

## MATH 36B: MATHEMATICAL STATISTICS

### 1. MLE

**Definition 1.1.** *The maximum likelihood estimator  $\theta_{MLE}$  of an estimator  $\theta$  is defined to be the value of  $\theta$  which makes the observed data most likely.*

**1.1. Intuitive examples.** Suppose you flip a coin twice and you get two heads. What prior conditions would make this outcome the most likely? If the coin were fair, the probability of getting two heads is  $1/4$ . If the coin were two-headed, the probability of getting two heads is 100%. So, the likelihood is the maximum if the coin is two-headed.

The *likelihood*  $L(\theta)$  of the prior condition  $\theta$  is defined to be the probability of the outcome assuming this prior condition. So, the likelihood that the coin is two-headed is 100%

$$L(\text{two-headed coin}) = \mathbb{P}(HH \mid \text{coin is two-headed}) = 1$$

**1.2. Computational examples.** Suppose that  $T$  is exponentially distributed. Thus the density function is

$$f_T(t) = \lambda e^{-\lambda t} \text{ for } t \geq 0 \quad (\text{and } f_T(t) = 0 \text{ for } t < 0)$$

If we have a sample of  $n$ :  $T_1 = t_1, T_2 = t_2, \dots, T_n = t_n$  what is the MLE for  $\lambda$ ?

The density function for independent  $T_i$  is the product of density functions:

$$L(\lambda) = \prod_{i=1}^n f(t_i) = \lambda^n e^{-\lambda \sum t_i}$$

You always take  $\ln L(\theta)$ :

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum t_i$$

Maximizing  $L$  is the same as maximizing  $\ln L$ . So, take the derivative:

$$\frac{\partial}{\partial \lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum t_i$$

Set the derivative equal to zero to get the max. (The second derivative  $-1/\lambda^2$  is always negative so we will get a max.)

$$\frac{n}{\lambda} = \sum t_i$$

Solve for  $\lambda$  to get the estimate for  $\lambda$ :

$$\hat{\lambda} = \frac{1}{\frac{1}{n} \sum t_i} = 1/\bar{t}$$

The estimator is

$$\hat{\lambda}_{MLE} = 1/\bar{T}$$

## 2. MOM

**Definition 2.1.** For any random variable  $X$  the  $n$ th moment of  $X$  is defined to be the expected value of  $X^n$ . The method of moments is to equate this with the experimental value

$$\overline{X^n} = \frac{1}{n} \sum_{i=1}^n X_i^n$$

for  $n = 1, 2, \dots, k$  if there are  $k$  parameters to estimate. (Notice that the bar goes over the exponent  $n$ .  $\overline{X^n}$  is not equal to  $\overline{X}^n$ .)

**2.1. Examples with one parameter.** When there is only one parameter  $\theta$ , the MOM says to put:

$$\mathbb{E}(X) = \overline{X}$$

and solve for  $\theta$ .

**2.1.1. uniform distribution.** Here  $X$  is between 0 and  $\theta$  with equal probability ( $f(x) = 1/\theta$  is constant). The expected value of  $X$  is

$$\mathbb{E}(X) = \frac{\theta}{2}$$

Set this equal to the experimental mean  $\overline{X}$  and solve for  $\theta$  to get

$$\theta_{MOM} = 2\overline{X}$$

(The MLE is  $X_{max}$ .)

**2.1.2. exponential distribution.** The expected value of  $T$  is

$$\mathbb{E}(T) = \int_0^{\infty} t f(t) dt = \int_0^{\infty} t \lambda e^{-\lambda t} dt = \frac{1}{\lambda}$$

Set this equal to  $\overline{T}$  and solve for  $\lambda$  to get

$$\hat{\lambda}_{MOM} = 1/\overline{T}$$

(same as  $\lambda_{MLE}$ )

We also need the second moment of  $T$  to do the next problem:

$$\mathbb{E}(T^2) = \int_0^{\infty} t^2 \lambda e^{-\lambda t} dt = \frac{2}{\lambda^2}$$

**2.2. two parameters.** When there are two parameters you need to solve the equations:

$$\mathbb{E}(X) = \bar{X}$$

and

$$\mathbb{E}(X^2) = \frac{1}{n} \sum X_i^2$$

simultaneously for the two parameters. The example we did was the *Gamma distribution* with parameters  $\lambda$  and  $r$ . Then  $T$  is the amount of time you have to wait for the  $r$ th occurrence of a Poisson event with rate  $\lambda$ . In other words

$$T = T_1 + T_2 + \cdots + T_r$$

where  $T_i$  are iid exponential variables.

$$\mathbb{E}(T) = \mathbb{E}\left(\sum T_i\right) = \sum \mathbb{E}(T_i) = \frac{r}{\lambda}$$

$$\begin{aligned} \mathbb{E}(T^2) &= \mathbb{E}\left(\sum T_i^2 + \sum_{i \neq j} T_i T_j\right) \\ &= r\mathbb{E}(T_i^2) + r(r-1)\mathbb{E}(T_i)^2 = \frac{2r}{\lambda^2} + \frac{r^2 - r}{\lambda^2} = \frac{r^2 + r}{\lambda^2} \end{aligned}$$

Set these equal to  $\bar{T}$  and  $\bar{T}^2$  and we get

$$\begin{aligned} \frac{\hat{r}}{\hat{\lambda}} &= \bar{T} \Rightarrow \frac{\hat{r}^2}{\hat{\lambda}^2} = \bar{T}^2 \\ \frac{\hat{r}^2 + \hat{r}}{\hat{\lambda}^2} &= \bar{T}^2 \end{aligned}$$

So,

$$\frac{\hat{r}}{\hat{\lambda}^2} = \bar{T}^2 - \bar{T}^2$$

Now we can solve for  $\hat{r}$  and  $\hat{\lambda}$ :

$$\hat{r}_{MOM} = \frac{\bar{T}^2}{\bar{T}^2 - \bar{T}^2} \quad (\text{unitless})$$

$$\hat{\lambda}_{MOM} = \frac{\bar{T}}{\bar{T}^2 - \bar{T}^2} \quad (1/\text{time units})$$

For the 36 eruption of MaunaLoa we got

$$\hat{r} = 1.4867$$

which seems to be very far from the value  $r = 1$  required for a Poisson process. So, it looks like the Gamma distribution fits better. But we will come back to this again.

## 3. UNBIASED ESTIMATORS

**Definition 3.1.** An estimator  $\hat{\theta}$  for  $\theta$  is called unbiased if

$$\mathbb{E}(\hat{\theta}) = \theta$$

For example, no matter what the distribution is,  $\bar{X}$  is always an unbiased estimator for  $\mu := \mathbb{E}(X)$  (provided that  $\mathbb{E}(X)$  is defined by a converging integral). Here is the proof:

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n} \sum \mathbb{E}(X_i) = \frac{1}{n} n\mu = \mu$$

For the normal distribution we found (using the Pythagorean theorem) that

$$\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

is a biased estimator for the variance  $\sigma^2$  since

$$\mathbb{E}(\hat{\sigma}^2) = \frac{n}{n-1} \sigma^2$$

This means an unbiased estimator for  $\sigma^2$  is the *sample variance*

$$S^2 := \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

## 4. EFFICIENCY

We want an estimator  $\hat{\theta}$  to be unbiased with as small a variance as possible. So, I defined the *efficiency* of an unbiased estimator  $\hat{\theta}$  to be the inverse of the variance:

$$\text{Eff}(\hat{\theta}) := \frac{1}{\text{Var}(\hat{\theta})}$$

This means we want the efficiency of  $\hat{\theta}$  to be as large as possible. The book only defines relative efficiency:

**Definition 4.1.** *Given two unbiased estimators  $\hat{\theta}_1, \hat{\theta}_2$  for the same parameter  $\theta$ , we say that  $\hat{\theta}_1$  is more efficient than  $\hat{\theta}_2$  if  $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$ . The relative efficiency is defined to be the ratio*

$$\text{Eff}(\hat{\theta}_1; \hat{\theta}_2) := \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)} = \frac{\text{Eff}(\hat{\theta}_1)}{\text{Eff}(\hat{\theta}_2)}$$

There is a theoretical lower bound for the variance of any unbiased estimator called the *Cramer-Rao* lower bound:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where  $I(\theta)$  is the amount of “information about  $\theta$ ” contained in the sample.

## 4.1. Fisher information.

**Definition 4.2.** *The Fisher information given by the sample is defined by*

$$I(\theta) := \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \ln L(\theta) \right)^2 \right] = \mathbb{E} \left[ - \left( \frac{\partial^2}{\partial \theta^2} \ln L(\theta) \right) \right]$$

I should have pointed out that this definition only makes sense when the support of the density function is independent of  $\theta$ .

4.1.1. *example: exponential distribution.* Here  $L(\lambda) = f(t) = \lambda e^{-\lambda t}$ . So,

$$\begin{aligned} \ln L(\lambda) &= \ln \lambda - \lambda T \\ \frac{\partial}{\partial \lambda} \ln L(\lambda) &= \frac{1}{\lambda} - T \\ \frac{\partial^2}{\partial \lambda^2} \ln L(\lambda) &= -\frac{1}{\lambda^2} \end{aligned}$$

So,

$$\mathbb{E} \left[ - \left( \frac{\partial^2}{\partial \lambda^2} \ln L(\lambda) \right) \right] = \frac{1}{\lambda^2}$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\ln L(\lambda)\right)^2\right] &= \mathbb{E}\left(\left(\frac{1}{\lambda} - T\right)^2\right) = \frac{1}{\lambda^2} - \frac{2}{\lambda}\mathbb{E}(T) + \mathbb{E}(T^2) \\ &= \frac{1}{\lambda^2} - \frac{2}{\lambda^2} + \frac{2}{\lambda^2} = \frac{1}{\lambda^2}\end{aligned}$$

4.1.2. *counter-example: uniform distribution.* For the uniform distribution (or for the triangle distribution given in the practice quiz) none of this is true!! (i.e., the two formulas for the Fisher information do not agree and there is an unbiased estimator which is more efficient than Cramér-Rao predicts is possible. This is because the uniform distribution has a distribution function whose support varies with  $\theta$ .)

The density function is  $f(y) = L(\theta) = 1/\theta$ . So,

$$\ln L(\theta) = -\ln \theta$$

$$\frac{\partial}{\partial\theta}\ln L(\theta) = -\frac{1}{\theta}$$

$$\frac{\partial^2}{\partial\theta^2}\ln L(\theta) = \frac{1}{\theta^2}$$

The two definitions of  $I(\theta)$  do not agree:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln L(\theta)\right)^2\right] = \frac{1}{\theta^2}$$

$$\mathbb{E}\left[-\left(\frac{\partial^2}{\partial\theta^2}\ln L(\theta)\right)\right] = -\frac{1}{\theta^2}$$

These are supposed to be equal but they have opposite signs. (For the triangle distribution they have different signs and they differ by a factor of 2.)

4.1.3. *proof.* Here is the proof that the two definitions of the Fisher information agree (assuming the density function is differentiable and has support independent of  $\theta$ ).

First of all,

$$\int f(x) dx = \int L(\theta) dx = 1$$

So, the derivative is zero:

$$\frac{\partial}{\partial\theta}\int L(\theta) dx = \int L'(\theta) dx = 0$$

This means the expected value of  $\frac{\partial}{\partial \theta} \ln L(\theta)$  is zero:

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial}{\partial \theta} \ln L(\theta) \right] &= \int \frac{\partial}{\partial \theta} \ln L(\theta) f(x) dx = \int \frac{L'(\theta)}{L(\theta)} f(x) dx \\ &= \int \frac{L'(\theta)}{f(x)} f(x) dx = \int L'(\theta) dx = 0 \end{aligned}$$

Differentiate by  $\theta$  again and we will still have 0:

$$\begin{aligned} \frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta} \ln L(\theta) f(x) dx &= \int \frac{\partial^2}{\partial \theta^2} \ln L(\theta) f(x) dx + \int \frac{\partial}{\partial \theta} \ln L(\theta) \frac{\partial}{\partial \theta} f(x) dx \\ &= \int \frac{\partial^2}{\partial \theta^2} \ln L(\theta) f(x) dx + \int \frac{\partial}{\partial \theta} \ln L(\theta) \frac{L'(\theta)}{f(x)} f(x) dx \\ &= \mathbb{E} \left[ \left( \frac{\partial^2}{\partial \theta^2} \ln L(\theta) \right) \right] + \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \ln L(\theta) \right)^2 \right] = 0 \end{aligned}$$

So, these two expected values have opposite sign.

#### 4.2. Cramér-Rao.

**Theorem 4.3.** *Suppose that  $\hat{\theta}$  is an unbiased estimator of  $\theta$  and suppose that the support of the density function does not depend on  $\theta$  then*

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

where  $I(\theta)$  is the Fisher information discussed above and  $n$  is the sample size.

$I(\theta)$  is the Fisher information given by a sample of 1. For a sample of size  $n$  you have  $n$  times as much information.

## 5. SUFFICIENCY

**Definition 5.1.** A statistic  $T = h(W_1, \dots, W_n)$  is defined to be sufficient for  $\theta$  if the conditional probability distribution of  $(W_1, \dots, W_n)$  given  $T$  is independent of  $\theta$ .

5.1. **Examples.** Take a deck of cards and delete all 13 cards of one suit. Pick any card. Then the suit is a sufficient statistic. (Given the suit any number in that suit is equally likely.)

The example we did in class was the uniform distribution. Then  $Y_{\max}$  is a sufficient statistic because, given the maximum value, the  $n - 1$  other values of  $Y$  must be between 0 and  $Y_{\max}$  with equal probability (making their expected value equal to  $\frac{1}{2}Y_{\max}$ ).

5.2. **Rao-Blackwell.** If you have an unbiased estimator and a sufficient estimator you can combine them to get an unbiased and sufficient estimator by the following formula.

**Theorem 5.2** (Rao-Blackwell). Suppose that  $\hat{\theta}$  is an unbiased estimator and  $T(X)$  is a sufficient statistic for  $\theta$ . Then the new estimator

$$\hat{\theta}^* := \mathbb{E}(\hat{\theta}|T)$$

is unbiased, sufficient and relatively efficient in the sense that

$$\text{Eff}(\hat{\theta}^*; \hat{\theta}) := \frac{\text{Var}(\hat{\theta})}{\text{Var}(\hat{\theta}^*)} \geq 1.$$

For example, take the uniform distribution

$$f_Y(y) = \frac{1}{\theta}, \quad 0 \leq y \leq \theta$$

Then an unbiased estimator is

$$\hat{\theta} = 2\bar{Y} = \frac{2}{n} \sum_{i=1}^n Y_i$$

and a sufficient statistic is  $T(Y) = Y_{\max}$ . By Rao-Blackwell we should combine these to get:

$$\hat{\theta}^* = \mathbb{E}(\hat{\theta}|Y_{\max}) = \frac{2}{n} \sum_{i=1}^n \mathbb{E}(Y_i|Y_{\max})$$

However, one of the  $Y_i$  must be equal to the maximum and the other  $n - 1$   $Y_i$  are, on the average, half the maximum. Thus

$$\hat{\theta}^* = \frac{2}{n} \left( Y_{\max} + (n - 1) \frac{1}{2} Y_{\max} \right) = \frac{n + 1}{n} Y_{\max}.$$

We verified in class that this is unbiased and more efficient than  $\hat{\theta}$ .