7.3: Maximum Likelihood

Basic Theory

The Method

Suppose again that we have an observable random variable $\boldsymbol{X}$ for an experiment, that takes values in a set S. Suppose also that distribution of $\boldsymbol{X}$ depends on an unknown parameter θ, taking values in a parameter space Θ. Of course, our data variable $\boldsymbol{X}$ will almost always be vector valued. The parameter θ may also be vector valued. We will denote the probability density function of $\boldsymbol{X}$ on S by f_θ for $\theta \in \Theta$. The distribution of $\boldsymbol{X}$ could be discrete or continuous.

The likelihood function is the function obtained by reversing the roles of $\boldsymbol{x}$ and θ in the probability density function; that is, we view θ as the variable and $\boldsymbol{x}$ as the given information (which is precisely the point of view in estimation).

The *likelihood function* at $\boldsymbol{x} \in S$ is the function $L_{\boldsymbol{x}} : \Theta \to [0, \infty)$ given by

$$L_{\boldsymbol{x}}(\theta) = f_\theta(\boldsymbol{x}), \quad \theta \in \Theta \tag{7.3.1}$$

In the method of *maximum likelihood*, we try to find the value of the parameter that maximizes the likelihood function for each value of the data vector.

Suppose that the maximum value of $L_{\boldsymbol{x}}$ occurs at $u(\boldsymbol{x}) \in \Theta$ for each $\boldsymbol{x} \in S$. Then the statistic $u(\boldsymbol{X})$ is a *maximum likelihood estimator* of θ.

The method of maximum likelihood is intuitively appealing—we try to find the value of the parameter that would have most likely produced the data we in fact observed.

Since the natural logarithm function is strictly increasing on $(0, \infty)$, the maximum value of the likelihood function, if it exists, will occur at the same points as the maximum value of the logarithm of the likelihood function.

The *log-likelihood function* at $\boldsymbol{x} \in S$ is the function $\ln L_{\boldsymbol{x}}$:

$$\ln L_{\boldsymbol{x}}(\theta) = \ln f_\theta(\boldsymbol{x}), \quad \theta \in \Theta \tag{7.3.2}$$

If the maximum value of $\ln L_{\boldsymbol{x}}$ occurs at $u(\boldsymbol{x}) \in \Theta$ for each $\boldsymbol{x} \in S$. Then the statistic $u(\boldsymbol{X})$ is a maximum likelihood estimator of θ

The log-likelihood function is often easier to work with than the likelihood function (typically because the probability density function $f_\theta(\boldsymbol{x})$ has a product structure).

Vector of Parameters

An important special case is when $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is a vector of k real parameters, so that $\Theta \subseteq \mathbb{R}^k$. In this case, the maximum likelihood problem is to maximize a function of several variables. If Θ is a continuous set, the methods of calculus can be used. If the maximum value of $L_{\boldsymbol{x}}$ occurs at a point $\boldsymbol{\theta}$ in the interior of Θ, then $L_{\boldsymbol{x}}$ has a local maximum at $\boldsymbol{\theta}$. Therefore, assuming that the likelihood function is differentiable, we can find this point by solving

$$\frac{\partial}{\partial \theta_i} L_{\boldsymbol{x}}(\boldsymbol{\theta}) = 0, \quad i \in \{1, 2, \ldots, k\} \tag{7.3.3}$$

or equivalently

$$\frac{\partial}{\partial \theta_i} \ln L_{\boldsymbol{x}}(\boldsymbol{\theta}) = 0, \quad i \in \{1, 2, \ldots, k\} \tag{7.3.4}$$

On the other hand, the maximum value may occur at a boundary point of Θ, or may not exist at all.

Random Samples

The most important special case is when the data variables form a random sample from a distribution.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the distribution of a random variable X taking values in R, with probability density function g_θ for $\theta \in \Theta$. Then $\boldsymbol{X}$ takes values in $S = R^n$, and the likelihood and log-likelihood functions for $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in S$ are

$$L_{\boldsymbol{x}}(\theta) = \prod_{i=1}^n g_\theta(x_i), \quad \theta \in \Theta$$

$$\ln L_{\boldsymbol{x}}(\theta) = \sum_{i=1}^n \ln g_\theta(x_i), \quad \theta \in \Theta$$

Extending the Method and the Invariance Property

Returning to the general setting, suppose now that h is a one-to-one function from the parameter space Θ onto a set Λ. We can view $\lambda = h(\theta)$ as a new parameter taking values in the space Λ, and it is easy to re-parameterize the probability density function with the new parameter. Thus, let $\hat{f}_\lambda(\boldsymbol{x}) = f_{h^{-1}(\lambda)}(\boldsymbol{x})$ for $\boldsymbol{x} \in S$ and $\lambda \in \Lambda$. The corresponding likelihood function for $\boldsymbol{x} \in S$ is

$$\hat{L}_{\boldsymbol{x}}(\lambda) = L_{\boldsymbol{x}}\left[h^{-1}(\lambda)\right], \quad \lambda \in \Lambda \tag{7.3.5}$$

Clearly if $u(\boldsymbol{x}) \in \Theta$ maximizes $L_{\boldsymbol{x}}$ for $\boldsymbol{x} \in S$. Then $h\left[u(\boldsymbol{x})\right] \in \Lambda$ maximizes $\hat{L}_{\boldsymbol{x}}$ for $\boldsymbol{x} \in S$. It follows that if U is a maximum likelihood estimator for θ, then $V = h(U)$ is a maximum likelihood estimator for $\lambda = h(\theta)$.

If the function h is not one-to-one, the maximum likelihood function for the new parameter $\lambda = h(\theta)$ is not well defined, because we cannot parameterize the probability density function in terms of λ. However, there is a natural generalization of the method.

Suppose that $h : \Theta \to \Lambda$, and let $\lambda = h(\theta)$ denote the new parameter. Define the *likelihood function* for λ at $\boldsymbol{x} \in S$ by

$$\hat{L}_{\boldsymbol{x}}(\lambda) = \max\left\{L_{\boldsymbol{x}}(\theta) : \theta \in h^{-1}\{\lambda\}\right\}; \quad \lambda \in \Lambda \tag{7.3.6}$$

If $v(\boldsymbol{x}) \in \Lambda$ maximizes $\hat{L}_{\boldsymbol{x}}$ for each $\boldsymbol{x} \in S$, then $V = v(\boldsymbol{X})$ is a *maximum likelihood estimator* of λ.

This definition extends the maximum likelihood method to cases where the probability density function is not completely parameterized by the parameter of interest. The following theorem is known as the *invariance property*: if we can solve the maximum likelihood problem for θ then we can solve the maximum likelihood problem for $\lambda = h(\theta)$.

In the setting of the previous theorem, if U is a maximum likelihood estimator of θ, then $V = h(U)$ is a maximum likelihood estimator of λ.

Proof

As before, if $u(\boldsymbol{x}) \in \Theta$ maximizes $L_{\boldsymbol{x}}$ for $\boldsymbol{x} \in S$. Then $h\left[u(\boldsymbol{x})\right] \in \Lambda$ maximizes $\hat{L}_{\boldsymbol{x}}$ for $\boldsymbol{x} \in S$.

Examples and Special Cases

In the following subsections, we will study maximum likelihood estimation for a number of special parametric families of distributions. Recall that if $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from a distribution with mean μ and variance σ^2, then the method of moments estimators of μ and σ^2 are, respectively,

$$M = \frac{1}{n}\sum_{i=1}^n X_i \tag{7.3.7}$$

$$T^2 = \frac{1}{n}\sum_{i=1}^n (X_i - M)^2 \tag{7.3.8}$$

Of course, M is the sample mean, and T^2 is the biased version of the sample variance. These statistics will also sometimes occur as maximum likelihood estimators. Another statistic that will occur in some of the examples below is

$$M_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \tag{7.3.9}$$

the second-order sample mean. As always, be sure to try the derivations yourself before looking at the solutions.

The Bernoulli Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the Bernoulli distribution with success parameter $p \in [0, 1]$. Recall that the Bernoulli probability density function is

$$g(x) = p^x (1 - p)^{1 - x}, \quad x \in \{0, 1\} \tag{7.3.10}$$

Thus, $\boldsymbol{X}$ is a sequence of independent indicator variables with $\mathbb{P}(X_i = 1) = p$ for each i. In the usual language of reliability, X_i is the outcome of trial i, where 1 means success and 0 means failure. Let $Y = \sum_{i=1}^n X_i$ denote the number of successes, so that the proportion of successes (the sample mean) is $M = Y / n$. Recall that Y has the binomial distribution with parameters n and p.

The sample mean M is the maximum likelihood estimator of p on the parameter space $(0, 1)$.

Proof

Note that $\ln g(x) = x \ln p + (1 - x) \ln(1 - p)$ for $x \in \{0, 1\}$ Hence the log-likelihood function at $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$ is

$$\ln L_{\boldsymbol{x}}(p) = \sum_{i=1}^n [x_i \ln p + (1 - x_i) \ln(1 - p)], \quad p \in (0, 1) \tag{7.3.11}$$

Differentiating with respect to p and simplifying gives

$$\frac{d}{dp} \ln L_{\boldsymbol{x}}(p) = \frac{y}{p} - \frac{n - y}{1 - p} \tag{7.3.12}$$

where $y = \sum_{i=1}^n x_i$. Thus, there is a single critical point at $p = y / n = m$. The second deriviative is

$$\frac{d^2}{dp^2} \ln L_{\boldsymbol{x}}(p) = -\frac{y}{p^2} - \frac{n - 1}{(1 - p)^2} < 0 \tag{7.3.13}$$

Hence the log-likelihood function is concave downward and so the maximum occurs at the unique critical point m.

Recall that M is also the method of moments estimator of p. It's always nice when two different estimation procedures yield the same result. Next let's look at the same problem, but with a much restricted parameter space.

Suppose now that p takes values in $\left\{\frac{1}{2}, 1\right\}$. Then the maximum likelihood estimator of p is the statistic

$$U = \begin{cases} 1, & Y = n \\ \frac{1}{2}, & Y < n \end{cases} \tag{7.3.14}$$

1. $\mathbb{E}(U) = \begin{cases} 1, & p = 1 \\ \frac{1}{2} + \left(\frac{1}{2}\right)^{n+1}, & p = \frac{1}{2} \end{cases}$
2. U is positively biased, but is asymptotically unbiased.
3. $\mathrm{mse}(U) = \begin{cases} 0 & p = 1 \\ \left(\frac{1}{2}\right)^{n+2}, & p = \frac{1}{2} \end{cases}$
4. U is consistent.

Proof

Note that the likelihood function at $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$ is $L_{\boldsymbol{x}}(p) = p^y (1 - p)^{n - y}$ for $p \in \left\{\frac{1}{2}, 1\right\}$ where as usual, $y = \sum_{i=1}^n x_i$. Thus $L_{\boldsymbol{x}}\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^y$. On the other hand, $L_{\boldsymbol{x}}(1) = 0$ if $y < n$ while $L_{\boldsymbol{x}}(1) = 1$ if $y = n$. Thus, if $y = n$ the maximum occurs when $p = 1$ while if $y < n$ the maximum occurs when $p = \frac{1}{2}$.

1. If $p = 1$ then $\mathbb{P}(U = 1) = \mathbb{P}(Y = n) = 1$, so trivially $\mathbb{E}(U) = 1$. If $p = \frac{1}{2}$,

$$\mathbb{E}(U) = 1\mathbb{P}(Y = n) + \frac{1}{2}\mathbb{P}(Y < n) = 1\left(\frac{1}{2}\right)^n + \frac{1}{2}\left[1 - \left(\frac{1}{2}\right)^n\right] = \frac{1}{2} + \left(\frac{1}{2}\right)^{n+1} \tag{7.3.15}$$

2. Note that $\mathbb{E}(U) \geq p$ and $\mathbb{E}(U) \to p$ as $n \to \infty$ both in the case that $p = 1$ and $p = \frac{1}{2}$.
3. If $p = 1$ then $U = 1$ with probability 1, so trivially $\mse(U) = 0$. If $p = \frac{1}{2}$,

$$\mse(U) = \left(1 - \frac{1}{2}\right)^2 \mathbb{P}(Y = n) + \left(\frac{1}{2} - \frac{1}{2}\right)^2 \mathbb{P}(Y < n) = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^{n+2} \tag{7.3.16}$$

4. From (c), $\mse(U) \to 0$ as $n \to \infty$.

Note that the Bernoulli distribution in the last exercise would model a coin that is either fair or two-headed. The last two exercises show that the maximum likelihood estimator of a parameter, like the solution to any maximization problem, depends critically on the domain.

U is uniformly better than M on the parameter space $\left\{\frac{1}{2}, 1\right\}$.

Proof

Recall that $\mse(M) = \var(M) = p(1-p)/n$. If $p = 1$ then $\mse(M) = \mse(U) = 0$ so that both estimators give the correct answer. If $p = \frac{1}{2}$, $\mse(U) = \left(\frac{1}{2}\right)^{n+2} < \frac{1}{4n} = \mse(M)$.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the Bernoulli distribution with unknown success parameter $p \in (0, 1)$. Find the maximum likelihood estimator of $p(1-p)$, which is the variance of the sampling distribution.

Answer

By the invariance principle, the estimator is $M(1-M)$ where M is the sample mean.

The Geometric Distribution

Recall that the *geometric distribution* on $\mathbb{N}_+$ with success parameter $p \in (0, 1)$ has probability density function

$$g(x) = p(1-p)^{x-1}, \quad x \in \mathbb{N}_+ \tag{7.3.17}$$

The geometric distribution governs the trial number of the first success in a sequence of Bernoulli trials.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the geometric distribution with unknown parameter $p \in (0, 1)$. The maximum likelihood estimator of p is $U = 1/M$.

Proof

Note that $\ln g(x) = \ln p + (x-1)\ln(1-p)$ for $x \in \mathbb{N}_+$. Hence the log-likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}_+^n$ is

$$\ln L_{\boldsymbol{x}}(p) = n \ln p + (y - n)\ln(1-p), \quad p \in (0, 1) \tag{7.3.18}$$

where $y = \sum_{i=1}^n x_i$. So

$$\frac{d}{dp} \ln L(p) = \frac{n}{p} - \frac{y - n}{1-p} \tag{7.3.19}$$

The derivative is 0 when $p = n/y = 1/m$. Finally, $\frac{d^2}{dp^2} \ln L_{\boldsymbol{x}}(p) = -n/p^2 - (y-n)/(1-p)^2 < 0$ so the maximum occurs at the critical point.

Recall that U is also the method of moments estimator of p. It's always reassuring when two different estimation procedures produce the same estimator.

The Negative Binomial Distribution

More generally, the *negative binomial distribution* on $\mathbb{N}$ with shape parameter $k \in (0, \infty)$ and success parameter $p \in (0, 1)$ has probability density function

$$g(x) = \binom{x+k-1}{k-1} p^k (1-p)^x, \quad x \in \mathbb{N} \tag{7.3.20}$$

If k is a positive integer, then this distribution governs the number of failures before the kth success in a sequence of Bernoulli trials with success parameter p. However, the distribution makes sense for general $k \in (0, \infty)$. The negative binomial distribution is studied in more detail in the chapter on Bernoulli Trials.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the negative binomial distribution on $\mathbb{N}$ with known shape parameter k and unknown success parameter $p \in (0, 1)$. The maximum likelihood estimator of p is

$$U = \frac{k}{k + M} \tag{7.3.21}$$

Proof

Note that $\ln g(x) = \ln\binom{x+k-1}{k-1} + k \ln p + x \ln(1-p)$ for $x \in \mathbb{N}$. Hence the log-likelihood function corresponding to $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}^n$ is

$$\ln L_{\boldsymbol{x}}(p) = nk \ln p + y \ln(1-p) + C, \quad p \in (0, 1) \tag{7.3.22}$$

where $y = \sum_{i=1}^n x_i$ and $C = \sum_{i=1}^n \ln\binom{x_i+k-1}{k-1}$. Hence

$$\frac{d}{dp} \ln L_{\boldsymbol{x}}(p) = \frac{nk}{p} - \frac{y}{1-p} \tag{7.3.23}$$

The derivative is 0 when $p = nk/(nk+y) = k/(k+m)$ where as usual, $m = y/n$. Finally, $\frac{d^2}{dp^2} \ln L_{\boldsymbol{x}}(p) = -nk/p^2 - y/(1-p)^2 < 0$, so the maximum occurs at the critical point.

Once again, this is the same as the method of moments estimator of p with k known.

The Poisson Distribution

Recall that the *Poisson distribution* with parameter $r > 0$ has probability density function

$$g(x) = e^{-r} \frac{r^x}{x!}, \quad x \in \mathbb{N} \tag{7.3.24}$$

The Poisson distribution is named for Simeon Poisson and is widely used to model the number of random "points" in a region of time or space. The parameter r is proportional to the size of the region. The Poisson distribution is studied in more detail in the chapter on the Poisson process.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the Poisson distribution with unknown parameter $r \in (0, \infty)$. The maximum likelihood estimator of r is the sample mean M.

Proof

Note that $\ln g(x) = -r + x \ln r - \ln(x!)$ for $x \in \mathbb{N}$. Hence the log-likelihood function corresponding to $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}^n$ is

$$\ln L_{\boldsymbol{x}}(r) = -nr + y \ln r - C, \quad r \in (0, \infty) \tag{7.3.25}$$

where $y = \sum_{i=1}^n x_i$ and $C = \sum_{i=1}^n \ln(x_i!)$. Hence $\frac{d}{dr} \ln L_{\boldsymbol{x}}(r) = -n + y/r$. The derivative is 0 when $r = y/n = m$. Finally, $\frac{d^2}{dr^2} \ln L_{\boldsymbol{x}}(r) = -y/r^2 < 0$, so the maximum occurs at the critical point.

Recall that for the Poisson distribution, the parameter r is both the mean and the variance. Thus M is also the method of moments estimator of r. We showed in the introductory section that M has smaller mean square error than S^2, although both are unbiased.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the Poisson distribution with parameter $r \in (0, \infty)$, and let $p = \mathbb{P}(X = 0) = e^{-r}$. Find the maximum likelihood estimator of p in two ways:

1. Directly, by finding the likelihood function corresponding to the parameter p.
2. By using the result of the last exercise and the invariance property.

Answer

e^{-M} where M is the sample mean.

The Normal Distribution

Recall that the *normal distribution* with mean μ and variance σ^2 has probability density function

$$g(x) = \frac{1}{\sqrt{2\,\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in \mathbb{R} \tag{7.3.26}$$

The normal distribution is often used to model physical quantities subject to small, random errors, and is studied in more detail in the chapter on Special Distributions

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the normal distribution with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$. The maximum likelihood estimators of μ and σ^2 are M and T^2, respectively.

Proof

Note that

$$\ln g(x) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2, \quad x \in \mathbb{R} \tag{7.3.27}$$

Hence the log-likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is

$$\ln L_{\boldsymbol{x}}(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2, \quad (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty) \tag{7.3.28}$$

Taking partial derivatives gives

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln L_{\boldsymbol{x}}(\mu, \sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^n x_i - n\mu\right) \\ \frac{\partial}{\partial \sigma^2} \ln L_{\boldsymbol{x}}(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \mu)^2 \end{aligned}$$

The partial derivatives are 0 when $\mu = \frac{1}{n}\sum_{i=1}^n x_i$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2$. Hence the unique critical point is (m, t^2). Finally, with a bit more calculus, the second partial derivatives evaluated at the critical point are

$$\frac{\partial^2}{\partial \mu^2} \ln L_{\boldsymbol{x}}(m, t^2) = -n/t^2, \; \frac{\partial^2}{\partial \mu \partial \sigma^2} \ln L_{\boldsymbol{x}}(m, t^2) = 0, \; \frac{\partial^2}{\partial (\sigma^2)^2} \ln L_{\boldsymbol{x}}(m, t^2) = -n/t^4 \tag{7.3.29}$$

Hence the second derivative matrix at the critical point is negative definite and so the maximum occurs at the critical point.

Of course, M and T^2 are also the method of moments estimators of μ and σ^2, respectively.

Run the Normal estimation experiment 1000 times for several values of the sample size n, the mean μ, and the variance σ^2. For the parameter σ^2, compare the maximum likelihood estimator T^2 with the standard sample variance S^2. Which estimator seems to work better in terms of mean square error?

Suppose again that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the normal distribution with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 \in (0, \infty)$. Find the maximum likelihood estimator of $\mu^2 + \sigma^2$, which is the second moment about 0 for the sampling distribution.

Answer

By the invariance principle, the estimator is $M^2 + T^2$ where M is the sample mean and T^2 is the (biased version of the) sample variance.

The Gamma Distribution

Recall that the *gamma distribution* with shape parameter $k > 0$ and scale parameter $b > 0$ has probability density function

$$g(x) = \frac{1}{\Gamma(k)\, b^k} x^{k-1} e^{-x/b}, \quad 0 < x < \infty \tag{7.3.30}$$

The gamma distribution is often used to model random times and certain other types of positive random variables, and is studied in more detail in the chapter on Special Distributions

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the gamma distribution with known shape parameter k and unknown scale parameter $b \in (0, \infty)$. The maximum likelihood estimator of b is $V_k = \frac{1}{k} M$.

Proof

Note that for $x \in (0, \infty)$,

$$\ln g(x) = -\ln \Gamma(k) - k \ln b + (k-1) \ln x - \frac{x}{b} \tag{7.3.31}$$

and hence the log-likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in (0, \infty)^n$ is

$$\ln L_{\boldsymbol{x}}(b) = -nk \ln b - \frac{y}{b} + C, \quad b \in (0, \infty) \tag{7.3.32}$$

where $y = \sum_{i=1}^n x_i$ and $C = -n \ln \Gamma(k) + (k-1) \sum_{i=1}^n \ln x_i$. It follows that

$$\frac{d}{db} \ln L_{\boldsymbol{x}}(b) = -\frac{nk}{b} + \frac{y}{b^2} \tag{7.3.33}$$

The derivative is 0 when $b = y / nk = 1 / km$. Finally, $\frac{d^2}{db^2} \ln L_{\boldsymbol{x}}(b) = nk / b^2 - 2y / b^3$. At the critical point $b = y / nk$, the second derivative is $-(nk)^3 / y^2 < 0$ so the maximum occurs at the critical point.

Recall that V_k is also the method of moments estimator of b when k is known. But when k is unknown, the method of moments estimator of b is $V = \frac{T^2}{M}$.

Run the gamma estimation experiment 1000 times for several values of the sample size n, shape parameter k, and scale parameter b. In each case, compare the method of moments estimator V of b when k is unknown with the method of moments and maximum likelihood estimator V_k of b when k is known. Which estimator seems to work better in terms of mean square error?

The Beta Distribution

Recall that the *beta distribution* with left parameter $a \in (0, \infty)$ and right parameter $b = 1$ has probability density function

$$g(x) = a x^{a-1}, \quad x \in (0, 1) \tag{7.3.34}$$

The beta distribution is often used to model random proportions and other random variables that take values in bounded intervals. It is studied in more detail in the chapter on Special Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the beta distribution with unknown left parameter $a \in (0, \infty)$ and right parameter $b = 1$. The maximum likelihood estimator of a is

$$W = -\frac{n}{\sum_{i=1}^n \ln X_i} = -\frac{n}{\ln(X_1 X_2 \cdots X_n)} \tag{7.3.35}$$

Proof

Note that $\ln g(x) = \ln a + (a-1) \ln x$ for $x \in (0, \infty)$ Hence the log-likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in (0, \infty)^n$ is

$$\ln L_{\boldsymbol{x}}(a) = n \ln a + (a-1) \sum_{i=1}^n \ln x_i, \quad a \in (0, \infty) \tag{7.3.36}$$

Therefore $\frac{d}{da}\ln L_{\boldsymbol{x}}(a) = n/a + \sum_{i=1}^n \ln x_i$. The derivative is 0 when $a = -n\big/\sum_{i=1}^n \ln x_i$. Finally, $\frac{d^2}{da^2}\ln L_{\boldsymbol{x}}(a) = -n/a^2 < 0$, so the maximum occurs at the critical point.

Recall that when $b = 1$, the method of moments estimator of a is $U_1 = M\big/(1-M)$, but when $b \in (0,\infty)$ is also unknown, the method of moments estimator of a is $U = M(M-M_2)\big/(M_2 - M^2)$. When $b = 1$, which estimator is better, the method of moments estimator or the maximum likelihood estimator?

In the beta estimation experiment, set $b = 1$. Run the experiment 1000 times for several values of the sample size n and the parameter a. In each case, compare the estimators U, U_1 and W. Which estimator seems to work better in terms of mean square error?

Finally, note that $1/W$ is the sample mean for a random sample of size n from the distribution of $-\ln X$. This distribution is the exponential distribution with rate a.

The Pareto Distribution

Recall that the *Pareto distribution* with shape parameter $a > 0$ and scale parameter $b > 0$ has probability density function

$$g(x) = \frac{ab^a}{x^{a+1}}, \quad b \le x < \infty \tag{7.3.37}$$

The Pareto distribution, named for Vilfredo Pareto, is a heavy-tailed distribution often used to model income and certain other types of random variables. It is studied in more detail in the chapter on Special Distribution.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the Pareto distribution with unknown shape parameter $a \in (0,\infty)$ and scale parameter $b \in (0,\infty)$. The maximum likelihood estimator of b is $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$, the first order statistic. The maximum likelihood estimator of a is

$$U = \frac{n}{\sum_{i=1}^n \ln X_i - n \ln X_{(1)}} = \frac{n}{\sum_{i=1}^n \left(\ln X_i - \ln X_{(1)}\right)} \tag{7.3.38}$$

Proof

Note that $\ln g(x) = \ln a + a \ln b - (a+1)\ln x$ for $x \in [b,\infty)$. Hence the log-likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is

$$\ln L_{\boldsymbol{x}}(a,b) = n \ln a + na \ln b - (a+1)\sum_{i=1}^n \ln x_i; \quad 0 < a < \infty,\ 0 < b \le x_i \text{ for each } i \in \{1,2,\ldots,n\} \tag{7.3.39}$$

Equivalently, the domain is $0 < a < \infty$ and $0 < b \le x_{(1)}$. Note that $\ln L_{\boldsymbol{x}}(a,b)$ is increasing in b for each a, and hence is maximized when $b = x_{(1)}$ for each a. Next,

$$\frac{d}{da}\ln L_{\boldsymbol{x}}\left(a, x_{(1)}\right) = \frac{n}{a} + n \ln x_{(1)} - \sum_{i=1}^n \ln x_i \tag{7.3.40}$$

The derivative is 0 when $a = n\big/\left(\sum_{i=1}^n \ln x_i - n \ln x_{(1)}\right)$. Finally, $\frac{d^2}{da^2}\ln L_{\boldsymbol{x}}\left(a, x_{(1)}\right) = -n/a^2 < 0$, so the maximum occurs at the critical point.

Recall that if $a > 2$, the method of moments estimators of a and b are

$$1 + \sqrt{\frac{M_2}{M_2 - M^2}}, \quad \frac{M_2}{M}\left(1 - \sqrt{\frac{M_2 - M^2}{M_2}}\right) \tag{7.3.41}$$

Open the the Pareto estimation experiment. Run the experiment 1000 times for several values of the sample size n and the parameters a and b. Compare the method of moments and maximum likelihood estimators. Which estimators seem to work better in terms of bias and mean square error?

Often the scale parameter in the Pareto distribution is known.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the Pareto distribution with unknown shape parameter $a \in (0, \infty)$ and known scale parameter $b \in (0, \infty)$. The maximum likelihood estimator of a is

$$U = \frac{n}{\sum_{i=1}^n \ln X_i - n \ln b} = \frac{n}{\sum_{i=1}^n (\ln X_i - \ln b)} \tag{7.3.42}$$

Proof

Modifying the previous proof, the log-likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is

$$\ln L_{\boldsymbol{x}}(a) = n \ln a + na \ln b - (a+1) \sum_{i=1}^n \ln x_i, \quad 0 < a < \infty \tag{7.3.43}$$

The derivative is

$$\frac{d}{da} \ln L_{\boldsymbol{x}}(a) = \frac{n}{a} + n \ln b - \sum_{i=1}^n \ln x_i \tag{7.3.44}$$

The derivative is 0 when $a = n \big/ \left(\sum_{i=1}^n \ln x_i - n \ln b\right)$. Finally, $\frac{d^2}{da^2} \ln L_{\boldsymbol{x}}(a) = -n/a^2 < 0$, so the maximum occurs at the critical point.

Uniform Distributions

In this section we will study estimation problems related to the uniform distribution that are a good source of insight and counterexamples. In a sense, our first estimation problem is the continuous analogue of an estimation problem studied in the section on Order Statistics in the chapter Finite Sampling Models. Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the uniform distribution on the interval $[0, h]$, where $h \in (0, \infty)$ is an unknown parameter. Thus, the sampling distribution has probability density function

$$g(x) = \frac{1}{h}, \quad x \in [0, h] \tag{7.3.45}$$

First let's review results from the last section.

The method of moments estimator of h is $U = 2M$. The estimator U satisfies the following properties:

1. U is unbiased.
2. $\operatorname{var}(U) = \frac{h^2}{3n}$ so U is consistent.

Now let's find the maximum likelihood estimator

The maximum likelihood estimator of h is $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$, the nth order statistic. The estimator $X_{(n)}$ satisfies the following properties:

1. $\mathbb{E}\left(X_{(n)}\right) = \frac{n}{n+1} h$
2. $\operatorname{bias}\left(X_{(n)}\right) = -\frac{h}{n+1}$ so that $X_{(n)}$ is negatively biased but asymptotically unbiased.
3. $\operatorname{var}\left(X_{(n)}\right) = \frac{n}{(n+2)(n+1)^2} h^2$
4. $\operatorname{mse}\left(X_{(n)}\right) = \frac{2}{(n+1)(n+2)} h^2$ so that $X_{(n)}$ is consistent.

Proof

The likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is $L_{\boldsymbol{x}}(h) = 1/h^n$ for $h \geq x_i$ for each $i \in \{1, 2, \ldots n\}$. The domain is equivalent to $h \geq x_{(n)}$. The function $h \mapsto 1/h^n$ is decreasing, and so the maximum occurs at the smallest value, namely $x_{(n)}$. Parts (a) and (c) are restatements of results from the section on order statistics. Parts (b) and (d) follow from (a) and (c).

Since the expected value of $X_{(n)}$ is a known multiple of the parameter h, we can easily construct an unbiased estimator.

Let $V = \frac{n+1}{n} X_{(n)}$. The estimator V satisfies the following properties:

1. V is unbiased.
2. $\var(V) = \frac{h^2}{n(n+2)}$ so that V is consistent.
3. The asymptotic relative efficiency of V to U is infinite.

Proof

Parts (a) and (b) follow from the previous result and basic properties of the expected value and variance. For part (c),

$$\frac{\var(U)}{\var(V)} = \frac{h^2/3n}{h^2/n(n+2)} = \frac{n+2}{3} \to \infty \text{ as } n \to \infty \tag{7.3.46}$$

The last part shows that the unbiased version V of the maximum likelihood estimator is a much better estimator than the method of moments estimator U. In fact, an estimator such as V, whose mean square error decreases on the order of $\frac{1}{n^2}$, is called *super efficient*. Now, having found a really good estimator, let's see if we can find a really bad one. A natural candidate is an estimator based on $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$, the *first* order statistic. The next result will make the computations very easy.

The sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ satisfies the following properties:

1. $h - X_i$ is uniformly distributed on $[0, h]$ for each i.
2. $(h - X_1, h - X_2, \ldots, h - X_n)$ is also a random sample from the uniform distribution on $[0, h]$.
3. $X_{(1)}$ has the same distribution as $h - X_{(n)}$.

Proof

1. This is a simple consequence of the fact that uniform distributions are preserved under linear transformations on the random variable.
2. This follows from (a) and that the fact that if $\boldsymbol{X}$ is a sequence of independent variables, then so is $(h - X_1, h - X_2, \ldots, h - X_n)$.
3. From part (b), $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ has the same distribution as $\min\{h - X_1, h - X_2, \ldots, h - X_n\} = h - \max\{X_1, X_2, \ldots, X_n\} = h - X_{(n)}$.

Now we can construct our really bad estimator.

Let $W = (n+1)X_{(1)}$. Then

1. W is an unbiased estimator of h.
2. $\var(W) = \frac{n}{n+2}h^2$, so W is not even consistent.

Proof

These results follow from the ones above:

1. $\mathbb{E}(X_{(1)}) = h - \mathbb{E}(X_{(n)}) = h - \frac{n}{n+1}h = \frac{1}{n+1}h$ and hence $\mathbb{E}(W) = h$.
2. $\var(W) = (n+1)^2\var(X_{(1)}) = (n+1)^2\var(h - X_{(n)}) = (n+1)^2\frac{n}{(n+1)^2(n+2)}h^2 = \frac{n}{n+2}h^2$.

Run the uniform estimation experiment 1000 times for several values of the sample size n and the parameter a. In each case, compare the empirical bias and mean square error of the estimators with their theoretical values. Rank the estimators in terms of empirical mean square error.

Our next series of exercises will show that the maximum likelihood estimator is not necessarily unique. Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the uniform distribution on the interval $[a, a+1]$, where $a \in \mathbb{R}$ is an unknown parameter. Thus, the sampling distribution has probability density function

$$g(x) = 1, \quad a \le x \le a + 1 \tag{7.3.47}$$

As usual, let's first review the method of moments estimator.

The method of moments estimator of a is $U = M - \frac{1}{2}$. The estimator U satisfies the following properties:

1. U is unbiased.

2. $\text{var}(U) = \frac{1}{12n}$ so U is consistent.

However, as promised, there is not a unique maximum likelihood estimatr.

Any statistic $V \in \left[X_{(n)} - 1, X_{(1)}\right]$ is a maximum likelihood estimator of a.

Proof

The likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is $L_{\boldsymbol{x}}(a) = 1$ for $a \le x_i \le a + 1$ and $i \in \{1, 2, \ldots, n\}$. The domain is equivalent to $a \le x_{(1)}$ and $a \ge x_{(n)} - 1$. Since the likelihood function is constant on this domain, the result follows.

For completeness, let's consider the full estimation problem. Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the uniform distribution on $[a, a + h]$ where $a \in \mathbb{R}$ and $h \in (0, \infty)$ are both unknown. Here's the result from the last section:

Let U and V denote the method of moments estimators of a and h, respectively. Then

$$U = 2M - \sqrt{3}T, \quad V = 2\sqrt{3}T \tag{7.3.48}$$

where $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean, and $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2$ is the biased version of the sample variance.

It should come as no surprise at this point that the maximum likelihood estimators are functions of the largest and smallest order statistics.

The maximum likelihood estimators or a and h are $U = X_{(1)}$ and $V = X_{(n)} - X_{(1)}$, respectively.

1. $E(U) = a + \frac{h}{n+1}$ so U is positively biased and asymptotically unbiased.
2. $E(V) = h\frac{n-1}{n+1}$ so V is negatively biased and asymptotically unbiased.
3. $\text{var}(U) = h^2 \frac{n}{(n+1)^2(n+2)}$ so U is consistent.
4. $\text{var}(V) = h^2 \frac{2(n-1)}{(n+1)^2(n+2)}$ so V is consistent.

Proof

The likelihood function corresponding to the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is $L_{\boldsymbol{x}}(a, h) = \frac{1}{h^n}$ for $a \le x_i \le a + h$ and $i \in \{1, 2, \ldots, n\}$. The domain is equivalent to $a \le x_{(1)}$ and $a + h \ge x_{(n)}$. Since the likelihood function depends only on h in this domain and is decreasing, the maximum occurs when $a = x_{(1)}$ and $h = x_{(n)} - x_{(1)}$. Parts (a)–(d) follow from standard results for the order statistics from the uniform distribution.

The Hypergeometric Model

In all of our previous examples, the sequence of observed random variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from a distribution. However, maximum likelihood is a very general method that does not require the observation variables to be independent or identically distributed. In the *hypergeometric model*, we have a population of N objects with r of the objects *type 1* and the remaining $N - r$ objects *type 0*. The *population size* N, is a positive integer. The *type 1 size* r, is a nonnegative integer with $r \le N$. These are the basic parameters, and typically one or both is unknown. Here are some typical examples:

1. The objects are devices, classified as *good* or *defective*.
2. The objects are persons, classified as *female* or *male*.
3. The objects are voters, classified as *for* or *against* a particular candidate.
4. The objects are wildlife or a particular type, either *tagged* or *untagged*.

We sample n objects from the population at random, without replacement. Let X_i be the type of the ith object selected, so that our sequence of observed variables is $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. The variables are identically distributed indicator variables, with $P(X_i = 1) = r/N$ for each $i \in \{1, 2, \ldots, n\}$, but are dependent since the sampling is without replacement. The number of type 1 objects in the sample is $Y = \sum_{i=1}^{n} X_i$. This statistic has the *hypergeometric distribution* with parameter N, r, and n, and has probability density function given by

$$P(Y=y)=\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}=\binom{n}{y}\frac{r^{(y)}(N-r)^{(n-y)}}{N^{(n)}},\quad y\in\{\max\{0,N-n+r\},\ldots,\min\{n,r\}\}\tag{7.3.49}$$

Recall the *falling power* notation: $x^{(k)}=x(x-1)\cdots(x-k+1)$ for $x\in\mathbb{R}$ and $k\in\mathbb{N}$. The hypergeometric model is studied in more detail in the chapter on Finite Sampling Models.

As above, let $\boldsymbol{X}=(X_1,X_2,\ldots,X_n)$ be the observed variables in the hypergeometric model with parameters N and r. Then

1. The maximum likelihood estimator of r with N known is $U=\lfloor NM\rfloor=\lfloor NY/n\rfloor$.
2. The maximum likelihood estimator of N with r known is $V=\lfloor r/M\rfloor=\lfloor rn/Y\rfloor$ if $Y>0$.

Proof

By a simple application of the multiplication rule, the PDF f of $\boldsymbol{X}$ is

$$f(\boldsymbol{x})=\frac{r^{(y)}(N-r)^{(n-y)}}{N^{(n)}},\quad \boldsymbol{x}=(x_1,x_2,\ldots,x_n)\in\{0,1\}^n\tag{7.3.50}$$

where $y=\sum_{i=1}^n x_i$.

1. With N known, the likelihood function corresponding to the data $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)\in\{0,1\}^n$ is

$$L_{\boldsymbol{x}}(r)=\frac{r^{(y)}(N-r)^{(n-y)}}{N^{(n)}},\quad r\in\{y,\ldots,\min\{n,y+N-n\}\}\tag{7.3.51}$$

   After some algebra, $L_{\boldsymbol{x}}(r-1)<L_{\boldsymbol{x}}(r)$ if and only if $(r-y)(N-r+1)<r(N-r-n+y+1)$ if and only if $r<Ny/n$. So the maximum of $L_{\boldsymbol{x}}(r)$ occurs when $r=\lfloor Ny/n\rfloor$.
2. Similarly, with r known, the likelihood function corresponding to the data $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)\in\{0,1\}^n$ is

$$L_{\boldsymbol{x}}(N)=\frac{r^{(y)}(N-r)^{(n-y)}}{N^{(n)}},\quad N\in\{\max\{r,n\},\ldots\}\tag{7.3.52}$$

   After some algebra, $L_{\boldsymbol{x}}(N-1)<L_{\boldsymbol{x}}(N)$ if and only if $(N-r-n+y)/(N-n)<(N-r)/N$ if and only if $N<rn/y$ (assuming $y>0$). So the maximum of $L_{\boldsymbol{x}}(r)$ occurs when $N=\lfloor rn/y\rfloor$.

In the reliability example (1), we might typically know N and would be interested in estimating r. In the wildlife example (4), we would typically know r and would be interested in estimating N. This example is known as the *capture-recapture* model.

Clearly there is a close relationship between the hypergeometric model and the Bernoulli trials model above. In fact, if the sampling is *with* replacement, the Bernoulli trials model with $p=r/N$ would apply rather than the hypergeometric model. In addition, if the population size N is large compared to the sample size n, the hypergeometric model is well approximated by the Bernoulli trials model, again with $p=r/N$.