STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 4

Quick review

- The three-step procedure of *all* Bayesian inference.
- We have seen two examples—political poll (binomial) and measure pollutant (normal).
- In these analysis, we used a “trick” to figure out what the posterior distribution is without carrying out the integration.

When does this “trick” work?

Political poll revisited

- Instead of a Beta(α, β) prior for θ, what if we want to choose $\xi(\theta) \propto e^{-(\theta-0.5)^4}\mathbf{1}_{0<\theta<1}$?
- Bayes' Theorem still applies.

$$\begin{aligned}
\xi(\theta|x) &\propto f(x|\theta)\xi(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot e^{-(\theta-0.5)^4} \\
&\propto \theta^x(1-\theta)^{n-x}e^{-(\theta-0.5)^4} \quad \text{for } 0 < \theta < 1.
\end{aligned}$$

This is not a standard distribution, and the normalizing constant must be evaluated through integration.

$$\xi(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}e^{-(\theta-0.5)^4}}{\int_0^1 u^x(1-u)^{n-x}e^{-(u-0.5)^4}du}.$$

This is *perfectly* valid. In fact if one has good reason to believe that $e^{-(\theta-0.5)^4}$ is the best reflection of prior knowledge then one can certainly choose this density as the prior.

- In the old days, this created a serious computational challenge.
- It is very difficult to calculate the normalizing constant, especially when θ is multi-dimensional.
- So it is desirable to use prior distributions that lead to simple posterior distributions, to give quick solutions.
- This is not as much a problem today due to the development of new methodology and faster computers. For simple problems, however, simple solutions may still be desirable.

So what is so special about our earlier examples that made them particularly neat?

Conjugate families

- Let Ψ be a family of probability distributions.
- Let $f(x|\theta)$ be the p.d.f of each data point conditional on the parameter θ.

If no matter which member in Ψ we choose as the prior distribution $\xi(\theta)$, the posterior distribution given an i.i.d. sample $X_1, X_2, \ldots, X_n$, $\xi(\theta|x_1, x_2, \ldots, x_n)$ is also a member of Ψ,

- then the family Ψ is said to be *conjugate* to the distribution $f(x|\theta)$.

Remark: Conjugacy by itself is not a useful concept.

- E.g. let Ψ be the family of all probability distributions.
- It is useful only when Ψ is a small enough family so that its members share interesting distributional properties.

Examples

- In the political poll example, the Beta(α, β) family is conjugate to the binomial distributions.
- In the air pollutant example, the Normal(μ, σ^2) family is conjugate to Normal(θ, τ^2) distributions with known τ^2.

Two other commonly used conjugate families

- Poisson-Gamma conjugacy.
- Exponential-Gamma conjugacy.

Number of phone calls to a customer service line

A company is deciding whether it should expand its customer service division and therefore wants to have an estimate of the average number of phone calls, denoted by θ, it receives between 9am-5pm.

- The data are the number of phone calls received between 9am-5pm on n different days—$X_1, X_2, \ldots, X_n$.
- *What assummptions are already involved? Are they reasonable?*

How to carry out a Bayesian inference on θ.

Again, we need the two pieces to feed into Bayes' Theorem

1. How do we model the distribution of the number of calls received on each day?
    - Given θ, one can model the number of calls received on different days as i.i.d. Poisson(θ) random variables.

2. What prior distribution $\xi(\theta)$ can we choose to characterize our *a priori* knowledge about θ.
    - A useful family for this purpose is the Gamma(α, β) family.

The Gamma(α, β) distribution

- Its pdf is

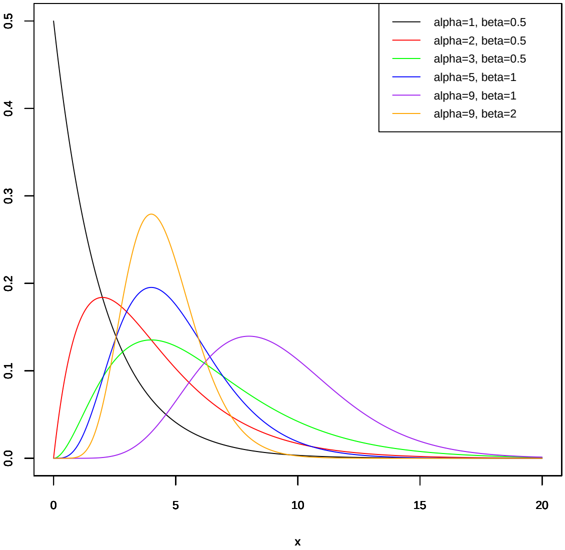
$$\begin{aligned} f(x|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

- Its mean and variance are

$$E[X] = \frac{\alpha}{\beta} \quad \text{and} \quad \mathrm{Var}[X] = \frac{\alpha}{\beta^2} = \frac{E[X]}{\beta}$$

- So for the same mean larger β gives smaller variance.
- We can use its relationship to the *exponential distribution* to remember these.

Gamma(α, β) p.d.fs



Applying Bayes' Theorem

$$\begin{aligned}
\xi(\theta|x_1,x_2,\ldots,x_n) &\propto f(x_1,x_2,\ldots,x_n|\theta)\xi(\theta) \\
&= \prod_{i=1}^{n} f(x_i|\theta)\cdot\xi(\theta) \\
&= \prod_{i=1}^{n} \frac{\theta^{x_i}e^{-\theta}}{x_i!}\cdot\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\
&\propto \theta^{\alpha+\sum_{i=1}^{n}x_i-1}e^{-(\beta+n)\theta}.
\end{aligned}$$

- This is exactly the variable part of a *Gamma*$(\alpha+\sum_{i=1}^{n}x_i,\beta+n)$ distribution.
- So $\xi(\theta|x_1,x_2,\ldots,x_n)$ must be this distribution.

The posterior expectation of θ is

$$\begin{aligned} E(\theta|x_1,x_2,\ldots,x_n) &= \frac{\alpha+\sum_{i=1}^n x_i}{\beta+n} \\ &= \left(\frac{\beta}{\beta+n}\right)\left(\frac{\alpha}{\beta}\right)+\left(\frac{n}{\beta+n}\right)\bar{x} \\ &= \left(\frac{\beta}{\beta+n}\right)\mu_\theta+\left(\frac{n}{\beta+n}\right)\bar{x} \\ &:= \mu_{\theta|x_1,x_2,\ldots,x_n}. \end{aligned}$$

The posterior variance is

$$\mathrm{Var}(\theta|x_1,x_2,\ldots,x_n) = \frac{\mu_{\theta|x_1,x_2,\ldots,x_n}}{\beta+n}.$$

- β measures the amount of information in the prior, while n measures the amount of information in the data.

Exponential-Gamma conjugacy

- Exercise: Show that the Gamma(α, β) family is also conjugate to the exponential distribution.
- The textbook covers that in Example 7.3.11, Theorem 7.3.4, and Example 7.3.12.
- Exercise 23 and 24 in Chapter 7 gives the most general results about common conjugate families that include all the cases we have seen. (The so-called *exponential* families. Don't confuse it with the family of exponential distributions.)

Posterior distribution as the ultimate inference goal

- The posterior distribution summarizes all information about the state of nature or parameter θ, *given* the data.
- This is the ideal goal of inference—all statistical questions regarding θ can be answered with this posterior distribution.
- Let us next look at one useful example—the estimation problem.

Point estimation

- A very common statistical problem is to "guess" the value of a parameter θ *based on observed data* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.
- *Functions of the data* that are used for guessing the values of a parameter are called *estimators* for the parameter. Common notations: $\hat{\theta}(\mathbf{X})$, $\delta(\mathbf{X})$, etc.
- If the observed data is $\mathbf{X} = \mathbf{x}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the realized value of an estimator $\delta(\mathbf{X})$ is $\delta(\mathbf{x})$, which is called an *estimate*.

- In other words, *estimators* are rules that specify how to “guess” for the parameter based on the data. So they are *functions of the data*.
- *Estimates* are the specific guesses of the parameter generated after observing the data according to the rules. That is, the are the corresponding *functions evaluated at the actual observed data*.

How to make “good” estimates/estimators?

- What is a criterion for *good* estimators/estimates?
- A good estimator should be such that the estimate and the actual parameter θ are *“likely to be close”*.

What does “likely to be close” mean?

1. The Bayesian view (the “*after-the-experiment*” view):
   - Both the parameter θ and data $\mathbf{X}$ are random variables.
   - *After* we have observed the data $\mathbf{X} = \mathbf{x}$, only θ is random, and its distribution is the posterior distribution $\xi(\theta|\mathbf{x})$.
   - In this case, we want to pick an estimate $\delta(\mathbf{x})$ such that *a posteriori* the parameter θ, which is random, will likely take values close to the estimate $\delta(\mathbf{x})$.

Note that here the parameter is random while the estimate $\delta(\mathbf{x})$ is a fixed number given the observed data $\mathbf{X} = \mathbf{x}$.

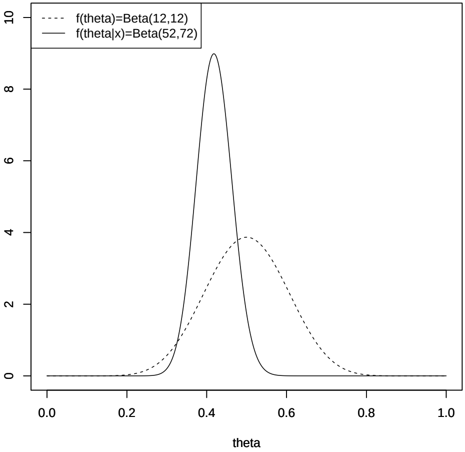
What does “likely to be close” mean?

- Later we will study the sampling view (the “*before-the-experiment*” view).

The Bayesian estimation problem

- Come back to the political poll example.
- With a Beta(α, β) prior on θ, after observing $X = x$, the posterior distribution of θ is Beta($\alpha + x, \beta + n - x$).
- What would be a good estimate for θ based on this posterior distribution?

Example: Under the *Beta*$(12, 12)$ prior



- Given $X = 40$, the (posterior) distribution of θ is $\text{Beta}(52, 72)$.
- Which value will you pick as a guess of θ?
- How about the mean, the median, or the mode of the posterior distribution?

- For example, if we choose the mean as the estimate. With $\alpha = 12$ and $\beta = 12$, given $X = 40$, this estimate is

$$\frac{52}{52+72} = \frac{13}{31}.$$

- If we had observed $X = 50$ instead of $X = 40$, then we would have had a different posterior distribution, namely Beta(62,62) distribution.
- The estimate $\delta(50)$ would instead be

$$\frac{62}{62+62} = \frac{1}{2}.$$

Our first estimator based on the posterior distribution

- We choose the estimate depending on the value of the observed data x.
- More generally, for any observed $X = x$, we can estimate θ by

$$E(\theta|x) = \frac{\alpha + x}{\alpha + x + \beta + (n - x)} = \frac{\alpha + x}{\alpha + \beta + n}.$$

- What is the corresponding *estimator*?
  - That is, if we repeat the experiment, what is function that maps the data X to our estimate?

$$\delta(X) = E(\theta|X) = \frac{\alpha + X}{\alpha + \beta + n}.$$

- What is the posterior mode estimate/estimator?