STAT 3503/8109 Lecture 11 Notes

Edoardo Airoldi

*Scribe: Srikar Katta**

Fall 2020: November 9, 2020

1 Introduction

Previously, we discussed methods of estimating unknown constants in a frequentist setting using maximum likelihood estimation and method of moments. We considered a scenario in which we track salary outcomes over time using a sequence of independent Bernoulli trials; then, we added complexity by including a covariate that indicated whether students had previously taken this class or not. Then, instead of considering only the estimation of unknown constants, we wanted to utilize prior information. One way to do this was by collection all of the outcomes of previous students. Using that, we could fit this model to a larger data set of observations. Another approach is to utilize recapitulations of prior outcomes and describe these outcomes with a *prior* distribution (i.e., a probability distribution that summarizes historical beliefs). First, we *calibrate* (i.e., set) the prior distribution by estimating the unknown constants underlying the prior distribution using past data and maximum likelihood estimation. Then, we computed the *posterior* distribution, the distribution that is a result of updating the prior. Then, using *Bayesian Analysis* (i.e., utilizing Bayes' Theorem), we can estimate the summaries of the posterior distribution. Here, we extend this discussion by introducing formal notation and demonstrate techniques to calculate the posterior.

*Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

1.1 Bayesian Analysis with Notation

Assume we have some observed data and an unobserved constant θ: $x_1...x_n \overset{\text{iid}}{\sim} f(X|\theta)$. Using maximum likelihood estimation (MLE), we can identify the argument that maximizes the likelihood (and equivalently the log likelihood). So,

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_\theta logL(\theta) \\ &= \arg\max_\theta f(x_1...x_n|\theta).\end{aligned}$$

Now, suppose we have observed historical data $y_1...y_m$. We can calibrate the prior $y_1...y_n \overset{\text{iid}}{\sim} \xi(y|a,b)$ with unknown constants a, b. Using MLE with all the previous data, we find $\hat{a}_{MLE}, \hat{b}_{MLE} = \arg\max_{a,b} L(a,b) = \arg\max_{a,b} \xi(y_1...y_m|a,b)$. Now, we can find a posterior distribution $f(\theta|x_1...x_n, \hat{a}_{MLE}, \hat{b}_{MLE})$. Then, we find our estimates for θ using the posterior mean

$$\hat{\theta}_{PM} = \mathbb{E}[\theta|x_1...x_n, \hat{a}_{MLE}, \hat{b}_{MLE}]$$

or the maximum a posteriori

$$\hat{\theta}_{MAP} = \arg\max_\theta f(\theta|x_1...x_n, \hat{a}_{MLE}, \hat{b}_{MLE}).$$

Finding the posterior distribution is essential to Bayesian analysis and will motivate our discussions today.

2 The Posterior

Assume we observe $x_1...x_n \overset{\text{iid}}{\sim} f(x|\theta)$ and $\theta \sim \xi(\theta|a,b)$. This is known as a *two-level model*. First, we find the likelihood:

$$\begin{aligned}Likelihood &= \mathbb{P}(\text{observed random variables}|\text{ constants}) \\ &= f(x_1...x_n|a,b).\end{aligned}$$

Now, we can define the prior as the the probability of the latent random variable of interest given unknown constants,

$$prior = \xi(\theta|a,b).$$

Now, recall the rules of conditional probability: given two sets A and B, $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\&B)}{P(B)}$. And from Bayes' Theorem, we know that $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)}{P(A)}$ as well. Applying that idea here, the posterior is then,

$$\begin{aligned} \textit{posterior} &= \mathbb{P}(\text{ latent random variables } | \text{ unknown constants, observed random variables}) \\ &= \frac{\mathbb{P}(\text{observed and latent random variables } | \text{ constants})}{\mathbb{P}(\text{observed random variables } | \text{ constants})}. \end{aligned}$$

So, in our situation,

$$\begin{aligned} \textit{posterior} &= \mathbb{P}(\theta|a, b, x_1...x_n) \\ &= \frac{\mathbb{P}(x_1...x_n, \theta|a, b)}{\mathbb{P}(x_1...x_n|a, b)} \\ &= \frac{\mathbb{P}(x_1...x_n|\theta)\mathbb{P}(\theta|a, b)}{\mathbb{P}(x_1...x_n|a, b)}, \text{ from Bayes' Theorem} \end{aligned}$$

Notice, this denominator is essentially the proper likelihood while the numerator is the complete likelihood. Since the proper likelihood is just the complete likelihood with the latent random variables integrated out, one way of estimating the posterior is to compute the following:

$$\begin{aligned} \textit{posterior} &= \frac{\text{complete likelihood}}{\text{proper likelihood}} \\ &= \frac{\text{complete likelihood}}{\int_{\text{latent rv}} \text{complete likelihood } d\text{latent rv}}. \end{aligned}$$

If we can compute the likelihood proper, then there is no issue and we can proceed. However, this may be difficult to identify. To overcome this, we can use a "trick" to use only the numerator in our calculations.

3 Full Calculation of the Posterior

Here, we will detail a method of calculating the posterior and finding the best Bayesian estimate of the unknown, utilizing the complete and proper likelihoods. First, consider a simple example with only one data point.

Example 3.1. *Assume we have one observed data point $x_1 \sim Bernoulli(\theta)$ with unknown*

parameter $\theta \sim Uniform[0,1] = Beta(a,b)$ *where* $a = b = 1$ *is unknown. Estimate* $\hat{\theta}_{PM}$*, the posterior mean of* θ*.*

Solution: In the previous notation, $Uniform[0,1]$ was equivalent to saying $\xi(\theta)$. Now, we can write the complete likelihood:

$$\begin{aligned} L(\theta)^{complete} &= \mathbb{P}(x_1, \theta) \\ &= \mathbb{P}(x_1|\theta)\xi(\theta) \\ &= \theta^{x_1}(1-\theta)^{1-x_1}, \text{ the joint probability of } x_1 \text{ and } \theta. \end{aligned}$$

Then, we can find the likelihood proper easily by integrating out θ, so

$$\begin{aligned} L(\theta) &= \int_\theta \theta^{x_i}(1-\theta)^{1-x_1} d\theta \\ &= \int_0^1 \theta^{x_i}(1-\theta)^{1-x_1} d\theta. \end{aligned}$$

In any probability calculation with integrals, we hope to bring the inside of the integral to a known probability distribution, because if the inside is a probability distribution, then this integral must equal 1 and our calculations simplify significantly. So, notice that $\theta \sim Beta(a,b)$ implies that

$$f(\theta|a,b) = \theta^{a-1}(1-\theta)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)},$$

where $\Gamma(\cdot)$ is a characterization function for the Beta distribution (the exact formula for which is not necessary here). Notice, $x_1 = x_1 + 0 = x_1 + 1 - 1$. So, I will take advantage of this and rewrite $L(\theta)$:

$$\begin{aligned} L(\theta)^{proper} &= \int_0^1 \theta^{x_1}(1-\theta)^{1-x_1} d\theta \\ &= \int_0^1 \theta^{x_1+1-1}(1-\theta)^{1-(x_1+1-1)} d\theta \\ &= \int_0^1 \theta^{x_1+1-1}(1-\theta)^{1-x_1-1+1} d\theta \\ &= \int_0^1 \theta^{x_1+1-1}(1-\theta)^{2-x_1-1} d\theta \\ &= \int_0^1 \theta^{x_1+1-1}(1-\theta)^{2-x_1-1} \cdot 1 d\theta \end{aligned}$$

$$
\begin{aligned}
&= \int_0^1 \theta^{x_1+1-1}(1-\theta)^{2-x_1-1} \cdot \frac{\Gamma(x_1+1+2-x_1)}{\Gamma(x_1+1)\Gamma(2-x_1)} \frac{\Gamma(x_1+1)\Gamma(2-x_1)}{\Gamma(x_1+1+2-x_1)} d\theta \\
&= \int_0^1 \theta^{x_1+1-1}(1-\theta)^{2-x_1-1} \cdot \frac{\Gamma(3)}{\Gamma(x_1+1)\Gamma(2-x_1)} \frac{\Gamma(x_1+1)\Gamma(2-x_1)}{\Gamma(3)} d\theta \\
&= \frac{\Gamma(x_1+1)\Gamma(2-x_1)}{\Gamma(3)} \int_0^1 \theta^{x_1+1-1}(1-\theta)^{2-x_1-1} \cdot \frac{\Gamma(3)}{\Gamma(x_1+1)\Gamma(2-x_1)} d\theta.
\end{aligned}
$$

Since $\frac{\Gamma(x_1+1)\Gamma(2-x_1)}{\Gamma(3)}$ does not depend on θ, we were able to pull it outside of the integral. Now, inside the integral, we have a full specified Beta distribution, $Beta(x_1+1, 2-x_1)$. So,

$$
\begin{aligned}
L(\theta)^{proper} &= \frac{\Gamma(x_1+1)\Gamma(2-x_1)}{\Gamma(3)} \int_0^1 \theta^{x_1+1-1}(1-\theta)^{2-x_1-1} \cdot \frac{\Gamma(3)}{\Gamma(x_1+1)\Gamma(2-x_1)} d\theta \\
&= \frac{\Gamma(x_1+1)\Gamma(2-x_1)}{\Gamma(3)} \cdot 1.
\end{aligned}
$$

Now, because we have both the complete and proper likelihoods, we can compute the posterior for θ given x_1:

$$
\begin{aligned}
f(\theta|x_1) &= \frac{\theta^{x_1}(1-\theta)^{1-x_1}}{\frac{\Gamma(x_1+1)\Gamma(2-x_1)}{\Gamma(3)}} \\
&= \theta^{x_1}(1-\theta)^{1-x_1} \frac{\Gamma(3)}{\Gamma(x_1+1)\Gamma(2-x_1)} \\
&= \theta^{x_1+1-1}(1-\theta)^{1-(x_1+1-1)} \frac{\Gamma(3)}{\Gamma(x_1+1)\Gamma(2-x_1)} \\
&= \theta^{x_1+1-1}(1-\theta)^{2-x_1} \frac{\Gamma(3)}{\Gamma(x_1+1)\Gamma(2-x_1)} \\
&= Beta(x_1+1, 2-x_1),
\end{aligned}
$$

so the posterior distribution of $\theta \sim Uniform[a,b]$ is $Beta(x_1+1, 2-x_1)$.

Now, we can compute the posterior mean,

$$
\begin{aligned}
\hat{\theta}_{PM} &= \mathbb{E}[\theta|x_1] \\
&= \frac{x_1+1}{3},
\end{aligned}
$$

so $\frac{x_1+1}{3}$ is the best Bayesian estimate for θ. □

Now that we have seen a simple case, we can move to greater abstraction and demonstrate this calculation with multiple data points.

Example 3.2. *Suppose* $x_1 ... x_n \overset{\text{iid}}{\sim} Bernoulli(\theta)$ *where* $\theta \sim Beta(a,b)$. *Assume* $x_1 ... x_n, a, b$

are all observed while θ is not.

Solution: First we can identify the likelihood, prior, and posterior in notation for this situation:

- The likelihood proper is $f(x_1...x_n|a,b)$
- The prior is $f(\theta|a,b)$
- The posterior is $f(\theta|x_1...x_n,a,b)$.

The difficult part here will be computing the posterior, but we can still detail how. First, from Bayes' Theorem, we know that

$$f(\theta|x_1...x_n,a,b) = \frac{f(x_1...x_n|\theta)f(\theta|a,b)}{f(x_1...x_n|a,b)},$$

and notice that the denominator is the proper likelihood, which is the same as the complete likelihood with the latent variables integrated out. So,

$$\begin{aligned} f(\theta|x_1...x_n,a,b) &= \frac{f(x_1...x_n|\theta)f(\theta|a,b)}{f(x_1...x_n|a,b)} \\ &= \frac{f(x_1...x_n|\theta)f(\theta|a,b)}{\int_\theta f(x_1...x_n,\theta|a,b)d\theta}. \end{aligned}$$

Now, from our IID assumption, we know that $f(x_1...x_n|\theta) = f(x_1|\theta)...f(x_n|\theta)$. So,

$$\begin{aligned} f(\theta|x_1...x_n,a,b) &= \frac{f(x_1...x_n|\theta)f(\theta|a,b)}{\int_\theta f(x_1...x_n,\theta|a,b)d\theta} \\ &= \frac{\prod_{i=1}^n f(x_i|\theta)f(\theta|a,b)}{\int_\theta \prod_{i=1}^n f(x_i,\theta|a,b)d\theta}, \end{aligned}$$

and then we can implement the same trick to identify the posterior distribution. In fact, we will still identify a Beta distribution as the posterior. □

There is one important note to consider: we had a Bernoulli model and Beta prior in both situations, and it yielded a Beta posterior. This is no coincidence. In fact, every time we have this set up, we will find a Beta posterior. Because of this, we say that the Bernoulli model is "conjugate" to the Beta prior.

The first layer of modeling is always done using conjugate priors because no matter what your model is, by utilizing a conjugate prior, you know that – by definition – the posterior

will also be in the same family as the prior. So, all that really needs to happen is the update step; the actual calculation of the posterior has become incredibly simplified. In our example with a prior $Uniform[0,1] = Beta(0,1)$, we used a Bernoulli model. So the posterior was $Beta(x_1+1, 2-x_1)$. So, in general, the Beta posterior will be $Beta(1+\sum_i x_i, 1+\sum_i(1-x_i))$, where the first parameter represents the number of successes and the second represents the number of failures.

4 Partial Calculation of The Posterior

We detailed a way of calculating a posterior using the definitions of likelihood. However, we can bypass the need to use integrals altogether, and we will outline how to do that in the subsequent examples.

Example 4.1. *Suppose* $x_1...x_n \sim Bernoulli(\theta)$ *where* $\theta \sim Beta(92, 100)$. *Find* $f(\theta|x_1...x_n, 92, 100)$, *the posterior and the Bayesian estimator.*

Solution: We begin this by outlining the posterior distribution's structure:

$$f(\theta|x_1...x_n, 92, 1000) = \frac{f(x_1...x_n|\theta)f(\theta|a,b)}{f(x_1...x_n|92,100)}.$$

Here, we will take advantage of the fact that θ does not exist in the denominator of the posterior calculations, so we can simply consider what the explicit posterior calculation is proportional to identify the posterior distribution itself. So,

$$\begin{aligned}
f(\theta|x_1...x_n, 92, 1000) &= \frac{f(x_1...x_n|\theta)f(\theta|a,b)}{f(x_1...x_n|92,100)} \\
&\propto f(x_1...x_n|\theta)f(\theta|a,b) \\
&= f(x_1|\theta)...f(x_n|\theta)f(\theta|a,b), \text{ because of the IID assumption} \\
&= \prod_{i=1}^{n} f(x_i|\theta)f(\theta|a,b) \\
&= \prod_{i=1}^{n} \left(\theta^{x_i}(1-\theta)^{1-x_i}\right)\left(\theta^{(92-1)}(1-\theta)^{(100-1)}\frac{\Gamma(100+92)}{\Gamma(100)\Gamma(92)}\right) \\
&= \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}\theta^{91}(1-\theta)^{99}\frac{\Gamma(192)}{\Gamma(100)\Gamma(92)}.
\end{aligned}$$

Again, since $\frac{\Gamma(192)}{\Gamma(100)\Gamma(92)}$ does not contain θ, it is not essential to our posterior distribution calculation, so we can take advantage of this proportionality idea. So,

$$\begin{aligned}
f(\theta|x_1...x_n, 92, 1000) &\propto \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}\theta^{91}(1-\theta)^{99}\frac{\Gamma(192)}{\Gamma(100)\Gamma(92)} \\
&\propto \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}\theta^{91}(1-\theta)^{99} \\
&= \theta^{91+\sum x_i}(1-\theta)^{99+\sum(1-x_i)} \\
&= \theta^{91+\sum x_i}(1-\theta)^{99+n-\sum x_i} \\
&= \theta^{1-1+91+\sum x_i}(1-\theta)^{1-1+99+n-\sum x_i} \\
&= \theta^{1-1+91+\sum x_i}(1-\theta)^{1-1+99+n-\sum x_i} \\
&= \theta^{92+\sum x_i-1}(1-\theta)^{100+n-\sum x_i-1}.
\end{aligned}$$

Now, we can again take advantage of the proportionality concept not to remove quantities but rather to *add* quantities. Namely, we multiply our posterior calculation so far by $\frac{\Gamma(92+\sum x_i+100+n-\sum x_i)}{\Gamma(92+\sum x_i)\Gamma(100+n-\sum x_i)}$ because this is the probability distribution function for $Beta(92+\sum x_i, 100+n-\sum x_i)$. Now, using this, we can again find the Bayesian estimator for θ using the posterior mean. So,

$$\begin{aligned}
\hat{\theta}_{PM} &= \mathbb{E}[\theta|x_1...x_n, 92, 100] \\
&= \frac{92+\sum x_i}{92+\sum x_i+100+n-\sum x_i} \\
&= \frac{92+\sum x_i}{192+n}.
\end{aligned}$$

□

There is one major note of caution. Some people write that the "posterior is proportional to the prior times the likelihood" because it involves two models. In one, you only have $x_1...x_n \overset{\text{iid}}{\sim} f(x|\theta)$ in model 1; in model 2, $x_1...x_n \overset{\text{iid}}{\sim} f(x|\theta)$ *and* $\theta \sim f(\theta|a,b)$. The correct statement would be that the posterior in model 2 is proportional to the prior in model 2 times the likelihood in model 1. So people who make the naive statement confuse the fact that the likelihood is the likelihood in model 1, not model 2. It is also possible that the likelihood in the product that people claim to be proportional to your prior is also not model 1, so this claim breaks under such situations and careful attention is needed.

5 Conclusion

The big picture idea here rests on understanding the sources of variation in the analyses. In the Frequentist perspective, our observed data has empirical variation and the model has theoretical variation with unknown *constants* that describe our model. In the Bayesian setting, our observed data has empirical variation and the model has theoretical variation. Now additionally, we have a second layer in which the information summarizing our model also has its own theoretical variation. These differences can be seen using "E.V.V.E," a way to find $\mathbb{V}(x)$:

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}_{x|\theta}(x|\theta)] + \mathbb{V}_{\theta}[\mathbb{E}_{x|\theta}(x|\theta)].$$

In the Frequentist setting, $\mathbb{E}_{x|\theta}(x|\theta)$ is a constant and does not have variability, so $\mathbb{V}[X] = \mathbb{E}[\mathbb{V}_{x|\theta}(x|\theta)]$. Essentially, the likelihood represents the empirical variation since the likelihood and model are the same in Frequentist outlook. However, in the Bayesian setting, $\mathbb{V}_{\theta}[\mathbb{E}_{x|\theta}(x|\theta)]$ is no longer a constant and has its own variability from the second layer. The model goes beyond the likelihood now.