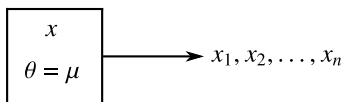


Lecture 23: How to find estimators §6.2

We have been discussing the problem of estimating on unknown parameter θ 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 μ . \bar{X} is an unbiased estimator of μ .

Also we had the more subtle problem of estimators B in $U(0, B)$

$$\boxed{X \sim \bigcup_{\theta=B} (0, B)} \quad \text{-----} \rightarrow$$

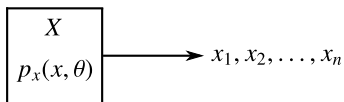
$$W = \frac{n+1}{n} \max(x_1, x_2, \dots, x_n)$$

is an unbiased estimators of $\theta = B$.

We discussed two desirable properties of estimators

- (i) unbiased
- (ii) minimum variance

the general problems. Given



How do you find an estimator $\hat{\theta} = h(x_1, x_2, \dots, x_n)$ for θ ?

There are two methods.

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

The Method of Moments

Definition 1

Let k be a non negative 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 $m_0(X) = 1$

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

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

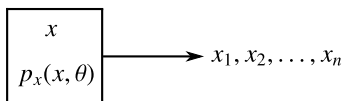
Definition 2

Let x_1, x_2, \dots, x_n be a sample from X . Then the k -th sample moment S_k is

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

Key Point

Given



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

The method of moments says

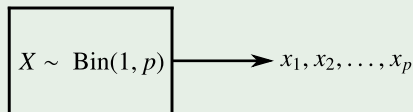
- (i) Equate the k -th population moment $m_k(X)$ to the k -th sample moment S_k .
- (ii) Solve the resulting system of equations for θ .

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

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

Example 1

Estimating P in a Bernoulli distribution



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

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

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

gives us the sample mean as an estimator for p

Example 1 (Cont.)

Recall that because the x 's are all either 1 or zero $x_1 + \dots + x_n = \#$ of successes and

$$\begin{aligned}\bar{x} &= \frac{\# \text{ of successes}}{n} \\ &= \text{the sample proportion} \\ \hat{p}_{mme} &= \bar{X}\end{aligned}$$

Example 2

The method of moments works well when you have several unknown parameters. Suppose we want to estimate *both* the mean μ and the variance σ^2 from a normal distribution (or any distribution)

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

Example 2 (Cont.)

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 μ for free, $\hat{\mu}_{mme} = \bar{X}$)

$$\begin{aligned}\sigma^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \mu^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{\sum x_i}{n} \right)^2 \\ &= \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum x_i)^2 \right)\end{aligned}$$

Example 2 (Cont.)

So

$$\widehat{\sigma^2}_{mme} = \frac{1}{n} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right)$$

Actually the best estimator for σ^2 is the sample variance

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

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

Example 3

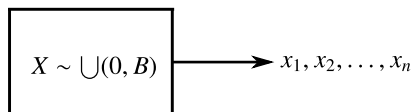
Estimating B in $U(0, B)$

Recall that we come up with the unbiased estimator

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

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

What do we get from the Method of Moments ?



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

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

$$\begin{aligned} \frac{B}{2} &= \bar{x} \\ \text{so } B &= 2\bar{x} \text{ and } \hat{B}_{mme} = 2\bar{X} \end{aligned}$$

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

$$\hat{B}_1 = \hat{B}_{mme} = 2\bar{X}.$$

Recall the other was

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

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”?

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

Recall from the Distribution Hand out 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

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

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

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

$$V(B_2) = V\left(\frac{n+1}{n} \text{Max}(X_1, \dots, X_n)\right)$$

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

We have from Problem 32, pg 252

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

and

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

Hence

$$\begin{aligned} E(W^2) &= \int_0^B w^2 \frac{nw^{n-1}}{B^n} dw = \frac{n}{B^n} \int_0^B w^{n+1} dw \\ &= \frac{n}{B^n} \left(\frac{W^{n+2}}{n+2} \right) \Big|_{w=0}^{w=B} = \frac{n}{n+2} B^2 \end{aligned}$$

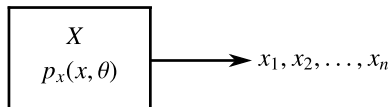
Hence

$$\begin{aligned}V(W) &= E(W^2) - E(W)^2 \\&= \frac{n}{n+2}B^2 - \left(\frac{n}{n+1}B\right)^2 \\&= B^2 \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} \right) \\&= B^2 \left(\frac{n(n+1)^2 - n^2(n+2)}{(n+1)^2(n+2)} \right) \\&= B^2 \left(\frac{n^3 + 2n^2 + n - n^3 - 2n^2}{(n+1)^2(n+2)} \right) \\&= \frac{n}{(n+1)^2(n+2)}B^2 \\V(\hat{B}_2) &= V\left(\frac{n+1}{n}W\right) = \frac{(n+1)^2}{n^2}V(W) \\&= \frac{\cancel{(n+1)^2}}{n^2} \frac{n}{\cancel{(n+1)^2}(n+2)}B^2 = \frac{1}{n(n+2)}B^2\end{aligned}$$

\hat{B}_2 is the winner because $n \geq 1$. If $n = 1$ they tie but of course $n \gg 1$ so \hat{B}_2 is a lot better.

The Method of Maximum 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 pdf $p_X(x, \theta)$ depends on an unknown parameter θ .



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 sample proof's as X so

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

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

$$\vdots$$

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

Hence

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

P is a function of θ , it is called the likelihood function and denoted L_θ -it is the likelihood of getting the sample we actually obtained.

Note, θ is unknown but x_1, x_2, \dots, x_n are known (given). So what is the best guess for θ - the number that maximizes the probability of getting the sample we actually observed. *This is the value of θ that is most compatible with the observed data.*

Bottom Line

Find the value of θ 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 $\hat{\theta}_{\text{mle}}$.

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 $L(\theta)$ ie. 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

$$p_X(x) = p^x(1-p)^{1-x}, x = 0, 1$$

Now since p is our unknown parameter θ we write

$$p_X(x, \theta) = \theta^x(1-\theta)^{1-x}, x = 0, 1$$

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

$$L(\theta) = p_X(x_1, \theta) \dots p_X(x_n, \theta)$$

and hence

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

Now we want to

$$\left. \begin{array}{l} 1. \text{ Compute } L'(\theta) \\ 2. \text{ Set } L'(\theta) = 0 \text{ and solve for } \\ \quad \theta \text{ in terms of } x_1, x_2, \dots, x_n \end{array} \right\} \quad (*)$$

We can make things much simpler by using the following trick. Suppose $f(x)$ is a real valued function that only takes positive value.

Put $h(x) = \ln f(x)$

Then $h'(x) = \frac{d}{dx} \ln f(x) \overset{\substack{\text{chain rule} \\ \swarrow}}{=} \frac{1}{f(x)} \frac{df}{dx} = \frac{f'(x)}{f(x)}$

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

$$h^1(x) = 0 \Leftrightarrow \frac{f'(x)}{f(x)} = 0 \Leftrightarrow f'(x) = 0$$

Also h takes a maximum value of x_* $\Leftrightarrow f$ takes a maximum value at x_* . This is because \ln is an increasing function so it preserves order relations.

($a < b \Leftrightarrow \ln a < \ln b$, have we assume $a > 0$ and $b > 0$)

Bottom Line Change (*) to (**)

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

Now back to $\text{Bin}(l, p)$

$$L(\theta) = \theta^{x_1} (1 - \theta)^{1-x_1} \dots \theta^{x_n} (1 - \theta)^{1-x_n}$$

rearrange

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

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

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

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 λ in $\text{Exp}(\lambda)$

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

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

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

and procede as before.

$L(\theta)$ nolonger 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 μ so it is reasonable to guess the best estimator for $\lambda = \frac{1}{\mu}$ will be $\frac{1}{\bar{x}}$. This is for from correct logically but *it helps to know where you are going*.
Away we go -let's not bother changing λ to θ .

$$\begin{aligned} L(\lambda) &= \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \dots \lambda e^{-\lambda x_n} \\ &= \lambda^n e^{-\lambda x_1} e^{-\lambda x_2} e^{-\lambda x_n} \\ L(\lambda) &= \lambda^n e^{-\lambda(x_1 + \dots + x_n)} \end{aligned}$$

Now we suspect we are looking for a function of \bar{x} so let's 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

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

Now

$$\begin{aligned} h'(\lambda) &= \frac{n}{\lambda} - n\bar{x} \text{ so} \\ h'(\lambda) &= 0 \Leftrightarrow \frac{n}{\lambda} = n\bar{x} \Leftrightarrow \lambda = \frac{1}{\bar{x}} \end{aligned}$$

Hence

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

Problem What if we wanted the mle of λ^2 instead of. The answer would be

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

by the

In variance Principle

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 $(\theta_1)_{mle}, \dots, (\theta_k)_{mle}$ of these parameters in terms of x_1, x_2, \dots, x_n . Then the mle of $h(\theta_1, \theta_2, \dots, \theta_k)$ 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

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

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

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