STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 9

Two interpretations of the likelihood function

From Bayes' Theorem

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

we know that if our prior $\xi(\theta)$ is "flat", then

$$\xi(\theta|x) \propto f(x|\theta) = L(\theta).$$

Under this interpretation, the MLE, $\hat{\theta}$, is the posterior mode of the parameter with a "flat" prior.

Another interpretation of $L(\theta)$

- Without taking the Bayesian perspective, $L(\theta) = p(x|\theta)$ or $f(x|\theta)$, is the conditional probability mass or density of the data given a value of θ:

$$L(\theta) = P(\text{observed data} \mid \text{state of nature } \theta).$$

- For two values of θ, θ_1 and θ_2, we can think of $L(\theta)$ as giving the relative probability of observing ("likelihood") of observing the data under θ_1 and θ_2.
- If $L(\theta_1)/L(\theta_2) = 2$, we are twice as likely to observe the data given θ_1 as given θ_2.
- Under this interpretation, the MLE for θ, i.e. the maximizer of $L(\theta)$ is essentially the value of θ *that best explains our data*.

Political poll example

Our observed data is that $X = 40$ out of 100 interviewees express their support for the governor.

- The likelihood under $\theta = 0.4$ is

$$L(0.4) = \binom{100}{40} 0.4^{40} 0.6^{60} \approx 0.081.$$

- The likelihood under $\theta = 0.3$ is

$$L(0.3) = \binom{100}{40} 0.3^{40} 0.7^{60} \approx 0.0085.$$

- We have

$$L(0.4)/L(0.3) \approx 9.6.$$

- We say that $X = 40$ is 9.6 times as likely to occur under $\theta = 0.4$ as it is under $\theta = 0.3$.
- This is different from saying that θ is 9.6 times as likely to be 0.4 as to be 0.3!

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 $\mathrm{Var}(X) = \theta^2$.

- 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?

We have seen that the MLE for θ is

$$\hat{\theta} = \bar{X}.$$

Suppose the problem is stated another way:

- If we let $\lambda = \frac{1}{\theta}$, X has an Exponential(λ) distribution with p.d.f

$$f(x|\lambda) = \lambda e^{-\lambda x} \quad \text{for } x > 0.$$

- λ has the meaning as the average number of bulb replacements per unit time.
- If we are interested in estimating λ, what is the MLE for it, $\hat{\lambda}$?

- We can write down the likelihood function in terms of λ

$$L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

- *Note that there is no Jacobian term. Distinguish reparametrization from change-of-variable.*
- We can then take a log of this and again maximize over λ.
- But there is an easier way …
- Because $\lambda = 1/\theta$, we must have

$$\hat{\lambda} = \frac{1}{\hat{\theta}} = \frac{1}{\bar{X}}.$$

- Why?

- Suppose this is not the case, let $\lambda_0 = 1/\hat{\theta}$.
- So we must have another λ_1 such that

$$L(\lambda_1) > L(\lambda_0).$$

  But that means that the likelihood under $\theta_1 = 1/\lambda_1$ is larger than than that under $\hat{\theta}$, contradicting the fact that $\hat{\theta}$ is the MLE for θ.
- Therefore we must have

$$\hat{\lambda} = \frac{1}{\hat{\theta}}.$$

This property is call the *invariance* property of MLEs.

More generally

- Let $h(\cdot)$ be any one-to-one function. If the MLE for θ is $\hat{\theta}$, then the MLE for $\psi = h(\theta)$ is

$$\hat{\psi} = h(\hat{\theta}).$$

- If $h(\cdot)$ is not one-to-one, we simply define the MLE for ψ to be

$$\hat{\psi} = h(\hat{\theta}).$$

- For example, the MLE for $\sqrt{\theta} + \theta^2$ is $\sqrt{\hat{\theta}} + \hat{\theta}^2$.
- *However, properties of $\hat{\theta}$ do not necessarily carry over to $\hat{\psi}$!*

For example, unbiasedness may not carry over.

- In our current example, $\hat{\theta} = \bar{X}$ is an *unbiased* estimator of θ.

$$E(\hat{\theta}) = E(\bar{X}) = \frac{n\theta}{n} = \theta.$$

- However, $\hat{\lambda}$ is typically not unbiased:

$$E(\hat{\lambda}) = E\left(\frac{1}{\hat{\theta}}\right) = E\left(\frac{1}{\bar{X}}\right) \neq \frac{1}{E(\bar{X})} = \frac{1}{\theta} = \lambda.$$

- Question: When does unbiasedness carry over?

Normal distribution with unknown mean and variance

- We observe n i.i.d. observations $X_1, X_2, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution.
- Both the mean μ and the variance σ^2 are unknown.
- How do we estimate the parameters μ and σ^2.

- The p.d.f of a single observation is

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \quad \text{for } -\infty < x < \infty.$$

- The likelihood function is

$$\begin{aligned} L(\mu,\sigma^2) = f_n(\mathbf{x}|\mu,\sigma^2) &= \prod_{i=1}^{n} f(x_i|\mu,\sigma^2) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

- To find the MLEs, we maximize $L(\mu,\sigma^2)$, or better yet ...
- We maximize $\log L(\mu,\sigma^2)$.

- After taking the log transformation,

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

- For simplicity, let $\phi = \sigma^2$.

$$\log L(\mu, \phi) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\phi - \frac{1}{2\phi}\sum_{i=1}^{n}(x_i - \mu)^2.$$

- To find the maximum, we differentiate

$$\frac{d}{d\mu}\log L(\mu, \phi) = -\frac{1}{2\phi}\sum_{i=1}^{n}(-2(x_i - \mu)) = \frac{n}{\phi}(\bar{x} - \mu)$$

and

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

- Setting the two derivatives to zero we get

$$\frac{n}{\hat{\phi}}(\bar{x}-\hat{\mu}) = 0$$

and

$$-\frac{n}{2\hat{\phi}} + \frac{1}{2\hat{\phi}^2}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = 0.$$

- Solving these two equations, we get

$$\hat{\mu} = \bar{x}$$

$$\hat{\phi} = \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

- To verify that this is indeed a maximum, we need to check that the Hessian matrix of $\log L(\mu, \phi)$

$$\begin{pmatrix} \frac{d^2}{d\mu^2} \log L(\mu, \phi) & \frac{d^2}{d\mu d\phi} \log L(\mu, \phi) \\ \frac{d^2}{d\phi d\mu} \log L(\mu, \phi) & \frac{d^2}{d\phi^2} \log L(\mu, \phi) \end{pmatrix}$$

is negative definite at $(\mu, \phi) = (\hat{\mu}, \hat{\phi})$.
- To see this, note that

$$\frac{d^2}{d\mu^2} \log L(\mu, \phi)|_{(\hat{\mu}, \hat{\phi})} = -\frac{n}{\hat{\phi}} < 0, \tag{1}$$

$$\begin{aligned} \frac{d^2}{d\phi^2} \log L(\mu, \phi)|_{(\hat{\mu}, \hat{\phi})} &= \frac{n}{2\hat{\phi}^2} - \frac{1}{\hat{\phi}^3} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \\ &= \frac{n}{2\hat{\phi}^2} - \frac{n}{\hat{\phi}^3} \hat{\phi} = -\frac{n}{2\hat{\phi}^2} < 0, \end{aligned}$$

- Finally,

$$\frac{d^2}{d\mu d\phi}\log L(\mu,\phi)|_{(\hat{\mu},\hat{\phi})} = \frac{-n}{\hat{\phi}^2}(\bar{x}-\hat{\mu}) = 0$$

Therefore

$$\det\left(\begin{array}{cc} \frac{d^2}{d\mu^2}\log L(\mu,\phi) & \frac{d^2}{d\mu d\phi}\log L(\mu,\phi) \\ \frac{d^2}{d\phi d\mu}\log L(\mu,\phi) & \frac{d^2}{d\phi^2}\log L(\mu,\phi) \end{array}\right)\Bigg|_{(\hat{\mu},\hat{\phi})} > 0. \qquad (2)$$

- (1) and (2) together show that the Hessian is indeed negative definite.
- So $(\hat{\mu},\hat{\phi})$ gives a (local) maximum of $L(\mu,\phi)$.
- How do we show it gives a global maximum? Can we prove it without matrix algebra?
  - One argument can go like this: First, for any given ϕ, $\hat{\mu}=\bar{x}$ maximimizes $\log L(\mu,\phi)$. Then maximize $\log L(\bar{x},\phi)$ over ϕ ...

- So we have found the MLEs

$$\hat{\mu} = \bar{X}$$
$$\hat{\phi} = \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Let's examine their properties.

Bias

First,

$$E(\hat{\mu}) = E(\bar{X}) = \frac{n\mu}{n} = \mu.$$

- Therefore $\hat{\mu}$ is an unbiased estimator of μ.
- As for $\hat{\phi} = \widehat{\sigma^2}$, we have

$$\begin{aligned}
\hat{\phi} = \widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)
\end{aligned}$$

$$E(\sum_{i=1}^{n} X_i^2) = \sum_{i=1}^{n} E(X_i^2) = \sum_{i=1}^{n} \Big(\mathrm{Var}(X_i) + (E(X_i))^2\Big)$$
$$= \sum_{i=1}^{n} (\sigma^2 + \mu^2) = n\sigma^2 + n\mu^2.$$

In addition,

$$E\left(\Big(\sum_{i=1}^{n} X_i\Big)^2\right) = \mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) + \left(E\Big(\sum_{i=1}^{n} X_i\Big)\right)^2$$
$$= n\sigma^2 + (n\mu)^2.$$

Therefore,

$$E(\widehat{\sigma^2}) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)$$
$$= \frac{1}{n}\left(n\sigma^2 + n\mu^2 - \frac{n\sigma^2 + n^2\mu^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Therefore the MLE for σ^2,

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is *not* unbiased, and its bias is

$$B_{\widehat{\sigma^2}}(\sigma^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

This bias converges to 0 with more and more data, i.e. as $n \uparrow \infty$.

- Can we eliminate this bias altogether? In other words, can we find another estimator for σ^2 that is unbiased?
- Hint:

$$E(\widehat{\sigma^2}) = \frac{n-1}{n}\sigma^2.$$

How about we let

$$s^2 = \left(\frac{n}{n-1}\right)\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then we know

$$E(s^2) = \frac{n}{n-1}E(\widehat{\sigma^2}) = \sigma^2.$$

- s^2 is called the *sample variance*.
- It is an unbiased estimator for σ^2.
- When there are a lot of data, i.e. n is very large, $\frac{n}{n-1} \approx 1$, in which case there is hardly any difference between s^2 and $\widehat{\sigma^2}$.

- Mostly due to *tradition*, s^2 is the estimator of choice for the normal variance σ^2.
- But practically the *unbiasedness* of s^2 really isn't that important. Why?
- Because we are typically interested in the standard deviation σ, not σ^2.
- What is the MLE for σ?

- By the *invariance* property of MLEs

$$\hat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

- We could also estimate σ by

$$s = \sqrt{s^2} = \sqrt{\frac{n}{n-1}}\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

- The estimator s is called the *sample standard deviation*.

It turns out that both $\hat{\sigma}$ and s are *biased* estimators for σ. More specifically, one can show that

$$E(s) = b_n \sigma$$

and thus

$$E(\hat{\sigma}) = \sqrt{\frac{n-1}{n}} b_n \sigma,$$

where

$$b_n = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}.$$

Again, can we eliminate this bias altogether?

- Yes! But this is rarely done in practice.