STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 8

Summary of what we have learned so far

- The general inference procedure based on Bayes Theorem.
- How to construct point estimates/estimators based on the posterior distribution. (The Bayes estimate/estimator).
- The sampling view of inference—parameters are fixed unknown quantities.
- Measures of good estimators under the sampling view (the "before-the-experiment" view).
    - Risk functions: e.g. mean error risk, mean square error risk, etc.
    - They represent the average distance between the estimate and the true θ under repeated experiments.

The natural question is: How do we construct good point estimators that tend to have good risk measures?

- In principle, any statistic can be used as an estimator.
- Of course such arbitrary estimators are unlikely to be "good", that is close to the true parameter.

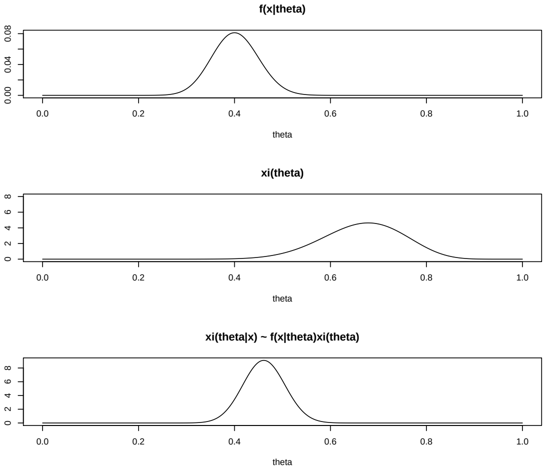
The principle of maximum likelihood

Recall Bayes' theorem

$$\xi(\theta|x) \propto \xi(\theta)f(x|\theta)$$

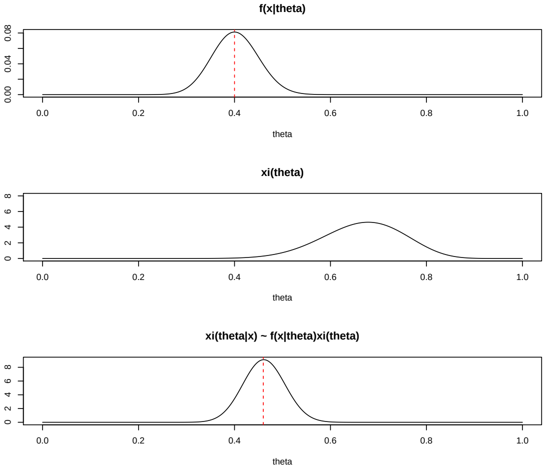
- One good candidate for an estimator based on the posterior distribution is the posterior mode. That is the value of θ that maximizes $\xi(\theta)f(x|\theta)$.
- Now let us take a skeptics' view and drop the prior $\xi(\theta)$ from the right hand side. How about we find the value of θ that maximizes $f(x|\theta)$?
- *In doing this, we are viewing $f(x|\theta)$ as a function of θ, not of x!*
- If $\xi(\theta)$ is relatively constant for θ values near the value that maximizes $\xi(\theta)f(x|\theta)$, then the maximizer of $f(x|\theta)$ should be very close to the maximizer of the posterior pdf.

The maximizer of $f(x|\theta)$ and the posterior mode.



Note that the horizontal axes are all θ, not x!

The maximizer of $f(x|\theta)$ and the posterior mode.



Note the closeness between the maximizer of $L(\theta) = f(x|\theta)$ and that of $\xi(\theta|x)$. This holds for many different choices of $\xi(\theta)$.

The likelihood function

Let us define

$$L(\theta) = f(x|\theta)$$

as the *likelihood function*. The notation suggests that it is viewed as *a function in θ*.

- It is the "probability" for the data to arise under θ. So we use it as the *empirical evidence* to support the value θ.
  - *Note that as we repeat the experiment, with different observed data values, $L(\theta)$ changes.*
  - $L(\theta)$ represents the empirical evidence for different θ values corresponding to the observed data.
- For example, if $L(0.8) \gg L(0.2)$, then the data seems to support $\theta = 0.8$ much more than $\theta = 0.2$.

In the case of multiple observations $X_1, X_2, \ldots, X_n$,

$$L(\theta) = f_n(\mathbf{x}|\theta) = f(x_1, x_2, \ldots, x_n|\theta).$$

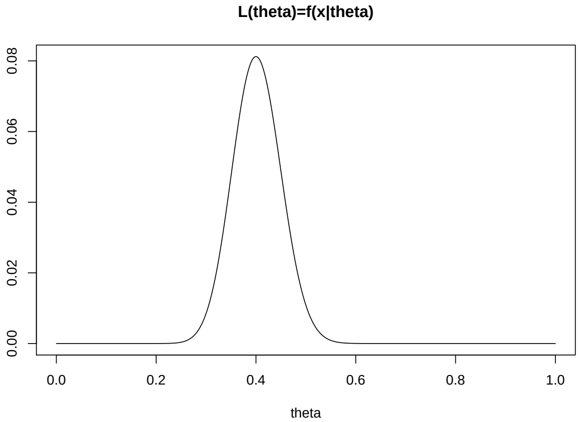
Likelihood function for the political poll example

For $X = x$, the likelihood function is

$$L(\theta) = p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad \text{for } 0 \le \theta \le 1.$$

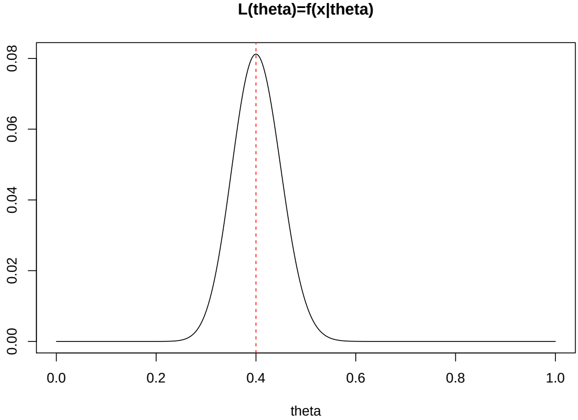
Now let us look at it as a function of θ.

Example: $X = 40$, $n = 100$.



$$L(\theta) = \binom{100}{40} \theta^{40} (1 - \theta)^{60} \quad \text{for } 0 < \theta < 1.$$

Example: $X = 40$, $n = 100$.



$$L(\theta) = \binom{100}{40} \theta^{40} (1 - \theta)^{60} \quad \text{for } 0 < \theta < 1.$$

The maximum likelihood estimator

- Given data $\mathbf{X} = \mathbf{x}$, the value of θ that maximizes the likelihood function $L(\theta)$ is said to be the *maximum likelihood estimate*.
- It is often denoted by $\hat{\theta}(\mathbf{x})$.
- The corresponding estimator constructed this way is denoted by $\hat{\theta}(\mathbf{X})$, and is called the *maximum likelihood estimator* or *MLE*.
- $\hat{\theta}$ is often used as a short notation for either the estimator $\hat{\theta}(\mathbf{X})$, or the estimate $\hat{\theta}(\mathbf{x})$, depending on the context.

How to find the MLE?

The usual way to find the maximizer of a function $L(\theta)$ is the following two steps:

1. We find a value θ such that
$$\frac{d}{d\theta}L(\theta) = 0.$$
2. Then we verify that this value of θ indeed is a maximizer by checking that for this value of θ
$$\frac{d^2}{d\theta^2}L(\theta) < 0.$$

Note that

1. The solution to the first equation may not be unique.
2. Strictly speaking these two steps will only find *local* maxima, while the MLE is the *global* maximum.

Political poll

$$\begin{aligned}
\frac{d}{d\theta}L(\theta) &= \frac{d}{d\theta}\binom{n}{x}\theta^{x}(1-\theta)^{n-x} \\
&= \binom{n}{x}\theta^{x-1}(1-\theta)^{n-x-1}\left(x(1-\theta)-(n-x)\theta\right) \\
&= 0
\end{aligned}$$

This equation has three roots 0, 1, or x/n. Generally x/n is the maximizer while the other two are minimizers.

- So the maximum likelihood estimate of θ is $\hat{\theta} = x/n$.
- The maximum likelihood estimator (MLE) is $\hat{\theta} = X/n$.

- Strictly speaking one should check $\frac{d^2}{d\theta^2}L(\theta)$ to make sure that our claim above is correct.
- But the computation seems to get very tedious.
- Is there an easier way?

- Yes! It's often much easier to maximize $\log L(\theta)$ then directly maximizing $L(\theta)$ itself.
- Since log is a monotone function the θ that maximizes $L(\theta)$ also maximizes $\log L(\theta)$ and vice versa.

Back to the polititcal poll example

We have

$$\log L(\theta) = \log \binom{n}{x} + x \log \theta + (n-x) \log(1-\theta).$$

Let's take a derivative and set it to zero

$$\frac{d}{d\theta} \log L(\theta) = 0 + \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0.$$

The solution to this equation is

$$\hat{\theta} = \frac{x}{n}.$$

It is also easier to check the second-order derivative

$$\frac{d^2}{d\theta^2} \log L(\theta) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} < 0$$

for all $\theta \in (0,1)$. Therefore we know $L(\hat{\theta})$ is a global maximum.

Another way to see that $\hat{\theta}$ is the *global* maximum

$$\frac{d}{d\theta}\log L(\theta) = \frac{(1-\theta)x - \theta(n-x)}{\theta(1-\theta)} = \frac{n}{\theta(1-\theta)}(\frac{x}{n} - \theta) = \frac{n}{\theta(1-\theta)}(\hat{\theta} - \theta).$$

So

- $\frac{d}{d\theta}\log L(\theta) > 0$ for $0 < \theta < \hat{\theta}$
- $\frac{d}{d\theta}\log L(\theta) < 0$ for $\hat{\theta} < \theta < 1$.

Therefore $\hat{\theta}$ must be the *global* maximum of $\log L(\theta)$ and hence of $L(\theta)$.

- So $\hat{\theta} = X/n$ is indeed the MLE.
- What is its MSE?

Example: Estimating average failure time of light bulbs

A particular type of light bulb will last time X, which can be modeled as an Exponential$(1/\theta)$ random variable

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta} \quad \text{for } x > 0.$$

Under this distribution $E(X) = \theta$ and $\text{Var}(X) = \theta^2$.

- Suppose θ is unknown, and we want to estimate it.
- We observe the life time of n such light bulbs $X_1, X_2, \ldots, X_n$.
- What is the MLE of the expected life time θ?

The likelihood function is

$$L(\theta) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}} \quad \text{for } \theta > 0.$$

Again before finding the maximizer, let us take the log

$$\log L(\theta) = -n \log \theta - \frac{\sum_{i=1}^{n} x_i}{\theta}.$$

Now let us find the maximizer of $\log L(\theta)$. The derivative is

$$\frac{d}{d\theta}\log L(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}.$$

Setting this to zero and solve for θ, we get

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

Again, we can verify that this is indeed a maximum by checking the second-order derivative

$$\frac{d^2}{d\theta^2}\log L(\theta) = \frac{n}{\theta^2} - 2\frac{\sum_{i=1}^{n} x_i}{\theta^3} = \frac{n}{\theta^2}(1 - 2\frac{\hat{\theta}}{\theta}).$$

At $\theta = \hat{\theta}$, this is equal to $-n/\hat{\theta}^2 < 0$. So this is indeed a maximum.

Strictly speaking, up to this point we only know this is a *local* maximum. We don't know whether this is a global maximum.

$$\frac{d}{d\theta}\log L(\theta) = \frac{n}{\theta^2}(\bar{x} - \theta) = \frac{n}{\theta^2}(\hat{\theta} - \theta).$$

So

- $\frac{d}{d\theta}\log L(\theta) > 0$ for $0 < \theta < \hat{\theta}$, and
- $\frac{d}{d\theta}\log L(\theta) < 0$ for $\theta > \hat{\theta}$.

Thus $\hat{\theta}$ is a *global* maximum of $\log L(\theta)$ and hence of $L(\theta)$.

- The MLE of θ is $\hat{\theta} = \bar{X}$.
- What is the MSE of $\hat{\theta} = \bar{X}$?

Another example

Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from the following distribution

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta} \quad \text{for } 0 < x < \theta \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

What is the MLE for θ?

$$
\begin{aligned}
L(\theta) &= \frac{1}{\theta^n} \quad \text{for } \theta > x_1, x_2, \ldots, x_n \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

What is the global maximizer for $L(\theta)$?

- The closer θ is to $\max\{x_1, x_2, \ldots, x_n\}$, the larger the likelihood is.
- *MLE for θ doesn't even exist!*
- Question: What is the MLE if the data are uniform on $[0, \theta]$ rather than on $(0, \theta)$?
  - Is it unbiased? What is its MSE?