STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 5

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?
- 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 $\mathbf{X} = \mathbf{x}$.

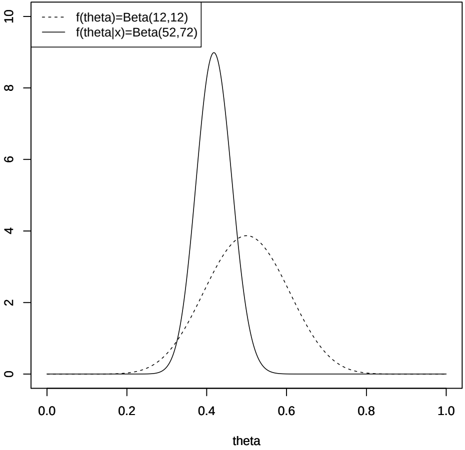
What does “likely to be close” mean?

- 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 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

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

- We have just constructed an *estimator*

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

- What is the posterior mode estimate/estimator?

Constructing estimates and estimators by minimizing posterior expected distance

- Can we make a formal rule in building estimators to achieve the *"likely closeness"* between the parameter and the estimate?
- Yes! How about choosing an estimate such that the expected *distance* between θ and the estimate is *as small as possible*.
- In particular, given the posterior distribution $\xi(\theta|\mathbf{x})$, we can choose an estimate a such that the expected distance between θ and a

$$E(|\theta - a|\,\big|\,\mathbf{x}) = \int_{-\infty}^{\infty} |\theta - a|\xi(\theta|\mathbf{x})d\theta.$$

  is minimized.
- That is, we can define an estimate $\delta^*(\mathbf{x})$ such that

$$\delta^*(\mathbf{x}) = \text{argmin}_a E(|\theta - a|\,\big|\,\mathbf{x})$$

- Estimates constructed this way are called *Bayes estimates*.

More generally (a decision theory setup)

- Different notions of distance can be adopted. We introduce a distance (or *loss*) function

$$L(\theta, a).$$

- Examples of common *loss* functions include:
  1. $L(\theta, a) = |\theta - a|$ is called the *absolute error loss*.
  2. $L(\theta, a) = (\theta - a)^2$ is called the *squared error loss*.
  3. $L(\theta, a) = \mathbf{1}(|\theta - a| > \Delta)$ is called the *step error loss*.
- The Bayes estimate is the value of a that minimizes the *posterior expectation* of the loss

$$\delta^*(\mathbf{x}) = \text{argmin}_a E\left(L(\theta, a)|\mathbf{x}\right) = \text{argmin}_a \int_{-\infty}^{\infty} L(\theta, a)\xi(\theta|\mathbf{x})d\theta$$

- For example, the Bayes *estimate* under the squared error loss is

$$\delta^*(\mathbf{x}) = \text{argmin}_a E((\theta - a)^2 \big| \mathbf{x})$$

Loss as the "cost" in decision making

- One can think of the loss function as characterizing the cost of choosing a as the estimate for θ. (Draw a graph.)
  - Consider the situation in which the statistician is making certain *decisions* based on the estimates.
  - The loss function characterizes the cost of choosing a as the estimate for a parameter θ.
  - So one can design custom-made losses for specific problems.
  - Think about the political poll example. What might be a realistic loss function for that?
  - The above simple loss functions are mostly chosen for their mathematical simplicity, especially the squared error loss.

The steps in Bayes estimation (or other decision problems)

1. Choose the distribution of the data given the parameter, $f(\mathbf{x}|\theta)$.
2. Specify a prior distribution for the parameter, $\xi(\theta)$.
3. After observing the data $\mathbf{X} = \mathbf{x}$, apply Bayes Theorem to get the poterior distribution of the parameter, $\xi(\theta|\mathbf{x})$.
4. Choose a loss function that specifies the distance between the parameter and the estimates.
5. Choose a number a that minimizes the expected distance $E(L(\theta,a)\big|\mathbf{x})$. This a is our *Bayes estimate* given data $\mathbf{X} = \mathbf{x}$, $\delta^*(\mathbf{x})$.
6. The corresponding estimator $\delta^*(\mathbf{X})$ is called the *Bayes estimator*. It describes the rule we will use to map data to the estimate had the experiment been repeated.

Bayes estimator under squared error loss

It turns out that with *squared error loss*, the Bayes estimate given $\mathbf{X} = \mathbf{x}$ is exactly the posterior mean of θ. That is the mean of the posterior distribution:

$$\delta^*(\mathbf{x}) = E(\theta|\mathbf{x}),$$

as long as this expecation is well-defined and finite.

The corresponding Bayes estimator is

$$\delta^*(\mathbf{X}) = E(\theta|\mathbf{X}).$$

Example: Political poll revisited

- Let us go back to our political poll example and find the Bayes estimator for the fraction θ under squared error loss.
- With a Beta(α, β) prior on θ, the Bayes estimator is

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

- That is, it minimizes the posterior expected squared error loss for any observed data $X = x$.
- Now let's see why the Bayes estimate for squared error loss is the posterior mean.

Bayes estimate under squared error loss

Let Y be a random variable with a finite mean $\mu_Y = E[Y]$. Then for any number a,

$$\begin{aligned} E(L(Y,a)) &= E(Y-a)^2 \\ &= E(Y-\mu_Y+\mu_Y-a)^2 \\ &= E(Y-\mu_Y)^2 + 2E(Y-\mu_Y)(\mu_Y-a) + (\mu_Y-a)^2 \\ &= \mathrm{Var}(Y) + (\mu_Y-a)^2. \end{aligned}$$

This is minimized when $a = \mu_Y$.

- Now let the random variable Y be our parameter θ.
- Given $\mathbf{X} = \mathbf{x}$, its distribution is the posterior distribution $\xi(\theta|\mathbf{x})$.
- Therefore the value a that minimizes $E(L(\theta,a)|\mathbf{x})$ is $E(\theta|\mathbf{x})$.

- One can show through more complex arguments that when $L(\theta, a)$ is the absolute error loss, the number a that minimizes $E(L(\theta, a)|\mathbf{x})$ is the median of posterior distribution $\xi(\theta|\mathbf{x})$.
- Thus the Bayes estimate

$$\delta^*(\mathbf{x}) = \text{the median of } \xi(\theta|\mathbf{x}).$$

- The Bayes estimator is

$$\delta^*(\mathbf{X}) = \text{the median of } \xi(\theta|\mathbf{X}).$$

- For the political poll example, given $X = 40$,
  - The Bayes estimate $\delta^*(40)$ is the median of $\text{Beta}(\alpha + 40, \beta + 60)$.
  - The Bayes estimator $\delta^*(X)$ is the median of $\text{Beta}(\alpha + X, \beta + n - X)$.

Question: What is the corresponding Bayes estimator for the step error loss?*

$$L(\theta, a) = \begin{cases} 1 & \text{if } |\theta - a| > \Delta \\ 0 & \text{otherwise.} \end{cases}$$

What happens when $\Delta \downarrow 0$?

The air pollutant example with a single reading

- The posterior distribution of θ, given a single measurement $X = x$ is $\mathrm{N}(\tilde{\mu}, \tilde{\sigma}^2)$ with

$$E(\theta|X = x) = \tilde{\mu} = \left(\frac{1/\sigma^2}{1/\sigma^2 + 1/\tau^2}\right)\mu + \left(\frac{1/\tau^2}{1/\sigma^2 + 1/\tau^2}\right)x.$$

- This is both the mean and the median of the posterior distribution.
- Bayes estimator under squared error loss is

$$\delta(X) = \left(\frac{1/\sigma^2}{1/\sigma^2 + 1/\tau^2}\right)\mu + \left(\frac{1/\tau^2}{1/\sigma^2 + 1/\tau^2}\right)X.$$

- What is the Bayes estimator under absolute error loss?
- How about under the step error loss?

Bayesian vs. sampling theory

- Up until now, we have been addressing the inference problem using Bayes' Theorem.
- Bayes' Theorem provides a recipe for getting a distribution of the state of nature or parameter θ *after observing the data*—$\xi(\boldsymbol{\theta}|\mathbf{x})$.
- This posterior distribution summarizes all of the uncertainty in the parameter in light of the data.
- This is exactly how humans think everyday, and is the *ideal* goal one can hope to get from any statistical inference procedure.
  - "Given that it is so cloudy, what is the chance for rain?"

The two modeling requirements for Bayesian inference are that

1. We need to choose a probability model for the data given the parameter θ: $f(x|\theta)$ or $p(x|\theta)$.
2. We must treat θ as a random quantity and choose a prior distribution for it: $\xi(\theta)$.

Virtually everyone is okay with the first requirement—e.g. modeling the political poll as a Binomial experiment, etc. But some have a problem with the second.

- Some statisticians stick to a strict “frequentist” view of probabilities in which probability must be interpretable as long-run relative frequencies rather than quantifying subjective belief or knowledge.
- Some others (more than the previous category) think that it’s too difficult to choose an appropriate prior distribution, especially in very complex problems.
- They are looking to solutions for inference without the second modeling requirement.

We will next start our study of *sampling theory*, which treats the parameter as a fixed unknown number, and bases inference entirely on $f(x|\theta)$—the *sampling distribution* of the data.

- Now let us consider θ as a *fixed* but *unknown* quantity.
- Bayes' Theorem tells us that we *need to* treat θ as random variable, and assign a prior distribution $\xi(\theta)$ to it, in order to be able to summarize the uncertainty about θ after observing data also as a probability distribution $\xi(\theta|x)$.
- We can no longer take this "*after-the-experiment*" perspective in our inference, because after the experiment, nothing is random.

- We *have to give up* that ideal goal of summarizing our knowledge about the parameter in light of data in terms a probability distribution.
- However, we can still try to achieve less ambitious goals, such as
    1. constructing good estimators for the parameter θ or a function of the parameter $g(\theta)$. (Point estimation)
    2. comparing two or more hypotheses e.g. $\theta = 2$ vs. $\theta = 3$. (Hypothesis testing.)
    3. These two topics will be the focus of much of the rest of this course.
- Note that because we can no longer take the "after-the-experiment" point of view, evaluating the performances of the corresponding statistical procedure must be done differently—under the repeated experiment point of view.