14.30 Introduction to Statistical Methods in Economics

Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 19

Konrad Menzel

April 28, 2009

1 Maximum Likelihood Estimation: Further Examples

Example 1 *Suppose $X \sim N(\mu_0, \sigma_0^2)$, and we want to estimate the parameters μ and σ^2 from an i.i.d. sample $X_1, \ldots, X_n$. The likelihood function is*

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

It turns out that it's much easier to maximize the log-likelihood,

$$\begin{aligned}
\log \mathcal{L}(\theta) &= \sum_{i=1}^{n} \log \left\{ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \right\} \\
&= \sum_{i=1}^{n} \left\{ \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(X_i-\mu)^2}{2\sigma^2} \right\} \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2
\end{aligned}$$

In order to find the maximum, we take the derivatives with respect to μ and σ^2, and set them equal to zero:

$$0 = \frac{1}{2\widehat{\sigma^2}} \sum_{i=1}^{n} 2(X_i - \hat{\mu}) \Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Similarly,

$$0 = -\frac{n}{2} \frac{2\pi}{2\pi\widehat{\sigma^2}} + \frac{1}{2\left(\widehat{\sigma^2}\right)^2} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \Leftrightarrow \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Recall that we already showed that this estimator is not unbiased for σ_0^2, so in general, Maximum Likelihood Estimators need not be unbiased.

Example 2 *Going back to the example with the uniform distribution, suppose $X \sim U[0, \theta]$, and we are interested in estimating θ. For the method of moments estimator, you can see that*

$$\mu_1(\theta) = \mathbb{E}_\theta[X] = \frac{\theta}{2}$$

so equating this with the sample mean, we obtain

$$\hat{\theta}_{MoM} = 2\bar{X}_n$$

What is the maximum likelihood estimator? Clearly, we wouldn't pick any $\hat{\theta} \leq \max\{X_1, \ldots, X_n\}$ *because a sample with realizations greater than* $\hat{\theta}$ *has zero probability under* $\hat{\theta}$*. Formally, the likelihood is*

$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } 0 \leq X_i \leq \theta \text{ for all } i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

We can see that any value of $\theta \leq \max\{X_1, \ldots, X_n\}$ *can't be a maximum because* $L(\theta)$ *is zero for all those points. Also, for* $\theta \geq \max\{X_1, \ldots, X_n\}$ *the likelihood function is strictly decreasing in* θ*, and therefore, it is maximized at*

$$\hat{\theta}_{MLE} = \max\{X_1, \ldots, X_n\}$$

Note that since $X_i < \theta_0$ *with probability 1, the Maximum Likelihood estimator is also going to be less than* θ_0 *with probability one, so it's not unbiased. More specifically, the p.d.f. of* $X_{(n)}$ *is given by*

$$f_{X_{(n)}}(y) = n[F_X(y)]^{n-1} f_X(y) = \begin{cases} \frac{n}{\theta_0}\left(\frac{1}{\theta_0}\frac{y}{\theta_0}\right)^{n-1} & \text{if } 0 \leq y \leq \theta_0 \\ 0 & \text{otherwise} \end{cases}$$

so that

$$\mathbb{E}[X_{(n)}] = \int_{-\infty}^{\infty} y f_{X_{(n)}}(y) dy = \int_0^{\theta_0} n\left(\frac{y}{\theta_0}\right)^n dy = \frac{n}{n+1}\theta_0$$

We could easily construct an unbiased estimator $\tilde{\theta} = \frac{n+1}{n} X_{(n)}$*.*

1.1 Properties of the MLE

The following is just a summary of main theoretical results on MLE (won't do proofs for now)

- If there is an efficient estimator in the class of consistent estimators, MLE will produce it.
- Under certain regularity conditions, MLE's will have an asymptotically normal distribution (this comes essentially from an application of the Central Limit Theorem)

Is Maximum Likelihood always the best thing to do? - not necessarily

- may be biased
- often hard to compute
- might be sensitive to incorrect assumptions on underlying distribution

2 Confidence Intervals

In order to combine information about the value of the estimate and its precision (as e.g. given by its standard error), what is often done is to report an interval around the estimate which is likely going to contain the actual value.

Example 3 *Suppose the captain of a Navy gunboat has to establish a beachhead on a stretch of shoreline, but first has to make sure that a battery on the beach - which can't be seen directly from the sea - is destroyed, or at least severely damaged.*
The boat already took some fire from the coast, and based on the direction the projectiles came from, the captain has an estimate $\hat{\theta}$ of the position of the battery, which is normally distributed with variance σ_β^2 around the true position θ_0.
The captain can fire a volley of missiles on a range of the beach, making sure that everything in that range gets destroyed. How can the captain determine what range of the shore to fire at so that he can be 95% sure that the battery will be destroyed so that it will be safe to land troops?

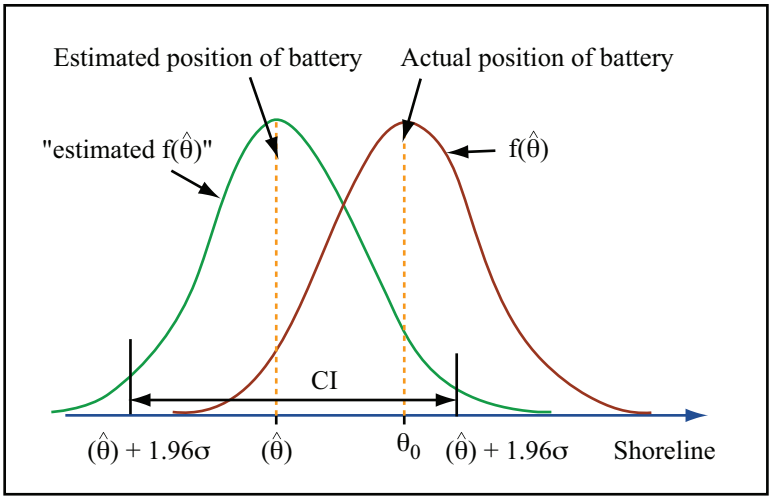

Image by MIT OpenCourseWare.

For the normal distribution, we know that 95% of the probability mass is within 1.96 standard deviations on either side of the mean. So if the captain orders to fire at the range $CI = [\hat{\theta} - 1.96\sigma, \hat{\theta} + 1.96\sigma]$, the probability that $\hat{\theta}$ will be such that $\theta_0 \in CI$ equals 95%.

So while earlier on we were only looking for a single function $\hat{\theta}(X_1, \ldots, X_n)$ which gives a value close to the actual parameter value θ_0, we will now try to construct two functions $A(X_1, \ldots, X_n) < B(X_1, \ldots, X_n)$ such that the two functions enclose the true parameter with probability greater or equal to some prespecified level.

Definition 1 *A $1-\alpha$ confidence interval for the parameter θ_0 is an interval $[A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)]$ depending on two data-dependent functions $A(\cdot)$ and $B(\cdot)$ such that*

$$P_{\theta_0}\left(A(X_1, \ldots, X_n) \leq \theta_0 \leq B(X_1, \ldots, X_n)\right) = 1 - \alpha$$

Typically, these functions are not unique, but by convention, we choose A and B such that $\frac{\alpha}{2}$ probability falls on each side of the interval.

For a *realization* of the confidence interval, $[A(x_1, \ldots, x_n), B(x_1, \ldots, x_n)]$, it doesn't make sense to say that $P(A(x_1, \ldots, x_n) \leq \theta_0 \leq B(x_1, \ldots, x_n)) = 1 - \alpha$, since now both the limits of the interval and the true parameter are just real numbers, so given a realization of the sample, the estimated interval either covers θ_0 (with probability 1, if you will), or it doesn't. We should be clear that it is the confidence interval (i.e. the functions $A(\cdot)$ and $B(\cdot)$) which are random given the true parameter, not θ_0.

The following is the most common case for which we'd like to construct a confidence interval.

Example 4 *Suppose $\hat{\theta} \sim N(\theta_0, \sigma^2)$, and we want to construct a $1-\alpha$ confidence interval. If $z_{1-\alpha/2}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution, i.e. $\Phi(z_{1-\alpha/2}) = 1-\frac{\alpha}{2}$, then we can check that*

$$CI = [\hat{\theta} - \sigma z_{1-\alpha/2}, \hat{\theta} + \sigma z_{1-\alpha/2}]$$

covers θ_0 with probability

$$\begin{aligned} P_{\theta_0}\left(\hat{\theta} - \sigma z_{1-\alpha/2} \leq \theta_0 \hat{\theta} + \sigma z_{1-\alpha/2}\right) &= P_{\theta_0}\left(-z_{1-\alpha/2} \leq \frac{\theta_0 - \hat{\theta}}{\sigma} \leq z_{1-\alpha/2}\right) \\ &= \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha \end{aligned}$$

since $\frac{\theta_0 - \hat{\theta}}{\sigma}$ is the standardization of $\hat{\theta}$, and therefore follows a standard normal distribution.
So if we want a 95% confidence interval, $z_{1-\alpha/2} = z_{0.975} = 1.96$, so the confidence interval is given by $\hat{\theta} \pm 1.96\sigma$.
This is the most commonly used way of obtaining confidence intervals, so you should make sure that you understand how this works.

Example 5 *Poll results are often reported with a "margin of error". E.g. the Gallup report for April 18* [1] *says that 46% of voters would vote for Clinton over McCain, 44% would vote for McCain, and 10% would choose neither or had no opinion. These results were based on 4,385 interviews, and the report goes on to state that "For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is 2 percentage points."*
What does this mean? - if the true vote share of a candidate is p, the variance of the average share in a sample of n voters would be $\mathrm{Var}(\bar{X}_n) = \frac{p(1-p)}{n}$, and you can verify that this variance is highest for $p = 0.5$. So for a sample of 4,385 interviewees, the maximal standard deviation is $\sqrt{\mathrm{Var}(\bar{X}_n)} \leq \sqrt{\frac{0.5 \cdot 0.5}{4385}} \approx 0.76\%$. By the Central Limit Theorem, $\bar{X}_n$ is approximately normally distributed, and we already saw that for a normal distribution, 95% of the probability mass is within 1.96 standard deviations of the mean. Therefore the interval $[\bar{X}_n - 1.96 \cdot 0.76\%, \bar{X}_n - 1.96 \cdot 0.76\%]$ will cover the true vote share with probability greater than 95%. For smaller subgroups of voters, the margin of error becomes larger.

Example 6 *A lab carries out a chemical analysis on blood to be used as evidence in a trial. To be acceptable as evidence, a 90% confidence interval for the amount of some substance should have length less than $0.001 g/ml$. The machine used for the analysis gives readings which are normally distributed around the true value with standard deviation $\sigma = 0.005 g/ml$. How many readings do we need in order to make sure that the 90% confidence interval is shorter than $0.001 g/ml$?*
The width of a 90% CI is

$$l = 2\frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.95) \approx 2\frac{0.005}{\sqrt{n}}1.645 = \frac{0.01645}{\sqrt{n}}$$

Therefore, in order for $l \leq 0.001$, we need $n \geq 16.45^2 = 270.6025$, so we'd need at least 271 (independent) readings.

The following example illustrates one way of constructing a confidence interval when the distribution of the estimator is not normal.

[1] http://www.gallup.com/poll/106630/Gallulp-Daily-Clinton-Moves-Within-Points-Obama.aspx

Example 7 *Suppose $X_1, \ldots, X_n$ are i.i.d. with $X_i \sim U[0, \theta]$, and we want to construct a 90% confidence interval for θ_0. Let*

$$\hat{\theta} = \max\{X_1, \ldots, X_n\} = X_{(n)}$$

the nth order statistic (as we showed last time, this is also the maximum-likelihood estimator). Even though, as we saw, $\hat{\theta}$ is not unbiased for θ, we can use it to construct a confidence interval for θ.
From results for order statistics, we saw that the c.d.f. of $\hat{\theta}$ is given by the c.d.f. of $\hat{\theta}$ is given by

$$F_{\hat{\theta}}(\theta) = \begin{cases} 0 & \theta \leq 0 \\ \left(\frac{\theta}{\theta_0}\right)^n & \text{if } 0 < \theta \leq \theta_0 \\ 1 & \text{if } \theta > \theta_0 \end{cases}$$

where we plugged in the c.d.f. of a $U[0, \theta_0]$ random variable, $F(x) = \frac{x}{\theta_0}$.
In order to obtain the functions for A and B, let us first find constants a and b such that

$$P_{\theta_0}(a \leq \hat{\theta} \leq b) = F_{\hat{\theta}}(b) - F_{\hat{\theta}}(b) = 0.95 - 0.05 = 0.9$$

We can find a and b by solving

$$F_{\hat{\theta}}(a) = 0.05 \text{ and } F_{\hat{\theta}}(b) = 0.95$$

so that we obtain $a = \sqrt[n]{0.05}\theta_0$ and $b = \sqrt[n]{0.95}\theta_0$. This doesn't give us a confidence interval yet, since looking at the definition of a CI, we want the true parameter θ_0 in the middle of the inequalities, and the functions on either side depend only on the data and other known quantities.
However, we can rewrite

$$0.9 = P_{\theta_0}(a \leq \hat{\theta} \leq b) = P_{\theta_0}\left(\sqrt[n]{0.05}\theta_0 \leq \hat{\theta} \leq \sqrt[n]{0.95}\theta_0\right) = P_{\theta_0}\left(\frac{\hat{\theta}}{\sqrt[n]{0.95}} \leq \theta_0 \leq \frac{\hat{\theta}}{\sqrt[n]{0.05}}\right)$$

Therefore

$$[A, B] = [A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)] = \left[\frac{\max\{X_1, \ldots, X_n\}}{\sqrt[n]{0.95}}, \frac{\max\{X_1, \ldots, X_n\}}{\sqrt[n]{0.05}}\right]$$

is a 90% confidence interval for θ_0. Notice that in this case, the bounds of the confidence intervals depend on the data only through the estimator $\hat{\theta}(X_1, \ldots, X_n)$. This need not be true in general.

Let's recap how we arrived at the confidence interval:

1. first get estimator $\hat{\theta}(X_1, \ldots, X_n)$ and the distribution of $\hat{\theta}$.
2. find $a(\theta), b(\theta)$ such that
$$P(a(\theta) \leq \hat{\theta} \leq b(\theta)) = 1 - \alpha$$
3. rewrite the event by solving for θ
$$P(A(X) \leq \theta \leq B(X)) = 1 - \alpha$$
4. evaluate $A(X), B(X)$ for the observed sample $X_1, \ldots, X_n$
5. the $1 - \alpha$ confidence interval is then given by
$$\widehat{CI} = [A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)]$$

2.1 Important Cases

1. $\hat{\theta}$ is normally distributed, $\text{Var}(\hat{\theta}) \equiv \sigma_{\hat{\theta}}^2$ is known: can form confidence interval

$$[A(X), B(X)] = \left[\hat{\theta} - \sqrt{\sigma_{\hat{\theta}}^2}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \hat{\theta} + \sqrt{\sigma_{\hat{\theta}}^2}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

2. $\hat{\theta}$ is normally distributed, $\text{Var}(\hat{\theta})$ unknown, but have estimator $\hat{S}^2 = \widehat{\text{Var}(\hat{\theta})}$: confidence interval is given by

$$[A(X), B(X)] = \left[\hat{\theta} - \sqrt{\hat{S}^2}t_{n-1}\left(1 - \frac{\alpha}{2}\right), \hat{\theta} + \sqrt{\hat{S}^2}t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

   where $t_{n-1}(p)$ is the pth percentile of a t-distribution with $n-1$ degrees of freedom.

3. $\hat{\theta}$ is not normal, but $n > 30$ or so: it turns out that all estimators we've seen (except for the maximum of the sample for the uniform distribution) will be asymptotically normal by the central limit theorem (it is not always straightforward how we apply the CLT in a given case). So we'll construct confidence intervals the same way as in case 2.

4. $\hat{\theta}$ not normal, n small: if the p.d.f. of $\hat{\theta}$ is known, can form confidence intervals from first principles (as in the last example). If the p.d.f. of $\hat{\theta}$ is not known, there is nothing we can do.

The reason for using the t-distribution in the second case is the following: since $\hat{\theta} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$,

$$\frac{\hat{\theta} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

On the other hand, we can check that

$$\frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi_{n-1}^2$$

since in this setting, $\hat{S}$ can usually be written as a sum of squared normal residuals with mean zero and variance σ^2. Therefore,

$$\frac{\hat{\theta} - \mu}{\sqrt{\hat{S}^2/n}} = \frac{\frac{\hat{\theta}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\hat{S}^2}{\sigma^2}}/\sqrt{n-1}} \sim \frac{N(0,1)}{\sqrt{\chi_{n-1}^2}} \sim t_{n-1}$$

Also note that in the general case 4 (and in the last example involving a uniform), we did not require that the statistic $\hat{\theta}(X_1, \ldots, X_n)$ be an unbiased or consistent estimator of anything, but it just had to be strictly monotonic in the true parameter. However, the way we constructed confidence intervals for the normal cases (with or without knowledge of the variance of $\hat{\theta}$, the estimator has to be unbiased, and in case 3 (n large), it would have to be consistent.