Module 1: A History of Bayesian Statistics

Rebecca C. Steorts

Agenda

- Origin of Bayes
- Bayes' Theorem
- A brief history

Origins of Bayes

To answer the question "Why Bayesian statistics?", we start by looking at the history of statistics from a Bayesian point of view.

Statistics has been around since the beginning of the 19th century, but even before that, probability with a Bayesian (subjective) flavor was being studied.

Origins of statistics

The word “statistics” is of Italian origin.

It is derived from “stato” (state), and a “statista” is a man who deals with affairs of the state.

The original meaning of statistics is thus a collection of facts of interest to a statesman. (Hald, 2003)

Thomas Bayes

Thomas Bayes, a reverend who was interested in probability, lived in England in the 18th century.

At the time, there was no distinction between descriptive and inferential probability although probability used for inferential purposes would eventually come to be known as inverse probability.

Pierre-Simon Laplace

During this same time period, Pierre-Simon Laplace, also worked in this area. Both Bayes and Laplace were aware of a relation that is now known as Bayes Theorem.

Notation

- $x \in X$ is observable, where X is the sample space (meaning it has a probability structure).
- Also $\theta \in \Theta$. While θ is just an index to a frequentist, a Bayesian requires that Θ has a probability structure.

In the equation for Bayes' Theorem, $p(\theta|x)$ suggests a density, and in this class we'll mostly work with densities, although p could also represent a mass function in the discrete case.

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

$p(\theta\|x)$	posterior
$p(x\|\theta)$	likelihood
$p(\theta)$	prior
$p(x)$	marginal likelihood

Bayes Theorem (continued)

- The proportionality $\propto$ in Eq. (1) signifies that the $1/p(x)$ factor may be ignored for the purpose of inference on θ.

We can see the origin of the phrase "inverse probability" by noting that Eq. (1) inverts the relationship between x and θ from $p(x|\theta)$ on the righthand side to $p(\theta|x)$ on the lefthand side.

Objective versus Subjective Bayes

- Bayes and Laplace were "objective Bayesians" in that they viewed the prior, $p(\theta)$, with suspicion.
- They asked why one could trust any particular prior and, from there, why one could trust the resulting inference.
- We will later meet "subjective Bayesians," who embrace priors; the latter paradigm arose in the 1940's and 1950's.
- The question confronting Bayes and Laplace, though, was how to choose the prior so as not to bias their inference.

Laplace's answer was to choose a (uniform) prior. Then $p(\theta)$ is constant, and Eq. (1) yields the relation posterior $\propto$ likelihood.

Objective versus Subjective Bayes (continued)

- For a century, only the Bayesian paradigm existed. Then, in the middle of the 19th century, there was a strong reaction against the prior.
- One objection was the observation that one-to-one transformations of the index may yield different prior densities; in particular, a uniform density for θ might yield a non-uniform density for some one-to-one function of θ. For example, consider the odds $\rho = \theta/(1-\theta)$ or the log-odds $r = \log[\theta/(1-\theta)]$.
- A uniform prior on θ yields densities on ρ and r that are each not uniform. Under Laplace's criterion above, we are not then encoding ignorance in the priors of ρ and r. Since these transformations are one-to-one, the choice to use θ, ρ, or r seems arbitrary. Therefore, it makes little sense to have ignorance in one case but not in the others.

Frequentism

In the 20th century, there was a search for a way to practice statistics without priors.

- Among the prominent figures of this period were Sir Ronald A. Fisher, Jerzy Neyman, and Abraham Wald, who each tried to create new principles upon which to found statistics.
- Fisher took the approach of studying the likelihood (and maximizing it as a function of the index).

Frequentism (continued)

- Neyman founded frequentism. Frequentism takes an entirely different approach from what we've encountered so far in that it stipulates a way of evaluating procedures, where a procedure can be any inference method (even a likelihood-based method or a method with a prior).
- A frequentist asks how the results would change if you ran a procedure over and over again, with the data changing each time.
- Type I and II errors are popular evaluations in this approach.
- This approach is particularly relevant for, e.g., considering software failure rates.
- Wald was a mathematician who, inspired by game theory, developed decision theory. Decision theory formalizes what it means to do inference. In defining such quantities as loss and risk, it quantifies how "good" a method is.

Back to Bayes

- After this intensive effort to circumvent the prior, the pendulum swang back even further in the Bayesian direction to subjective Bayesianity with the work of Leonard J. Savage and Bruno de Finetti.
- These two were uncomfortable with p-values and Type I/II errors. (Review these at home if needed).
- They found paradoxes and incoherencies in the frequentist framework, and, in reaction, embraced priors.
- Emphasizing the subjectivity of the prior, a subjective frequentist will, in practice, sit down with a domain expert to and an appropriate prior for any problem.
- Note: there were massive conflicts between Fisher, Neyman, Wald, Savage, and de Finetti.

Bayes (continued)

In a move that may be seen as coming full circle, the emphasis on subjective Bayesianity was followed by the rise of objective Bayes.

- This new movement, featuring physicist-turned-statistician Harold Jeffreys, was about going back to Laplace's work but trying to improve upon it.
- Notably, objective Bayesians are willing to use frequentist analytic tools to guide their choice of priors. The appeal of these tools is that they are automatic (especially useful when there are many parameters) and do not require a domain expert.

Bayes (continued)

Examples of such tools, include:

- consistency: Does an estimator (random variable) converge, in a probabilistic sense, to the right answer?
- rates of convergence: Sometimes rates slower than $n^{-1/2}$ are not good.
- unbiasedness: All Bayesian procedures are biased, but most frequentists now.
- admissibility: Is there another procedure that dominates the one in question everywhere in the parameter space?

Modern Bayesian Statitical Methods

In this course, we'll go through the following topics:

- Bayesian modeling
- conjugacy
- decision theory
- non-conjugate Bayesian models
- computational tools for Bayesian models
- hierachical Bayesian models (multivariate)
- advanced Bayesian models (regression, logistic regression, and others)
- special topics