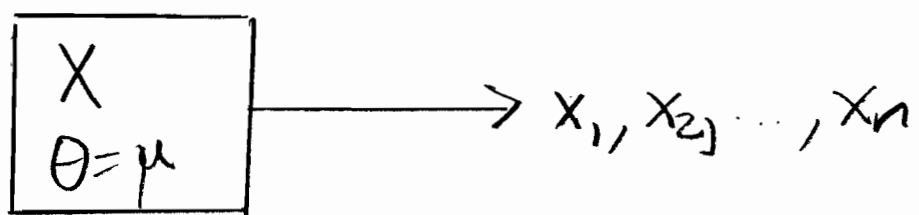


# Lecture 23

## How to find estimators §6.2

We have been discussing the problem of estimating an unknown parameter  $\theta$  in a probability distribution if we are given a sample  $x_1, x_2, \dots, x_n$  from that distribution. We introduced two examples.



Use the sample mean  $\bar{x} = \frac{x_1 + \dots + x_n}{n}$  to estimate population mean  $\mu$ .  
 $\bar{x}$  is an unbiased estimator of  $\mu$ .

Also we had the more subtle problem of estimating  $\theta$  in  $U(0, \theta)$

$$\boxed{X \sim U(0, \theta)} \quad \theta = \theta \quad \rightarrow$$

$$W = \frac{n+1}{n} \max(x_1, x_2, \dots, x_n)$$

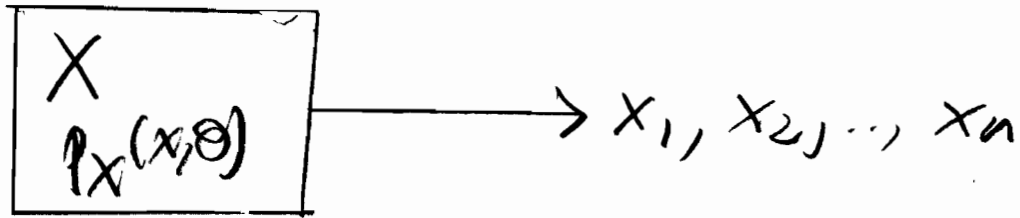
is an unbiased estimator of  $\theta$ .

We discussed two desirable properties of estimators

(i) unbiased

(ii) minimum variance..

the general problem. Given



How do you find an estimator

$$\hat{\theta} = h(x_1, x_2, \dots, x_n) \text{ for } \theta ?$$

There are two methods.

- (i) The method of moments
- (ii) The method of maximum likelihood.

# The Method of Moments

4

Definition 1 Let  $k$  be a nonnegative integer, and  $X$  be a random variable. Then the  $k$ -th moment  $m_k(X)$  of  $X$  is given by

$$m_k(X) = E(X^k), \quad k \geq 0.$$

so  $m_0(X) = 1$

$$m_1(X) = E(X) = \mu$$

$$m_2(X) = E(X^2) = \sigma^2 + \mu^2$$

## Definition 2

Let  $x_1, x_2, \dots, x_n$  be a sample from  $X$ .

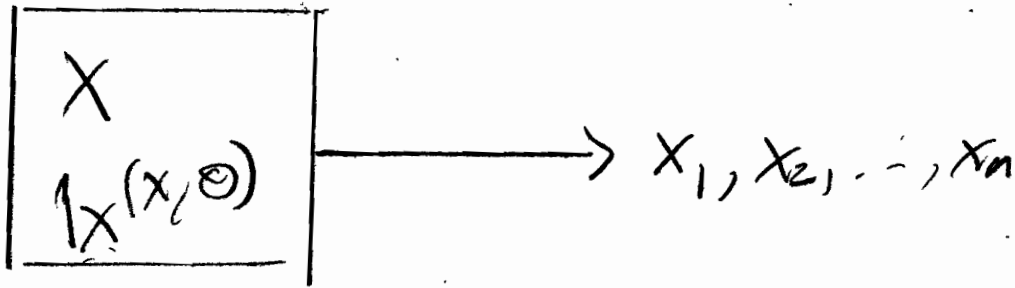
Then the  $k$ -th sample moment  $S_k$  is

$$S_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad \text{so } S_1 = \bar{x}$$

# Key Point

Given

5



The  $k$ -th moment  $m_k(X)$  (k-th population moment) depends on  $\theta$  whereas the  $k$ -th sample moment does not - it is just the average sum of powers of the  $x_i$ 's.

The method of moments says

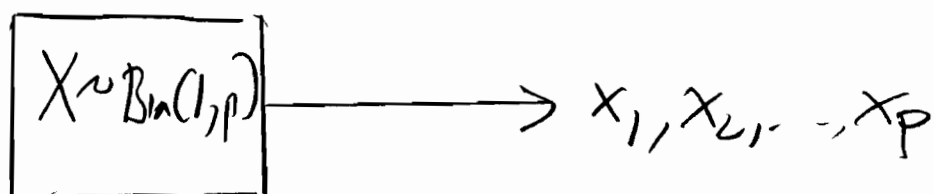
- (i) Equate the  $k$ -th population moment  $m_k(X)$  to the  $k$ -th sample moment  $S_k$ .

(ii) Solve the resulting system of equations for  $\theta$

(\*)  $m_k(X) = S_k, 1 \leq k < \infty$

We will denote the answer by  $\hat{\theta}_{mme}$

Example Estimating p in a Bernoulli distribution



The first population moment  $m_1(X)$  is the mean  $E(X) = p = \theta$

The first sample moment  $S_1$  is the sample mean so looking at the first equation of (\*)

$m_1(X) = S_1, \text{ so } p = \bar{x}$

gives us the sample mean as an estimator for p

7

Recall that because the  $x_i$ 's are all either 1 or zero

$$x_1 + \dots + x_n = \# \text{ of successes}$$

$$\text{and } \bar{X} = \frac{\# \text{ of successes}}{n}$$

$$\hat{p}_{\text{mme}} = \bar{X} = \text{the sample proportion}$$

## Example 2

The method of moments works well when you have several unknown parameters

Suppose we want to estimate both

the mean  $\mu$  and the variance  $\sigma^2$

from a normal distribution (or any distribution)

$$X \sim N(\mu, \sigma^2)$$

8

We equate the first two population moments to the first two sample moments

$$m_1(X) = S_1$$

$$m_2(X) = S_2$$

so

$$\mu = \bar{X}$$

$$\sigma^2 + \mu^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$$

Solving (we get  $\mu$  for free,  $\hat{\mu}_{MLE} = \bar{X}$ )

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \mu^2$$

$$= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left( \frac{\sum x_i}{n} \right)^2$$

$$= \frac{1}{n} \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum x_i)^2 \right)$$



So

$$\hat{\sigma}_{\text{mme}}^2 = \frac{1}{n} \left( \sum X_i^2 - \frac{(\sum X_i)^2}{n} \right)$$

Actually the best estimator for  $\sigma^2$  is the sample variance

$$S^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - \frac{(\sum X_i)^2}{n} \right)$$

$\hat{\sigma}_{\text{mme}}^2$  is a biased estimator.

### Example 3

Estimating B in  $U(0, B)$

Recall that we come up with the unbiased estimator

$$\hat{B} = \frac{n+1}{n} \max(x_1, x_2, \dots, x_n)$$

Put  $W = \max(x_1, \dots, x_{n+1})$

What do we get from the Method of Moments?

$$\boxed{X \sim U(0, B)} \rightarrow x_1, x_2, \dots, x_n$$

$$\text{Then } E(X) = \frac{0+B}{2} = \frac{B}{2}$$

So equating the first population moment  $m_1(X) = \mu$  to the first sample moment  $S_1 = \bar{x}$  we get

$$\frac{B}{2} = \bar{x}$$

$$\text{so } B = 2\bar{x} \quad \text{and } \hat{B}_{\text{mme}} = 2\bar{X}$$

This is unbiased because

$$E(\bar{X}) = \text{population mean} = \frac{B}{2}$$

$$\text{so } E(2\bar{X}) = B$$

So we have a new unbiased estimator

$$\hat{\beta}_1 = \hat{\beta}_{\text{MLE}} = 2\bar{X}.$$

Recall the other was

$$\hat{\beta}_2 = \frac{n+1}{n} W.$$

where  $W = \text{Max}(X_1, \dots, X_n)$

Which one is better?

We will interpret this to mean "which one has the smaller variance"?

$$\underline{V(\hat{B}_1) = V(2\bar{X})}$$

12

Recall from the Distribution

Handout that  $X \sim U(A, B)$

$$\Rightarrow V(X) = \frac{(B-A)^2}{12}$$

Now  $X \sim U(0, B)$  so

$$V(X) = \frac{B^2}{12}$$

This is the population variance

We also know

$$V(\bar{X}) = \frac{\sigma^2}{n} = \frac{\text{Population Variance}}{n}$$

$$\text{so } V(\bar{X}) = \frac{B^2}{12n}$$

$$\text{Then } V(\hat{B}_1) = V(2\bar{X}) = 4 \frac{B^2}{12n} = \frac{B^2}{3n}$$

$$V(\hat{B}_2) = V\left(\frac{n+1}{n} \text{Max}(X_1, \dots, X_n)\right) \quad 13$$

We have  $W = \text{Max}(X_1, X_2, \dots, X_n)$

We have from Problem 32, pg 252

$$E(W) = \frac{n}{n+1} B$$

or d

$$f_W(w) = \begin{cases} \frac{nw^{n-1}}{B^n}, & 0 \leq w \leq B \\ 0, & \text{otherwise} \end{cases}$$

Hence

$$E(W^2) = \int_0^B w^2 \frac{nw^{n-1}}{B^n} dw = \frac{n}{B^n} \int_0^B w^{n+1} dw$$

$$= \frac{n}{B^n} \left( \frac{w^{n+2}}{n+2} \right) \Big|_{w=0}^{w=B} = \frac{n}{n+2} B^2$$

Hence

14

$$\begin{aligned}V(W) &= E(W^2) - E(W)^2 \\&= \frac{n}{n+2} B^2 - \left(\frac{n}{n+1} B\right)^2 \\&= B^2 \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) \\&= B^2 \left(\frac{n(n+1)^2 - n^2(n+2)}{(n+1)^2(n+2)}\right) \\&= B^2 \left(\frac{n^3 + 2n^2 + n - n^3 - 2n^2}{(n+1)^2(n+2)}\right) \\&= \frac{n}{(n+1)^2(n+2)} B^2\end{aligned}$$

$$\begin{aligned}V(\hat{B}_2) &= V\left(\frac{n+1}{n} W\right) = \frac{(n+1)^2}{n^2} V(W) \\&= \frac{(n+1)^2}{n^2} \frac{n}{(n+1)^2(n+2)} B^2 = \frac{1}{n(n+2)} B^2\end{aligned}$$

$\hat{B}_2$  is the winner because 15

$n \geq 1$  . If  $n=1$  they tie

but of course  $n \gg 1$  so  $\hat{B}_2$

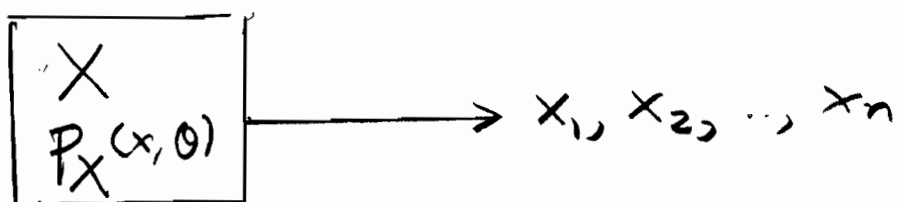
is a lot better.

# The Method of Maximum

16

## Likelihood (a brilliant idea)

Suppose we have an actual sample  $x_1, x_2, \dots, x_n$  from the space of a discrete random variable  $X$  whose pmf  $P_X(x, \theta)$  depends on an unknown parameter  $\theta$ .



What is the probability  $P$  of getting the sample  $x_1, x_2, \dots, x_n$  that we actually obtained. It is

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

by independence

$$= P(X_1 = x_1) P(X_2 = x_2) \dots P(X_n = x_n)$$



But since  $X_1, X_2, \dots, X_n$   
 are samples from  $X$  they have  
 the same pmf's as  $X$  so

$$P(X_1 = x_1) = P(X = x_1) = P_X(x_1, \theta)$$

$$P(X_2 = x_2) = P(X = x_2) = P_X(x_2, \theta)$$

⋮

$$P(X_n = x_n) = P(X = x_n) = P_X(x_n, \theta)$$

Hence

$$P = P_X(x_1, \theta) P_X(x_2, \theta) \dots P_X(x_n, \theta)$$

$P$  is a function of  $\theta$ , it is  
 called the likelihood function

and denoted  $L(\theta)$  - it is the  
 likelihood of getting the sample  
 we actually obtained.

Note,  $\theta$  is unknown but  $x_1, x_2, \dots, x_n$  are known (given).

So what is the best guess for  $\theta$

- the number that maximizes the probability of getting the sample we actually observed. This is the value of  $\theta$  that is most compatible with the observed data,

### Bottom Line

Find the value of  $\theta$  that maximizes the likelihood function  $L(\theta)$

This is the "method of maximum likelihood".

The resulting estimator will be called the maximum likelihood estimator, abbreviated mle and denoted  $\hat{\theta}_{mle}$ .

### Remark (We will be lazy)

In doing problems, following the text, we won't really maximize  $L(\theta)$  we will just find a critical point of  $L(\theta)$  i.e. a point where  $L'(\theta)$  is zero. Later in your career if you have to do this you should check that the critical point is indeed a maximum.

# Examples

1. The mle for  $p$  in  $\text{Bin}(1, p)$

$X \sim \text{Bin}(1, p)$  means the

pmf of  $X$  is

$x$	0	1
$P(X=x)$	$1-p$	$p$

There is a simple formula for this

$$P_X(x) = p^x (1-p)^{1-x}, \quad x=0, 1$$

Now since  $p$  is our unknown parameter  $\theta$  we write

$$P_X(x, \theta) = \theta^x (1-\theta)^{1-x}, \quad x=0, 1.$$

so

$$P_X(x_1, \theta) = \theta^{x_1} (1-\theta)^{1-x_1}$$

$\vdots$

$$P_X(x_n, \theta) = \theta^{x_n} (1-\theta)^{1-x_n}$$

Hence

$$L(\theta) = P_X(x_1, \theta) \cdots P_X(x_n, \theta)$$

and hence

$$L(\theta) = \underbrace{\theta^{x_1} (1-\theta)^{1-x_1} \theta^{x_2} (1-\theta)^{1-x_2} \cdots \theta^{x_n} (1-\theta)^{1-x_n}}_{\text{positive number}}$$

Now we want to

1. Compute  $L'(\theta)$

2. Set  $L'(\theta) = 0$  and solve for  $\theta$  in terms of  $x_1, x_2, \dots, x_n$  } (\*)

We can make things much simpler by using the following trick.

Suppose  $f(x)$  is a real valued function that only takes positive values

Put  $h(x) = \ln f(x)$  ↙ chain rule

$$\text{Then } h'(x) = \frac{d}{dx} \ln f(x) = \frac{1}{f(x)} \frac{df}{dx} = \frac{f'(x)}{f(x)}$$

So the critical points of  $h$  are the same points as those of  $f$

$$h'(x) = 0 \Leftrightarrow \frac{f'(x)}{f(x)} = 0 \Leftrightarrow f'(x) = 0$$

(Also  $h$  takes a maximum value at  $x_*$   $\Leftrightarrow$   $f$  takes a maximum value at  $x_*$ . This is because  $\ln$  is an increasing function so it preserves order relations. ( $a < b \Leftrightarrow \ln a < \ln b$ , here we assume  $a > 0$  and  $b > 0$ )

Bottom Line

Change  $(x)$  to  $(x_*)$ .

1. Compute  $h(\theta) = \ln L(\theta)$ .
2. Compute  $h'(\theta)$
3. Set  $h'(\theta) = 0$  and solve for  $\theta$  in terms of  $x_1, x_2, \dots, x_n$

Now back to Bin(1, p)

$$\begin{aligned}
 L(\theta) &= \theta^{x_1} (1-\theta)^{1-x_1} \dots \theta^{x_n} (1-\theta)^{1-x_n} \\
 &\stackrel{\text{rearrange}}{=} \theta^{x_1} \theta^{x_2} \dots \theta^{x_n} (1-\theta)^{1-x_1} (1-\theta)^{1-x_2} \dots (1-\theta)^{1-x_n} \\
 &= \theta^{x_1+x_2+\dots+x_n} (1-\theta)^{n-(x_1+x_2+\dots+x_n)}
 \end{aligned}$$

Now take the natural logarithm

$$h(\theta) = \ln L(\theta) = (x_1 + \dots + x_n) \ln \theta + (n - (x_1 + \dots + x_n)) \ln(1-\theta)$$

Now apply  $\frac{d}{d\theta}$  to each side using

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

so

$$h'(\theta) = \frac{x_1 + \dots + x_n}{\theta} - \frac{n - (x_1 + \dots + x_n)}{1 - \theta}$$

So we have to solve  $h'(\theta) = 0$  or

$$\frac{x_1 + \dots + x_n}{\theta} = \frac{n - (x_1 + \dots + x_n)}{1 - \theta}$$

$$(1 - \theta)(x_1 + \dots + x_n) = \theta(n - (x_1 + \dots + x_n))$$

$$x_1 + \dots + x_n - \theta(x_1 + \dots + x_n) = n\theta - \theta(x_1 + \dots + x_n)$$

$$x_1 + \dots + x_n = n\theta$$

$$\theta = \frac{x_1 + \dots + x_n}{n} = \bar{x}$$

so  $\hat{\theta}_{mle} = \bar{X}$



2. The mle for  $\lambda$  in  $\text{Exp}(\lambda)$

$$\boxed{\begin{array}{l} X \sim \text{Exp}(\lambda) \\ \lambda = \theta \end{array}} \longrightarrow x_1, x_2, \dots, x_n$$

We have

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Now we have a continuous distribution

We define  $L(\theta)$  by

$$L(\theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$

and proceed as before.

$L(\theta)$  no longer has a nice interpretation

Let's try to guess the answer. We have  $E(X) = \mu = \frac{1}{\lambda}$  and we know that  $\bar{x}$  is the best estimator for  $\mu$  so it is reasonable to guess the best estimator for  $\lambda = \frac{1}{\mu}$  will be  $\frac{1}{\bar{x}}$ .

This is far from correct logically but it helps to know where you are going.

Away we go. - let's not bother changing  $\lambda$  to  $\theta$ .

$$L(\lambda) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \dots \lambda e^{-\lambda x_n}$$

$$= \lambda^n e^{-\lambda x_1 - \lambda x_2 - \dots - \lambda x_n}$$

$$L(\lambda) = \lambda^n e^{-\lambda(x_1 + \dots + x_n)}$$

Now we suspect we are looking for a function of  $\bar{x}$  so lets use

$$x_1 + x_2 + \dots + x_n = n\bar{x}$$

(sum = n average)

to obtain

$$L(\lambda) = \lambda^n e^{-\lambda n \bar{x}}$$

Once again it helps to take the natural logarithm

$$\begin{aligned} h(\lambda) &= \ln L(\lambda) = \ln(\lambda^n e^{-\lambda n \bar{x}}) \\ &= \ln \lambda^n + \ln e^{-\lambda n \bar{x}} \end{aligned}$$

$$h(\lambda) = n \ln \lambda - \lambda n \bar{x}$$

Now  $h'(\lambda) = \frac{n}{\lambda} - n\bar{x}$  so

$$h'(\lambda) = 0 \Leftrightarrow \frac{n}{\lambda} = n\bar{x} \Leftrightarrow \lambda = \frac{1}{\bar{x}}$$

Hence  $\hat{\lambda}_{mle} = \frac{1}{\bar{X}}$

### Problem

What if we wanted the mle of  $\lambda^2$  instead of. The answer would be

$$\hat{\lambda^2}_{mle} = \frac{1}{\bar{X}^2} \lambda^2$$

by the

# Invariance Principle

29

Suppose we are given a sample  $x_1, x_2, \dots, x_n$  from a probability distribution whose pdf (or pmf) depends on  $k$  unknown parameters  $\theta_1, \theta_2, \dots, \theta_k$ . Suppose we have computed the mle's  $(\hat{\theta}_1)_{mle}$

---  $(\hat{\theta}_k)_{mle}$  of these parameters in terms of  $x_1, x_2, \dots, x_n$ . Then the mle of

$$h(\theta_1, \theta_2, \dots, \theta_k) \text{ is } h((\hat{\theta}_1)_{mle}, \dots, (\hat{\theta}_k)_{mle})$$

or

$$h(\theta_1, \dots, \theta_k)_{mle} = h((\hat{\theta}_1)_{mle}, \dots, (\hat{\theta}_k)_{mle})$$

## One more example

In Example 6.17 of the text it is shown that

$$\hat{\sigma}^2_{mle} = \frac{1}{n} \left( \sum X_i^2 - \frac{(\sum X_i)^2}{n} \right) = \hat{\sigma}^2_{mme}$$

Hence 
$$\hat{\sigma}_{mle} = \sqrt{\frac{1}{n} \sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

(here  $h(\theta) = \sqrt{\theta}$  and  $\theta = \sigma^2$ )