509: Intro to Bayesian Analysis

Scott Oser
Lecture #5

Thomas Bayes

Outline

Previously: we now have introduced all of the basic probability distributions and some useful computing tools for dealing with them.

Today: we return to Bayes theorem and see how Bayesian analysis is used to test hypotheses and especially to “fit” for model parameters.

Bayes' Theorem

$$P(H|D,I)=\frac{P(H|I)P(D|H,I)}{P(D|I)}$$

This just follows from laws of conditional probability---even frequentists agree, but they give it a different interpretation.

H = a hypothesis (e.g. “this material is a superconductor”)
I = prior knowledge or data about H
D = the data

P(H|I) = the “prior probability” for H

P(D|H,I) = the probability of measuring D, given H and I. Also called the “likelihood”

P(D|I) = a normalizing constant: the probability that D would have happened anyway, whether or not H is true.

Note: you can only calculate P(D|I) if you have a “hypothesis space” you're comparing to. A hypothesis is only “true” relative to some set of alternatives.

Frequentist vs. Bayesian Comparison

Bayesian Approach

- "The probability of the particle's mass being between 1020 and 1040 MeV is 98%."
- Considers the data to be known and fixed, and calculates probabilities of hypotheses or parameters.
- Requires *a priori* estimation of the model's likelihood, naturally incorporating prior knowledge.
- Well-defined, automated "recipe" for handling almost all problems.
- Requires a model of the data.

Frequentist Approach

- "If the true value of the particle's mass is 1030 MeV, then if we repeated the experiment 100 times only twice would we get a measurement smaller than 1020 or bigger than 1040."
- Considers the model parameters to be fixed (but unknown), and calculates the probability of the data given those parameters.
- Uses "random variables" to model the outcome of unobserved data.
- Many "ad hoc" approaches required depending on question being asked. Not all consistent!
- Requires a model of the data.

An example with parameter estimation: coin flip

Someone hands you a coin and asks you to estimate the *p* value for the coin (probability of getting heads on any given flip).

You flip the coin 20 times and get 15 heads.

What do you conclude?

Bayesian coin flipping

Someone hands you a coin and asks you to estimate the *p* value for the coin (probability of getting heads on any given flip).

You flip the coin 20 times and get 15 heads.

What do you conclude?

$$P(H|D,I)=\frac{P(H|I)P(D|H,I)}{P(D|I)}$$

Here H is the hypothesis that p has some particular value. To proceed we must evaluate each term.

Evaluating the terms in the Bayesian coin flip

$$P(H|D,I)=\frac{P(H|I)P(D|H,I)}{P(D|I)}$$

First, some notation. Let me use *p* in place of H.

Prior: let's assume a uniform prior for *p*. So P(H|I) = P(*p*) = 1.

Likelihood factor: P(D|*p*). This is the probability of observing our data, given H. We model this as a binomial distribution:

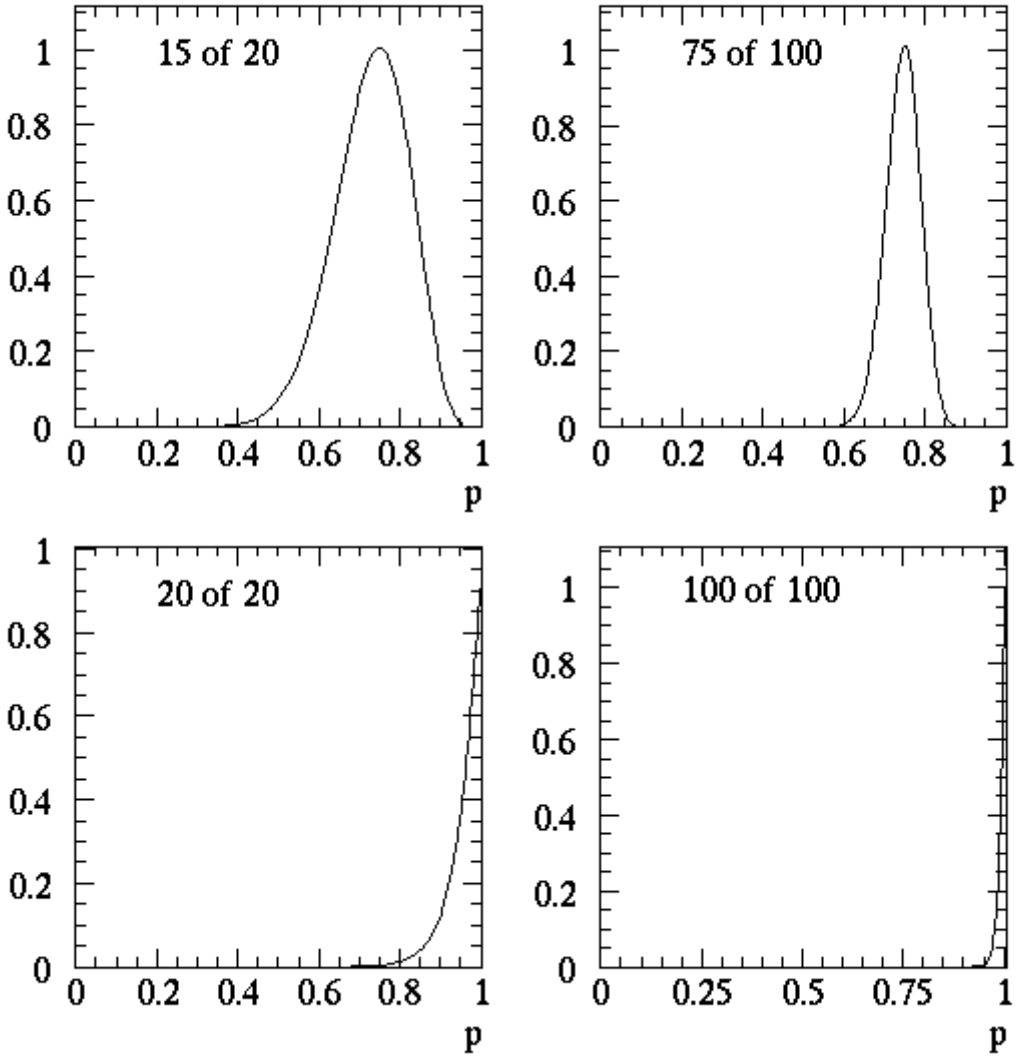
$$P(D|p)=\frac{N!}{m!(N-m)!}p^m(1-p)^{N-m}$$

Finally P(D|I). This is the probability of observing the data, summed over all hypotheses (here, all possible values of *p*).

$$P(D|I)=\int_0^1 dp\,P(p)\frac{N!}{m!(N-m)!}p^m(1-p)^{N-m}$$

Solution for P(*p*|D,I): uniform prior

$$P(p|D) \propto P(p)\,P(D|p) = \frac{N!}{m!(N-m)!}\,p^{m}(1-p)^{N-m}$$



Bayesian coin flip: alternate prior

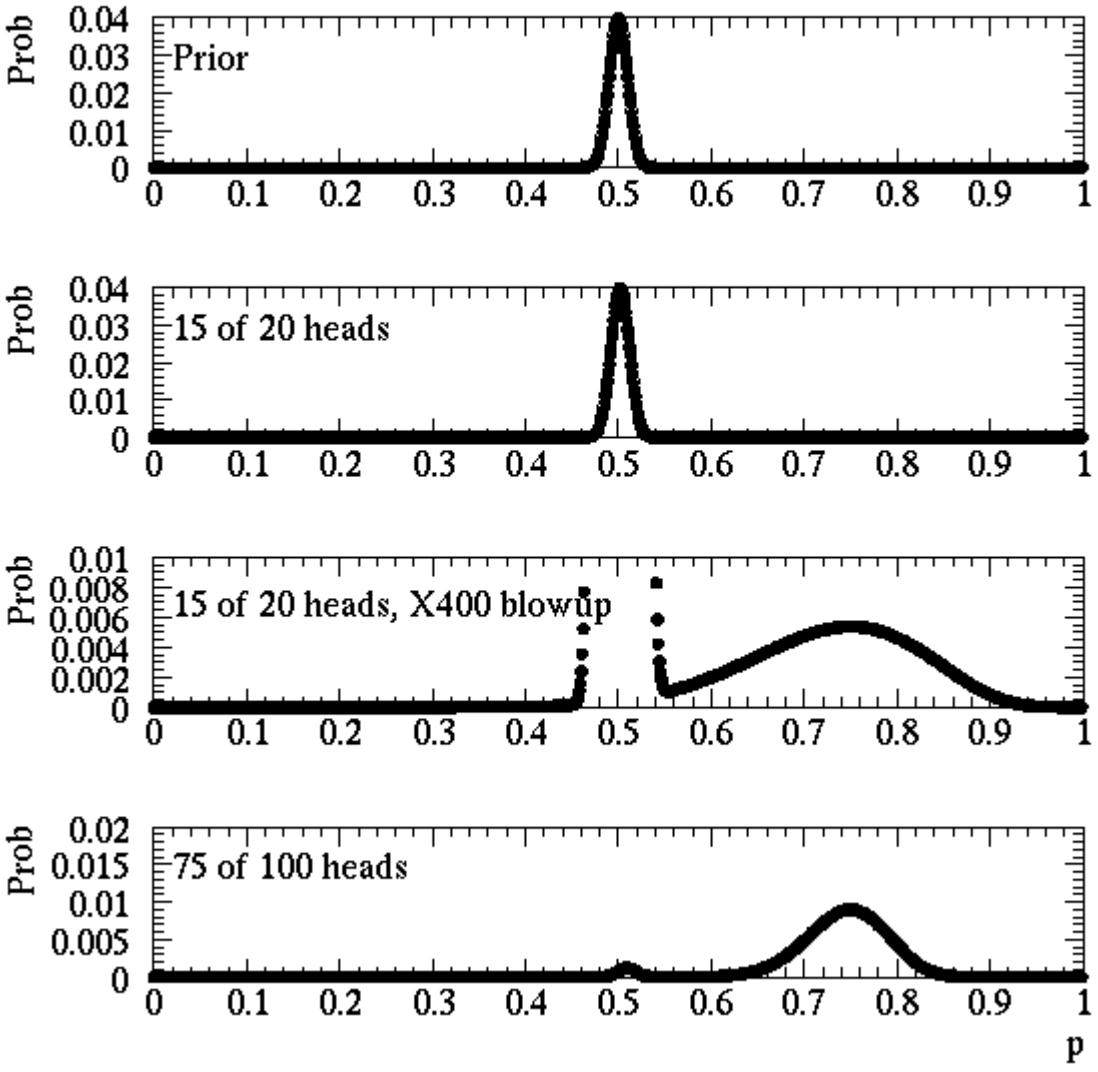
If a friend hands you a coin in the lunchroom, is it really reasonable to assume a uniform prior for *p*? Unbalanced coins must be really rare!

Consider a more plausible prior:

1) You're 99.9% sure this is a normal coin. A normal coin has *p=0.5*. But even normal coins might be a little off-kilter, so model its distribution as a Gaussian with mean 0.5 and width σ=0.01.
2) There's a 0.1% chance this is a trick coin. If so, you have no idea what its true *p* value would be, so use a uniform distribution.

$$P(p) = 0.999 \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(p-0.5)^2}{2\sigma^2}} + 0.001 \times 1$$

Solution for P(*p*|D,I): more realistic prior



Prob in peak at 0.5 = 0.999

Prob in peak at 0.5 = 0.997

Prob in peak at 0.5 = 0.030

Dependence on choice of prior

Clearly you get a different answer depending on which prior you choose! This is a big point of controversy for critics.

A Bayesian's reply: “Tough.”

In Bayesian analysis, dependence on choice of priors is a feature, not a bug. The prior is a quantitative means of incorporating external information about the quantities being measured. If the answer depends strongly on the choice of prior, this just means that the data is not very constraining.

In contrast, classical frequentist analysis doesn't require you to spell out assumptions so clearly---what are you implicitly assuming or ignoring?

Good habits for Bayesian analysis:

- be explicit about your choice of prior, and justify it
- try out different priors, and show how result changes

Contrast with frequentist approach

A frequentist would use the data to directly estimate *p* from the data, without invoking prior. Best estimate is *p*=15/20=0.75.

Frequentist would probably try to assign an "error bar" to this value. Perhaps noting that variance of binomial is *Np(1-p)*, we could calculate Var=20(0.75)(0.25)=3.75, or σ=sqrt(Var)=1.94. So the error on *p* might be 1.94/20 = 0.097, so *p*=0.75 ± 0.10. (What would a frequentist do if he observed 20/20 heads?)

But interpretation is very different. Frequentist would not speak of the probability of various *p* values being true. Instead we talk about whether the data is more likely or less likely given any specific *p* value. Very roundabout way of speaking!

Note that the *p* value estimation did not:

- yield a probability distribution for *p*
- did not incorporate any background information (eg. the fact that almost any coin you regularly encounter will be a fair coin)

Advantages of a Bayesian approach

If you start with some probability distribution for the value of a parameter, or an estimate of the likelihood of a hypothesis, and then you learn some new piece of information (“the data”), Bayes' theorem immediately tells you how to update your distribution.

The strongest benefit of Bayesian statistics is that it directly answers the question you're really asking: how likely is your hypothesis? For example, you can calculate probabilities for things like: what is the probability that there's an undiscovered planet in the outer solar system?

You can ONLY directly calculate the odds of a hypothesis being true if you assume some prior, and if your interpretation of probability allows you to think of probability as a measure of credibility (rather than just frequency).

Practical advantages of a Bayesian approach

Using Bayes theorem has a number of practical advantages:

1) It's conceptually simple. Every problem amounts to:
   A. list all of the possible hypotheses
   B. assign a prior to each hypothesis based upon what you already know
   C. calculate the likelihood of observing the data for each hypothesis, and then use Bayes' theorem
2) It gives an actual probability estimate for each hypothesis
3) It makes it easy to combine different measurements and to include background information
4) It's guaranteed to be self-consistent and in accord with "common sense"
5) It makes handling systematic errors very easy

But the whole thing fails if you don't know how to do A or B. Our next lecture will be spent almost entirely on this question. In that case, you probably fall back on frequentist alternatives. These use only C, but at a cost: they cannot directly tell you the relative probabilities of different hypotheses.

The Normalization Term (aka the denominator)

$$P(H|D,I)=\frac{P(H|I)P(D|H,I)}{P(D|I)}$$

The term P(D|I) is first of all a normalization term. It's the probability of the data summed over all considered hypotheses. (Really it's the integral of the numerator over all values of H).

It's also a check on the validity of your assumptions. If P(D|I) is very, very small, then either you got unlucky, or your prior was far off, or your hypothesis set (denoted by I) doesn't include the true hypothesis.

Sherlock Holmes on hypotheses

"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?"

Bayesian analysis enforces this, since the renormalization procedure demands that one of the hypothesis explicitly under consideration must be the correct one.

If your set of hypotheses is incorrect, your analysis is too.

Writing P(H|D,I) instead of just P(H|D) is one reminder that background assumptions are *always* being made.

Odds Ratio

$$O_{12} = \frac{P(M_1|D,I)}{P(M_2|D,I)} = \frac{P(M_1|I)P(D|M_1,I)}{P(M_2|I)P(D|M_2,I)}$$

$$O_{12} = \frac{P(M_1|I)}{P(M_2|I)} \frac{P(D|M_1,I)}{P(D|M_2,I)} \equiv \frac{P(M_1|I)}{P(M_2|I)} B_{12}$$

The odds ratio is useful because the normalization factors cancel. It's the ratio of the prior probability estimates times the Bayes factor (ratio of the global likelihoods given the data D).

Odds ratios can be easily converted back into probabilities by restoring the normalization factors:

$$P(M_i|D,I) = \frac{O_{i1}}{\sum_{i=1}^{N} O_{i1}}$$

Bayesian justification of Occam's Razor

“Plurality ought never be imposed without necessity.”---William of Ockham

“Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred.”

“We are to admit no more causes of natural things than such are both true and sufficient to explain their appearances.”---Isaac Newton

“Never attribute to malice what can be explained by incompetence.”--Scott Oser

Consider these two hypotheses:

1) Model M_0 is true. It has no free parameters, but there is one parameter θ whose value is fixed by theory to θ_0.

2) Model M_1 is true. It has a single free parameter θ.

Which of these models should we favour given the data?

At first glance, M_1 is more powerful in a sense. After all, it includes $\theta=\theta_0$ as one special case, so shouldn't it always be more likely than the more restricted hypothesis?

Intuitively this can't be right, because this would say we should always favour the more complicated hypothesis, even when the data are perfectly consistent with both.

Odds ratio and the Occam factor

Let's calculate the Bayes factor for the two hypotheses.

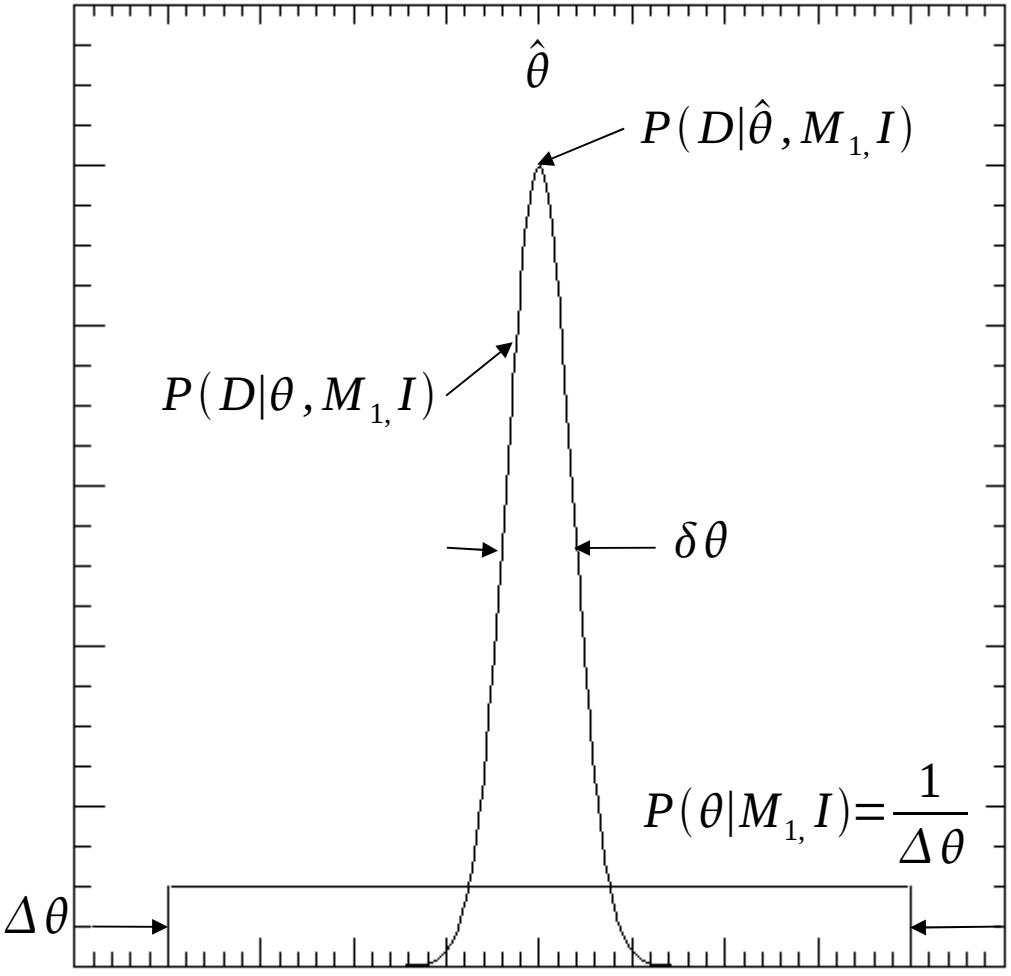
1) For M_0, $P(D|M_0,I)=P(D|\theta_0,M_1,I)=\mathscr{L}(\theta_0)$. Simple to evaluate.

2) For M_1 we need to marginalize over all possible values of θ, including the prior for θ.

$$P(D|M_1,I)=\int d\theta\, P(\theta|M_1,I)\,P(D|\theta,M_1,I)$$

We can approximate this integral. First, let's assume the data is actually pretty constraining compared to the prior, so that $P(\theta|M_1,I)$ is approximately flat over the range for which $P(D|\theta,M_1,I)$ is non-zero.

Odds ratio and the Occam factor



$$\int_{\Delta\theta} d\theta\, P(\theta|M_1,I) = P(\theta|M_1,I)\,\Delta\theta = 1$$

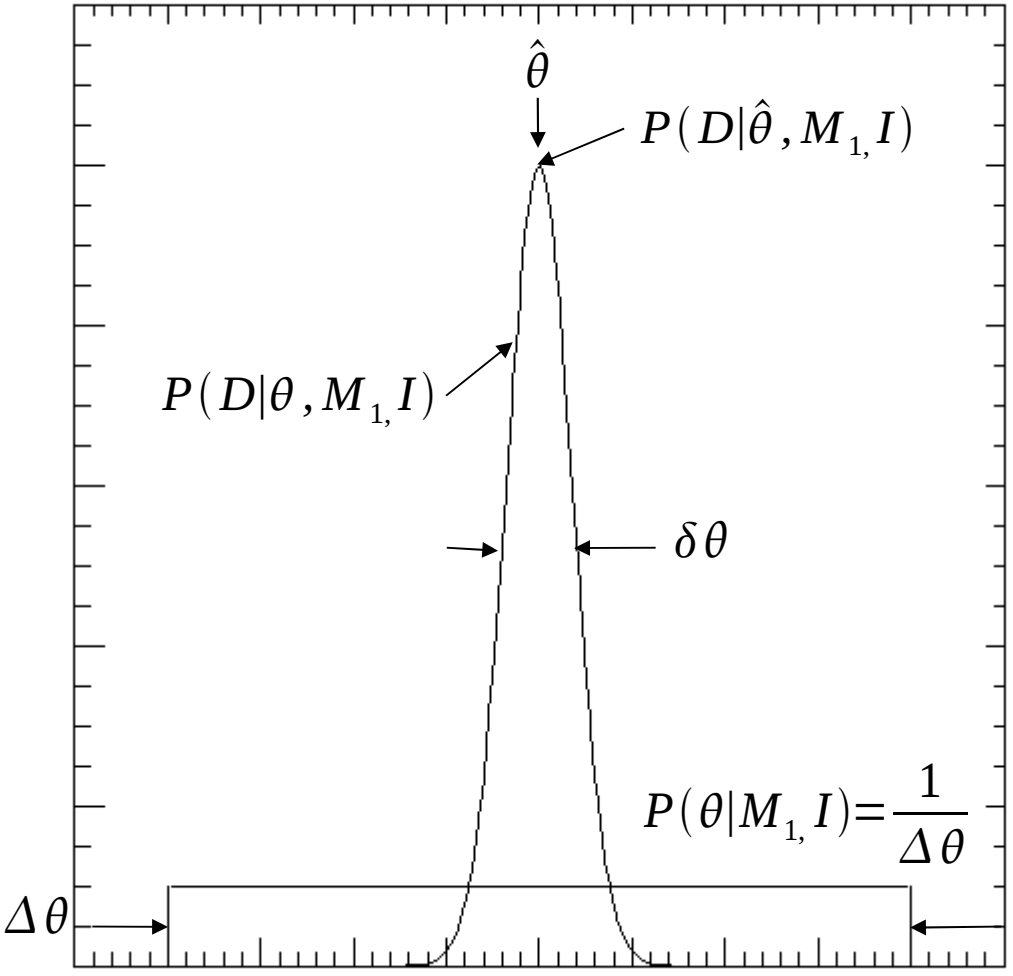
$$\int_{\Delta\theta} d\theta\, P(D|\theta,M_1,I) \equiv p(D|\hat{\theta},M_1,I)\,\delta\theta$$

$$P(D|M_1,I) = \int d\theta\, P(\theta|M_1,I)\,P(D|\theta,M_1,I)$$

$$= \frac{1}{\Delta\theta}\int d\theta\, P(D|\theta,M_1,I)$$

$$\approx \frac{\delta\theta}{\Delta\theta} P(D|\hat{\theta},M_1,I)$$

Odds ratio and the Occam factor



Bayes factor:

$$B \approx \frac{P(D|\hat{\theta}, M_1, I)}{P(D|\theta_0, M_1, I)} \frac{\delta\theta}{\Delta\theta}$$

The first part is always bigger than or equal to 1.

But the second factor is smaller than 1. Since the posterior width $\delta\theta$ is narrower than the prior width $\Delta\theta$, the second parameter is penalized for the "wasted" parameter space.

The Bayes factor will only favour the more complicated model if the likelihood factor (blue) is much better, overwhelming the Occam factor (red).

Summary of the Occam factor

Bayesian analysis automatically penalizes more complicated models in a quantitative way compared to simpler models.

This happens in the process of marginalizing over free parameters in the model. The more free parameters you have to marginalize over, the larger the penalty.

It is still of course possible that a more complicated model fits the data better. If the probability of the data under the simpler model is very small, but much larger under the more complicated model, then the complicated model will still be favoured in spite of penalty factor.

The penalty factor is perhaps intuitively obvious. The more free parameters you have, the more likely it is that your model matched the data just by blind luck. The model that makes more specific predictions (has *fewer* free parameters) will tend to be favoured, so long as it is consistent with data.

Nuisance parameters

A “nuisance parameter” is a parameter model that affects the probability distributions but which we don't care about for its own sake. An example would be a calibration constant of an apparatus---not the sort of thing you report in the abstract, but important nonetheless.

Bayesian analysis gives a simple procedure for handling these: assign priors to all parameters, calculate the joint posterior PDF for all parameters, then marginalize over the unwanted parameters.

If θ is an interesting parameter, while α is a calibration constant, we write:

$$P(\theta|D,I) = \int d\alpha\, P(\theta,\alpha|D,I) = \int d\alpha \left[\frac{P(\alpha|I)\,P(\theta|I)\,P(D|\theta,\alpha,I)}{P(D|I)} \right]$$

(I've assumed independent priors on α and θ, but this is not necessary.)

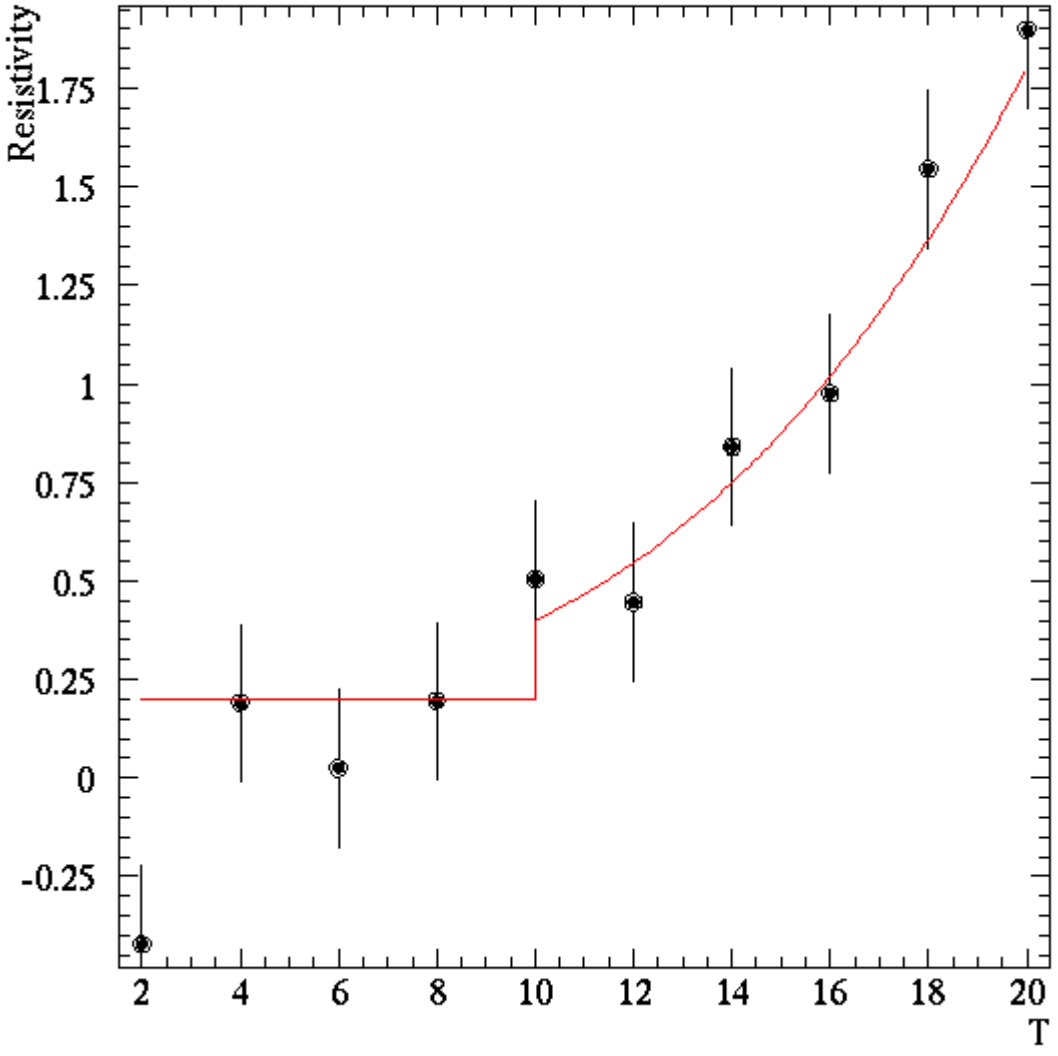
Systematic uncertainties

$$P(\theta|D,I)=\int d\alpha\, P(\theta,\alpha|D,I)=\int d\alpha \left[\frac{P(\alpha|I)P(\theta|I)P(D|\theta,\alpha,I)}{P(D|I)}\right]$$

Nuisance parameters provide an obvious way to include systematic uncertainties. Introduce a parameter characterizing the systematic, specify a prior for the true values of that systematic, then integrate over the nuisance parameter to get the PDF for the quantity you do care about.

The frequentist version is much nastier---without the language of a “prior”, the marginalization procedure, and the philosophy of treating the data as generating a PDF for the parameters, it's much harder to handle systematics. We'll discuss this in future sessions.

An involved example: estimating a superconductor's critical temperature



Superconductor has sudden drop in resistivity below its critical temperature. Model it as:

$R = B$ (if $T<T_c$)

$R = B+A(T/T_c)^3$ (if $T>T_c$)

Here B is a calibration offset, T_c is the critical temperature, and A is an uninteresting material parameter.

Data at right drawn from true distribution shown in red.

Superconductor: define the model

There are three parameters, only one of which we really care about. Let's assume uniform priors for each:

$P(B) = 1 \qquad (0<B<1)$
$P(A) = 1 \qquad (0<A<1)$
$P(T_c)=1/20 \qquad (0<T_c<20)$

And now we define the model. The model will be that the data are scattered around the theoretical curve

$R = B \qquad (\text{if } T<T_c)$
$R = B+A(T/T_c)^3 \qquad (\text{if } T>T_c)$

with Gaussian errors having $\sigma=0.2$ (we assume this is known from characterization of the apparatus).

Superconductor: the form of the likelihood

Need to write down a form for P(D|A,B,T_c,I)

$$P(D|A,B,T_c,I)=\prod_{i=1}^{N}\exp\left[-\frac{1}{2\sigma^2}\left(D_i-R(T_i)\right)^2\right]$$

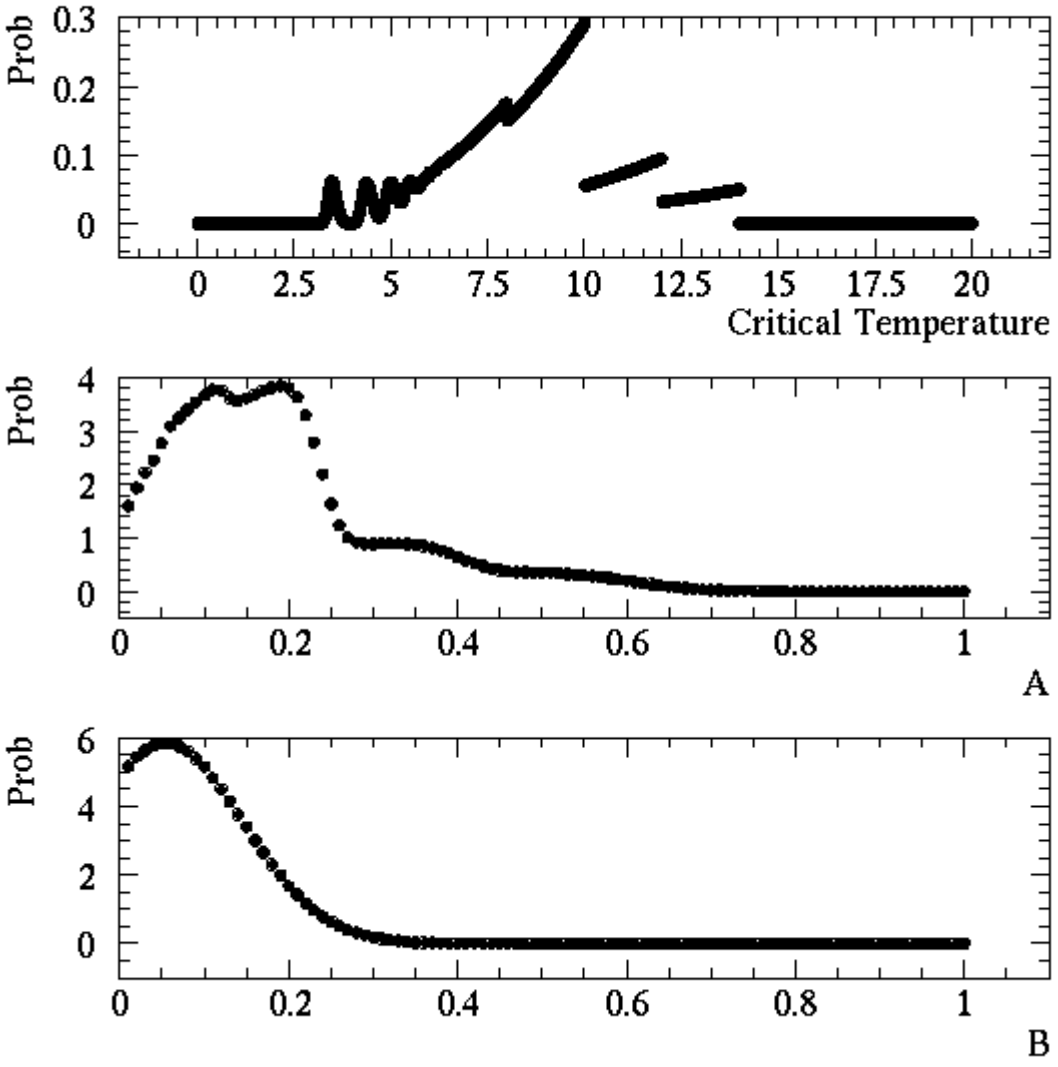
where R(T_i) is the piecewise-defined function given previously. All the dependence on model parameters is contained in R(T).

Bayes theorem now immediately defines a joint PDF for the parameters by

$$P(A,B,T_c|D,I)\propto P(A,B,T_c|I)\,P(D|A,B,T_c,I)$$

All there is left to do is to normalize the PDF, and marginalize over the unwanted variables to get the PDFs on any parameter you care about.

Superconductor: marginalized PDFs



Here I show the marginalized PDFs for T_c, A, and B. A is perhaps like you would have expected. B is OK---low, but data was quite a bit low as well.

(True values: A=0.2, B=0.2, T_c=10)

PDF for T_c puzzled me at first. It spikes near true value, but is not very smooth. The reason is that the model being fitted is discontinuous, so you get discontinuities at the data points.

Discussion: What prior should I use for θ_{13}?

My current experiment aims to measure the neutrino mixing parameter θ_{13}. It's small, and consistent with zero. We want to measure its true value if we can, which should be possible if it's not too small. Actually what we'll measure is more like $\sin^2 \theta_{13,}$ for which we hope to beat the world limit by a factor 20.

At a conference a theorist and I were discussing a competing experiment which will improve the current limit by a factor of 3. Two views were expressed:

A. "If you lower the current limit by a factor of 3, then Experiment X has a 2/3 chance of seeing something, since $\sin^2 \theta_{13}$ can have any value between zero and the current limit."
B. "But we don't have any idea of the scale of $\sin^2 \theta_{13.}$ It is as likely to be between 0.001-0.01 as it is between 0.01-0.1. We should use a prior that is uniform in the logarithm---there's a high probability for $\sin^2 \theta_{13}$ to be very small."

WHO IS RIGHT?

EXTRA MATERIAL

Probability as a generalization of Aristotlean logic

Compare:
1) If A is true, then B is true.
2) B is false.
3) Therefore A is false

with:
1) If A is true, then B is true
2) B is probably not true
3) Having learned that B is probably not true, I am less convinced than A is true.

The “subjective” interpretation of probability can be considered to be an attempt to generalize from deductive logic.

Desiderata of inductive reasoning

Here's how to “derive” the rules of probability. Demand that:

1) Degrees of plausibility are represented by real numbers
2) “Common sense”: as data supporting a hypothesis is accumulated, the plausibility of the hypothesis should increase continuously and monotonically.
3) Consistency: if there are two valid ways to calculate a probability, both methods should give the same answer!

An amazing result: if you try to construct a system of assigning degrees of plausibility to statements according to these requirements, the only unique way to do so results in the regular rules of probability (product rule, sum rule, Bayes' theorem)!

Conclusion: probability is the unique inductive generalization of “Boolean algebra”

Objective vs. Subjective

Ed Jaynes imagined a robot programmed to use a Bayesian interpretation of probability:

“Anyone who has the same information, but comes to a different conclusion than our robot, is necessarily violating one of those desiderata. While nobody has the authority to forbid such violations, it appears to us that a rational person, should he discover that he was violating one of them, would wish to revise his thinking ...”

In other words, Bayesian probability estimates are still objective provided that any two observers starting with the same information and looking at the same data will assign the same probability estimates to the result.