Chapter 7
Point Estimation
Maximum Likelihood II

Outline

Asymptotic Properties

Much of the attraction of maximum likelihood estimators is based on their properties for large sample sizes. We summarizes some important properties below.

1. Consistency. If θ_0 is the state of nature and $\hat{\theta}_n(X)$ is the maximum likelihood estimator based on n observations from a simple random sample, then

$$\hat{\theta}_n(X) \to \theta_0 \quad \text{as } n \to \infty.$$

in probability.

In words, as the number of observations increase, the distribution of the maximum likelihood estimator becomes more and more concentrated about the true state of nature.

Asymptotic Properties

If θ_0 is more likely that a another parameter value θ, then

$$\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X) \quad \text{if and only if} \quad \frac{1}{n}\sum_{i=1}^{n} \ln \frac{f(X_i|\theta_0)}{f(X_i|\theta)} > 0.$$

By the strong law of large numbers, this sum converges almost surely to

$$E_{\theta_0}\left[\ln \frac{f(X_1|\theta_0)}{f(X_1|\theta)}\right].$$

which is greater than 0. Thus, for a large number of observations and a given value of $\theta \neq \theta_0$, then with a probability approaching one as $n \to \infty$, $\mathbf{L}(\theta_0|X) > \mathbf{L}(\theta|X)$ and so the maximum likelihood estimator has a high probability of being very near θ_0. This is a statement of the consistency of the estimator.

Asymptotic Properties

2. Asymptotic normality and efficiency. Under some technical assumptions

$$\sqrt{n}(\hat{\theta}_n(X) - \theta_0).$$

converges in distribution as $n \to \infty$ to a normal random variable with mean 0 and variance $1/I(\theta_0)$, the Fisher information for one observation. Thus,

$$\mathrm{Var}_{\theta_0}(\hat{\theta}_n(X)) \approx \frac{1}{nI(\theta_0)},$$

the lowest variance possible under the information inequality. Let

$$Z_n = \frac{\hat{\theta}(X) - \theta_0}{1/\sqrt{nI(\theta_0)}}.$$

Then, by Slutsky's theorem, Z_n converges in distribution to a standard normal random variable.

Asymptotic Properties

Write the linear approximation of the score function about θ_0,

$$\frac{d}{d\theta} \ln L(\theta|X) \approx \frac{d}{d\theta} \ln L(\theta_0|X) + (\theta - \theta_0) \frac{d^2}{d\theta^2} \ln L(\theta_0|X).$$

Now substitute $\theta = \hat{\theta}_n(X)$ and note that $\frac{d}{d\theta} \ln L(\hat{\theta}_n(X)|X) = 0$. Then

$$\sqrt{n}(\hat{\theta}_n(X) - \theta_0) \approx -\sqrt{n} \frac{\frac{d}{d\theta} \ln L(\theta_0|X)}{\frac{d^2}{d\theta^2} \ln L(\theta_0|X)} = \frac{\frac{1}{\sqrt{n}} \frac{d}{d\theta} \ln L(\theta_0|X)}{-\frac{1}{n} \frac{d^2}{d\theta^2} \ln L(\theta_0|X)}.$$

Asymptotic Properties

Now assume that θ_0 is the true state of nature. Then, the random variables $d \ln f(X_i|\theta_0)/d\theta$ are independent with mean 0 and variance $I(\theta_0)$. Thus, the numerator

$$\frac{1}{\sqrt{n}} \frac{d}{d\theta} \ln L(\theta_0|X) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d}{d\theta} \ln f(X_i|\theta_0)$$

converges in distribution, by the central limit theorem, to a normal random variable with mean 0 and variance $I(\theta_0)$.

Asymptotic Properties

For the denominator, $-d^2 \ln f(X_i|\theta_0)/d\theta^2, i = 1, \ldots, n$ are independent with mean $I(\theta_0)$. Thus,

$$-\frac{1}{n}\frac{d^2}{d\theta^2} \ln L(\theta_0|X) = -\frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\theta^2} \ln f(X_i|\theta_0)$$

converges, by the law of large numbers, to $I(\theta_0)$. Thus, by Slutsky's theorem, the distribution of the ratio, $\sqrt{n}(\hat{\theta}_n(X) - \theta_0)$, converges to a normal random variable with variance

$$I(\theta_0)/I(\theta_0)^2 = 1/I(\theta_0).$$

Asymptotic Properties

3. Properties of the log likelihood surface. For large sample sizes, the variance of a maximum likelihood estimator is approximately the reciprocal of the Fisher information

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln L(\theta|X)\right].$$

The Fisher information can be approximated by the observed information based on the data $\mathbf{x}$,

$$J(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\ln L(\hat{\theta}(\mathbf{x})|\mathbf{x}),$$

giving the negative of the curvature of the log-likelihood surface at the maximum likelihood estimate $\hat{\theta}(\mathbf{x})$.

- If the curvature is small near the maximum likelihood estimator, then the likelihood surface is nearly flat and the variance is large.
- If the curvature is large, the likelihood decreases quickly at the maximum and thus the variance is small.

Major League Baseball

Let's model the proportion of victories for a Major League Baseball team by a $Beta(\alpha, \alpha)$ random variable. The mean of this random variable is $1/2$. Its variance is

$$\sigma^2 = \frac{1}{4(2\alpha+1)} \quad \text{and} \quad \alpha = \frac{1}{2}\left(\frac{1}{4\sigma^2} - 1\right).$$

Thus, the method of moments estimator uses the second moment (or the variance) is

$$\hat{\alpha} = \frac{1}{2}\left(\frac{1}{4s^2} - 1\right).$$

For the 2019 season,

```
p<-c(636,593,519,414,333,623,574,447,364,292,660,599,481,444,420,
   599,574,531,500,352,562,549,519,463,426,654,525,475,438,432)/1000
> (ahatmm<-(1/(4*var(p))-1)/2)
[1] 12.52599
```

Major League Baseball

The likelihood, the log-likelihood, and its derivatice

$$\begin{aligned}
\mathbf{L}(\alpha|p) &= \frac{\Gamma(2\alpha)^n}{\Gamma(\alpha)^{2n}}\left(\prod_{i=1}^{n} p_i(1-p_i)\right)^{\alpha-1} \\
\ln \mathbf{L}(\alpha|p) &= n(\ln\Gamma(2\alpha) - 2\ln\Gamma(\alpha)) + (\alpha-1)\sum_{i=1}^{n}\ln(p_i(1-p_i)) \\
\frac{d}{d\alpha}\ln \mathbf{L}(\alpha|p) &= 2n\left(\frac{d}{d\alpha}\ln\Gamma(2\alpha) - \frac{d}{d\alpha}\ln\Gamma(\alpha)\right) + \sum_{i=1}^{n}\ln(p_i(1-p_i))
\end{aligned}$$

This derivative is 0 at the value $\hat{\alpha}$ satisfying

$$0 = 2\left(\frac{d}{d\alpha}\ln\Gamma(2\hat{\alpha}) - \frac{d}{d\alpha}\ln\Gamma(\hat{\alpha})\right) + \overline{\ln(p(1-p))}$$

Major League Baseball

We can solve numerically for the maximum likelihood estimate using the `uniroot` command in R. The derivative of the logarithm of the gamma function

$$\psi(\alpha) = \frac{d}{d\alpha} \ln \Gamma(\alpha)$$

is known as the digamma function and is called in R with `digamma`.

```
ldata<-function(a) value<-2(digamma(2*a)-digamma(a)) + mean(log(p*(1-p)))
> (ahatmle<-uniroot(ldata,c(5,100))$root)
[1] 13.13876
```

The standard deviation of $\hat{\alpha}$ is approximately $1/\sqrt{nI(\alpha)}$ for n (here $n = 30$) observations and Fisher information

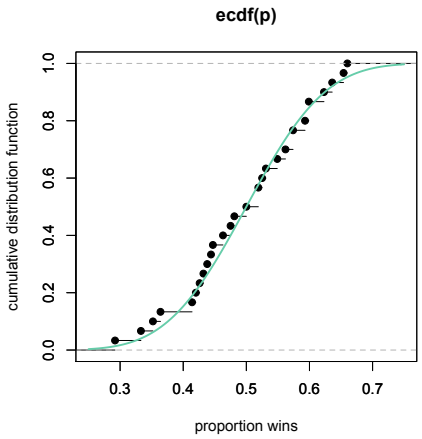
$$I(\alpha) = -\frac{d^2}{d\alpha^2} \ln \mathbf{L}(\alpha|p) = -\left(4\frac{d^2}{d\alpha^2} \ln \Gamma(2\alpha) - 2\frac{d^2}{d\alpha^2} \ln \Gamma(\alpha)\right).$$

Major League Baseball

$\psi_1(\alpha) = d^2 \ln \Gamma(\alpha)/d\alpha^2$ is known as the trigamma function and is called in R with `trigamma`.

```
> I<-function(a) -(4*trigamma(2*a)
  -2*trigamma(a))
> I(ahatmle)
[1] 0.003006483
> 1/sqrt(30*I(ahatmle))
[1] 3.329738
```

So, for Major League Baseball in 2019 the estimated values for α are $\hat{\alpha}_{mm} = 12.52599$ and $\hat{\alpha}_{mle} = 13.13876$. To compare the empirical distribution to the maximal likelihood estimate for the beta distribution



Fisher Information

For a multidimensional parameter space $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, the Fisher information $I(\theta)$ is a matrix. As with one-dimensional case, the ij-th entry has two alternative expressions, namely,

$$I(\theta)_{ij} = E_\theta \left[\frac{\partial}{\partial \theta_i} \ln L(\theta|X) \frac{\partial}{\partial \theta_j} \ln L(\theta|X) \right] = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln L(\theta|X) \right].$$

Rather than taking reciprocals to obtain an estimate of the variance, we find the matrix inverse $I(\theta)^{-1}$.

- The diagonal entries of $I(\theta)^{-1}$ gives estimates of variances.
- The off-diagonal entries of $I(\theta)^{-1}$ give estimates of covariances.

Fisher Information

To be precise, for n observations, let $\hat{\theta}_{i,n}(X)$ be the maximum likelihood estimator of the i-th parameter. Then

$$\mathrm{Var}_\theta(\hat{\theta}_{i,n}(X)) \approx \frac{1}{n} I(\theta)^{-1}_{ii} \qquad \mathrm{Cov}_\theta(\hat{\theta}_{i,n}(X), \hat{\theta}_{j,n}(X)) \approx \frac{1}{n} I(\theta)^{-1}_{ij}.$$

When the i-th parameter is θ_i, the asymptotic normality and efficiency can be expressed by noting that the z-score

$$Z_{i,n} = \frac{\hat{\theta}_i(X) - \theta_i}{\sqrt{I(\theta)^{-1}_{ii}/n}}.$$

is approximately a standard normal. As we saw in one dimension, we can replace the information matrix with the observed information matrix,

$$J(\hat{\theta})_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln L(\hat{\theta}(X)|X).$$

Monsoon Rains

We return to the model of the gamma distribution for the monsoon rainfall. To obtain the maximum likelihood estimate for the gamma family of random variables, write the likelihood

$$\begin{aligned}\mathbf{L}(\alpha,\beta|\mathbf{x}) &= \left(\frac{\beta^\alpha}{\Gamma(\alpha)}x_1^{\alpha-1}e^{-\beta x_1}\right)\cdots\left(\frac{\beta^\alpha}{\Gamma(\alpha)}x_n^{\alpha-1}e^{-\beta x_n}\right)\\ &= \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n(x_1x_2\cdots x_n)^{\alpha-1}e^{-\beta(x_1+x_2+\cdots+x_n)}.\end{aligned}$$

and its logarithm

$$\ln \mathbf{L}(\alpha,\beta|\mathbf{x}) = n(\alpha\ln\beta - \ln\Gamma(\alpha)) + (\alpha-1)\sum_{i=1}^n \ln x_i - \beta\sum_{i=1}^n x_i.$$

The score function is a vector $\left(\frac{\partial}{\partial\alpha}\ln\mathbf{L}(\alpha,\beta|\mathbf{x}), \frac{\partial}{\partial\beta}\ln\mathbf{L}(\alpha,\beta|\mathbf{x})\right)$.

Gamma Distribution

$$\ln \mathbf{L}(\alpha, \beta|\mathbf{x}) = n(\alpha \ln \beta - \ln \Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^{n} \ln x_i - \beta \sum_{i=1}^{n} x_i.$$

The zeros of the components of the score function determine the maximum likelihood estimators. Thus, to determine these parameters, we solve the equations

$$\frac{\partial}{\partial \alpha} \ln \mathbf{L}(\hat{\alpha}, \hat{\beta}|\mathbf{x}) = n(\ln \hat{\beta} - \frac{d}{d\alpha} \ln \Gamma(\hat{\alpha})) + \sum_{i=1}^{n} \ln x_i = 0$$

and $$\frac{\partial}{\partial \beta} \ln \mathbf{L}(\hat{\alpha}, \hat{\beta}|\mathbf{x}) = n\frac{\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^{n} x_i = 0, \quad \text{or} \quad \bar{x} = \frac{\hat{\alpha}}{\hat{\beta}}.$$

Substituting $\hat{\beta} = \hat{\alpha}/\bar{x}$ into the first equation results the following relationship for $\hat{\alpha}$.

$$n(\ln \hat{\alpha} - \ln \bar{x} - \frac{d}{d\alpha} \ln \Gamma(\hat{\alpha}) + \overline{\ln x}) = 0$$

Gamma Distribution

This can be solved numerically.

```
> ldata<-function(a) log(a)-digamma(a)-log(mean(x))+mean(log(x))
> (ahat<-uniroot(ldata,c(0.01,10))$root)
[1] 0.9333061
> (bhat<-ahat/mean(x))
[1] 0.07591485
```

Exercise. To determine the variance of these estimators, compute the appropriate second derivatives.

$$I(\alpha,\beta)_{11} = -\frac{\partial^2}{\partial\alpha^2}\ln \mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha), \quad I(\alpha,\beta)_{22} = -\frac{\partial^2}{\partial\beta^2}\ln \mathbf{L}(\alpha,\beta|\mathbf{x}) = n\frac{\alpha}{\beta^2},$$

$$I(\alpha,\beta)_{12} = -\frac{\partial^2}{\partial\alpha\partial\beta}\ln \mathbf{L}(\alpha,\beta|\mathbf{x}) = -n\frac{1}{\beta}.$$

Gamma Distribution

This give a Fisher information matrix

$$I(\alpha,\beta)=n\begin{pmatrix}\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha) & -\frac{1}{\beta}\\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2}\end{pmatrix}\quad I(0.933,0.076)=30\begin{pmatrix}1.821 & -13.173\\ -13.173 & 161.946\end{pmatrix}.$$

The inverse matrix

```
> I2<-matrix(c(1.821056,-13.17265,-13.17265,161.9461),ncol=2)
> solve(I2)
          [,1]       [,2]
[1,] 1.3340513 0.10851135
[2,] 0.1085114 0.01500118
```

Gamma Distribution

$$I(\alpha,\beta)^{-1} = \frac{1}{30}\begin{pmatrix} 1.334 & 0.109 \\ 0.109 & 0.015 \end{pmatrix}.$$

Thus,

$$\mathrm{Var}(\hat{\alpha}) \approx 0.0444 \qquad \sigma_{\hat{\alpha}} \approx 0.211$$
$$\mathrm{Var}(\hat{\beta}) \approx 0.0005 \qquad \sigma_{\hat{\beta}} \approx 0.0224$$

Compare this with the method of moments estimators

```
> t<-replicate(10000,g(rgamma(17,0.9045441,0.07357536)))
> sd(t[1,]^2/t[2,]) #standard deviation for alphahat
[1] 0.4866726
> sd(t[1,]/t[2,])    #standard deviation for betahat
[1] 0.0509116
```

Exercise. Estimate the correlation $\rho(\hat{\alpha},\hat{\beta})$. 0.767

Gamma Distribution

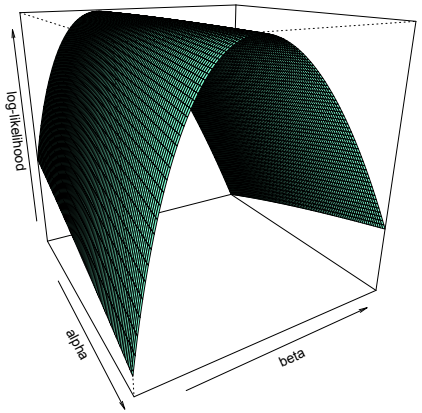


Figure: The log-likelihood surface. The graph shows $0.86 \leq \alpha \leq 1.00$ and $0.03 \leq \beta \leq 0.17$

The log-likelihood surface for the gamma distribution is shown in a neighborhood about the maximum likelihood estimate.

- The curvature in the α is very small, reflecting a relatively large variance.
- The curvature in the β is very large, reflecting a relatively small variance.

Exercise. Find the eigenvector and eigenvalues for the covariance matrix. What do they signify.