Markov Chain Monte Carlo: Approaches for Statistical Inference

Prof. Dr. Narayan Prasad Adhikari
Central Department of Physics
Tribhuvan University Kirtipur, Kathmandu, Nepal

This lecture note is based on Text Book

January 3, 2023

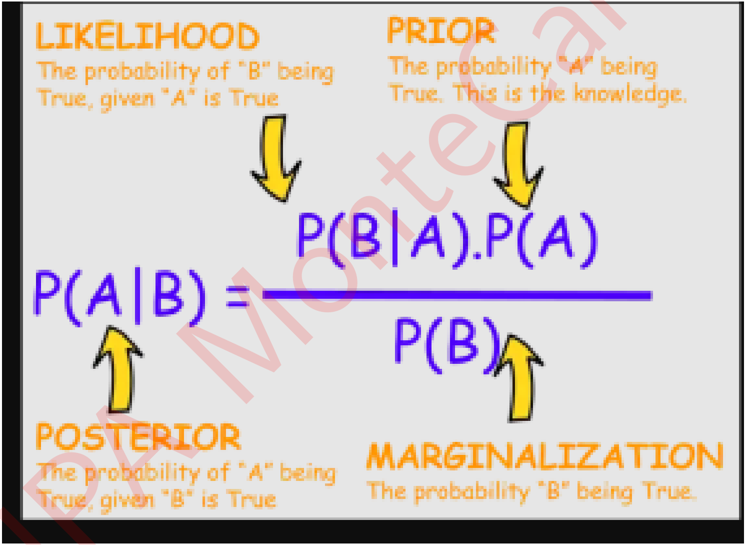
■Outline

- Introduction
- Motivating vignettes
- Defining the approaches
- Bayes vs frequentist approach
- Some basic Bayesian models

■Thomas Bayes

Figure: Thomas Bayes (1701 – 7 April 1761) was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem

■ Bayes Theorem



LIKELIHOOD

The probability of "B" being True, given "A" is True

PRIOR

The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR

The probability of "A" being True, given "B" is True

MARGINALIZATION

The probability "B" being True.

Introduction

- The practicing statistician faces a variety of challenges: designing complex studies, summarizing complex data sets, fitting probability models, drawing conclusions about the present, and making predictions for the future. Statistical studies play an important role in scientific discovery, in policy formulation, and in business decisions. Applications of statistics are ubiquitous, and include clinical decision making, conducting an environmental risk assessment, setting insurance rates, deciding whether (and how) to market a new product, and allocating federal funds. Currently, most statis- tical analyses are performed with the help of commercial software packages, most of which use methods based on a classical, or frequentist, statistical philosophy.

■Introduction

- The Bayesian approach to statistical design and analysis is emerging as an increasingly effective and practical alternative to the frequentist one. Indeed, due to computing advances that enable relevant Bayesian designs and analyses, the philosophical battles between frequentists and Bayesians that were once common at professional statistical meetings are being replaced by a single, more eclectic approach.

■Motivating vignettes

Personal Probability

- Suppose you have submitted your first manuscript to a journal and have assessed the chances of its being accepted for publication. This assessment uses information on the journal's acceptance rate for manuscripts like yours (let's say around 30%), and your evaluation of the manuscript's quality.
- Subsequently, you are informed that the manuscript has been accepted (congratulations!). What is your updated assessment of the probability that your next submission (on a similar topic) will be accepted?

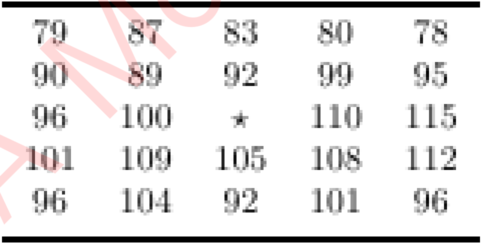
■Motivating vignettes

- The direct estimate is of course 100% (thus far, you have had one success in one attempt), but this estimate seems naive given what we know about the journal's overall acceptance rate (our external, or prior, information in this setting). You might thus pick a number smaller than 100%; if so, you are behaving as a Bayesian would because you are adjusting the (unbiased, but weak) direct estimate in the light of your prior information. This ability to formally incorporate prior information into an analysis is a hallmark of Bayesian methods, and one that frees the analyst from ad hoc adjustments of results that "don't look right."

■Motivating vignettes

Missing Data

- Consider Table (below), reporting an array of stable event prevalence or incidence estimates scaled per 10,000 population, with one value (indicated by “$\star$”) missing at random. We may think of them as geographically aligned disease prevalences, or perhaps as death rates cross-tabulated by clinic and age group.



79	87	83	80	78
90	89	92	99	95
96	100	⋆	110	115
101	109	105	108	112
96	104	92	101	96

■Motivating vignettes

- With no direct information for $\star$, what would you use for an estimate? Does 200 seem reasonable? Probably not, since the unknown rate is surrounded by estimates near 100. To produce an estimate for the missing cell you might fit an additive model (rows and columns) and then use the model to impute a value for $\star$, or merely average the values in surrounding cells. These are two examples of borrowing information. Whatever your approach, some number around 100 seems reasonable.

- Now assume that we obtain data for the $\star$ cell and the estimate is, in fact, 200, based on 2 events in a population of 100 ($200 = 10000 \times 2/100$). Would you now estimate $\star$ by 200 (a very unstable estimate based on very little information), when with no information a moment ago you used 100?

■Motivating vignettes

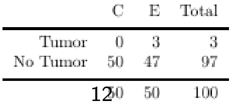
- While 200 is a perfectly valid estimate (though its uncertainty should be reported), some sort of weighted average of this direct estimate (200) and the indirect estimate you used when there was no direct information (100) seems intuitively more appealing. The Bayesian formalism allows just this sort of natural compromise estimate to emerge.

- Finally, repeat this mental exercise assuming that the direct estimate is still 200 per 10,000, but now based on 20 events in a population of 1000, and then on 2000 events in a population of 100,000. What estimate would you use in each case? Bayes and empirical Bayes methods structure this type of statistical decision problem, automatically giving increasing weight to the direct estimate as it becomes more reliable.

Motivating vignettes

Bioassay:measurement of the concentration or potency of a substance by its effect on living cells or tissues.

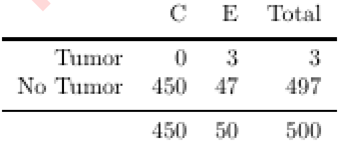
- Consider a carcinogen bioassay where you are comparing a control group (C) and an exposed group (E) with 50 rodents in each (see Table below). In the control group, 0 tumors are found; in the exposed group, there are 3, producing a non-significant, one-sided Fisher exact test p-value (p-values gives the probability that the null hypothesis is true) of approximately 0.125. However, your colleague, who is a veterinary pathologist, states, “I don’t know about statistical significance, but three tumors in 50 rodents is certainly biologically significant!”



	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
	50	50	100

■Motivating vignettes

- This belief may be based on information from other experiments in the same lab in the previous year in which the tumor has never shown up in control rodents. For example, if there were 400 historical controls in addition to the 50 concurrent controls, none with a tumor, the one-sided p-value becomes 0.001 (see Table below). Statistical and biological significance are now compatible. In general, it can be inappropriate simply to pool historical and concurrent information. However, Bayes and empirical Bayes methods may be used to structure a valid synthesis



	C	E	Total
Tumor	0	3	3
No Tumor	450	47	497
	450	50	500

■Motivating vignettes

Attenuation Adjustment

- In a standard errors-in-variables simple linear regression model, the least squares estimate of the regression slope (β) is biased toward 0, an example of attenuation (the reduction of the force, effect, or value of something). More formally, suppose the true regression is $Y = \beta x + \epsilon$, $\epsilon \sim N(0, \sigma_\epsilon^2)$, but Y is regressed not on x but on X $\equiv$ x + δ, where $\delta \sim N(0, \sigma_\delta^2)$. Then the least squares estimate $\hat{\beta}$ has expectation $E[\hat{\beta}] \simeq \rho\beta$, with

$$\rho = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\delta^2}$$

If ρ is known or well-estimated, one can correct for attenuation and produce an unbiased estimate by using $\hat{\beta}/\rho$ to estimate β.

■Motivating vignettes

- Though unbiasedness is an attractive property, especially when the standard error associated with the estimate is small, in general it is less important than having the estimate "close" to the true value. The expected squared deviation between the true value and the estimate (mean squared error, or MSE) provides an effective measure of proximity. Fortunately for our intuition, MSE can be written as the sum of an estimator's sampling variance and its squared bias,

$$MSE = variance + (bias)^2$$

- The unbiased estimate sets the second term to 0, but it can have a very large MSE relative to other estimators; in this case, because dividing $\hat{\beta}$ by ρ inflates the variance as the price of eliminating bias. Bayesian estimators typically strike an effective tradeoff between variance and bias.

■Defining Approaches

- Three principal approaches to inference guide modern data analysis: frequentist, Bayesian, and likelihood.
- The frequentist evaluates procedures based on imagining repeated sampling from a particular model (the likelihood), which defines the probability distribution of the observed data conditional on unknown parameters. Properties of the procedure are evaluated in this repeated sampling framework for fixed values of unknown parameters; good procedures perform well over a broad range of parameter values.

■Defining Approaches

- The Bayesian requires a sampling model and, in addition, a prior distribution on all unknown quantities in the model (parameters and missing data). The prior and likelihood are used to compute the conditional distribution of the unknowns given the observed data (the posterior distribution), from which all statistical inferences arise. Allowing the observed data to play some role in determining the prior distribution produces the empirical Bayes (EB) approach. The Bayesian evaluates procedures over repeated sampling of unknowns from the posterior distribution for a given data set. The empirical Bayesian may also evaluate procedures under repeated sam- pling of both the data and the unknowns from their joint distribution.

■Defining Approaches

- The likelihoodist (or Fisherian) develops a sampling model but not a prior, as does the frequentist. However, inferences are restricted to procedures that use the data only as reported by the likelihood, as a Bayesian would. Procedure evaluations can be from a frequentist, Bayesian, or EB point of view.

The Bayes-frequentist controversy

- While probability has been the subject of study for hundreds of years (most notably by mathematicians retained by rich noblemen to advise them on how to maximize their winnings in games of chance), statistics is a relatively young field. Linear regression first appeared in the work of Francis Galton in the late 1800s, with Karl Pearson adding correlation and goodness-of-fit measures around the turn of the last century. The field did not really blossom until the 1920s and 1930s, when R.A. Fisher developed the notion of likelihood for general estimation, and Jerzy Neyman and Egon Pearson developed the basis for classical hypothesis testing. A flurry of research activity was energized by the World War II, which generated a wide variety of difficult applied problems and the first substantive government funding for their solution in the United States and Great Britain.

■The Bayes-frequentist controversy

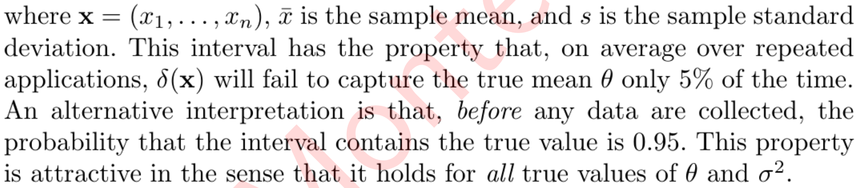
- By contrast, Bayesian methods are much older, dating to the original 1763 paper by the Rev. Thomas Bayes, a minister and amateur mathematician. The area generated some interest by Laplace, Gauss, and others in the 19th century, but the Bayesian approach was ignored (or actively opposed) by the statisticians of the early 20th century. Fortunately, during this period several prominent non-statisticians, most notably Harold Jeffreys (a physicist) and Arthur Bowley (an econometrician), continued to lobby on behalf of Bayesian ideas (which they referred to as "inverse probability"). Then, beginning around 1950, statisticians such as L.J. Savage, Bruno de Finetti, Dennis Lindley, and many others began advocating Bayesian methods as remedies for certain deficiencies in the classical approach. The following example discusses the case of interval estimation.

■The Bayes-frequentist controversy

Example 1.1 Suppose $X_i \overset{iid}{\sim} N(\theta, \sigma^2)$, $i = 1, \ldots, n$, where N denotes the normal (Gaussian) distribution and *iid* stands for "independent and identically distributed." We desire a 95% interval estimate for the population mean θ. Provided n is sufficiently large (say, bigger than 30), a classical approach would use the confidence interval

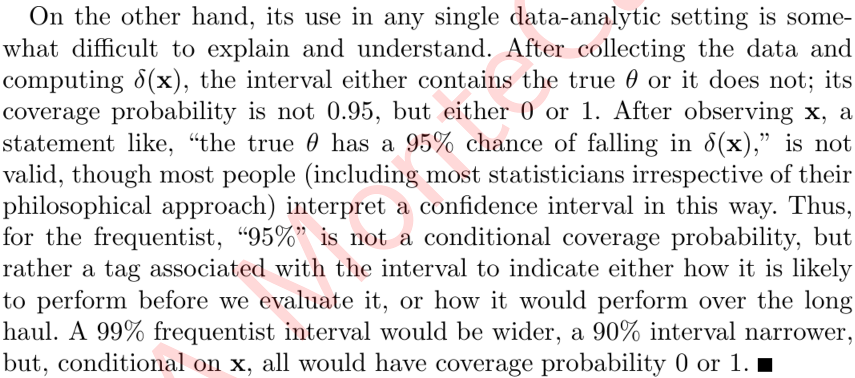
$$\delta(\mathbf{x}) = \bar{x} \pm 1.96 s/\sqrt{n} \ ,$$

■The Bayes-frequentist controversy



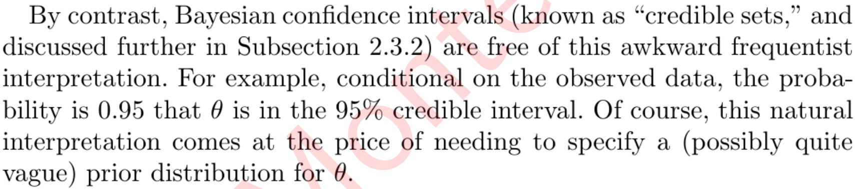
where $\mathbf{x} = (x_1, \ldots, x_n)$, $\bar{x}$ is the sample mean, and s is the sample standard deviation. This interval has the property that, on average over repeated applications, $\delta(\mathbf{x})$ will fail to capture the true mean θ only 5% of the time. An alternative interpretation is that, *before* any data are collected, the probability that the interval contains the true value is 0.95. This property is attractive in the sense that it holds for *all* true values of θ and σ^2.

■The Bayes-frequentist controversy



On the other hand, its use in any single data-analytic setting is somewhat difficult to explain and understand. After collecting the data and computing $\delta(\mathbf{x})$, the interval either contains the true θ or it does not; its coverage probability is not 0.95, but either 0 or 1. After observing $\mathbf{x}$, a statement like, "the true θ has a 95% chance of falling in $\delta(\mathbf{x})$," is not valid, though most people (including most statisticians irrespective of their philosophical approach) interpret a confidence interval in this way. Thus, for the frequentist, "95%" is not a conditional coverage probability, but rather a tag associated with the interval to indicate either how it is likely to perform before we evaluate it, or how it would perform over the long haul. A 99% frequentist interval would be wider, a 90% interval narrower, but, conditional on $\mathbf{x}$, all would have coverage probability 0 or 1. ■

■The Bayes-frequentist controversy



By contrast, Bayesian confidence intervals (known as “credible sets,” and discussed further in Subsection 2.3.2) are free of this awkward frequentist interpretation. For example, conditional on the observed data, the probability is 0.95 that θ is in the 95% credible interval. Of course, this natural interpretation comes at the price of needing to specify a (possibly quite vague) prior distribution for θ.

■The Bayes-frequentist controversy

Example 1.2 Consider the following simple experiment, originally suggested by Lindley and Phillips (1976), and reprinted many times. Suppose in 12 independent tosses of a coin, I observe 9 heads and 3 tails. I wish to test the null hypothesis $H_0 : \theta = 1/2$ versus the alternative hypothesis $H_a : \theta > 1/2$, where θ is the true probability of heads. Given only this much information, two choices for the sampling distribution emerge:

1. *Binomial:* The number $n = 12$ tosses was fixed beforehand, and the random quantity X was the number of heads observed in the n tosses. Then $X \sim Bin(12, \theta)$, and the likelihood function is given by

$$L_1(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{12}{9} \theta^9 (1-\theta)^3 \ . \tag{1.2}$$

2. *Negative binomial:* Data collection involved flipping the coin until the third tail appeared. Here, the random quantity X is the number of heads required to complete the experiment, so that $X \sim NegBin(r = 3,\ \theta)$,

■The Bayes-frequentist controversy

with likelihood function given by

$$L_2(\theta) = \binom{r+x-1}{x} \theta^x (1-\theta)^r = \binom{11}{9} \theta^9 (1-\theta)^3 \; . \qquad (1.3)$$

Under either of these two alternatives, we can compute the p-value corresponding to the rejection region, "Reject H_0 if $X \geq c$." Doing so using the binomial likelihood (1.2), we obtain

$$\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = .075 \; ,$$

while for the negative binomial likelihood (1.3),

$$\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{\infty} \binom{2+j}{j} \theta^j (1-\theta)^3 = .0325 \; .$$

■The Bayes-frequentist controversy

Thus, using the "usual" Type I error level $\alpha = .05$, we see that the two model assumptions lead to two different decisions: we would reject H_0 if X were assumed negative binomial, but not if it were assumed binomial. But there is no information given in the problem setting to help us make this determination, so it is not clear which analysis the frequentist should regard as "correct." In any case, assuming we trust the statistical model, it does not seem reasonable that how the experiment was monitored should have any bearing on our decision; surely only its *results* are relevant! Indeed, the likelihood functions tell a consistent story, since (1.2) and (1.3) differ only by a multiplicative constant that does not depend on θ. ■

■The Bayes-frequentist controversy

- A Bayesian explanation of what went wrong in the previous example would be that the Neyman-Pearson approach allows unobserved outcomes to affect the rejection decision. That is, the probability of X values "more extreme" than θ (the value actually observed) was used as evidence against H 0 in each case, even though these values did not occur. More formally, this is a violation of a statistical axiom known as the *Likelihood Principle*
- ***The Likelihood Principle*** states that once the data value x has been observed, the likelihood function $L(\theta|x)$ contains all relevant experimental information delivered by x about the unknown parameter θ.

■The Bayes-frequentist controversy

- In the previous example, L_1 and L_2 are proportional to each other as functions of θ, hence are equivalent in terms of experimental information (recall that multiplying a likelihood function by an arbitrary function $h(x)$ does not change the MLE $(\hat{\theta})$. Yet in the Neyman-Pearson formulation (power of hypothesis), these equivalent likelihoods lead to two dfferent inferences regarding θ. Put another way, frequentist test results actually depend not only on what x was observed, but on how the experiment was stopped.

■Advantages of Bayesian Approach

1. Bayesian methods provide the user with the ability to formally incorporate prior information.
2. Inferences are conditional on the actual data.
3. The reason for stopping the experimentation does not affect Bayesian inference
4. Bayesian answers are more easily interpretable by nonspecialists (a concern in Example 1.1)
5. All Bayesian analyses follow directly from the posterior; no separate the- ories of estimation, testing, multiple comparisons, etc. are needed.
6. Any question can be directly answered through Bayesian analysis.
7. Bayes procedure possesses numerous optimality properties.

■Some basic Bayesian models