CHAPTER 3 - ASYMPTOTIC THEORY OF MAXIMUM LIKELIHOOD ESTIMATORS

§3.1 Asymptotic theory

Theorem 3.1

Assume $X_1, \ldots, X_n$ are independently and identically distributed from distribution $f_\theta(x)$. Provided the Cramer-Rao lower bound exists, then asymptotically the Maximum Likelihood Estimates satisfy:

- $\hat{\theta}$ is consistent for θ.
- $\hat{\theta}$ is asymptotically unbiased for θ.
- $\hat{\theta}$ is asymptotically fully efficient.
- $\hat{\theta}$ is asymptotically Normally distributed.

$$\hat{\theta} \sim N(\theta, I_n^{-1}(\theta))$$

where $I_n(\theta)$ is the Fisher Information Matrix which has entries

$$[I_n(\theta)]_{i,j} = -\mathsf{E}\left[\frac{\partial^2 \ln f_\theta(x_1, \ldots, x_n)}{\partial\theta_i \partial\theta_j}\right].$$

We do not cover a proof of this result in this course (although there is a link to some related, non-examinable resources on the unit webpage).

Examples

1. Exponential Suppose $X_1, \ldots, X_n$ are iid $Exp(\lambda)$ random variables where λ is the rate parameter:

$$\begin{aligned}
f_\lambda(\mathbf{x}) &= \lambda^n \exp(-\lambda \sum_{i=1}^n x_i) \\
\ln f_\lambda(\mathbf{x}) &= n \ln \lambda - \lambda \sum_{i=1}^n x_i \\
\frac{d \ln f_\lambda(\mathbf{x})}{d\lambda} &= \frac{n}{\lambda} - \sum_{i=1}^n x_i \\
\frac{d^2 \ln f_\lambda(\mathbf{x})}{d\lambda^2} &= -\frac{n}{\lambda^2}
\end{aligned}$$

From this we learn that the MLE is $\hat{\lambda} = n / \sum_{i=1}^n X_i$ and $I_n(\lambda) = n/\lambda^2$. Although we do not know the exact distribution of $\hat{\lambda}$, Theorem 3.1 tells us that, asymptotically

$$\hat{\lambda} \sim N(\lambda, \frac{\lambda^2}{n}).$$

2. Normal Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ random variables.

$$\begin{aligned}
f_{\mu,\sigma^2}(\mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2) \\
\ln f_{\mu,\sigma^2}(\mathbf{x}) &= \frac{-n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 \\
\frac{\partial \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\mu) \\
\frac{\partial \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\sigma^2} &= \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (x_i-\mu)^2 \\
\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu^2} &= \frac{-n}{\sigma^2} \\
\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial(\sigma^2)^2} &= \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^n (x_i-\mu)^2 \\
\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu\partial\sigma^2} &= \frac{-1}{(\sigma^2)^2}\sum_{i=1}^n (x_i-\mu)
\end{aligned}$$

Simultaneously solving $\frac{\partial \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu} = 0$ with $\frac{\partial \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\sigma^2} = 0$, gives

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

at which point, $\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu^2} < 0$, $\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial(\sigma^2)^2} < 0$ and $\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu\partial\sigma^2} = 0$, so these are indeed the Maximum Likelihood Estimators. To find the asymptotic distribution of the MLEs, we need to use

$$\begin{aligned}
\mathsf{E}\left[\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu^2}\right] &= \frac{-n}{\sigma^2} \\
\mathsf{E}\left[\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial(\sigma^2)^2}\right] &= \frac{-n}{2(\sigma^2)^2} \\
\mathsf{E}\left[\frac{\partial^2 \ln f_{\mu,\sigma^2}(\mathbf{x})}{\partial\mu\partial\sigma^2}\right] &= 0 \\
\text{so that } I_n(\mu,\sigma^2) &= \left[\begin{array}{cc} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{array}\right] \\
\text{and thus, asymptotically} \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} &\sim N_2\left[\begin{array}{cc} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2(\sigma^2)^2}{n} \end{array}\right]
\end{aligned}$$

§3.2 Applications

The applications of the asymptotic theory will largely be in finding large sample approximate confidence intervals and/or hypothesis tests.

Example of an approximate confidence interval

Suppose $X_1, \ldots, X_n$ iid $Bern(p)$ and that we are interested in finding an approximate confidence interval for the odds ratio $\tau = \frac{p}{1-p}$.

$$
\begin{aligned}
f_p(\mathbf{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
L(p) &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \\
\ell(p) &= (\sum_{i=1}^{n} x_i)\ln p + (n - \sum_{i=1}^{n} x_i)\ln(1-p) \\
\frac{d\ell(p)}{dp} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} \\
\frac{d^2\ell(p)}{dp^2} &= -\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-p)^2} \\
\text{So } \hat{p} &= \frac{\sum_{i=1}^{n} X_i}{n} \\
\text{and } I_n(p) &= \frac{\sum_{i=1}^{n} \mathsf{E}(X_i)}{p^2} + \frac{n - \sum_{i=1}^{n} \mathsf{E}(X_i)}{(1-p)^2} \\
&= \frac{n}{p(1-p)} \\
\text{i.e. } \hat{p} &\sim N\left(p, \frac{p(1-p)}{n}\right) \text{ asymptotically.}
\end{aligned}
$$

Using the functional invariance of MLEs and the full generality of the CRLB, we can then see that

$$
\begin{aligned}
\hat{\tau} &= \frac{\hat{p}}{1-\hat{p}} \\
\frac{d\tau}{dp} &= \frac{1}{(1-p)^2} \\
\text{so } \hat{\tau} &\sim N\left(\tau, \frac{p(1-p)}{n} \frac{1}{(1-p)^4}\right) \text{ asymptotically} \\
&\sim N\left(\tau, \frac{p}{n(1-p)^3}\right) \text{ asymptotically.}
\end{aligned}
$$

To find an approximate confidence interval for τ, we would use the asymptotic distribution. However as that requires knowing the values of p for the mean and variance, an extra layer of approximation is made by substituting in the MLE of p. So, an approximate 95% confidence interval for τ would be

$$
\frac{\hat{p}}{1-\hat{p}} - 1.96\sqrt{\frac{\hat{p}}{n(1-\hat{p})^3}} \leq \tau \leq \frac{\hat{p}}{1-\hat{p}} + 1.96\sqrt{\frac{\hat{p}}{n(1-\hat{p})^3}} .
$$

Example of an approximate hypothesis test

Suppose $X_1, \ldots, X_n$ are iid from the double exponential distribution with positive parameters λ_+ and λ_-:

$$\begin{aligned} f_{\lambda_+,\lambda_-}(x) &= \frac{\lambda_+}{2}\exp(-\lambda_+ x) I_{[x>0]} + \frac{\lambda_-}{2}\exp(\lambda_- x) I_{[x<0]}, \quad -\infty < x < \infty \\ \text{then } f_{\lambda_+,\lambda_-}(\mathbf{x}) &= \frac{1}{2^n}\,\lambda_+^{n_+}\exp(-\lambda_+\sum_{i=1}^n x_i I_{[x_i>0]})\,\lambda_-^{n_-}\exp(\lambda_-\sum_{i=1}^n x_i I_{[x_i<0]}) \end{aligned}$$

where $n_+ = \sum_{i=1}^n I_{[x_i>0]}$ and $n_- = \sum_{i=1}^n I_{[x_i<0]}$ with $n_+ + n_- = n$. Taking logs and differentiating

$$\begin{aligned} \ln f_{\lambda_+,\lambda_-}(\mathbf{x}) &= -n\ln 2 + n_+\ln\lambda_+ - \lambda_+\sum_{i=1}^n x_i I_{[x_i>0]} + n_-\ln\lambda_- + \lambda_-\sum_{i=1}^n x_i I_{[x_i<0]} \\ \frac{\partial \ln f_{\lambda_+,\lambda_-}(\mathbf{x})}{\partial\lambda_+} &= \frac{n_+}{\lambda_+} - \sum_{i=1}^n x_i I_{[x_i>0]} \\ \frac{\partial \ln f_{\lambda_+,\lambda_-}(\mathbf{x})}{\partial\lambda_-} &= \frac{n_-}{\lambda_-} + \sum_{i=1}^n x_i I_{[x_i<0]} \\ \frac{\partial^2 \ln f_{\lambda_+,\lambda_-}(\mathbf{x})}{\partial\lambda_+^2} &= -\frac{n_+}{\lambda_+^2} \\ \frac{\partial^2 \ln f_{\lambda_+,\lambda_-}(\mathbf{x})}{\partial\lambda_-^2} &= -\frac{n_-}{\lambda_-^2} \\ \frac{\partial^2 \ln f_{\lambda_+,\lambda_-}(\mathbf{x})}{\partial\lambda_+\partial\lambda_-} &= 0 \end{aligned}$$

From this we conclude that, provided neither n_+ nor n_- are zero, the MLEs of the two parameters are

$$\hat\lambda_+ = \frac{n_+}{\sum_{i=1}^n x_i I_{[x_i>0]}} \quad \hat\lambda_- = -\frac{n_-}{\sum_{i=1}^n x_i I_{[x_i<0]}}.$$

To test for equality of λ_+ and λ_-, we will use the asymptotic distribution of $\hat\lambda_+$ and $\hat\lambda_-$:

$$\begin{aligned} I_n(\lambda_+,\lambda_-) &= \begin{bmatrix} -E\left(-\frac{n_+}{\lambda_+^2}\right) & 0 \\ 0 & -E\left(-\frac{n_-}{\lambda_-^2}\right) \end{bmatrix} \\ \mathsf{E}(n_+) &= \mathsf{E}(\sum_{i=1}^n I_{[x_i>0]}) \\ &= \sum_{i=1}^n \mathsf{P}([x_i>0]) \\ &= \frac{n}{2} \text{ as does } \mathsf{E}(n_-) \\ \text{so } I_n(\lambda_+,\lambda_-) &= \begin{bmatrix} \frac{n}{2\lambda_+^2} & 0 \\ 0 & \frac{n}{2\lambda_-^2} \end{bmatrix} \\ \text{and } \begin{pmatrix} \hat\lambda_+ \\ \hat\lambda_- \end{pmatrix} &\sim N_2\left(\begin{pmatrix} \lambda_+ \\ \lambda_- \end{pmatrix}, \begin{bmatrix} \frac{2\lambda_+^2}{n} & 0 \\ 0 & \frac{2\lambda_-^2}{n} \end{bmatrix}\right) \text{ asymptotically.} \end{aligned}$$

To test

$$H_0 : \lambda_+ = \lambda_- \quad \text{vs} \quad H_1 : \lambda_+ \neq \lambda_-$$

we note that this is equivalent to

$$\begin{aligned}
H_0 : \lambda_+ - \lambda_- = 0 \qquad \text{vs} \qquad H_1 : \lambda_+ - \lambda_- \neq 0 \\
\hat{\lambda}_+ - \hat{\lambda}_- \quad &\sim \quad N\left(\lambda_+ - \lambda_-, \frac{2\lambda_+ 2}{n} + \frac{2\lambda_-^2}{n}\right) \text{ asymptotically} \\
\text{so } \hat{\lambda}_+ - \hat{\lambda}_- \quad &\sim \quad N\left(0, \frac{2\lambda_+^2}{n} + \frac{2\lambda_-^2}{n}\right) \text{ asymptotically under } H_0.
\end{aligned}$$

As with the previous example, the variance term here can only be approximated (by plugging in the MLEs). We would then reject the null hypothesis at the 5% level (for example) if

$$\left| \frac{\hat{\lambda}_+ - \hat{\lambda}_-}{\sqrt{\frac{2\hat{\lambda}_+^2}{n} + \frac{2\hat{\lambda}_-^2}{n}}} \right| > 1.96.$$