Bayesian estimation

Chrysafis Vogiatzis

Lecture 19

Learning objectives

After this lecture, we will be able to:

- Use Bayesian estimation to find point estimators for unknown parameters.
- Propose new point estimators for unknown parameters based on prior information.

Motivation: Heads or Tails?

We flip a coin 10 times and we get 6 Heads and 4 Tails. Do you believe it is a fair coin? What does the method of moments and the maximum likelihood estimation method say about this situation?

Quick review

During these past two lectures, we discussed two methods to identify "good" estimators $\hat{\Theta}$ for a series of unknown parameters:

- **the method of moments**.
  1. Compute the moments of the population, calculated as $E\left[X^k\right]$.
  2. Compute the moments of the sample (empirical moments), calculated as $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.
  3. Equate the two and solve a system of equations for the unknown parameters.
- **maximum likelihood estimation**.
  1. Calculate the likelihood function as
  $$L(\theta) = f(X_1, \theta) \cdot f(X_2, \theta) \cdot \ldots \cdot f(X_n, \theta)$$
  2. Or the log-likelihood function as
  $$\ln(L(\theta)) = \ln(f(X_1, \theta)) + \ln(f(X_2, \theta)) + \ldots + \ln(f(X_n, \theta))$$
  3. Find the maximizer (usually found by setting the derivative per each unknown parameter equal to 0).

Both these methods have one thing in common: they require no prior information to work, but instead they base all of their observations on the obtained sample. What if I already know some more information about what is going on?

Bayesian estimation through an example

We begin in a slightly different way than usually. We begin with an example to help us build intuition! Assume I carry 3 coins with me:

1. One with both sides showing Heads.
2. One with both sides showing Tails.
3. One that is fair and has a side of Heads and a side of Tails.

Assume I randomly pick one coin and start flipping it. I report to you the number of tries (n) and the number of Heads (x). For example, I may tell you $n = 8$, $x = 5$ or $n = 2$, $x = 0$, and so on.

Flipping the coin: first take

I let you know that I flipped the coin three times and got Heads both times: $n = 3, x = 2$. What are the method of moments and the maximum likelihood estimators for p?

We will have $E[X] = p$ and $\frac{1}{3} \cdot (1+1+0) = \frac{2}{3}$, and equating will give $\hat{p} = \frac{2}{3}$. The likelihood function is $L(p) = p^2 \cdot (1-p)$, and maximizing will also give $\hat{p} = \frac{2}{3}$.

But... I carry three coins with me. Shouldn't I use this information somehow?

1. Can it be my "2-Heads" coin?
2. Can it be my "2-Tails" coin?
3. Does it have to be my "50-50" coin?

This is the key to realizing what Bayesian estimation brings to the table: extra information in the form of prior probabilities for the parameters that are unknown.

Bayesian estimation

We separate the discussion between discrete sets for the values the parameter can take (like in the previous example where I carried 3 distinct coins with me) and between continuous sets, where the parameter can be any real number in a range of values.

For discrete parameter values

Before describing the method, we provide some notation:

- **prior probabilities** ("priors"): the probability of seeing a certain parameter $P(\theta)$.
- **likelihood probabilities** ("likelihoods"): the likelihood of seeing an outcome *given* a certain parameter $P(X|\theta)$.
- **posterior probabilities** ("posteriors"): the multiplication of the two $P(\theta) \cdot P(X|\theta)$.

A quick note about the likelihoods: those are calculated in identical manner as the likelihood function in the maximum likelihood estimation method!
The Bayesian estimation method then states that:

"The higher the posterior probability,
the better the chance of having that parameter."

This is it! This is the whole method!

Flipping the coin: second take

Let us go back to the example where I carried three coins ("2-Heads"; "2-Tails"; and "50-50") and I picked one at random. After 3 tries, we got 2 Heads: $n = 3, x = 2$. Let's see what we have for these three distinct cases of $p = 1, p = 0, p = 0.5$:

- priors $P(p)$: probability of picking a certain coin, that is $P(p = 1) = P(p = 0) = P(p = 0.5) = \frac{1}{3}$.
- likelihoods $P(X = 2|p)$: likelihood function of seeing two Heads for each coin. For example, the likelihood function for the $p = 0.5$ coin with $x = 2$ Heads in $n = 3$ tries would be: $p^2 \cdot (1 - p) = 0.5^2 \cdot 0.5 = 0.125$.
- posteriors $P(p) \cdot P(X = 2|p)$: we will need to calculate this for each coin.

Let us put this in table format.

parameter p	prior $P(p)$	likelihood $P(X = 2\|p)$	posterior $P(p) \cdot P(X = 2\|p)$
0	$\frac{1}{3}$	$0^2 \cdot 1^1 = 0$	$\frac{1}{3} \cdot 0 = 0$
1	$\frac{1}{3}$	$1^2 \cdot 0^1 = 0$	$\frac{1}{3} \cdot 0 = 0$
0.5	$\frac{1}{3}$	$0.5^2 \cdot 0.5^1 = 0.125$	$\frac{1}{3} \cdot 0.125 = 0.04166\overline{6}$

The maximum value (and only non-zero probability!) is achieved for the "50-50" coin so it must be this!

See? It is pretty intuitive. Of course, we may complicate things by making the probability of picking a coin a little more general.

Flipping the coin: third take

I still have three types of coins on me. But, given that I am an adult that carries money wherever I go, I carry more actual coins ("50-50") than novelty coins ("2-Heads", "2-Tails"). More specifically, I carry 8 real coins and 1 of each novelty coin. I take a coin out and toss it twice and get two Heads! Which coin is it?

Let us try the table format again:

parameter p	prior $P(p)$	likelihood $P(X=2\|p)$	posterior $P(p)\cdot P(X=2\|p)$
0	$\frac{1}{8}$	$0^2=0$	$\frac{1}{8}\cdot 0=0$
1	$\frac{1}{8}$	$1^2=1$	$\frac{1}{8}\cdot 1=0.125$
0.5	$\frac{3}{4}$	$0.5^2=0.25$	$\frac{3}{4}\cdot 0.25=0.1875$

The maximum value is still achieved for a "50-50" coin, so we are inclined to think we picked one. Note how much closer the posteriors are, though..

It would actually take one more Heads to change our parameter estimation towards the "2-Heads" novelty coin! Why is that?

Normalizing may also be useful. Instead of looking at the posterior values as they are in the end, we may turn them into actual "%" values to help compare them. To normalize simply take each posterior and divide it by the summation of all posterior probabilities. For example, in our third take (see above) we would end up with probabilities:

- $P(p=0)=\frac{0}{0+0.125+0.1875}=0.$
- $P(p=1)=\frac{0.125}{0+0.125+0.1875}=0.4.$
- $P(p=0.5)=\frac{0.1875}{0+0.125+0.1875}=0.6.$

This helps us quantify our parameter estimation even more. There is a 40% chance we picked the "2-Heads" coin and a 60% chance we picked one of the "50-50" coins.
Let's work one more example before we move to the continuous case.

A computer vision system: third take

A machine learning algorithm for computer vision is trained to observe the first vehicle that passes from an intersection at or after 8am every day. Then, it reports the time from that vehicle to the next one again and again. We assume this time is exponentially distributed but with unknown λ.

- If the first vehicle of the day was a personal car, then $\lambda_1 = 1$ per minute.
- If the first vehicle of the day was a motorcycle, then $\lambda_2 = 1$ per 5 minutes.
- If the first vehicle was a truck, then $\lambda_3 = 1$ per 10 minutes.
- If the first vehicle was a bike, then $\lambda_4 = 1$ per 12 minutes.

We observe a sample of 5 times: $X_1 = 9$ minutes, $X_2 = 8.5$ minutes, $X_3 = 8$ minutes, $X_4 = 10.5$ minutes. What is the probability of each parameter λ?

First, we need to calculate the prior probabilities $P(\lambda)$. The first vehicle of the day is:

- a personal car with probability $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.732$ (why?),
- a motorcycle with probability $\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.146$,
- a truck with probability $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.073$,
- or a bike with probability $\frac{\lambda_4}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = 0.049$.

With these in hand, we calculate the likelihood functions as being $\lambda \cdot e^{-\lambda \cdot X_1} \cdot \lambda \cdot e^{-\lambda \cdot X_2} \cdot \ldots \cdot \lambda \cdot e^{-\lambda \cdot X_n}$ since we have an exponentially distributed random variable.
For example, if $\lambda = \lambda_1 = 1$, then we would have $1 \cdot e^{-1 \cdot 9} \cdot 1 \cdot e^{-1 \cdot 8.5} \cdot 1 \cdot e^{-1 \cdot 8} \cdot 1 \cdot e^{-1 \cdot 10.5} = e^{-36} = 2.32 \cdot 10^{-16}$.
Finally:

parameter λ	prior $P(\lambda)$	likelihood $P(X_1, X_2, X_3, X_4 \mid \lambda)$	posterior $P(\lambda) \cdot P(X_1, X_2, X_3, X_4 \mid \lambda)$
$\lambda_1 = 1$	0.732	$2.32 \cdot 10^{-16}$	$1.70 \cdot 10^{-16}$
$\lambda_2 = 0.2$	0.146	$1.19 \cdot 10^{-6}$	$1.74 \cdot 10^{-7}$
$\lambda_3 = 0.1$	0.073	$2.73 \cdot 10^{-6}$	$1.99 \cdot 10^{-7}$
$\lambda_4 = 0.06\overline{6}$	0.049	$1.79 \cdot 10^{-6}$	$8.77 \cdot 10^{-8}$

From the results it seems that the vehicle that first passed today is more likely a truck!

If we wanted to assign probability values to each type of vehicle, we would report:

- personal car: $\frac{6.24 \cdot 10^{-17}}{6.24 \cdot 10^{-17} + 1.43 \cdot 10^{-7} + 1.81 \cdot 10^{-7} + 8.18 \cdot 10^{-8}} = 1.54 \cdot 10^{-10} \approx 0.$
- motorcycle: 0.3527.
- truck: 0.4458.
- bike: 0.2015.

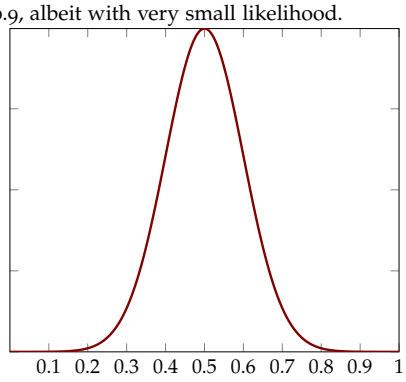
We can now move to the continuous case.

For continuous parameter values

Let us again begin with an example. The method is largely still the same; but the definitions of some of the items change slightly to accommodate the continuous nature of the unknown parameter(s).

Say we have a coin that is made with the goal of being fair; that is, "50-50". But, materials fail and get deposited more on one side than the other resulting in different compositions for the probability of Heads and Tails. Say, in the end, the probability of Heads is normally distributed with $\mathcal{N}(0.5, 0.01)$, that is a mean of $\mu = 0.5$ and a variance $\sigma^2 = 0.01 \implies \sigma = 0.1$. Visually, we would get the distribution of Figure 1

Figure 1: The distribution of the probability of getting Heads in the continuous version of the problem. We see how $p = 0.5$ is more likely, but we can get values as low as 0.1, 0.2, or as high as 0.8, 0.9, albeit with very small likelihood.

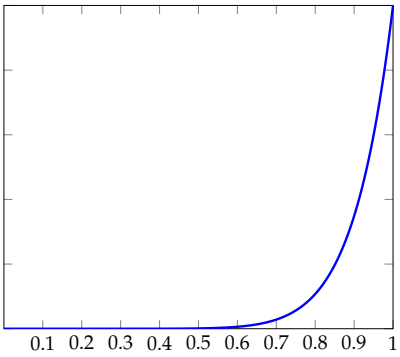


Now that we know this, say we tossed a coin 10 times, and got 10 straight times Heads! Recall that both the method of moments and the maximum likelihood estimation method would simply assume that the coin has $p = 1$ and proceed.

Getting 10 Heads in 10 tosses would be highly improbable for a coin that is "50-50", but it could mean that I have a biased coin towards Heads. So, what should our estimate be?

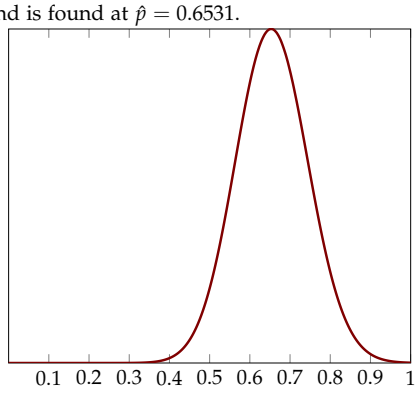
First, calculate the likelihood function, the way we did during the maximum likelihood estimation calculations. In this case, it would be $L(p) = p^{10}$. Let's plot that (see Figure 2).

Figure 2: The likelihood function of getting 10 Heads after tossing a coin 10 times. It is maximized at $p = 1$, which would then be our maximum likelihood estimator.



In Bayesian estimation for discrete-valued parameters earlier, we calculated $P(\theta)$ (priors) with $P(X|\theta)$ (likelihoods) to obtain a series of posteriors that we would compare. In the continuous version, we calculate $f(\theta)$ (prior distribution) with $L(\theta)$ (likelihood function) to obtain a posterior distribution that we would then find the maximizer at! Confused? Let's look at this visually again in Figure 3.

Figure 3: The posterior distribution, found by multiplying $f(\theta)$ (the pdf of the normal distribution $\mathcal{N}(0.5, 0.01)$) with the likelihood function $L(\theta)$. The maximizer here is the Bayesian estimator and is found at $\hat{p} = 0.6531$.



Let us define the notation for the method then:

- **prior distribution**: the distribution of the real-valued and continuous parameter θ, $f(\theta)$.
- **likelihood function**: the likelihood function, built just as in the maximum likelihood estimation metho, $L(\theta)$.

- **posterior distribution**: the multiplication of the two functions $f(\theta) \cdot L(\theta)$.

The Bayesian estimation method for continuous parameters states that:

> "The Bayesian estimator is found by maximizing the posterior distribution."

And, yes! This sums it up. Let us view one example from beginning to end using the method.

Mortality risk

We call mortality risk of a hospital the probability of death occurring for any patient admitted to the hospital. The mortality risk in US hospitals is in general **exponentially distributed** with a mean at 1.5% (that is, $\lambda = \frac{1}{1.5\%}$). You have been observing a hospital and have seen 25 deaths in the first 150 patient admissions. What is the Bayesian estimator for the true mortality rate of the hospital?

Right away, distinguish between two items:

1. the prior distribution that we believe the mortality risk to be distributed as (exponential)
2. the mortality rate itself is a Bernoulli random variable (p and $1 - p$); in our case, we have a sample we have collected (25 deaths in 150 admissions) to help us estimate p.

So, let us start collecting what we need one-by-one.

Prior distribution:

$$f(p) = \frac{1}{1.5} e^{-\frac{1}{1.5}p}.$$

Likelihood function:

$$L(p) = p^{25} \cdot (1-p)^{125}.$$

Posterior distribution:

$$f(p) \cdot L(p) = \frac{1}{1.5} e^{-\frac{1}{1.5}p} \cdot p^{25} \cdot (1-p)^{125}.$$

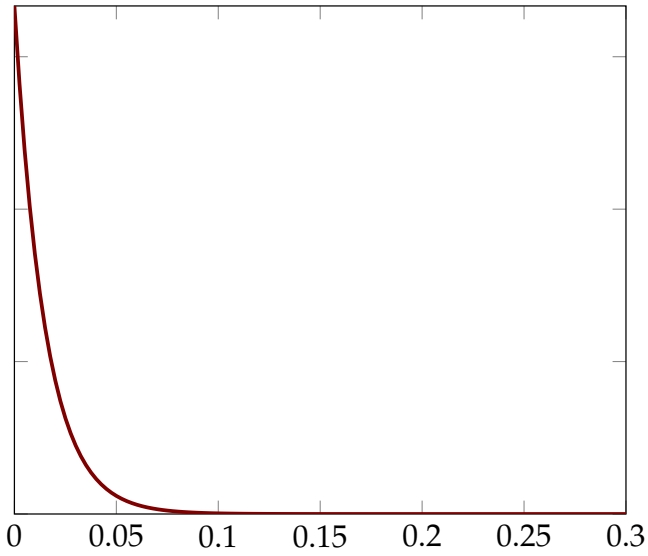
Mortality risk (cont'd)

To maximize, get the derivative and equate to o to get:

$$\frac{\partial f(p)\cdot L(p)}{\partial p} = \frac{4}{9}e^{-\frac{2}{3}p}\,(p-1)^{124}\,p^{24}\,((p-226)\,p+37.5) = 0 \implies$$
$$\implies ((p-226)\,p+37.5) = 0 \implies p = 0.16605.$$

The maximizer is, then, at $\hat{p} = 0.16605$.

If we want to, we can see the same result visually. First, plot our prior beliefs/distribution:



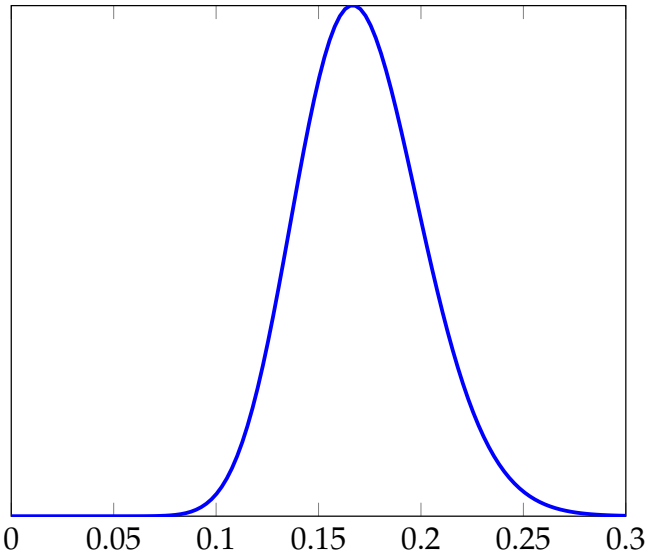
Then, plot our likelihood function based on the sample collected:

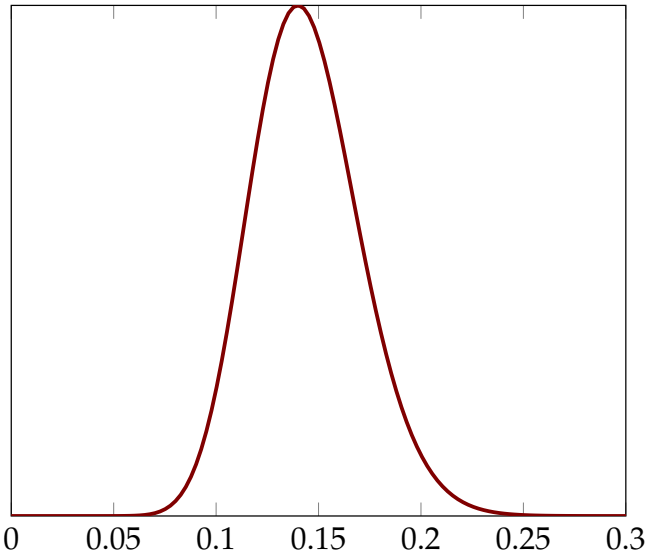
And finally plot the posterior distribution, and check that the maximizer is indeed at $\hat{p} = 0.16605$:

One last example

Let us work on one more example for continuously distributed parameters. Assume we have a population distribution with pdf $f(x) = (\theta+1)\,x^{\theta}$, for $0 \leq x \leq 1$. Moreover, assume that θ is not totally random, but is instead distributed with pdf $f(\theta) = \frac{1}{12}\,(3-\theta)$, defined over $-2 \leq \theta \leq 2$. Assume we have collected a sample of $X_1 = 0.9, X_2 = 0.89, X_3 = 0.76, X_4 = 0.96$. What is the Bayesian estimator for θ?

You may inspect the solution visually as a homework assignment. Algebraically, though, we would multiply the prior distribution





($f(\theta)$) with the likelihood function ($L(\theta)$) to obtain the posterior distribution. In mathematical terms:

$$
\begin{aligned}
f(\theta) &= \frac{1}{12}\,(3-\theta) \\
L(\theta) &= (\theta+1)\,X_1^{\theta}\cdot(\theta+1)\,X_2^{\theta}\cdot(\theta+1)\,X_3^{\theta}\cdot(\theta+1)\,X_4^{\theta} = \\
&= (\theta+1)^4\,(X_1\cdot X_2\cdot X_3\cdot X_4)^{\theta} = (\theta+1)^4\,0.5844096^{\theta} \\
f(\theta)\cdot L(\theta) &= \frac{1}{12}\,(3-\theta)\cdot(\theta+1)^4\,0.5844096^{\theta}
\end{aligned}
$$

Getting the derivative of the posterior, and equating it to 0, we get:

$$\frac{\partial f(\theta)L(\theta)}{\partial \theta} = 0 \implies 0.0447628 \cdot 0.58441^{\theta}(1+\theta)^3(17.4783 + \theta(-11.3083 + \theta)) = 0.$$

We get three possible solution: $\theta = -1$, $\theta = 1.85$, or $\theta = 9.46$. We note that the last one cannot happen as θ is between -2 and 2. Between the two remaining possible solutions, we compare their posterior distribution values:

- $f(-1) \cdot L(-1) = \frac{1}{12}(3 - (-1)) \cdot ((-1) + 1)^4 \, 0.5844096^{-1} = 0.$
- $f(1.85) \cdot L(1.85) = \frac{1}{12}(3 - 1.85) \cdot (1.85 + 1)^4 \, 0.5844096^{1.85} = 2.34.$

Hence, $\hat{\theta} = 1.85$ is the maximizer and the Bayesian estimator.

I lied.. Here is the visual version of the posterior also. It is clear that 1.85 is indeed the maximizer!