Chapter 8 Estimation of Parameters and Fitting of Probability Distributions

8.5 The Method of Maximum Likelihood

Point Estimation: Let a random variable X have a pdf that is of known functional form, but depends on an unknown parameter θ. We assume that θ may take any value in a set Ω.

We can denote members of this family of pdf's by $f(x;\theta)$ for $\theta \in \Omega$. The entire family of pdf's can be described by $\{f(x;\theta) : \theta \in \Omega\}$. Here the set Ω is referred to as the **parameter space**.

For example, consider the family of normal distributions with variance equal to 1, but with unknown mean, $\{N(\theta, 1) : \theta \in \Omega\}$, where Ω is the set $-\infty < \theta < \infty$.

The objective of point estimation is to estimate the **true** value of θ by using a random sample of observations from the distribution.

One method of point estimation is called **maximum likelihood estimation**.

Example 1: Let $X_1, X_2, ..., X_n$ denote a random sample from a distribution with pdf

$$f(x) = \theta^x (1-\theta)^{1-x}$$

for $x = 0, 1$ and $\Omega = [0, 1]$. Let $L(\theta)$ denote the **likelihood function**. (Your book uses lik(θ).)

$$L(\theta) = P[X_1 = x_1, X_2 = x_2, ..., X_n = x_n]$$

$$= \prod_{i=1}^{n} P[X_i = x_i] = \prod_{i=1}^{n} f(x_i; \theta)$$

$$= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

A reasonable method for constructing a point estimate of θ would be to select the value of $\theta \in \Omega$ for which the probability of the observed data is the greatest. Thus, we want to maximize $L(\theta)$ as a function of θ.

This is equivalent to finding the value at which the **log-likelihood function** reaches its maximum.

$$l(\theta) = ln[L(\theta)] = \left(\sum_{i=1}^{n} x_i\right) ln(\theta) + \left(n - \sum_{i=1}^{n} x_i\right) ln(1-\theta)$$

To find this point, take the derivative of the log-likelihood with respect to θ, and set it equal to 0.

$$\frac{dln[L(\theta)]}{d\theta} = \frac{\Sigma_{i=1}^{n} x_i}{\theta} - \frac{n - \Sigma_{i=1}^{n} x_i}{1-\theta} = 0$$

for $\theta \in (0,1)$. By solving for the root of this equation we see that the solution is

$$\theta = \frac{\Sigma\, x_i}{n} = \bar{x}$$

Thus, for this family of distributions distribution the statistic

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

is called the **maximum likelihood estimator** of θ.

In general, let $X_1, X_2, ..., X_n$ be a random sample from a distribution having pdf $f(x;\theta)$ for $\theta \in \Omega$. The likelihood function is given by

$$L(\theta; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

and is defined for $\theta \in \Omega$.

Suppose the statistic

$$\hat{\theta} = u(X_1, .., X_n)$$

has the property that

$$L(\hat{\theta}; x_1, ..., x_n) \geq L(\theta; x_1, ..., x_n)$$

for all $\theta \in \Omega$.

Then $\hat{\theta}$ is called a maximum likelihood estimator of θ.

Definition Any statistic whose mathematical expectation is equal to a parameter θ is called an **unbiased** estimator of θ. Otherwise, the statistic is said to be biased.

In some cases, the maximum likelihood estimator is unbiased, but that is not the case in general.

Example 2: Consider the family of Bernoulli distributions of Example 1. Is the mle unbiased?

$$\hat{\theta} = \bar{X}$$

$$E[\hat{\theta}] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = E[X]$$

$$= \sum_{x=0}^{1} f(x)x = (0)(1-\theta) + (1)(\theta) = \theta$$

So, we can see that $\hat{\theta}$ is an unbiased estimate of θ.

Example 3: Let

$$f(x;\theta) = \frac{1}{\theta}$$

for $0 < x \leq \theta$, and $\Omega = (0, \infty)$.

Let $X_1, X_2, ..., X_n$ be a random sample from this distribution, and let $X_{(1)} < X_{(2)} < ... < X_{(n)}$ denote the corresponding order statistics. Then

$$L(\theta; x_1, ..., x_n) = \frac{1}{\theta^n} I[0 < x_{(n)} \leq \theta]$$

This is maximized by taking θ as small as possible subject to $I[0 < x_{(n)} \leq \theta] = 1$.

Clearly, the solution is then

$$\hat{\theta} = X_{(n)}$$

One might expect $\hat{\theta}$ to be slightly biased, because $X_{(n)}$ will always be somewhat less than θ with probability 1, and can never be greater than θ.

Let $Y_n = X_{(n)}$ for easier notation.

$$E[\hat{\theta}] = E[Y_n] = \int_0^\theta y_n f_n(y_n) dy_n$$

Recall how to find $f_n(y_n)$. For $0 < y_n \leq \theta$,

$$F_n(y_n) = [F(y_n)]^n = \left(\frac{y_n}{\theta}\right)^n$$

and

$$f_n(y_n) = F'(y_n) = n\left(\frac{y_n}{\theta}\right)^{n-1}$$

$$E[\hat{\theta}] = E[Y_n] = \frac{n}{\theta^n}\int_0^\theta y_n^n dy_n = \frac{n}{\theta^n}\frac{(\theta^{n+1})}{(n+1)} = \frac{n\theta}{n+1}$$

We can see that $\hat{\theta}$ is biased in estimation of θ.
However, notice that

$$\lim_{n\to\infty} \frac{n\theta}{n+1} = \theta.$$

The bias disappears as n becomes larger. In that sense, we can say that $\hat{\theta}$ is **asymptotically unbiased**. This is closely related but not identical to the concept of **consistency**.

Definition A statistic that converges in probability to a parameter θ is called a **consistent** estimator of θ.

Let's show that $\hat{\theta} = X_{(n)}$ in the previous example is a consistent estimator of θ.
Let $0 < \epsilon < \theta$.

$$P[|\hat{\theta} - \theta| > \epsilon] = P[|X_{(n)} - \theta| > \epsilon]$$

$$= P[X_{(n)} < \theta - \epsilon]$$

$$\left(\frac{\theta - \epsilon}{\theta}\right)^n$$

which approaches 0 as n approaches infinity. This proves that $\hat{\theta}$ is a consistent estimator of θ.

In many cases, probability distributions are determined by more than one parameter. For example, consider the normal distribution which is determined by the mean $\theta_1 = \mu$, and the variance $\theta_2 = \sigma^2$.

$$f(x; \theta_1, \theta_2) = \frac{1}{(2\pi\theta_2)^{1/2}} e^{-\frac{(x-\theta_1)^2}{2\theta_2}}$$

Let $X_1, X_2, ..., X_n$ denote a random sample from this distribution. Then we write the log-likelihood as a function of θ_1 and θ_2 by

$$l(\theta) = ln[L(\theta_1, \theta_2; x_1, x_2, ..., x_n)] = \sum_{i=1}^{n} ln[f(x_i; \theta_1, \theta_2)]$$

$$= -\frac{nln(2\pi\theta_2)}{2} - \frac{\Sigma_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2}$$

This is maximized by taking partial derivatives and setting them equal to 0.

$$\frac{\partial ln[L]}{\partial \theta_1} = \frac{\Sigma_{i=1}^{n}(x_i - \theta_1)}{\theta_2}$$

and

$$\frac{\partial ln[L]}{\partial \theta_2} = \frac{\Sigma_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2}$$

By setting these partial derivatives equal to 0 we see that a solution is obtained when $\theta_1 = \bar{x}$ and $\theta_2 = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2$.

$$\hat{\theta}_1 = \bar{X}$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

8.4 The Method of Moments

Sometimes it is impossible to find maximum likelihood estimators in a convenient closed form and numerical methods must be used. There are other ways to obtain point estimates.

Method of moments simply equates the moments of the distribution to the corresponding moments of the sample.

The expectation $E(X^k)$ is frequently called the kth moment of the distribution.

The sum $M_k = \Sigma_{i=1}^{n} \frac{X_i^k}{n} = (1/n)\,\Sigma_{i=1}^{n} X_i^k)$ is the kth moment of the sample.

Example: Find the MLE and MOM for a random sample of size n from the $Gamma(\lambda)$ distribution.

8.5.3 Confidence Intervals for Maximum Likelihood Estimates

Interval Estimation: Point estimation of a parameter is often accompanied by interval estimation, in which the goal is to identify a strategy for constructing intervals that have some prespecified probability of containing the value of the parameter. Such intervals are called **confidence intervals**.

Suppose $X_1, X_2, ..., X_n$ is a random sample from a distribution with unknown mean μ and unknown variance σ^2.

Our aim is to construct a confidence interval for the mean μ. This amounts to defining a procedure to compute a lower limit $A = u_A(X_1, X_2, ..., X_n)$ and an upper limit $B = u_B(X_1, X_2, ..., X_n)$ such that for a specified **confidence level** $1 - \alpha$,

$$P[\mu \in (A, B)] = 1 - \alpha$$

Notice that A and B are functions of the data, so they can be viewed as random variables. How can we do this?

First, consider a random sample from a distribution that is $N(\mu, \sigma^2)$, σ^2 known. Let's consider the MLE of μ, $\hat{\mu} = \bar{X}$. Know $\bar{X}$ is $N(\mu, \sigma^2/n)$, then

$$P[-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}] = 1 - \alpha$$

implies that

$$P[\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}] = 1 - \alpha.$$

Thus, we may take $A = \bar{X} - z_{\alpha/2}\sigma/\sqrt{n}$ and $B = \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$

so that our confidence interval for μ is

$$(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}).$$

We can construct a confidence interval that has exactly the desired coverage probability for each n

What if σ^2 is not known? We can use S^2 as an estimator.

In Section 6.3 we saw that when $X_1, X_2, ..., X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $n - 1$ degrees of freedom.
Then if we define $t_{a/2}$ in a similar way as $z_{\alpha/2}$, we know that

$$P[-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}] = 1 - \alpha$$

which implies that a confidence interval of level $1 - \alpha$ for μ is given by

$$(\bar{X} - t_{\alpha/2}S/\sqrt{n}, \bar{X} + t_{\alpha/2}S/\sqrt{n}).$$

Example 1: A sample of 25 birthweights is taken from the population of year 1998 birthweights. Assume that the *sample* standard deviation for birthweight is 20 oz.

Construct a 95% confidence interval for the population mean birthweight.

$\bar{x} = 116$ oz, is our point estimate of the population mean birthweight. To find a 95% confidence interval (1-α=.95), first find $t_{.05/2}$.
By using Table 4 and 24 degrees of freedom, we find $t_{.05/2} = 2.064$.
95% confidence interval for the mean birthweight is given by,

$$(116 - 2.064\frac{s}{\sqrt{25}}, 116 + 2.064\frac{s}{\sqrt{25}}).$$

Plugging in $s = 20$ we find that the 95 percent confidence interval is

$$(107.774, 124.256)$$

Next, consider a special case of the central limit theorem, which is the normal approximation to the binomial distribution. Let Y be $b(n, p)$ for some unknown p in the interval $(0, 1)$.

Let $\hat{p} = Y/n$ denote a point estimate of p.

$$E[\hat{p}] = E[Y/n] = p$$

and

$$Var[\hat{p}] = p(1-p)/n$$

By the central limit theorem,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

converges in distribution to a standard normal random variable. Thus, for large values of n,

$$P[-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}] \approx 1 - \alpha$$

Also, because $\hat{p}$ converges in probability to p we apply the Theorem to obtain that

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

converges in distribution to a standard normal random variable.

This implies that

$$P[-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{\alpha/2}] \approx 1-\alpha$$

From this we obtain a confidence interval for p of level $1-\alpha$ by

$$(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}).$$

Example 2: In a study of 100 patients with an ocular vascular disorder, 40 were found to show improved visual function after a one-month course of systemic steroid therapy.

Obtain a 90 percent confidence interval for p, the population improvement rate.

$\hat{p} = 40/100 = .4.$

Standard error= $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{.4 \times .6/100}$

$z_{.10/2} = 1.645.$

90% CI= $.4 \pm 1.645 \times \sqrt{.4 \times .6/100}$=(0.32,0.48).

Let $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$ denote independent random samples from two distributions, $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively.

Denote the means of the samples by $\bar{X}$ and $\bar{Y}$ and the sample variances by S_X^2 and S_Y^2. Our aim is to find a confidence interval for the difference $\mu_X - \mu_X$. The obvious point estimator of this difference is $\bar{X} - \bar{Y}$.

We know that $\bar{X} - \bar{Y}$ has a normal distribution with mean $\mu_1 - \mu_2$ and variance $\sigma^2/n + \sigma^2/m$. Thus,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}}$$

has a standard normal distribution.

In Chapter 6 we saw that $(n-1)S_X^2/\sigma^2$ and $(m-1)S_Y^2/\sigma^2$ have chi-square distributions with $n-1$ and $m-1$ degrees of freedom, respectively.

Because they are independent, we may infer that

$$((n-1)S_X^2 + (m-1)S_Y^2))/\sigma^2$$

has a chi-square distribution with $n+m-2$ degrees of freedom.

Because of the independence of $\bar{X}$, $\bar{Y}$, S_X^2 and S_X^2 we see that

$$T = \frac{(\bar{X}-\bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}}$$

has a t-distribution with $n+m-2$ degrees of freedom.

Then we can define a confidence interval with level $1-\alpha$ for $\mu_1 - \mu_2$ according to (A, B), where

$$A = (\bar{X}-\bar{Y}) - t_{\alpha/2}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}$$

and

$$B = (\bar{X}-\bar{Y}) + t_{\alpha/2}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}$$

Now suppose that X and Y are independent with distributions $b(n_1, p_1)$, and $b(n_2, p_2)$, respectively.

We can use our knowledge about the sampling distribution of $\hat{p}_1 - \hat{p}_2$, to construct confidence intervals for the difference between two population proportions $p_1 - p_2$.

Again, the central limit theorem is used, so we require that both n_1 and n_2 are large.

A confidence interval for $p_1 - p_2$ of level $1 - \alpha$ is as follows:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}$$