Bayesian Phylogenetics

Nick Matzke

I. Background: Philosophy of Statistics

What is the point of statistics? And what are you doing when you reach a statistical conclusion? These questions are basically never asked or answered in most introductory statistics classes, from middle school through many graduate courses.

The questions only became apparent to me when I began to realize that the field of statistics is not like basic mathematics, even though at first it seems like just an application of the math you learned in high school. In basic math, answers are either right or wrong, and that's it. In statistics, the "right" method (and thus answer) can often be a matter of opinion. In statistics, there are

- judgment calls,
- background philosophies,
- uncooperative data (e.g., data that don't fit ideal criteria, such as independence, or following a standard distribution – especially e.g. biological and spatial data),
- uncooperative calculations (e.g., non-integrable functions, calculations that take too long, problems that involve evaluating more possibilities than there are atoms in the universe)
- *important*, *practical* decisions that depend upon the conclusion reached, despite all of the above

Question: What are some practical decisions that rely upon statistical conclusions? In science? In biology? In phylogenetics?

E.g.: Ou *et al*. (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science* 256, 1165–1171. doi: 10.1126/science.256.5060.1165

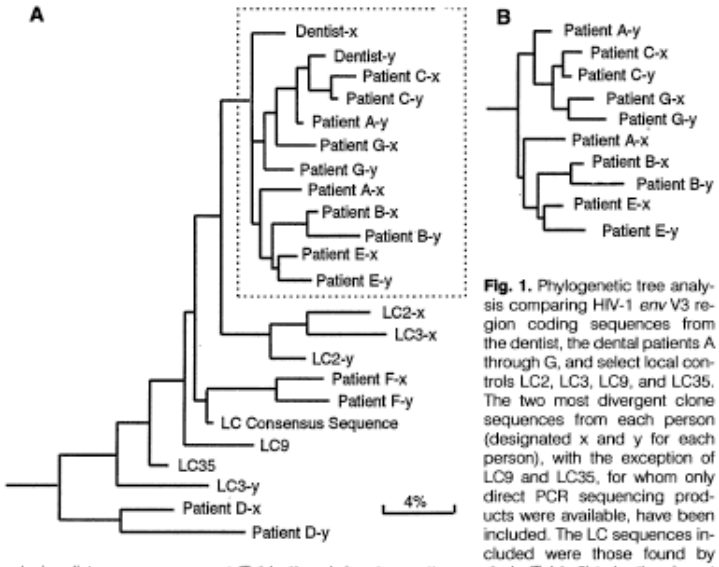


Fig. 1. Phylogenetic tree analysis comparing HIV-1 *env* V3 region coding sequences from the dentist, the dental patients A through G, and select local controls LC2, LC3, LC9, and LC35. The two most divergent clone sequences from each person (designated x and y for each person), with the exception of LC9 and LC35, for whom only direct PCR sequencing products were available, have been included. The LC sequences included were those found by pairwise distance measurement (Table 1) and signature pattern analysis (Table 2) to be the closest control sequences to the dental group sequences, which are enclosed by a box in (**A**). An LC consensus sequence has also been included, and the tree was rooted upon the African sample ELI. The PAUP parsimony algorithm was used to analyze 279 aligned sites (*25*), of which 146 sites were varied. When the dentist's viral sequences were withdrawn from the analysis, or required to cluster with the LC consensus sequence (*25*), the dental clade remained otherwise unaffected, as shown in (**B**). Vertical distances are for clarity only; the lengths of the horizontal branches are proportional to the single base changes and can be read as percentage differences with the scale bar.

de Oliveira *et al*. (2006). HIV-1 and HCV sequences from Libyan outbreak. *Nature*, 444, 836-837. doi:10.1038/444836a Received: 4 November 2006; Accepted 24 November 2006; Published online 6 December 2006. http://www.nature.com/nature/journal/v444/n7121/full/444836a.html

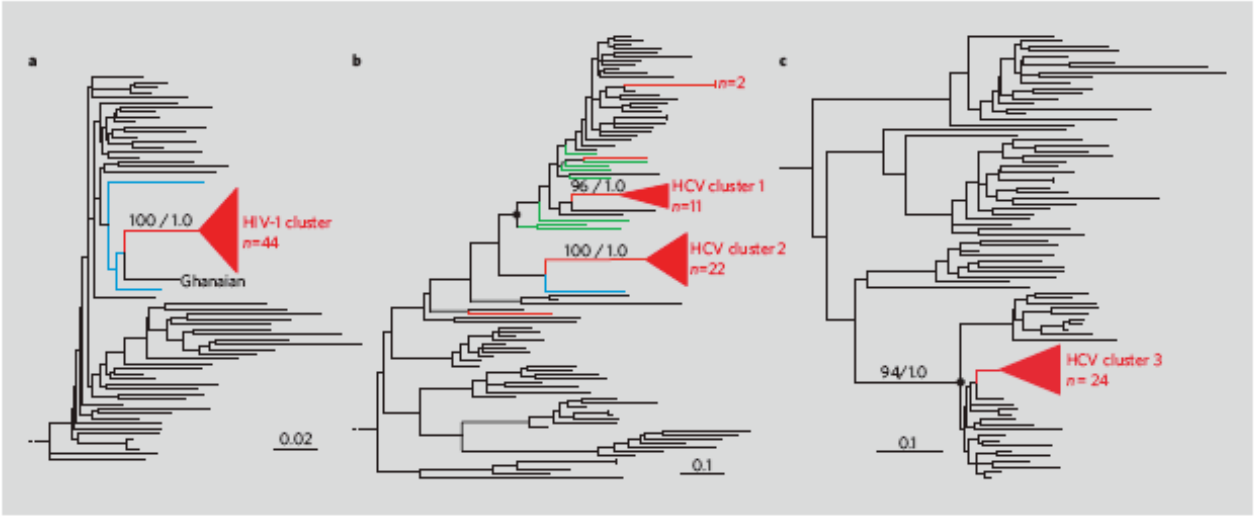


Figure 1 | HIV-1 and HCV sequences from 1998 Al-Fateh Hospital (AFH) outbreak. **a–c**, Estimated maximum-likelihood phylogenies for HIV-1 CRF02_AG (**a**), HCV genotype 4 (**b**) and HCV genotype 1 (**c**). Source of sequences used for analysis: AFH, red; Egypt, green; Cameroon, blue. Black circles mark the common ancestor of HCV subtype 4a and 1a; numbers above AFH lineages give clade support values using bootstrap and bayesian methods, respectively. Scale bar units are nucleotide substitutions per site. For visual clarity, AFH clusters are represented by triangles and some non-informative reference strains are excluded.

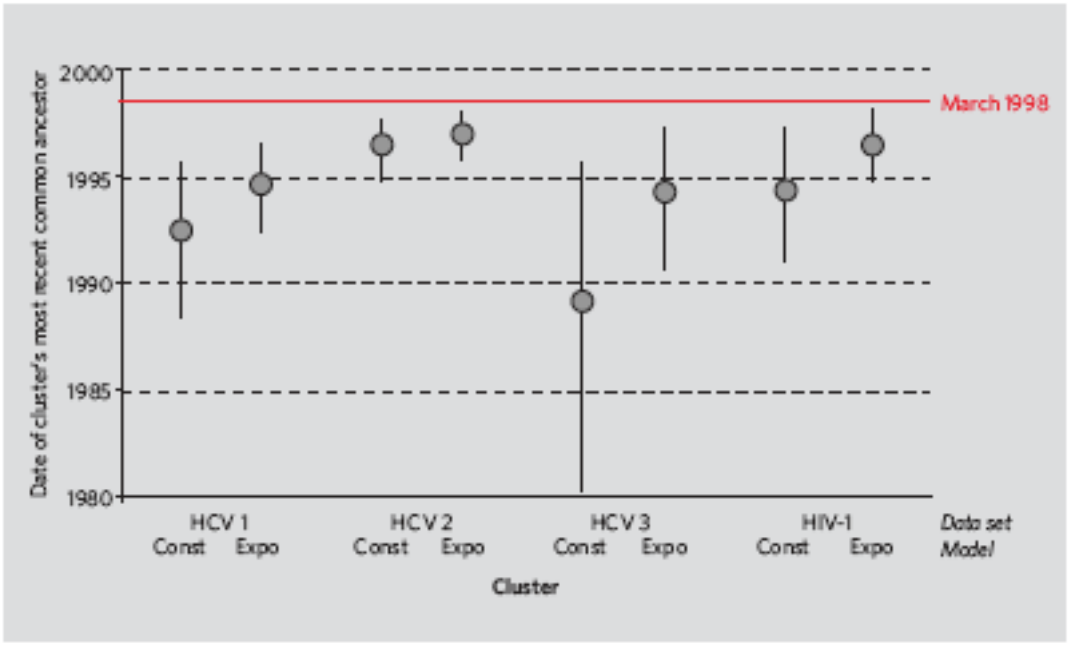
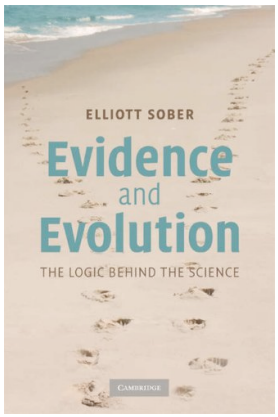


Figure 2 | Estimated dates of the most recent common ancestor for each cluster. Results obtained by using different evolutionary models. Vertical lines show the 95% highest posterior density intervals. Red line shows time of arrival of the foreign staff in March 1998. For further details, see supplementary information. 'Const', constant size; 'Expo', exponential growth.

An excellent, broad, and sophisticated-yet-introductory discussion of statistics and its application to evolutionary science is Elliot Sober's (2008) *Evidence and Evolution: The logic behind the science*. The chapters:

1. Evidence
2. Intelligent design
3. Natural selection
4. Common ancestry



Sober's main point is that it is extremely important to be extremely clear on what question, exactly, you are asking. The grand debates between different statistical "schools of thought" – Bayesian, Likelihoodism, Frequentism – and about what specific methods are appropriate are often much more resolvable if you think carefully about what question you have, and what information (data) you have or can get.

Sober (2008), p. 3:

> The statistician Richard Royall begins his excellent book on the concept of evidence (Royall 1997:4) by distinguishing three questions:
>
> (1) What does the present evidence say?
> (2) What should you believe?
> (3) What should you do?
>
> […] answering question (2) requires more than an answer to (1), and answering question (3) requires more than an answer to (2).

The best feature of Sober is the extremely clear introduction to three major statistical "schools of thought," and discussion of the strengths and weaknesses of each in numerous specific real-world situations (including inferring common ancestry versus separate ancestry, and inferring the action of natural selection).

II. Bayesianism, Likelihoodism, Frequentism

Except for basic probability, essentially all the statistics that any of you learned in high school and college was frequentist (without saying so). So for most people, frequentist statistics – ideas like chi-squared tests, t-tests, regression, ANOVA, and testing of null hypotheses – simply is "statistics."

Sir Ronald Aylmer Fisher (1890-1962)

Strangely, frequentist statistics is actually the youngest school of thought, and its dominance is a recent phenomenon, dating only to the early/mid-20th century. Frequentism definitely benefited from being the favored approach during the explosion of professional science over the last 100 years, and frequentism was particularly strong in biology, especially genetics and population genetics. The famous evolutionary biologist Sir Ronald A. Fisher was *also* probably the most important founder/promoter of frequentism. E.g. Wikipedia quotes Richard Dawkins calling him "the greatest of Darwin's successors," and someone else calling him "a genius who almost single-handedly created the foundations for modern statistical science." (Note equation of frequentism with "modern statistical science!")

Another Wikipedia gem: L.J. Savage, "I occasionally meet geneticists who ask me whether it is true that the great geneticist R.A. Fisher was also an important statistician" (*Annals of Statistics*, 1976).

(In addition to helping to found population genetics, frequentist statistics, inventing Fisher's Fundamental Theorem of Natural Selection, and being knighted, Fisher was also an avid lifelong eugenicist, and a lifelong devout Anglican; the concept linking all of this together is "Progress," but that is a different lecture…)

Bayes

Bayesianism is actually much older, dating back at least to the 1700s and discussions of games of chance and probabilities. The name comes from the Reverend Thomas Bayes (1702-1761), who proposed a special case of what came to be called "Bayes' theorem" in his posthumous *Essay Towards Solving a Problem in the Doctrine of Chances* (1764).

Bayes' theorem is easiest to understand by starting with basic probability and conditional probability.

Basic Probability

Let's first remember some basic probability.

- P(E) = P(event) = "Probability than an event occurs in a trial"

Often writers talk about the P(data) or P(observations) instead of P(event).

- Probabilities of exclusive events must sum to 1, so P(E) + P(not E) = 1

Discussion Questions:
- What is P(heads)?
- What is P(rolling a 1) = P(event = 1) = P(1)?

Conditional Probability

In reality, to answer the questions above, we need some model or hypothesis before we can calculate the probability. This is:

- P(event given some model/hypothesis) = P(event | hypothesis) = P(E | H)
- "model" and "hypothesis" get used interchangably

E.g., the probability of getting a 1 *on a 6-sided fair die* is

- P(event = 1 | "6-sided fair die") = 1/6, or
- P(E|H) = 1/6, where E="rolling a 1" and H="die is six-sided and fair"

What is the probability of rolling a 1 if the die is randomly picked from 2 dice, where 1 die is 6-sided and fair, and 1 die is 6-sided and all 1s?

- P(event=1 | "fair die") / P("fair die") + P(event=1 | "die w/ all 1s") / P("die w/all ones)

- = P(E=1 | H1) / P(H1) + P(E=1 | H2) / P(H2), where H1 = fair die, H2 = all ones
- = P(E|H1) / P(H1) + P(E|H2) / P(H2)
- = (1/6) / (1/2) + (6/6) / (1/2)
- = 1/12 + 1/2 = 7/12

P(E|H) is a *conditional probability*, i.e. the probability of E given H.

The above example, or thinking of probability in terms of proportions gives us Kolmogorov's (1950) definition of conditional probability (Sober 2008, p. 9):

- P(E | H) = P(E & H) / P(H)

Both E and H, while we call them "events" and "hypotheses", are really both just propositions. Randomly rolling a "1" is no different than randomly picking the fair or unfair die. So E and H can be switched:

- P(H | E) = P(H & E) / P(E)

Since P(H & E) and P(E & H) are the same thing, we can say something interesting:
- P(E & H) = P(E | H) P(H)
- P(H & E) = P(H | E) P(E)

So,
- P(H | E) P(E) = P(E | H) P(H)
- P(H | E) = P(E | H) P(H) / P(E)

Bayes' theorem

This is the standard version of Bayes' theorem. Let's write the same thing in a few different ways:

$$P(H \mid E) = \frac{P(E \mid H)\ P(H)}{P(E)}$$

$$P(\text{hypothesis} \mid \text{event}) = \frac{P(\text{event} \mid \text{hypothesis})\ P(\text{hypothesis})}{P(\text{event})}$$

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\ P(\text{model})}{P(\text{data})}$$

$$\text{Posterior probability} = \frac{\text{Likelihood} * \text{Prior probability of the model}}{\text{Unconditional probability of the data}}$$

Notes:
- *Prior probability* is probability of the model, before you look at the data
- *Posterior probability* is the probability of the model, after adding the data
- The *Likelihood* is the *probability that the model confers on the data*. Keep in mind that it is a probability *of the data*, *not of the model,* although one might prefer a model if it gives the observed data a higher likelihood than another model.
    - Statistical "likelihood" is very different from colloquial "likelihood"

- The *Unconditional probability of the data* is the probability of the data summed over all possible conditions, i.e. an integral.
    - If we think of probability as proportions, then it makes sense that we would need to *normalize* the numerator of Bayes' theorem, so that the posterior probability represents the probability (out of a maximum of 1) of the model.
    - Also known as the "nasty normalizing constant"
    - The integral that gives P(data) is P(data) = ∫ P(data | model) P(model) d(model)
    - This integral is very often impossible, except in simple cases, or certain families of distributions

Example: HIV tests

A classic application is to disease tests. Let's imagine the following:

- 1 in 1000 persons in a population has HIV
    - P(HIV+) = 1/1000
    - Therefore, P(HIV-) = 999/1000
- Doctors have an HIV test that has a 99% true positive rate (it is 99% likely to say "HIV positive" when someone is HIV positive).
    - P(HIV+ test | HIV+) = 0.99
    - Therefore, P(HIV-test | HIV+) = 0.01
- The test also has a 2% false positive rate (it is 2% likely to say "HIV positive" when someone is HIV negative).
    - P(HIV+ test | HIV-) = 0.02
    - Therefore, P(HIV- test | HIV-) = 0.98
- Note: These kinds of error rates are typical for many biochemical tests, relying on strength of antibody binding and like, due to natural variability, cross-reactions to other proteins, a disease being in early stages, or human misuse.

So, if you go to the doctor and get an HIV test, what is the probability that you have the disease:

- If you test negative?
- If you test positive?

If you test negative, you want to know P("actually HIV+" | "HIV negative test"):

P("actually HIV+" | "HIV negative test") = P(HIV- test | HIV+) P(HIV+) / P(HIV- test)

- Get normalizing constant, the unconditional probability of an HIV- test:
    - P(HIV- test) = P(HIV- test | HIV+) P(HIV+) + P(HIV- test | HIV-) P(HIV-)
    - P(HIV- test) = 0.02 * 0.001 + 0.98 * 0.999 = 0.97903
- Put it all together:
    - P(HIV+ | HIV- test) = 0.01 * 0.001 / 0.97903 = 0.0000102

P("actually HIV+" | "HIV positive test") = P(HIV+ test | HIV+) P(HIV+) / P(HIV+ test)

- Get normalizing constant, the unconditional probability of an HIV- test:
    - P(HIV+ test) = P(HIV+ test | HIV+) P(HIV+) + P(HIV+ test | HIV-) P(HIV-)
    - P(HIV+ test) = 0.99 * 0.001 + 0.02 * 0.999 = 0.02097
- Put it all together:
    - P(HIV+ | HIV+ test) = 0.99 * 0.001 / 0.02097 = 0.047

In other words, if you test positive, you still probably don't have the disease (only a 4% chance), given the error rate, and the low prevalence of the disease in the population (the low prior). This has all kinds of implications for cost/benefit analysis of widespread disease (and drug) testing, the ethics of basing legal and employment decisions on such tests, the protocols used for informing people of their test results, etc.

Bayes factors

The Bayes factor is the change in your beliefs from prior to posterior. E.g., if your HIV test is positive, your probability of actually being HIV+ has changed from 0.001 to 0.047. This is a Bayes factor of:

BF = 0.047 / 0.001 = 47

(Technically, the Bayes factor is ratio of the marginal likelihoods of your two models, but if the prior on your two models is the same (we are looking at the same model here, model=HIV+, we are just comparing P(HIV+) before data and P(HIV+) after data), then these are the same thing.

(So, full definition of Bayes Factor:

$BF = P(data \mid H_1) / P(data \mid H_0)$

So:

$P(H_1 \mid data) / P(H_0 \mid data) = BF * P(H_1) / P(H_0)$

…edit due to Michael Jordan's stats 260 class)

In other words, your estimate of the chance that you have HIV should be 47 times higher than it was before! But your overall chance of having HIV, given the assumptions (especially the strong, low prior probability), is still less than 5%.

If you were a member of a population with a high chance of having HIV (say, 1 in 10 = 0.1), you could re-run the numbers implied by a positive test. The BF would be lower (but still greater than 1), but your posterior probability of having HIV would be much higher.

Discrete probabilities versus probability densities

One thing that is confusing to introductory students is that Bayes' theorem is typically introduced using simple problems with discrete probabilities. I.e., with HIV, the model is discrete (HIV+ or HIV-), and so is the data (HIV+ test or HIV- test).

In real life, often we are trying to estimate continuous parameters like branch lengths and substitution rates. Here, the prior, likelihood, and posterior are represented by *continuous functions*. E.g. the probability density functions for *p*, the proportion of time a coin will turn up heads (where 0.5 = fair coin) is shown in Sober (2008), p. 22:

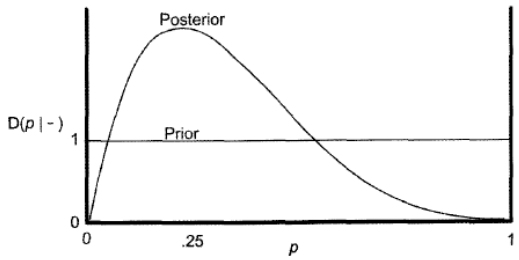


Figure 1.3 A flat prior density distribution for p and the non-flat posterior density occasioned by observing one head in four tosses. The prior expected value of p is 0.5; given this prior, the posterior expected value of p is 0.33.

Under this "flat prior", your initial guess is that p has an equal chance of being any value. After observing 1 head in 4 tosses, your posterior reflects that observed data.

The fact that probability densities are represented by functions means that employing Bayes' theorem involves multiplying, dividing, and integrating functions, which can get complex, although there are a number of useful reference works on the web on the relationships between statistical distributions.

We can, of course, pick a different prior on the proportion of heads, and update that with the likelihood instead to produce a posterior (Albert 2009, p. 25):

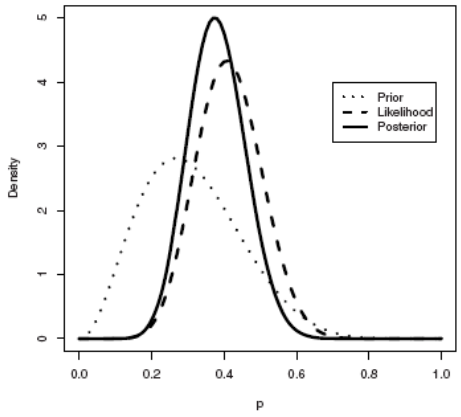


Fig. 2.3. The prior density $g(p)$, the likelihood function $L(p)$, and the posterior density $g(p|\text{data})$ for learning about a proportion p.

Discussion question: What are the differences between the maximum likelihood (ML) estimate for p, and the posterior density of p, in the figure above?

Bayesianism: Pros and Cons

The debates over Bayesianism vs. other approaches are fairly epic, here is a short summary:

Pros:

1. Your beliefs both before and after looking at the new data are explicit, available for all others to judge for themselves.
2. Posterior distributions are better than "point estimates" – i.e., instead of a (point) mean estimate of heights, with a standard deviation which *assumes* some distribution (e.g. the

normal distribution), posteriors can take any shape and are not necessarily controlled by some theoretical assumption.
    - This is why Bayesians talk about *credibility* intervals rather than *confidence* intervals.
3. Bayesian methods can be very flexible, taking into account quite complex models and datasets.
4. The interpretation of posterior probabilities seems fairly obvious and intuitive.
5. Others?

Cons:

1. Getting the nasty normalization constant can be very hard.
    - This is (or at least has been) a practical barrier, not necessarily a philosophical problem.
    - However, this has been ameliorated somewhat by:
        i. Theoretical work, e.g. conjugate priors are known in some situations – a conjugate prior for a certain likelihood function produces a posterior with the same distribution as the prior
        ii. Numerical integration can be attempted when exact mathematical integration is impossible; e.g. MCMC sampling approaches

2. People don't like it because they think all statistics is frequentist.
    - This is a sociological statement, not necessarily an argument, although it is a reason that you need to know the pros and cons of the different approaches, and be able to provide an argument.
    - Scientists tend to be practical, and will go with whatever works for their problem and data.

3. The biggie: how do you choose a prior? Isn't that arbitrary?
    - One response is that everyone is operating under some prior belief, whether or not they admit it, and that it is best to be explicit about it.
    - There are different schools of thought within Bayesianism about how to obtain priors, e.g.
        i. "Objective Bayesians" try to come up with "unbiased" priors that are maximally agnostic about what the true value is. E.g. Laplace justified the use of "flat priors" with the Principle of Indifference.
            1. The Principle of Indifference is flawed. Sober (2008): One might think a reasonable prior on the existence of God is 0.5 – 50/50 chance he exists or not! But there are other options, e.g. what about Zeus?
            2. However, uniform, flat, priors are not always truly agnostic, e.g. watch out for:
                a. The limits on the uniform distribution (ranges from 0 to 10? 0 to infinity?)
                b. Uniform in what space?

c. Depending on the shape of the likelihood curve, a flat prior may actually be putting a lot of weight in an unusual place.

3. Theoretical work indicates that sometimes you can find a "reference prior", which is a prior that has the minimum theoretical impact on the posterior compared to the data.
    a. E.g. Jeffries prior rather than flat for the coin-toss example
    b. However, reference priors are not necessarily conjugate or otherwise convenient
4. Hyperpriors – make the choice of prior itself a variable. (Huelsenbeck: "But the madness has got to stop somewhere!")

ii. Subjective Bayesians: they love putting prior knowledge into the prior, that is the whole point of having a prior.
    1. E.g., a statistician might consult domain experts to get a sense of what a reasonable prior is for a problem.
    2. One might even conduct a formal survey of experts. See e.g. Huelsenbeck et al. (2002), Systematic Biology, p. 678:

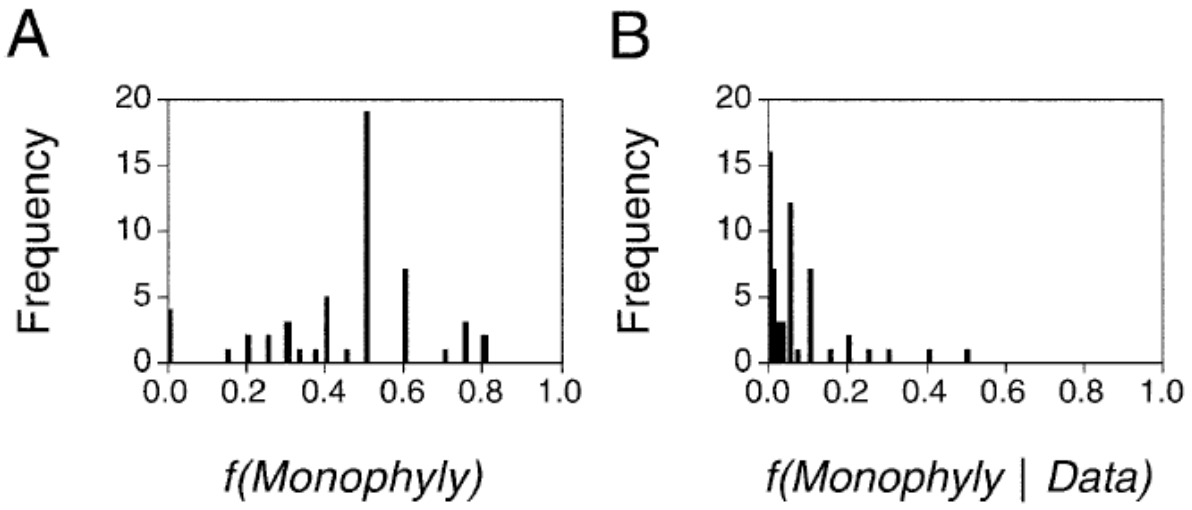


FIGURE 1. Frequency histograms of the responses concerning *Ipomoea* monophyly. (A) Prior beliefs. (B) Updated beliefs.

iii. Or, there are other criteria, e.g. convenience and calculation speed are reasons to prefer conjugate priors

- In general, Bayesians hope/expect that the data will "swamp the prior"

Alternatives to Bayesianism: Likelihoodism

What if you hate priors so much that you don't want to use a Bayesian technique? Or, more fairly, what if you find yourself in a situation where the prior is just intractable? E.g., Sober asks, what is the prior probability of the theory of evolution? What is the prior probability of the General Theory of Relativity?

One alternative is "likelihoodism," where you simply ignore the prior and use only the likelihood.

However, remember what likelihood means.

Discussion question: Sober (2008) asks: What's the likelihood that gremlins are in your attic, given that you hear noise coming from your attic?

Likelihood makes sense for comparing two well-specified hypotheses or models, e.g. comparing relativity to Newtonian physics, and seeing which model confers a higher probability on the data.

Alternatives to Bayesianism: Frequentism

Sober (2008), p. 49: "Frequentists assess a rule of inference by examining the (expected) frequencies of good and bad outcomes when the rule is applied repeatedly."

Frequentism developed as an attempt to make judgments without dealing with priors at all. So other mathematical criteria are used: admissibility (one estimator always works better), optimal unbiased estimator, minimum variance/least squares, maximum likelihood, minimize risk / worse case analysis.

Frequentist methods always rely on taking the expectation of all of the data you might have gotten – i.e., the null hypothesis.

Bayesians ask: how do you pick that null hypothesis, and thereby decide what data you might have gotten? Wouldn't it be better to just use the data directly?

Also: what does rejecting a null actually tell you? Only "not null." Not very useful for building explanatory models.

Generalizations:

- Bayesians integrate (to find the proportion of the probability occupied by a model parameter, out of the total probability of all possible values of the model parameter)

- Frequentists differentiate (to find the maximum likelihood or some other optimum estimator), then find the probability of their estimate according to data they might have gotten under the null model

Take home messages:

- Be pragmatic and have an argument supporting your method for a particular problem

- Sober (2008), p. 2:

  "The debate between Bayesians and frequentists has come to resemble the trench warfare of

World War I. Both sides have dug in well; they have their standard arguments, which they lob like grenades across the no-man's-land that divides the two armies. The arguments have become familiar and so have the responses. Neither side views the situation as a stalemate, since each regards its own arguments as compelling. And yet the warfare continues.

Fortunately, the debate has not brought science to a standstill, since scientists frequently find themselves in the convenient situation of not having to care which of the two approaches they should use. Often, when a Bayesian and a frequentist consider a biological theory in the light of a body of evidence, they both give the theory high marks. This allows biologists to walk away happy; they've got their answer to the biological question of interest and don't need to worry whether Bayesianism or frequentism is the better statistical philosophy. Biologists care about making discoveries about *organisms*; the *nature of reasoning* is not their subject, and they are usually content to leave such "philosophical" disputes for statisticians and philosophers to ponder.

Scientists are *consumers* of statistical methods, and their attitude towards methodology often resembles the attitude that most of us have towards consumer products like cars and computers. We read *Consumer Reports* and other magazines to get expert advice on what to buy, but we rarely delve deeply into what makes cars and computers tick. Empirical scientists often use statisticians, and the "canned" statistical packages they provide, in the same way that consumers use *Consumer Reports*. This is why the trench warfare just described is not something in which most biologists feel themselves to be engulfed. They live, or try to live, in neutral Switzerland; the Battle of the Marne (they hope) involves others, far from home."

- One possible philosophy: "I'm a Bayesian in principle, a likelihoodist in practice, and a frequentist in public." (courtesy Doug Theobald)

- Consider Sober's 3 questions. *Very* roughly:
    1. What does the present evidence say? ← Use likelihoodism
    2. What should you believe? ← Use Bayesianism
    3. What should you do? ← Use frequentism (with a good risk function)

III. Bayesian phylogenetics

We have expended a lot of time and effort going through the background of statistical schools of thought. First, this important general background (I wish someone had given me this lecture when I was a freshman in college). Second, the various complicated debates among phylogeneticists often turn out to be expressions of the fundamental debate between different statistical schools of thought, whether this is realized or not. Third, the Bayesian phylogenetic method makes a lot more sense if you have been introduced to Bayes' theorem first.

The goal of Bayesian phylogenetics

The goal is to find the posterior probability density of a hypothesis/model/model parameters of interest, e.g. a phylogenetic tree, a clade, a substitution rate, etc.:

$$P(\text{tree, parameters} \mid \text{data}) = P(\text{data} \mid \text{tree, parameters})\, P(\text{tree, parameters}) / P(\text{data})$$

tree = tree topology and branch lengths

parameters = parameters of substitution model (e.g. GTR + I + gamma), other parameters
data = sequence alignment or character matrix

Unfortunately, finding the unconditional P(data) requires finding ∫ P(data | tree, parameters) P(tree, parameters) d(tree, parameters), which is impossible.

Sampling the posterior distribution with Markov Chain, Monte Carlo (MCMC) method

Monte Carlo: stochastic draws from distributions
Markov Chain: the values of parameters are explored in a series of steps (a "chain")

MCMC method:

1. Initialize tree + parameters (= "the model") to some values (e.g. drawn from prior).
2. Propose a new model, model'
3. Calculate the probability of the chain moving to the new model, using the following rule:

   $$R = \min\left[1,\ \frac{P(\text{model'} \mid \text{data})}{P(\text{model} \mid \text{data})} \times \frac{P(\text{model} \mid \text{model'})}{P(\text{model'} \mid \text{model})}\right]$$

   The beauty here is that the P(data) cancels out: (X = data, psi or ψ = tree + parameters)

$$R = \min\left[1, \frac{f(\Psi' \mid X)}{f(\Psi \mid X)} \times \frac{f(\Psi \mid \Psi')}{f(\Psi' \mid \Psi)}\right]$$

$$= \min\left[1, \frac{f(X \mid \Psi')f(\Psi')/f(X)}{f(X \mid \Psi)f(\Psi)/f(X)} \times \frac{f(\Psi \mid \Psi')}{f(\Psi' \mid \Psi)}\right]$$

$$= \min\left[1, \underbrace{\frac{f(X \mid \Psi')}{f(X \mid \Psi)}}_{\text{likelihood ratio}} \times \underbrace{\frac{f(\Psi')}{f(\Psi)}}_{\text{prior ratio}} \times \underbrace{\frac{f(\Psi \mid \Psi')}{f(\Psi' \mid \Psi)}}_{\text{proposal ratio}}\right],$$

4. Draw a uniform random number from Uniform(0,1). It is less than R, then switch the chain to the new model.
5. Return to step #2 and repeat.

This procedure is repeated many times. This gradually explores the space of trees and parameters, spending more time on trees/models which confer a higher likelihood on the data. Once the approximate maximum likelihood is reached, the algorithm will wander around. Trees are saved at

regular intervals, and that collection of trees (from the wandering stage) is considered "a sample from the posterior distribution".

The collection of trees represents your estimate of the phylogeny. As when you have multiple equally parsimonious trees, you can produce a consensus tree by strict consensus, majority rule, etc. You can also find the proportion of trees which contain a particular clade of interest, which is the posterior probability of that clade, and this can be compared to the prior probability of the clade by computing a Bayes Factor.

Details, sometimes problematic:

- There is a bit of an art to thoroughly exploring a large, complex space, and not getting stuck on a local optimum. "Proposal mechanisms"

- MrBayes uses "metropolis coupled MCMC", or MCMCMC, to help explore the space. Here, there are (by default) 4 chains, 1 cold and 3 hot. The hot chains are allowed to vary more freely. If a chain happen to find a region with higher likelihood, then (again at a certain rate determined by R), then the chain will become the new cold chain which gets sampled.

- Convergence: typically 2 analyses are run independently, and their convergence is measured. The runs will initially be far apart, and approach each other. This is the "burn-in" period. As a rule of thumb, you want convergence of the average standard deviation of split frequencies to get below 0.01.

- You can have convergence problems especially when searching a really big space, e.g. >150 taxa, or when your data just doesn't have enough signal to resolve the phylogeny.

Some pros and cons

1. Posterior probabilities of clades ("clade credibility values") have a natural interpretation, unlike bootstraps. (Bootstraps also tend to be "too conservative".)

2. Very flexible method to incorporate lots of different data

3. But can be quite slow (hours-days)

4. Various technical issues

More discussion of MrBayes in lab…