STA 111: Probability & Statistical Inference

Lecture Twelve – Bayesian Inference

D.S. Sections 7.2, 7.3 & 7.4

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Questions from Last Lecture
- Bayesian Inference
- Conjugacy
- Bayesian Estimators
- Recap

Introduction

– So far we have talked about point estimators, desirable properties of point estimators and one way to derive point estimators – the maximum likelihood method.

– In statistics, there are two major paradigms, the Bayesian paradigm and the classical or frequentist paradigm and our discussions on statistics so far fall under the classical paradigm.

– The objective of this lecture is to simply introduce you to the Bayesian way of thinking about statistics.

– Lastly, we will see how to derive Bayesian estimators.

Bayesian Inference vs. Classical Inference

In the previous lecture we discussed maximum likelihood inference. A maximum likelihood estimate is the parameter value which has the greatest chance of generating the data that were observed (assuming that the analyst has correctly specified the probability model for the data, say exponential or normal or uniform).

This is in line with the frequentist paradigm, where we treat parameters as unknown constants and try to estimate them (use the observed data to take an educated guess about what the population parameter should be).

Bayesian Inference vs. Classical Inference

Under the Bayesian paradigm, parameters are treated as random variables, and we rely o Bayes' rule for inference.

Here treating the parameters as random variables mean we need to find the distribution over all possible parameter values. The distribution of the parameter given the observed data is called the posterior distribution. Again, we have to assume that the probability model has been correctly specified.

Interpretation

One key distinction between the methods is that a Bayesian uses probability to describe their personal uncertainty about the world, whereas a frequentist does not.

For example, a lawyer might want to know whether a client is guilty of murder. If she were Bayesian, she could say something like **"Given the evidence, I think the probability that the client is guilty is at least 0.8."**

A frequentist lawyer on the other hand first assumes that either the client did or didn't – we just don't know which. The frequentist lawyer makes a different statement: **"If the client is innocent, then the probability of having so much evidence against him/her is at most 0.05."**

There are important philosophical and mathematical distinctions between these perspectives.

History and Background

Bayesian inference was invented by the Reverend Thomas Bayes (remember Bayes' rule?), and published posthumously in 1763. The difficulty in calculating most integrals kept it from being widely used until 1990 when a new algorithm was invented (by Alan Gelfand of the Duke statistics department).

Before the data are collected, the Bayesian has a prior opinion about the value of a parameter θ. This prior expresses her uncertainty, and provides a prior density on the parameter, or $\pi(\theta)$.

Then the Bayesian observes data $x_1, \ldots, x_n$ where the data are a random sample from some specified probability model with density $f(x; \theta)$.

Now the Bayesian sees how the data has changed her prior opinion about θ and uses Bayes' rule to find his/her posterior density $\pi^*(\theta \mid x_1, \ldots, x_n)$.

Formula

Recall Bayes' Rule: For a finite partition $A_1, \ldots, A_n$ and an event B,

$$\mathbb{P}[A_i|B] = \frac{\mathbb{P}[B|A_i] \times \mathbb{P}[A_i]}{\sum_{j=1}^{n} \mathbb{P}[B|A_j] \times \mathbb{P}[A_j]}.$$

In the context of Bayesian inference, B is the observed data, the A_i's are all possible parameter values. However, since the possible parameter values are usually continuous, we need to rewrite Bayes' Rule in the language of densities:

$$\pi^*(\theta \,|\, x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \,|\, \theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(x_1, \ldots, x_n \,|\, \theta)\pi(\theta)\, d\theta}.$$

Here $\pi(\theta)$ is one's belief about the parameter before seeing the data, and $\pi^*(\theta \,|\, x_1, \ldots, x_n)$ is one's belief after seeing the data.

Note that the numerator contains the likelihood function and the denominator is just some constant in terms of the x_i's, since we integrate θ out of the picture.

Conjugate Distributions

As mentioned, it is usually hard to solve the integrals that arise in Bayesian statistics. Specifically, it is difficult to evaluate the integral in the denominator of the density version of Bayes' Rule.

But there are a handful of exceptions (called conjugate families or distributions), and fortunately these cover some important and practical situations. These entail three pairs of distributions:

- the Normal-Normal case
- the Beta-Binomial case
- the Gamma-Poisson case.

In each pair the first distribution describes the statistician's prior belief about θ, and the second distribution is the model for how the data are generated for a specific value of θ.

Conjugate Distributions

In the Normal-Normal case, one thinks the data are normally distributed with some unknown mean μ and known variance σ^2. You don't know μ, but your prior belief is that μ is normally distributed with a mean ν and variance τ^2. Then you observe data $x_1, \ldots, x_n$ and apply Bayes Rule to find the posterior distribution of μ. It turns out that the posterior density $\pi^*(\mu|x_1, \ldots, x_n)$ is

$$\mathsf{N}\left(\frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

You could prove all this using the density version of Bayes Rule.

If you attempt this, a good trick is to treat the denominator as some constant c. On multiplying the numerator terms, you can recognize the product as being, up to a constant, the density function of a normal distribution. Then just take c to be whatever is needed to ensure the density integrates to 1. We will derive the Beta-Binomial case to see how the math works out.

Examples

Example 1: Suppose you believe that chest measurements in inches are normally distributed with unknown mean μ and variance $\sigma^2 = 4$.

You do not know μ, but before you begin, you believe it is probably near 41, and you are pretty confident (say 95% probability) that the mean is within plus/minus 6 inches of 41.

If you express this uncertainty as a normal distribution, then $\nu = 41$ and $\tau^2 = 9$ (since two standard deviations on each side is 6 inches, then one sd is 3 inches, and so the variance is 9).

Suppose you observe $\bar{x} = 39.85$ inches and $n = 5732$. Thus Bayes' Rule implies you should now believe that the true average chest circumference is normally distributed with mean

$$\nu^* = \frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2} = \frac{(41 \times 4) + (5732 \times 39.85 \times 9)}{4 + (5732 \times 9)} = 39.85009.$$

Note that the posterior mean is very close to the sample mean.

Examples

Similarly, your uncertainty about the location of μ has gotten very much smaller. The variance of your posterior distribution is

$$\tau^{*2} = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} = \frac{4 * 9}{4 + 5732 * 9} = 0.0007.$$

The large sample size has dramatically reduced your uncertainty about the average chest circumference.

If someone asks you what you think the mean chest circumference is, you can answer $39.85009 \pm 2\sqrt{0.0007}$ (with 95% probability).

Note that the posterior mean is the weighted average of the prior mean ν and the sample mean $\bar{x}$. One can re-write the formula as:

$$\nu^* = \frac{\sigma^2}{\sigma^2 + n\tau^2}\nu + \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{x}.$$

So when n is large, most of the weight goes on $\bar{x}$, the data. But when n is small, most of the weight goes on your prior belief ν.

Conjugate Distributions

In the Beta-Binomial case, you think that your data come from a binomial distribution with an unknown probability of success θ.

You do not know the value of θ, but you have a prior distribution on it. Specifically, your prior is a beta distribution.

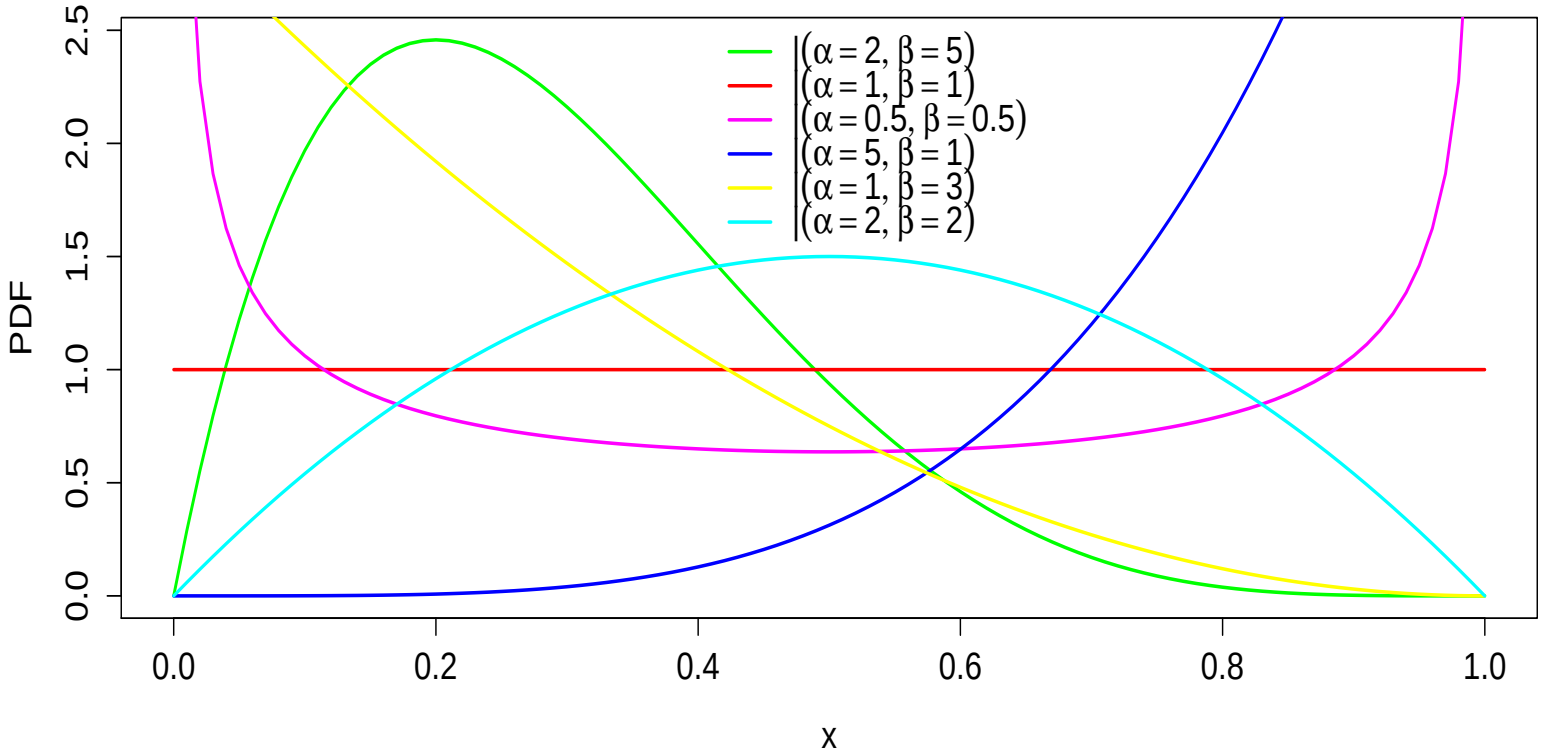
The beta family has two parameters, $\alpha > 0$ and $\beta > 0$, and the beta density on θ is

$$f(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1} \text{ for } 0 \leq \theta \leq 1.$$

where $\Gamma(n) = (n - 1)!$.

One could pick some other distribution with support on $[0, 1]$, if it expressed your personal beliefs about θ. But the beta family is flexible (conjugate to the Binomial likelihood) and it makes the Bayesian mathematics easy.

Conjugate Distributions



These plots show the densities of the beta distribution for different choices of α and β. Which choices would make sense in a coin tossing context?

Conjugate Distributions

Suppose your prior on θ is Beta(α, β). And your data are binomial, so the likelihood function for x successes in n trials is $\binom{n}{x} \theta^x (1-\theta)^{n-x}$. Then Bayes' Rule shows that the posterior on θ is Beta$(\alpha + x, \beta + n - x)$.

$$
\begin{aligned}
\pi^*(\theta \mid x) &= \frac{f(x \mid \theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(x \mid \theta)\pi(\theta)\, d\theta} \\
&= \frac{\left[\binom{n}{x} \theta^x (1-\theta)^{n-x}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\right]}{\int_0^1 \left[\binom{n}{x} \theta^x (1-\theta)^{n-x}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\right] d\theta} \\
&= \ldots \textbf{ some algebra } \ldots \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}
\end{aligned}
$$

which we recognize as the Beta$(\alpha + x, \beta + n - x)$ density.

Examples

Example 2: Suppose you want to find the Bayesian estimate of the probability θ that a coin comes up Heads. Before you see the data, you express your uncertainty about θ as a beta distribution with $\alpha = \beta = 2$. Then you observe 10 tosses, of which only 1 was Heads. Now the posterior density $\pi^*(\theta \mid x, n)$ is Beta$(3, 11)$.

The mean of Beta(α, β) is $\alpha/(\alpha + \beta)$. So before you saw the data, you thought the mean for θ was $2/(2+2) = 0.5$. After seeing the data, you believe it is $3/(3+11) = 0.214$.

The variance of Beta(α, β) is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. So before you saw data, your uncertainty about θ (i.e., your standard deviation) was $\sqrt{4/[4^2 * 5]} = 0.22$. But after seeing 1 Heads in 10 tosses, your uncertainty is 0.106.

As the number of tosses goes to infinity, your uncertainty goes to zero.

Conjugate Distributions

For the Gamma-Poisson case, you believe that the data come from a Poisson distribution with parameter λ, and your uncertainty about λ is expressed by a gamma distribution.

The gamma distribution has two parameters, $\alpha > 0$ and $\beta > 0$. Its density function is

$$f(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Using Bayes' Rule, one can show that if $x_1, \ldots x_n$ are are an observed random sample from a $\text{Po}(\lambda)$ distribution, and if your prior $\pi(\lambda)$ on λ is $\text{Gamma}(\alpha, \beta)$, then your posterior $\pi^*(\lambda \,|\, x_1, \ldots x_n)$ is $\text{Gamma}(\alpha + \sum_{i=1}^{n} x_i, \beta + n)$.

Examples

Example 3 (to be done in class:) Suppose you want to do inference on λ, the mean number of customers that arrive at a store per hour. Before you observe data, you believe that λ has a gamma distribution with $\alpha = 8$, $\beta = 2$. If you observe a total of 50 customers in 10 hours and assume the number of customers per hour has a Poisson distribution, what is your posterior density on λ.

Examples

Example 4 (to be done in class – D.S. Section 7.3 Exercises, Question 10:) Suppose that a random sample is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, and the prior distribution of θ is a normal distribution for which the standard deviation is 1. What is the smallest number of observations that must be included in the sample in order to reduce the standard deviation of the posterior distribution of θ to the value 0.1?

Examples

Example 5 (to be done in class:) Suppose that income per hour for white collar jobs in North Carolina has a normal distribution with unknown mean μ and variance 5. Prior to seeing the data, suppose I believe that μ is at least \$25 with 0.4 probability but at most \$27 with 0.8 probability. If I observe a sample mean of \$25 from a random sample of 500 workers. What is the posterior distribution of μ.

Bayesian Estimators

The result of a Bayesian inference is a posterior distribution over the entire parameter space. That distribution completely expresses your belief about the probabilities for all possible values of the parameter.

Often one needs to have a summary of that belief. Two standard choices are the mean of the posterior distribution and the median of the posterior distribution.

The posterior mean is your best one-number guess when your penalty for being wrong is proportional to $(\hat{\theta} - \theta)^2$, where θ is the parameter of interest. So large mistakes are heavily penalized.

The posterior median is your best one-number guess when your penalty for being wrong is proportional to $|\hat{\theta} - \theta|$. Here large mistakes are not so heavily penalized.

Recap

Today we covered:

- The difference between Bayesian and frequentist/classical paradigms
- Conjugacy
- Bayesian Estimators