COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Bayesian Inference

- In this lecture we will introduce the concept of using probability to express information.
- Then we will define the basic terminology used in Bayesian inference.
- We will also consider some examples to illustrate the new concepts.

Statistical Inference

- Statistical inference or induction is the process of learning about the general characteristics of a population from a subset of members of that population.
- Population characteristics, like mean, variance, … are typically expressed in terms of a parameter θ. We don't know the value of the parameter.
- Very often the parameter θ is a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.
- After a dataset *x* is obtained, the information it contains can be used to decrease our uncertainty about the population characteristics.
- Quantifying this change in uncertainty is the purpose of Bayesian inference.
- The sample space *X* is the set of all possible datasets, from which a single dataset *x* will result.
- The parameter space Θ is the set of possible parameter values, from which we hope to identify the value that best represents the true population characteristics.

Maximum Likelihood Approach

Recall the maximum likelihood technique, in which $\theta \in \Theta$ is considered *fixed* and unknown, characterizing a population with density $f(x|\theta)$. Then an iid sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ is obtained from the population density.

The likelihood function, $L(\theta \mid \boldsymbol{x})$, is defined as

$$L(\theta \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

and the goal is to find a suitable estimate $\hat{\theta}$ of θ.

Most often the difficulties with this approach are optimization type problems, like finding MLE $= \underset{\theta \in \Theta}{\operatorname{argmax}} L(\theta \mid \boldsymbol{x})$, which in turn are often reduced to problems solving equations.

Bayesian Approach

- The main idea is that θ is now a *random variable* with a prior distribution $\pi(\theta)$ which describes our belief about the possible values of θ
- For each θ , the likelihood or sampling model $f(\boldsymbol{x} \mid \theta)$ describes our belief about the values of $\boldsymbol{x}$ for the given value of θ.
- Then the information brought by the sample $\boldsymbol{x}$ is combined with the prior to form the posterior distribution $\pi(\theta \mid \boldsymbol{x})$ by using Bayes formula:

$$f(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\,\pi(\theta)}{m(\boldsymbol{x})},$$

where

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \theta)\,\pi(\theta)\,d\theta$$

is the marginal distribution of $\boldsymbol{X}$.

It is useful to think that θ being random represents our uncertainty about its value. Before seeing the data, our uncertainty is represented by $\pi(\theta)$ and after that by $f(\theta \mid \boldsymbol{x})$.

Example: Bayes Theorem

For a set of mutually exclusive and exhaustive events A_1, A_2, … (i.e. $P(\bigcup_i A_i) = \sum_i P(A_i) = 1$), we have:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}$$

In this case the "parameter" is the indicator which one of the sets A_1, A_2, … occurs with prior probabilities $P(A_i)$, which are then updated to posterior probabilities $P(A_i|B)$, after new "data" is observed from event B.

Example – Diagnostic testing

A new HIV test is claimed to have "95% sensitivity and 98% specificity"
In a population with an HIV prevalence of 1/1000, what is the chance that a patient testing positive has HIV?
Let A be the event patient is truly positive, $\bar{A}$ be the event that they are truly negative. (Note A and $\bar{A}$ are the A_1 and A_2 from the theorem.)
Let B be the event that they test positive.

Diagnostic Testing ctd.

We want p(A|B)

"95% sensitivity" means that p(B|A) = 0.95

"98% specificity" means that $P(B|\bar{A}) = 0.02$

So, from Bayes Theorem

$$P(A|B) = \frac{P(A \cap B)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})}$$

$$= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} = 0.045$$

Thus over 95% of those testing positive will, in fact, not have HIV.

Being Bayesian

So, the vital issue in this example is how should this test result *change our prior* belief that the patient is HIV positive?

The disease prevalence ($p = 0.001$) can be thought of as a '*prior*' probability.

Observing a positive result causes us to modify this probability to $p = 0.045$ which is our '*posterior*' probability that the patient is HIV positive.

This use of Bayes theorem applied to *observables* is uncontroversial. However, its use in general statistical analyses where *parameters* are unknown quantities is more controversial.

Back to Bayesian Inference

Extend conditional probability to random variables:

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

And to conditional densities:

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

Then Bayes' Theorem becomes:

$$f(y|x) = \frac{f(x|y)f(y)}{\int f(x|y)f(y)dy}$$

When we apply it to y = θ we obtain the posterior

$$f(\theta|\,\boldsymbol{x}) = \frac{f(\boldsymbol{x}\mid\theta)\,\pi(\theta)}{m(\boldsymbol{x})}$$

Remarks

- Both classic and Bayesian statistics have advantages and disadvantages.
- Bayesian approach has been criticized for over-reliance on convenient priors and lack of robustness.
- The frequentist approach has been criticized for inflexibility (failure to incorporate prior information) and incoherence (failure to process information systematically).
- For large *n*, as well as when the prior is uniform, the Bayesian method will provide similar results to the classical likelihood approach method.

Example 1: Gaussian Model

It is easier to consider first a model with one unknown parameter and only one observation.

Suppose we have one observation such that:

$$X \mid \theta \sim N(\theta, \sigma^2)$$
$$\theta \sim N(\mu, \tau^2)$$

That is,

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}, x \in \mathbb{R}$$
$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}, \theta \in \mathbb{R}$$
$$f(x, \theta) = \frac{1}{2\pi\sigma\tau} e^{-\frac{1}{2}\left[\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2}\right]}$$

Find $m(x)$ and $f(\theta|x)$.

Solution (nontraditional)

Since $\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2}$ is a non-negative definite quadratic form, the joint distribution of (X, θ) must be a bivariate normal.

But

$$E(X) = E[E(X \mid \theta)] = E(\theta) = \mu$$
$$V(X) = E[V(X \mid \theta)] + V[E(X \mid \theta)] = E(\sigma^2) + V(\theta) = \sigma^2 + \tau^2$$

$$\Rightarrow X \sim N(\mu, \sigma^2 + \tau^2)$$

Now, using the general result that if $(X, \theta) \sim N_2$, then

$$\theta \mid x \sim N_1\left(\mu_\theta + \frac{\sigma_\theta}{\sigma_X}\rho(x-\mu_X), (1-\rho^2)\sigma_\theta^2\right)$$

We find $\text{cov}(X, \theta) = E(X\,\theta) - E(X)E(\theta) = E[E(X\theta \mid \theta)] - \mu^2 = \tau^2$ and

$$\theta \mid x \sim N\left(\frac{\mu\sigma^2 + x\tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$$

Exercise: Repeat the above using only algebra and calculus.

Example1: Normal distribution (continued)

Known variance, unknown mean

Now suppose we have a sample of Normal data:

$$X_1, \dots, X_n \sim N(\theta, \sigma^2)$$

Let us again assume we know the variance, σ^2, and we assume a prior distribution for the mean, θ, based on our prior beliefs:

$$\theta \sim N(\mu, \tau^2)$$

Now we wish to construct the posterior distribution $f(\theta \mid x_1, \dots, x_n)$

(Finish as exercise)

Posterior for the mean of a normal distribution (Chapter 5)

- This is just a preview. We will study this model in detail in Ch. 5.
- The unknown parameter is the mean θ
- The prior is $\theta \sim N(\mu, \tau^2)$. That is,

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$$

- Model for the data is $X_i \sim N(\theta, \sigma^2)$. That is,

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$$

- Conditional likelihood is:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$$

Posterior for the mean of a normal distribution

Hence, without the normalizing constant $m(x)$, the posterior is proportional to:

$$f(\theta|x) \propto \pi(\theta) f(x|\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$$

After using some algebra, you can prove that

$$\sum_{i=1}^{n} \left(\frac{x_i - \theta}{\sigma}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 - \frac{2\theta}{\sigma^2} \sum_{i=1}^{n} x_i + n \frac{\theta^2}{\sigma^2}$$

Then …

$$f(\theta|x) \propto e^{-\frac{1}{2}\theta^2\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right) + \theta\left(\frac{\mu}{\tau^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right) + const} \ldots$$

Precisions and means

When you finish this example as an exercise, you will show that the posterior is normally distributed with mean and variance:

$$E(\theta|\boldsymbol{x}) = \frac{\frac{\mu}{\tau^2} + \frac{\bar{x}}{\sigma^2/n}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \qquad Var(\theta|\boldsymbol{x}) = \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$$

In Bayesian statistics the precision = 1/variance is often more important than the variance.

For the Normal model we have that the posterior precision = sum of prior precision and data precision, and the posterior mean is a (precision weighted) average of the prior mean and data mean.

Heights Example

- Ten subjects had both their heights measured.
- Their heights (in cm) where as follows:

169.6,166.8,157.1,181.1,158.4,165.6,166.7,156.5,168.1,165.3

- We will assume the variance is known to be 50.

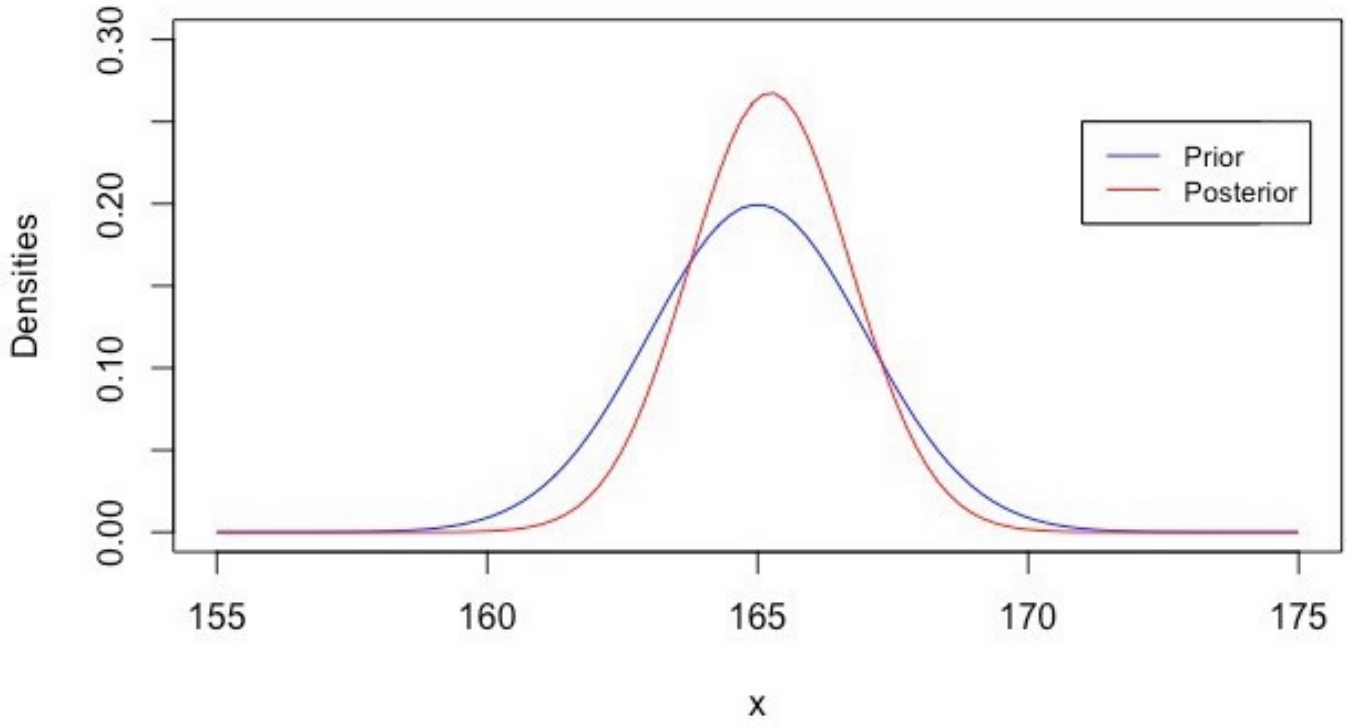
Two experts gave the following prior distributions for the mean height:

Individual 1: $\pi_1(\theta)$ = N(165, 2^2)

Individual 2: $\pi_2(\theta)$ = N(170, 3^2)

Use R to construct and plot the posterior using Individual 1 prior.

Prior and posterior comparison



Estimating θ under Bayesian Paradigm

Q: What do we do with $f(\theta|x)$?

A: It can be reported as a whole posterior likelihood of the parameter, but very often we report just a point estimator $\hat{\theta}$ .The most commonly used estimator is

$$\hat{\theta} = E(\theta \mid x) = \int \theta \, f(\theta|x) d\,\theta$$

In general, we can find estimate of $h(\theta)$ for any integrable function $h(\cdot)$ by

$$\hat{h}(\theta) = E[h(\theta) \mid x] = \int h(\theta) \, f(\theta|x) d\,\theta$$

Therefore, the Bayesian approach often results in integration problems. Some of these difficulties are:

- $f(\theta|x)$ might not be available in closed form or be partially available due to $m(x)$ being unavailable.
- the integration $\int h(\theta) \, f(\theta|x) d\,\theta$ cannot be done analytically.