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 $x_1, x_2, \ldots, x_n$ from that distribution. We introduced two examples.

$$\boxed{\begin{array}{c} x \\ \theta = \mu \end{array}} \longrightarrow x_1, x_2, \ldots, x_n$$

Use the *sample* mean $\overline{x} = \dfrac{x_1 + \ldots + x_n}{n}$ to estimate *population* mean μ. $\overline{X}$ is an unbiased estimator of μ.

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

$$\boxed{X \sim \bigcup_{\theta=B} (0, B)} \dashrightarrow$$

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

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

We discussed two desirable properties of estimators

(i) unbiased

(ii) minimum variance

the general problems. Given

$$\boxed{\begin{matrix} X \\ p_x(x, \theta) \end{matrix}} \longrightarrow x_1, x_2, \ldots, x_n$$

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

$$
\begin{aligned}
& m_k(X) = E(X^k),\ k \geq 0 \\
\text{so} \quad & m_0(X) = 1 \\
& m_1(X) = E(X) = \mu \\
& m_2(X) = E(X^2) = \sigma^2 + \mu^2
\end{aligned}
$$

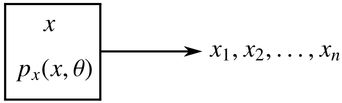
Definition 2

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

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

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

(ii) Solve the resulting system of equations for θ.

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

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

Example 1

Estimating P in a Bernoulli distribution

$$\boxed{X \sim \mathrm{Bin}(1, p)} \longrightarrow x_1, x_2, \ldots, x_p$$

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 $(*)$

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

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 $x_1 + \ldots + x_n = \neq$ of successes and

$$\overline{x} = \frac{\# \, ofsuccesses}{n}$$
$$= \text{ the sample proportion}$$
$$\hat{p}_{mme} = \overline{X}$$

Example 2

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

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

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

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

Solving (we get μ for free, $\hat{\mu}_{mme} = \overline{X}$)

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

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

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

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

What do we get from the Method of Moments ?

$$\boxed{X \sim \bigcup(0, B)} \longrightarrow x_1, x_2, \ldots, x_n$$

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

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

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

$$\text{so} \qquad B = 2\overline{x} \text{ and } \hat{B}_{mme} = 2\overline{X}$$

This is unbiased because

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

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

So we have a new unbiased estimator

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

Recall the other was

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

where $W = \text{ Max } (X_1, \ldots, X_n)$
Which one is better?
We will interpret this to mean “which one has the smaller variance”?

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

Recall from the Distribution Hard 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

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

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

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

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

We have $W = \ \text{Max}\ (X_1, X_2, \ldots, X_n)$
We have from Problem 32, pg 252

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

and

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

Hence

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

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

Hence

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

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

The Method of Maximum Likelihood (a brilliant idea)

Suppose we have an actual sample $x_1, x_2, \ldots, x_n$ from the space of a discrete random variable x whose proof $p_X(x, \theta)$ depends on an unknown parameter θ.

$$\boxed{\begin{array}{c} X \\ p_x(x,\theta) \end{array}} \longrightarrow x_1, x_2, \ldots, x_n$$

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

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

by independence

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

But since $X_1, X_2, \ldots, X_n$ are samples from X they have the sample proof's as X so

$$\begin{aligned}
P(X_1 = x_1) &= P(X = x_1) = P_X(x_1, \theta) \\
P(X_2 = x_2) &= P(X = x_2) = P_X(x_2, \theta) \\
&\vdots \\
P(X_n = x_n) &= P(X = x_n) = P_X(x_n, \theta)
\end{aligned}$$

Hence

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

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

Note, θ is unknown but $x_1, x_2, \ldots, x_n$ are known (given). So what is the nest guess for θ - the number that maximizes the probability of getting the sample use 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 cancer if your have to do this *you should check that the critical point is indeed a maximum*.

Examples

1. The mle for p in Bin$(1, p)$

$X \sim Bin(1, p)$ means the proof 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

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

$$\vdots$$

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

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

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

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

chain rule

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

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 $b > 0$)
<u>Bottom Line</u> 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 $x_1, x_2, \ldots, x_n$

Now back to $Bin(l, p)$

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

rearrange

$$= \theta^{x_1}\theta^{x_2} \ldots \theta^{x_n}(1-\theta)^{1-x_1}(1-\theta)^{1-x_2} \ldots (1-\theta)^{1-x_n}$$

$$= \theta^{x_1+x_2+\ldots+x_n}(1-\theta)^{n-(x_1+x_2+\ldots+x_n)}$$

Now take the natural logarithm

$$h(\theta) = lnL(\theta) = (x_1 + \ldots + x_n)ln\theta + (n - (x_1 + \ldots + x_n))ln(1-\theta)$$

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

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

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

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

$$\begin{aligned}
&(1 - \theta)(x_1 + \ldots + x_n) = \theta(n - (x_1 + \ldots + x_n)) \\
&x_1 + \ldots + x_n - \theta\cancel{(x_1 + \ldots + x_n)} = n\theta - \theta\cancel{(x_1 + \ldots + x_n)} \\
&x_1 + \ldots + x_n = n\theta \\
&\theta = \frac{x_1 + \ldots + x_n}{n} = \overline{x} \\
\text{so} \qquad &\hat{\theta}_{mle} = \overline{X}
\end{aligned}$$

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

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

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

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

and procede as before.
$L(\theta)$ nolonger has a nice interpretation

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

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

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

$$x_1 + x_2 + \ldots + x_n = n\overline{x}$$

(sum = n average)
to obtain

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

Once again it helps to take the notarial logarithm

$$\begin{aligned} h(\lambda) &= lnL(\lambda) = ln(\lambda^n e^{-\lambda n\overline{x}}) \\ &= ln\lambda^n + lne^{-\lambda n\overline{x}} \\ h(\lambda) &= nln\lambda - \lambda n\overline{x} \end{aligned}$$

Now

$$h'(\lambda) = \frac{n}{\lambda} - n\overline{x} \text{ so}$$

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

Hence

$$\widehat{\lambda}_{mle} = \frac{1}{\overline{X}}$$

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

$$\widehat{\lambda^2}_{mle} = \frac{1}{\overline{X}}2$$

by the

In variance Principle

Suppose we are given a sample $x_1, x_2, \ldots, x_n$ from a probability distribution whose pdf (or proof) depends on k unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$. Suppose we have computed the mle's $(\theta\theta_1)_{mle's} \ldots (\hat{\theta}_k)_{mle}$ of these parameters in terms of $x_1, x_2, \ldots, x_n$. Then the mle of $h(\theta_1, \theta_2, \ldots, \theta_n)$ is $h\left((\hat{\theta}_1)_{mles} \ldots, (\hat{\theta}_k)_{mle}\right)$ or

$$h(\theta_1, \widehat{\ldots, \theta}_k)_{mle} = h\left((\hat{\theta}_1)_{mle}, \ldots, (\hat{\theta}_k)_{mle}\right)$$

One more example

In Example 6.17 of the text if is shown that

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

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

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