1 *The Likelihood Function*

In the previous chapter, we introduced statistical models $\{P_\theta : \theta \in \Omega\}$ which describe the probability models that could generate the observed data. In this chapter we will develop inferences that depend only on the model $\{P_\theta : \theta \in \Omega\}$ and the data s.

The simplest setting has the statistical model resulting in discrete distribution. Suppose that we observe the data s and that the pmf is p_θ. The **likelihood function** $L(\cdot|s)$ is defined on the parameter space Ω by

$$L(\theta|s) = p_\theta(s), \qquad \theta \in \Omega.$$

For the observed data s, the likelihood is the probability of observing s when the true value of the parameter is θ. This induces an ordering on Ω in that we believe that θ_1 is more plausible as the true value of the parameter than θ_2 if

$$p_{\theta_1}(s) > p_{\theta_2}(s) \qquad \text{or} \qquad L(\theta_1|s) > L(\theta_2|s).$$

Remarks:

- We note that $L(\theta|s)$ is the probability of the value s given that the true value of the parameter is θ.
- The likelihood $L(\theta|s)$ is not the probability of θ given that we have observed s.
- In many cases the value of $L(\theta|s)$ is small for all θ. Thus, we are interested in the relative value of the likelihood.
- The above implies that we should consider likelihood ratios

$$\frac{L(\theta_1|s)}{L(\theta_2|s)}$$

in determining inferences for θ based on the likelihood function.

We now define the **likelihood function** for a random sample $X_1, \ldots, X_n$ from a distribution with pmf or pdf $f_\theta(\,\cdot\,)$. The joint pmf or pdf $X_1, \ldots, X_n$ is

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Now, suppose that $x_1, \ldots, x_n$ are the observed values of $X_1, \ldots, X_n$. Then the likelihood function is

$$L(\theta | x_1 \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

The notation $L(\theta | x_1, \ldots, x_n)$ indicates that we regard $L(\cdot | x_1, \ldots, x_n)$ as a function of θ. After the data are observed, $x_1 \ldots, x_n$ are viewed as constants.

Remark: In the case where we have several parameters $\theta_1, \ldots, \theta_k$, we define the **likelihood function** for a random sample $X_1, \ldots, X_n$ from a distribution with pmf or pdf as $f_{\theta_1,\ldots,\theta_k}(\,\cdot\,)$

$$L(\theta_1, \ldots, \theta_k | x_1 \ldots, x_n) = \prod_{i=1}^{n} f_{\theta_1,\ldots,\theta_k}(x_i).$$

For a random sample from a discrete distribution with one parameter θ,

$$L(\theta|x_1,\ldots,x_n) = P_\theta[X_1 = x_1,\ldots,X_n = x_n].$$

We compare the likelihood function at two parameter values, θ_1 and θ_2. If

$$L(\theta_1|x_1,\ldots,x_n) > L(\theta_2|x_1,\ldots,x_n),$$

then the observed data $X_1 = x_1,\ldots,X_n = x_n$ are more likely to have occured if $\theta = \theta_1$ than if $\theta = \theta_2$.

For a continuous rv X with pdf $f_\theta(x)$, we can write that

$$P_\theta[x-\epsilon < X < x+\epsilon] = \int_{x-\epsilon}^{x+\epsilon} f_\theta(x)dx \approx 2\epsilon f_\theta(x) = 2\epsilon L(\theta|x).$$

Thus, if

$$\frac{2\epsilon L(\theta_1|x)}{2\epsilon L(\theta_2|x)} = \frac{L(\theta_1|x)}{L(\theta_2|x)} > 1,$$

we feel that X is more likely to be near x when $\theta = \theta_1$.

Example 44: In a sample of 50 adult Americans, only 14 correctly described the Bill of Rights as the first ten amendments to the U. S. Constitution. Estimate the proportion of Americans that can give a correct description of the Bill of Rights.

Let $Y_i = 1$ if the person is correct and $Y_i = 0$ if the person is wrong. Then the pmf of a single observation is

$$p_\theta(y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$$

and the likelihood function is

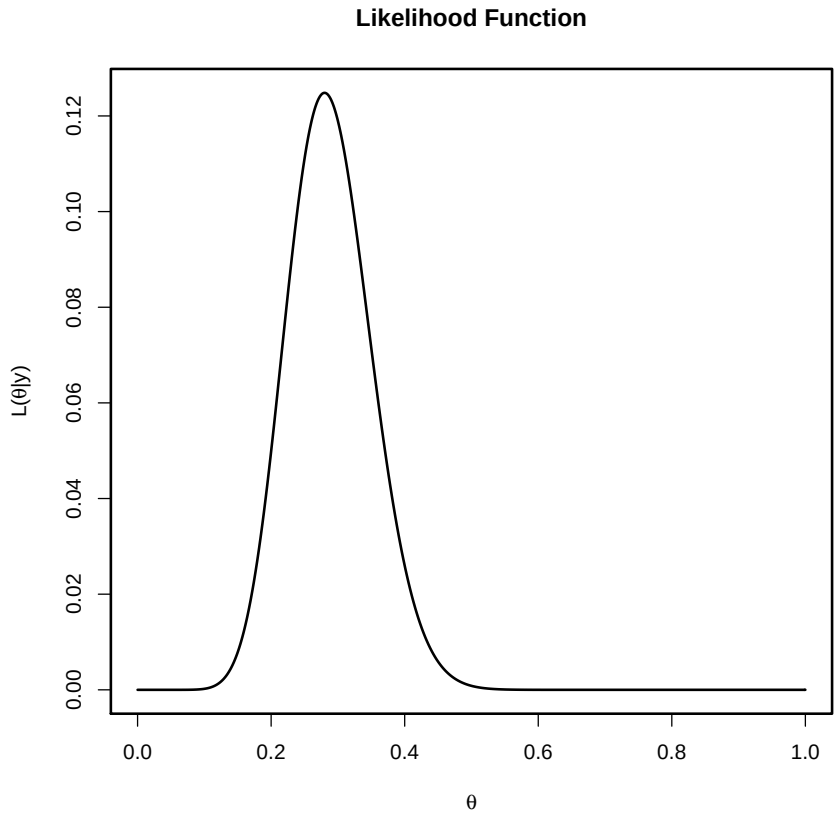
$$L(\theta|y_1,\ldots,y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{y}(1-\theta)^{n-y}.$$

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

For our data,

$$L(\theta|y) = \theta^y(1-\theta)^{n-y} = \theta^{14}(1-\theta)^{50-14}.$$

The following is a plot of the likelihood function:



1.1 Sufficient Statistics

In the previous example, the value of the likelihood function depended on the observed sample $\{y_1, \ldots, y_n\}$ only through $y = \sum_{i=1}^{n} y_i$. Such a simplification of the likelihood function will occur for many of the models that we use in this course.

A statistic $T(s)$ is said to be a sufficient statistic for the model $\{P_\theta : \theta \in \Omega\}$ (or simply for θ) if, whenever $T(s_1) = T(s_2)$, then

$$L(\theta|s_1) = c(s_1, s_2)L(\theta|s_2).$$

We will typically use the Factorization Theorem to check for a sufficient statistic:

Factorization Theorem Suppose that the density (or pmf) for the model P_θ is given by $f_\theta(s)$. Then T is a sufficient statistic for the model if the density factors

$$f_\theta(s) = h(s)g_\theta(T(s)),$$

where g_θ and h are nonnegative and h does not depend on θ.

Example 44:

The pmf of a single observation is $p_\theta(y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$, and the likelihood function (or joint pmf) is

$$f_\theta(y_1,\ldots,y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{T(y_1,\ldots,y_n)}(1-\theta)^{n-T(y_1,\ldots,y_n)} \times 1.$$

where $T(y_1,\ldots,y_n) = \sum_{i=1}^n y_i$. Set $g_\theta(T) = \theta^T(1-\theta)^{n-T}$ and $h(y_1,\ldots,y_n) = 1$. By the Factorization Theorem, T is sufficient for θ.

1.2 Maximum likelihood estimates

We now use the likelihood $L(\theta|s)$ for the data s to define a point estimate of θ.

For the observed data s, the *maximum likelihood estimates* are values $\hat{\theta}$ such that

$$L(\hat{\theta}(s)|s) \geq L(\theta|s).$$

for all θ in the parameter space Ω.

The intuition of maximum likelihood is that $\hat{\theta}$ is the parameter value such that the observed data s *is the most probable*. In a certain sense, then, the maximum likelihood estimates are those parameter values that *are the most consistent with the observed data*.

Often we use the log-likelihood function for inference:

$$\ell(\theta|s) = \log(L(\theta|s))$$

Since $\log(x)$ is a one-to-one and increasing function of x, $\ell(\theta_1|s) > \ell(\theta_2|s)$ iff $L(\theta_1|s) > L(\theta_2|s)$. Thus, we could have defined the mle as the value of the parameter(s) that maximizes the log-likelihood.

Log-likelihood for a Random Sample:

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f_\theta(\,\cdot\,)$. Then the likelihood function is

$$L(\theta|x_1 \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

and the log-likelihood function is

$$\ell(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{N} \log(f_\theta(x_i)).$$

Multiparameter Case: For the observed data s, the *maximum likelihood estimates* are values $\hat{\theta}_1(s), \ldots, \hat{\theta}_k(s)$ such that

$$L(\hat{\theta}_1(s), \ldots, \hat{\theta}_k(s)|s) \geq L(\theta_1, \ldots, \theta_k|s).$$

for all $(\theta_1, \ldots, \theta_k)$ in the parameter space Ω. As in the single parameter case, we can use log-likelihood function for inference:

$$\ell(\theta_1, \ldots, \theta_k|s) = \log(L(\theta_1, \ldots, \theta_k|s)).$$

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f_{\theta_1,\ldots,\theta_k}(\,\cdot\,)$. Then the likelihood function is

$$L(\theta_1, \ldots, \theta_k|x_1 \ldots, x_n) = \prod_{i=1}^{n} f_{\theta_1,\ldots,\theta_k}(x_i),$$

and the log-likelihood function is

$$\ell(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \sum_{i=1}^{N} \log(f_{\theta_1,\ldots,\theta_k}(x_i)).$$

1.3 Computation of the MLE

We now consider the case where θ is a single parameter ($k = 1$) and Ω is contained in the real line. The mle is the value $\hat{\theta}(s)$ that maximizes the likelihood $L(\theta|s)$ or equivalently, the log-likelihood $\ell(\theta|s) = \log(L(\theta|s))$.

The score function is defined to be the first partial derivative of the log-likelihood function with respect to θ :

$$S(\theta|s) = \frac{\partial \ell(\theta|s)}{\partial \theta}.$$

To obtain the MLE, we solve the score equation:

$$S(\theta|s) = \frac{\partial \ell(\theta|s)}{\partial \theta} = 0.$$

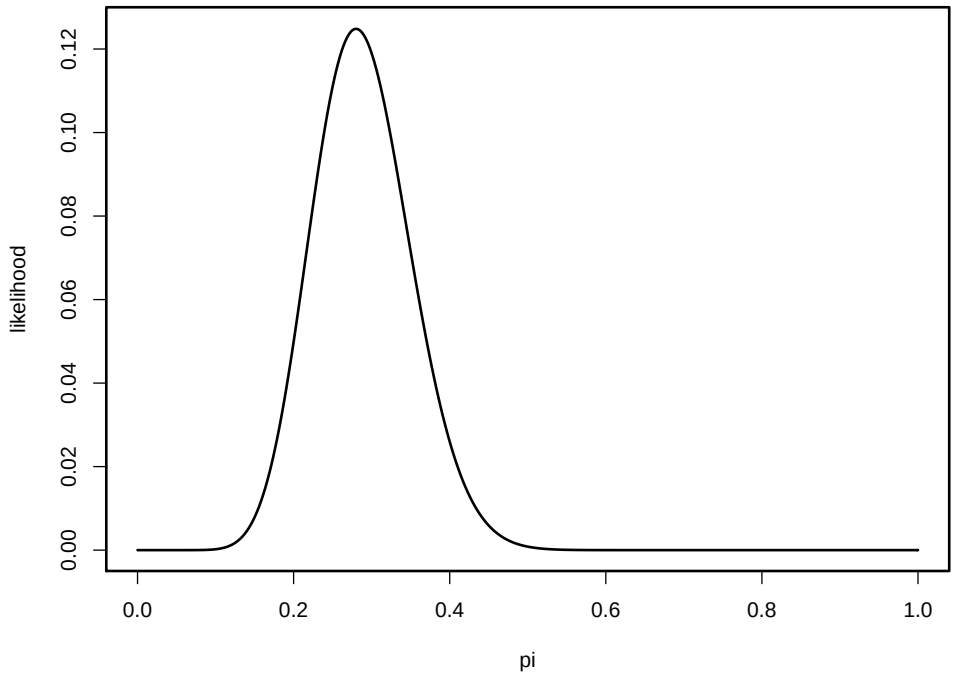
We also need to check that our solution is an global maximum. To check for a local maximum, we can check that

$$\left.\frac{\partial S(\theta|s)}{\partial \theta}\right|_{\theta=\hat{\theta}(s)} = \left.\frac{\partial^2 \ell(\theta|s)}{\partial \theta^2}\right|_{\theta=\hat{\theta}(s)} < 0.$$

Example 44: The likelihood function is

$$L(\theta|y_1,\ldots,y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{y}(1-\theta)^{n-y} = \theta^{14}(1-\theta)^{50-14}.$$

The following is a plot of the likelihood function:



It is an easier calculus problem to maximize the log-likelihood:

$$\begin{aligned} \ell(\theta|y) &= \log(L(\theta|y)) = \log\left[\theta^y(1-\theta)^{n-y}\right] \\ &= y\log(\theta) + (n-y)\log(1-\theta) \end{aligned}$$

To maximize, take the derivative and set $= 0$ to obtain the score equation:

$$\frac{\partial \ell(\theta|y)}{\partial \theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta} = 0$$

We obtain the mle:

$$\hat{\theta}(y) = \frac{y}{n} = \frac{\sum_i y_i}{n} = \frac{14}{50} = 0.28$$

Example 45 Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $f_{\mu,\sigma}(\,\cdot\,)$. The likelihood function is

$$L(\mu, \sigma | x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right].$$

We need to maximize the log-likelihood, $\ell = \log(L)$.

$$\begin{aligned} \ell(\mu, \sigma) &= \log(L(\mu, \sigma | x_1, \ldots, x_n)) \\ &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned}$$

To find the maximizer of $\ell(\mu, \sigma)$, take partial derivatives with respect to μ and σ. Then set them equal to 0.

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu).$$

Setting this equal to 0 yields $\mu = \bar{x}$. Now

$$\frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

$$\Longrightarrow \qquad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Substituting in $\mu = \bar{x}$, we obtain

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

By computing second partials, one may verify that, in fact,

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

maximize the likelihood. The quantity $\hat{\sigma}^2$ is a version of the *sample variance*.

Since we know that μ and σ^2 are the mean and variance of the normal distribution, it makes sense that we would use the sample mean and variance to estimate them.

We call the quantities $\bar{x}$ and $\hat{\sigma}^2$ maximum likelihood estimates. They are fixed values computed from the observed values $x_1, \ldots, x_n$.

If we replace $x_1, \ldots, x_n$ in the above expression by the rvs $X_1, \ldots, X_n$, we obtain maximum likelihood estimators which are random variables and have a probability distribution called a sampling distribution. We derived the sampling distributions of $\bar{X}$ and $\hat{\sigma}^2$ in Chapter 4.

Example 46 Let $X_1, \ldots, X_n$ be a random sample from the exponential(λ) distribution. Find the mle of λ.

The log-likelihood is

$$\ell(\lambda) = \log(L(\lambda|x_1, \ldots, x_n)) = \log\left(\lambda^n e^{-\lambda \sum_{i=1}^n x_i}\right) = n\log(\lambda) - \lambda \sum_{i=1}^n x_i.$$

We differentiate the log-likelihood and set $= 0$ to get the score equation:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0.$$

Now solve to get the mle of λ:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i}.$$

Example 47 Suppose $X_1, \ldots, X_n$ is a random sample from the gamma distribution

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{(0,\infty)}(x).$$

The log-likelihood is

$$\begin{aligned}
\ell(\alpha,\lambda) &= \log(L(\alpha,\lambda|x_1,\ldots,x_n)) = \log\left(\prod_{i=1}^n \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i}\right) \\
&= n\alpha\log(\lambda) + (\alpha-1)\sum_{i=1}^n \log(x_i) - \lambda\sum_{i=1}^n x_i - n\log(\Gamma(\alpha))
\end{aligned}$$

The score equations (also called likelihood equations) are

$$\frac{\partial\ell(\alpha,\lambda)}{\partial\alpha} = n\log\lambda + \sum_{i=1}^n \log(x_i) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0$$

and

$$\frac{\partial\ell(\alpha,\lambda)}{\partial\lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^n x_i = 0.$$

We solve the second equation for λ:

$$\hat{\lambda} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} x_i} = \frac{\hat{\alpha}}{\bar{x}}.$$

We substitute this into the first equation to obtain a nonlinear equation for $\hat{\alpha}$:

$$n\log(\hat{\alpha}) - n\log(\bar{x}) + \sum_{i=1}^{n}\log(x_i) - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0.$$

This equation cannot be solved in closed form. An iterative method of solving the equation must be used. This requires starting values such as the moment estimates which we will cover later.

Example 48 Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on the interval $[0, \theta],\ \theta > 0$. Find the mle of θ.

The likelihood is

$$L(\theta|x_1, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n}, & 0 \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

This is a decreasing function of θ for $\theta > 0$. (Why?)

Also, the constraints on the x_is imply that $x_i \leq \theta$, all i. Equivalently, $x_{(n)} = \max\{x_1, \ldots, x_n\} \leq \theta$.

To maximize a decreasing function, we take the smallest allowable value of θ to be the mle:

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

1.4 Invariance Property of MLEs

A basic property of the MLE is invariance. Suppose we let $X_1, \ldots, X_n$ form a random sample from a distribution with pmf or pdf $f_\theta(x)$ and let $\hat{\theta}$ be the mle of θ.

Consider the alternative parameterization using $\tau = \psi(\theta)$ where ψ is a one-to-one function of θ. The *plug-in estimate* of τ is given by $\hat{\tau} = \psi(\hat{\theta})$. Alternatively, we could find the mle of τ using the new parameterization.

The invariance property of the mle says that it makes no difference which parameterization we use for the finding the mle:

- If $\hat{\theta}$ is the mle of θ and $\psi(\theta)$ is a one-to-one function, then $\psi(\hat{\theta})$ is the mle of $\psi(\theta)$.
- Example 6.2.7 in the text shows that a plug-in estimate can behave badly when ψ is not one-to-one.

Example 44 again Consider a sequence of Bernoulli trials. Let $Y_i = 1$ if the person is correct and $Y_i = 0$ if the person is wrong. Then the pmf of a single observation is

$$f_\theta(y_i) = \theta^{y_i}(1-\theta)^{1-y_i},$$

and the likelihood function is

$$L(\theta|y) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{y}(1-\theta)^{n-y},$$

where $y = \sum_{i=1}^{n} y_i$. The mle of θ was found to be

$$\hat{\theta} = \frac{Y}{n}.$$

Consider the new parameter and its mle using invariance:

$$\xi = \frac{\theta}{1-\theta}, \quad \hat{\xi} = \frac{\hat{\theta}}{1-\hat{\theta}} = \frac{\frac{Y}{n}}{1-\frac{Y}{n}} = \frac{Y}{n-Y}.$$

The likelihood in terms of ξ is

$$\text{lik}(\xi) = \prod_{i=1}^{n} \frac{\xi^{y_i}}{\xi + 1} = \frac{\xi^{y}}{(\xi + 1)^{n}}.$$

The likelihood equation is

$$\frac{\partial \ell(\xi)}{\partial \xi} = \frac{\partial}{\partial \xi}[y \log(\xi) - n \log(\xi + 1)] = \frac{y}{\xi} - \frac{n}{\xi + 1}$$

Set this equal to zero and solve to get the mle of ξ:

$$\hat{\xi} = \frac{Y}{n - Y}.$$

2 *Method of Moments Estimators*

We now introduce another approach to estimation of a parameter θ when sampling from a distribution with pmf/pdf $f_\theta(x)$ where $\theta \in \Omega$. Recall that the first population moment of a rv X is

$$\mu_X = E_\theta(X).$$

Note: $E_\theta(X)$ denotes the expectation of X using the distribution with pmf/pdf $f_\theta(x)$. This expectation typically is a function of θ.

We now introduce sample moments. Given a random sample $X_1, \ldots, X_n$, the first sample moment is $m_1 = \bar{X}$, the sample mean.

Method of moments:

Express the population mean as a function of the unknown parameter θ, solve for θ, and substitute the sample mean for the population mean.

Example 46 again Let $X_1, \ldots, X_n$ be a random sample from the exponential(λ) distribution. Find the moment estimator of λ.

The exponential(λ) distribution has pdf

$$f_\lambda(x) = \lambda e^{-\lambda x},\ x > 0.$$

From the properties of the gamma distribution, the mean of an exponential rv is

$$E_\lambda(X) = \frac{1}{\lambda}.$$

Solve for λ:

$$\lambda = \frac{1}{E_\lambda(X)}.$$

Substitute $\bar{X}$ for $E_\lambda(X)$:

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

We can extend the method of moments to estimation of multiple parameters $\theta_1, \ldots, \theta_k$ when sampling from a distribution with pmf/pdf $f_{\theta_1,\ldots,\theta_k}(x)$ where $(\theta_1, \ldots, \theta_k) \in \Omega$. Recall that the j^{th} population moment of a rv X is

$$\mu_j = E_{\theta_1,\ldots,\theta_k}(X^j).$$

Note: $E_{\theta_1,\ldots,\theta_k}(X^j)$ denotes the expectation of X^j using the distribution with pmf/pdf $f_{\theta_1,\ldots,\theta_k}(x)$.

We now introduce sample moments. Given a random sample $X_1, \ldots, X_n$, the jth sample moment is

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad j = 1, 2, \ldots .$$

Multiparameter method of moments:

The population moments are functions of the unknown parameters. Express unknown parameters as functions of the population moments, and then substitute sample moments for population moments.

Example 47 again Suppose $X_1, \ldots, X_n$ is a random sample from the gamma distribution

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{(0,\infty)}(x).$$

Recall that the first two moments of the gamma distribution are

$$E_{\alpha,\lambda}(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad E_{\alpha,\lambda}(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

Now, solve these two equations for α and λ. This gives

$$\alpha = \frac{[E_{\alpha,\lambda}(X)]^2}{E_{\alpha,\lambda}(X^2) - [E_{\alpha,\lambda}(X)]^2}$$

and

$$\lambda = \frac{E_{\alpha,\lambda}(X)}{E_{\alpha,\lambda}(X^2) - [E_{\alpha,\lambda}(X)]^2}.$$

To obtain the method of moments estimators, we simply replace $E(X)$ by $\bar{X}$ and $E(X^2)$ by m_2. This gives

$$\hat{\alpha} = \frac{\bar{X}^2}{m_2 - \bar{X}^2}$$

and

$$\hat{\lambda} = \frac{\bar{X}}{m_2 - \bar{X}^2}.$$

We know that both α and λ must be positive. Are $\hat{\alpha}$ and $\hat{\lambda}$ positive?

Since each data value X_i is positive, $\bar{X}$ must be positive. This means the two estimates are positive if and only if $m_2 - \bar{X}^2 > 0$.

It's easy to show that

$$m_2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which obviously is positive.

3 Properties of Estimators

We will use the MLE $\hat{\theta}$ as an estimate of the true value of θ. If we are interested in estimating the characteristic of $\psi(\theta)$, a natural estimate is the plug-in estimate $\psi(\hat{\theta})$. More generally we can consider properties of a general estimator T of $\psi(\theta)$. Some important questions about estimators include the following:

- How close is the expected value of the estimator to the parameter it estimates?
- How close can we expect an estimate to be to the parameter it estimates?
- How is the behavior of an estimator related to sample size?
- How does the estimator compare to other estimators?

3.1 Bias

Let T be an estimator of a parametric quantity $\psi(\theta)$. A basic property of the estimator is its bias. The bias of an estimator T is defined as

$$\text{Bias}_\theta(T) = E_\theta(T) - \psi(\theta).$$

If $\text{Bias}_\theta(T) = 0$ for all $\theta \in \Omega$, we say that T is an unbiased estimator of $\psi(\theta)$.

Example 46 again Let $X_1, \ldots, X_n$ be a random sample from the exponential(λ) distribution. Check the bias of the maximum likelihood estimator as an estimator of λ:

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

We will use the fact the $Y = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$.

$$E_\lambda\left(\frac{1}{Y}\right) = \int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)}y^{n-1}e^{-\lambda y}dy$$

$$= \frac{\lambda^n}{\Gamma(n)}\frac{\Gamma(n-1)}{\lambda^{n-1}}\int_0^\infty \frac{\lambda^{n-1}}{\Gamma(n-1)}y^{n-2}e^{-\lambda y}dy = \frac{\lambda}{n-1}$$

Thus,

$$E_\lambda(\hat{\lambda}) = E_\lambda\left(\frac{n}{Y}\right) = \frac{n}{n-1}\lambda \quad \text{and} \quad \text{Bias}_\lambda(\hat{\lambda}) = \frac{\lambda}{n-1}.$$

Remark: Sometimes one can adjust a biased estimator to create an unbiased estimator. We will illustrate that here, but will reserve judgment on which estimator is better. In the above example, let

$$\tilde{\lambda} = \frac{n-1}{n}\hat{\lambda} = \frac{n-1}{Y}.$$

Then $E_\lambda(\tilde{\lambda}) = \lambda$.

Example 45 again Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $f_{\mu,\sigma}(\,\cdot\,)$. The maximum likelihood estimators were found to be

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Check the bias of these estimators.

We know from Chapter 3 that $E(\bar{X}) = \mu$ for any distribution. Thus, $\bar{X}$ is an unbiased estimator of μ.

From Chapter 4, we know that $U = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$. Thus,

$$E\left(\hat{\sigma}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{\sigma^2}{n}E(U) = \frac{(n-1)\sigma^2}{n}.$$

In fact, the previous result holds for any distribution with mean μ and variance σ^2:

$$\begin{aligned}
E[\hat{\sigma}^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}X_i^2 - n(\bar{X})^2\right] \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n}E(X_i^2) - nE[(\bar{X})^2]\right\} \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n}(\sigma^2+\mu^2) - n\left[\frac{\sigma^2}{n}+\mu^2\right]\right\} \\
&= \frac{1}{n}\left\{n\sigma^2 + n\mu^2 - (\sigma^2+n\mu^2)\right\} = \frac{1}{n}(n\sigma^2-\sigma^2) = \frac{n-1}{n}\sigma^2
\end{aligned}$$

Thus, $\hat{\sigma}^2$ has bias:

$$\text{Bias}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

3.2 Standard Error

To measure variation in the estimator T of $\psi(\theta)$, one can evaluate either the variance or the standard deviation of its sampling distribution. The most commonly reported quantity is the standard error:

$$SE_\theta(T) = \sqrt{\mathrm{Var}_\theta(T)}.$$

Example 46 again Let $X_1, \ldots, X_n$ be a random sample from the exponential(λ) distribution. Obtain the standard error of the maximum likelihood estimator,

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

As in finding the bias, let $Y = \sum_{i=1}^n X_i$. Then

$$E_\lambda\left(\frac{1}{Y^2}\right) = \int_0^\infty \frac{1}{y^2}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} dy = \frac{\lambda^2}{(n-1)(n-2)}.$$

Thus,

$$\begin{aligned}\mathrm{Var}_\lambda(\hat{\lambda}) &= n^2\mathrm{Var}_\lambda\left(\frac{1}{Y}\right) = n^2\lambda^2\left[\frac{1}{(n-1)(n-2)} - \left(\frac{1}{n-1}\right)^2\right] \\ &= \lambda^2\frac{n^2}{(n-2)(n-1)^2}\end{aligned}$$

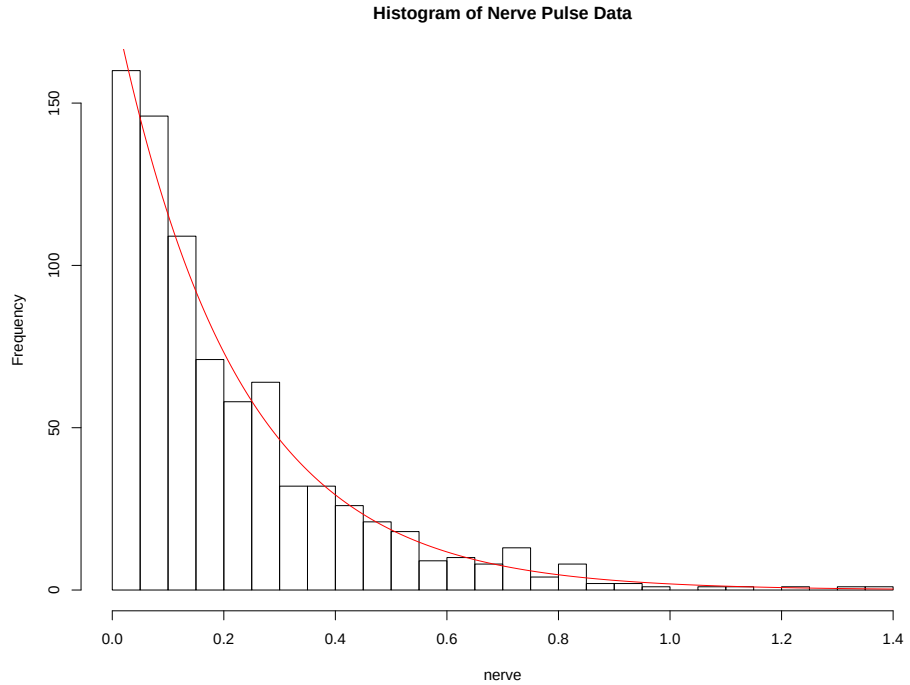
and

$$SE_\lambda(\hat{\lambda}) = \frac{\lambda n}{(n-1)\sqrt{n-2}}.$$

We often wish to estimate the standard error since it depends on the unknown parameter. Here, the estimated standard error of $\hat{\lambda}$ is

$$s_{\hat{\lambda}} = \frac{\hat{\lambda} n}{(n-1)\sqrt{n-2}}.$$

Example 43 again Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fiber. The data appear in the following histogram:



The mean is $\bar{x} = 0.2186$ and $\hat{\lambda} = 1/0.2186 = 4.575$. The exponential distribution with $\lambda = 4.575$ is superimposed on the histogram. The estimated standard error of $\hat{\lambda}$ is

$$s_{\hat{\lambda}} = \frac{\hat{\lambda} n}{(n-1)\sqrt{n-2}} = \frac{4.575(799)}{798\sqrt{797}} = 0.1623.$$

3.3 Mean Squared Error

A means of judging how well $\hat{\theta}$ estimates θ is to use the mean squared error. The mean squared error of an estimator T of a parametric quantity $\psi(\theta)$ is

$$\mathrm{MSE}_\theta(T) = E_\theta[(T - \psi(\theta))^2].$$

We may express the mean squared error of T as

$$\begin{aligned} \mathrm{MSE}_\theta(T) &= E_\theta\left[(T - E_\theta(T) + E_\theta(T) - \psi(\theta))^2\right] \\ &= \mathrm{Var}_\theta(T) + [\mathsf{Bias}_\theta(T)]^2. \end{aligned}$$

The mean squared error is particularly useful for comparing two or more estimators. If we can show that

$$\mathrm{MSE}_\theta(T_1) \leq \mathrm{MSE}_\theta(T_2)$$

no matter what the value of θ is, then this is a good reason to prefer T_1 to T_2.

Example 46 again Let $X_1, \ldots, X_n$ be a random sample from the exponential(λ) distribution. Obtain the mean squared error of the moment estimator and of the bias adjusted estimator.

Earlier we found the bias and variance of $\hat{\lambda}$:

$$\text{Bias}_\lambda(\hat{\lambda}) = \frac{\lambda}{n-1}$$

$$\text{Var}_\lambda(\hat{\lambda}) = \lambda^2 \frac{n^2}{(n-2)(n-1)^2}$$

Thus, the mean squared error is

$$\begin{aligned} \text{MSE}_\lambda(\hat{\lambda}) &= \text{Var}_\lambda(\hat{\lambda}) + \text{Bias}^2(\hat{\lambda}) \\ &= \lambda^2 \frac{n^2}{(n-2)(n-1)^2} + \left(\frac{\lambda}{n-1}\right)^2 \\ &= \lambda^2 \frac{n+2}{(n-1)(n-2)} \end{aligned}$$

The bias-adjusted estimator is

$$\tilde{\lambda} = \frac{n-1}{Y} = \frac{n-1}{n}\hat{\lambda}.$$

We then have

$$\text{Bias}_\lambda(\tilde{\lambda}) = 0$$

$$\begin{aligned} \text{Var}_\lambda(\tilde{\lambda}) &= \left(\frac{n-1}{n}\right)^2 \text{Var}_\lambda(\hat{\lambda}) \\ &= \left(\frac{n-1}{n}\right)^2 \lambda^2 \frac{n^2}{(n-2)(n-1)^2} = \frac{\lambda^2}{n-2} \end{aligned}$$

Thus,

$$\text{MSE}_\lambda(\tilde{\lambda}) = \text{Var}_\lambda(\tilde{\lambda}) = \frac{\lambda^2}{n-2} < \lambda^2 \frac{n+2}{(n-1)(n-2)} = \text{MSE}_\lambda(\hat{\lambda}).$$

Example 45 again Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $f_{\mu,\sigma}(\cdot)$. The mles were found to be

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

These were also found to be the moment estimators. We earlier obtained the expectations of these estimators and determined that the following properties:

- $\bar{X}$ is an unbiased estimator of μ.
- $\text{Bias}_{\mu,\sigma}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$.

Does this mean $\hat{\sigma}^2$ is a bad estimator?

We could correct the bias in estimating σ^2 by using the sample variance of Chapter 5:

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

We next obtain the variances and mean squared errors of the two estimators of σ^2, $\hat{\sigma}^2$ and S^2.

$$\begin{aligned}\mathrm{Var}_{\mu,\sigma}(\hat{\sigma}^2) &= \mathrm{Var}_{\mu,\sigma}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\ &= \frac{\sigma^4}{n^2}\mathrm{Var}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}\right) = \frac{2(n-1)\sigma^4}{n^2}\end{aligned}$$

The MSE of $\hat{\sigma}^2$ becomes

$$MSE(\hat{\sigma}^2) = \mathrm{Var}(\hat{\sigma}^2) + \mathrm{Bias}^2(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{-\sigma^2}{n}\right)^2 = \frac{(2n-1)\sigma^4}{n^2}$$

We next consider the sample variance, S^2.

$$\begin{aligned}\mathrm{Var}_{\mu,\sigma}(S^2) &= \mathrm{Var}_{\mu,\sigma}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\\ &= \frac{\sigma^4}{(n-1)^2}\mathrm{Var}\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\right) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}.\end{aligned}$$

If we compare the MSEs of the two estimators, we see that the MSE of the biased estimator, the MLE $\hat{\sigma}^2$, is smaller than that of the unbiased estimator, the sample variance S^2:

$$\frac{MSE(S^2)}{MSE(\hat{\sigma}^2)} = \frac{\frac{2\sigma^4}{n-1}}{\frac{(2n-1)\sigma^4}{n^2}} = \frac{2n^2}{(2n-1)(n-1)} > 1.$$

Example 43 again Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on the interval $(0, \theta),\ \theta > 0$. Check the bias and mse of the mle of θ. Compare the mle to the moment estimator.

First we need to find the pdf of $\hat{\theta} = X_{(n)} = \max\{X_1, \ldots, X_n\}$. It is most easily derived by using the cdf. For $0 \le x \le \theta$,

$$F_\theta(x) = P[X_{(n)} \le x] = P[X_1 \le x, \ldots, X_n \le x] = \prod_{i=1}^{n} P[X_i \le x] = \left(\frac{x}{\theta}\right)^n.$$

The pdf of $X_{(n)}$ is then

$$f(x) = \frac{nx^{n-1}}{\theta^n}, \quad 0 \le x \le \theta.$$

The first two moments are

$$\begin{aligned}
E(\hat{\theta}) &= \frac{1}{\theta^n}\int_0^\theta x\, nx^{n-1}dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{\theta^n}\frac{x^{n+1}}{n+1}\Big|_0^\theta = \frac{n}{n+1}\theta, \\
E(\hat{\theta}^2) &= \frac{1}{\theta^n}\int_0^\theta x^2\, nx^{n-1}dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1} dx = \frac{n}{\theta^n}\frac{x^{n+2}}{n+2}\Big|_0^\theta = \frac{n}{n+2}\theta^2, \\
\mathrm{Var}(\hat{\theta}) &= E(\hat{\theta}^2) - (E(\hat{\theta}))^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\right)^2\theta^2 \\
&= \frac{n}{(n+2)(n+1)^2}\theta^2.
\end{aligned}$$

The resulting bias and MSE of $\hat{\theta}$ are

$$\begin{aligned}
\mathrm{Bias}(\hat{\theta}) &= E(\hat{\theta}) - \theta = -\frac{\theta}{n+1}, \\
MSE(\hat{\theta}) &= \frac{n}{(n+2)(n+1)^2}\theta^2 + \left(-\frac{\theta}{n+1}\right)^2 = \frac{2\theta^2}{(n+1)(n+2)}.
\end{aligned}$$

We could adjust the mle to remove the bias:

$$\check{\theta} = \frac{n+1}{n}\hat{\theta}.$$

The resulting variance and MSE of $\check{\theta}$ are

$$\begin{aligned} MSE(\check{\theta}) &= \mathrm{Var}(\check{\theta}) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}(\hat{\theta}) \\ &= \left(\frac{n+1}{n}\right)^2 \frac{n}{(n+2)(n+1)^2}\theta^2 \\ &= \frac{\theta^2}{n(n+2)}, \end{aligned}$$

$$\frac{MSE(\hat{\theta})}{MSE(\check{\theta})} = \frac{\frac{2\theta^2}{(n+1)(n+2)}}{\frac{\theta^2}{n(n+2)}} = \frac{2n}{n+1} > 1.$$

Hence, the bias-adjusted estimator has smaller MSE for $n > 1$.

We now find the method of moments estimator and examine its properties.

$$E(X) = \frac{\theta}{2} \implies \text{the moment estimator is } \tilde{\theta} = 2\bar{X}.$$

The resulting bias and MSE are

$$\begin{aligned} \text{Bias}(\tilde{\theta}) &= E(\tilde{\theta}) - \theta = 2\,\frac{\theta}{2} - \theta = 0, \\ MSE(\tilde{\theta}) &= \text{Var}(\tilde{\theta}) = 4\text{Var}(\bar{X}) = 4\,\frac{\theta^2}{12n} = \frac{\theta^2}{3n}. \end{aligned}$$

We see that the MSE goes to zero as a function of $\frac{1}{n}$ for the moment estimator and as a function of $\frac{1}{n^2}$ for the other two estimators.

3.4 Consistency of Estimators

Definition: Let T_n be a estimator of $\psi(\theta)$ based on a sample $X_1, \ldots, X_n$ from $f_\theta(x)$. Then the sequence of estimators $\{T_n, n = 1, 2, \ldots\}$ is said to be *consistent for* $\psi(\theta)$ if

$$T_n \xrightarrow{P} \psi(\theta).$$

By the Weak Law of Large Numbers, the sample moments converge in probability to the population moments:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E(X^k).$$

If the functions that relate the estimates to the sample moments are continuous, then estimators will converge in probability to the parameters.

Remark: A convenient way to show that an estimator is consistent is the show that its MSE goes to zero as n tends to infinity. This result can be proved in the same manner as the Weak Law of Large Numbers:

$$\begin{aligned} 0 &\leq \lim_{n\to\infty} P_\theta[|T_n - \psi(\theta)| \geq \varepsilon] \\ &\leq \lim_{n\to\infty} \frac{E_\theta[(T_n - \psi(\theta))^2]}{\varepsilon^2} \\ &= \lim_{n\to\infty} \frac{\mathrm{MSE}_\theta(T_n)}{\varepsilon^2} = 0. \end{aligned}$$

Since $\mathrm{MSE}_\theta(T_n) = \mathrm{Var}_\theta(T_n) + [\mathrm{Bias}_\theta(T_n)]^2$, the sequence of estimators $\{T_n\}$ is consistent if

$$\mathrm{Var}_\theta(T_n) \longrightarrow 0, \quad \text{and} \quad \mathrm{Bias}_\theta(T_n) \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

Example 46 again Let $X_1, \ldots, X_n$ be a random sample from the exponential (λ) distribution. It is easy to show the consistency of the moment estimator,

$$\hat{\lambda}_n = \frac{1}{\bar{X}} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

Since

$$\lim_{n\to\infty} \mathrm{MSE}(\hat{\lambda}_n) = \lim_{n\to\infty} \lambda^2 \frac{n+2}{(n-1)(n-2)} = 0, \quad \hat{\lambda}_n \xrightarrow{P} \lambda.$$

Example 43 again Let $X_1, \ldots, X_n$ be a random sample from the uniform distribut ion on the interval $(0, \theta),\ \theta > 0$. Check the consistency the mle of θ.

Since

$$\lim_{n\to\infty} \mathrm{MSE}(\hat{\theta}) = \lim_{n\to\infty} \frac{2\theta^2}{(n+1)(n+2)} = 0, \quad \hat{\theta}_n \xrightarrow{P} \theta.$$

3.5 Asymptotic Distribution of Estimators

We need to obtain the sampling distribution of point estimators in order to derive the properties of the estimators and also confidence intervals and hypothesis tests based on the estimators. However, it is often not practical to obtain the exact sampling distribution of an estimator T_n calculated from a random sample $X_1, \ldots, X_n$. In such cases, we can obtain the large sample distribution of the statistic and use it to approximate the sampling distribution for finite sample size n. A basic result that is useful for statistics that are sample means is the Central Limit Theorem:

Let $X_1, \ldots, X_n$ be a random sample from a distribution having variance σ^2 (with $0 < \sigma^2 < \infty$) and mean μ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for each real number z,

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z).$$

Thus, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z$, where Z is a standard normal rv.

Another Method of Finding Asymptotic Distributions

In many situations, such as in using method of moments estimators, one is interested in obtaining an asymptotic distribution of an estimator that is a function of a statistic that we know is asymptotically normal. The so-called delta method (or propagation of errors) enables us to obtain the large sample distribution of the estimator of interest.

Suppose that we know that

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, V(\theta)),$$

and we are interested in the asymptotic distribution of the estimator $g(T_n)$ of $g(\theta)$.

We assume that $g(t)$ is a continuous function that has a sufficient number of derivatives. Then we can apply Taylor's theorem to g:

$$g(t) = g(\theta) + (t - \theta)g'(\theta) + \text{higher order terms.}$$

We now apply the Taylor expansion to T_n and ignore the higher order terms:

$$g(T_n) = g(\theta) + (T_n - \theta)g'(\theta).$$

We rearrange the terms to get

$$\sqrt{n}(g(T_n) - g(\theta)) = g'(\theta)\sqrt{n}(T_n - \theta).$$

Since $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, V(\theta))$ by assumption, we obtain the result that

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, [g'(\theta)]^2 V(\theta)).$$

Remark: We can apply this result to most method of moment estimators. The Central Limit Theorem implies that $\bar{X}_n$ is asymptotically normal. Since the MOM estimator is a smooth function of $\bar{X}_n$, we can use this result to obtain the asymptotic distribution of the MOM estimator.

Example 46 again Let $X_1, \ldots, X_n$ be a random sample from the exponential(λ) distribution. The moment estimator of λ was found to be $\hat{\lambda} = 1/\bar{X}_n$. From the properties of the gamma distribution, the mean and variance of an exponential rv are

$$E_\lambda(X) = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{Var}_\lambda(X) = \frac{1}{\lambda^2}.$$

By the Central Limit Theorem, $\sqrt{n}(\bar{X}_n - 1/\lambda) \xrightarrow{D} N(0, 1/\lambda^2)$.

Now $g(t) = 1/t$ and $g'(t) = -1/t^2$. Thus, we apply the delta method to $\hat{\lambda} = g(\bar{X}_n) = 1/\bar{X}_n$:

$$\sqrt{n}(\hat{\lambda} - \lambda) = \sqrt{n}(g(\bar{X}_n) - g(1/\lambda)) \xrightarrow{D} N(0, (1/\lambda^2)(g'(1/\lambda))^2).$$

Since

$$\frac{1}{\lambda^2}\left(g'\left(\frac{1}{\lambda}\right)\right)^2 = \left(\frac{1}{\lambda^2}\right)\left(\frac{-1}{(1/\lambda)^2}\right)^2 = \lambda^2,$$

we get the result

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{D} N(0, \lambda^2).$$

3.5.1 Another Approach to Estimating the Standard Error

Here we could easily find an exact expression for the standard error of our estimator. For other estimators, this may not be the case. We can approximate the standard error computationally using a technique known as the bootstrap. We will first discuss the parametric bootstrap.

Suppose that we take a random sample $X_1, \ldots, X_n$ from a distribution with parameter θ. Suppose that $\hat{\theta} = S(X_1, \ldots, X_n)$ for some statistic S whose sampling distribution is difficult to derive.

- If we knew the true value of the parameter, say θ_0, we could generate $x_1, \ldots, x_n$ from $f_{\theta_0}(x)$ and compute $s_1 = s(x_1, \ldots, x_n)$.
- If we repeat this a large number B of times, we obtain a random sample of values of S, $(s_1, \ldots, s_B)$.
- We use this random sample to estimate the sampling distribution of S.
- Since we do not know θ_0, we will replace it by $\hat{\theta}$ and carry out this procedure.

Example: We will use the parametric bootstrap to estimate the standard error of the maximum likelihood estimator of the rate parameter for the exponential distribution.

Here $\hat{\lambda} = 4.575$ and $n = 799$.

- We generate $B = 1000$ samples of size $n = 799$ from the exponential distribution with $\lambda = 4.575$.
- Compute $\hat{\lambda}$ for each sample.
- Find the standard deviation of the 1000 values of $\hat{\lambda}$.

Following is the R code for carrying out these calculations.

```
> mean(nerve)
[1] 0.2185732
> 1/mean(nerve)
[1] 4.575126
> temp=rep(0,1000)
> for (i in 1:1000)temp[i]= 1/mean(rexp(799,rate=1/mean(nerve)))
> sd(temp)
[1] 0.1629491
```

Here we were curiously coincident with our previous estimate of the standard error of $\hat{\lambda}$, $s_{\hat{\lambda}} = 0.1623$.

Other bootstrap samples of size $B = 1000$ resulted in estimated standard errors of

$0.1640392,\ 0.1633224,\ 0.1664940,\ 0.1648791,$ and 0.1573375.

Another approach does not use an assumed form for the pdf $f_\theta(x)$. The nonparametric bootstrap involves resampling from the observed data. We will suppose that we want to approximate the sampling distribution of a statistic, $S(X_1, \ldots, X_n)$.

- Take a sample $x_1^*, \ldots, x_n^*$ of size n with replacement from the observed data, $x_1, \ldots, x_n$.
- Compute the observed value of the statistic, $s_1^* = s(x_1^*, \ldots, x_n^*)$.
- If we repeat this a large number B of times, we obtain B bootstrap values of S, $(s_1^*, \ldots, s_B^*)$.
- We use these bootstrap values to estimate the sampling distribution of S.

Following is the R code to generate 1000 bootstrap estimates of λ and compute the estimated standard error of $\hat{\lambda}$:

```
> temp=rep(0,1000)
> for(i in 1:1000)temp[i]=1/mean(sample(nerve,replace=TRUE))
> sd(temp)
[1] 0.1594686
```

Other bootstrap estimates of the standard error of $\hat{\lambda}$ tended to be larger than those obtained using the parametric bootstrap.

3.6 Large Sample Theory for Maximum Likelihood Estimators

The MLE has excellent large sample properties under certain regularity conditions. We suppose that $X_1, \ldots, X_n$ is a random sample from a population with pdf or pmf $f_\theta(x)$.

- The density $f_\theta(x)$ is a smooth function of θ.
- The support of the distribution, $\{x : f_\theta(x) > 0\}$ does not depend on the parameter, θ.
- The parameter space Ω satisfies certain conditions.
- $\mathrm{Var}\left(\frac{\partial \log(f_\theta(X)}{\partial \theta}\right)$ is finite.

We denote the "true" value of θ by θ_0.

Important properties of the MLE include the following:

1. $\hat{\theta}_n$ is consistent. That is

$$\hat{\theta}_n \xrightarrow{P} \theta_0 \quad \text{as} \quad n \longrightarrow \infty$$

2. The MLE is asymptotically normal:

$$\frac{\hat{\theta}_n - \theta_0}{\sqrt{V_n(\theta_0)}} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$

We usually interpret this to mean

$$\hat{\theta}_n \overset{\text{approx}}{\sim} N(\theta_0, \hat{V}_n)$$

where

$$\hat{V}_n = V_n(\hat{\theta}_n)$$

We need to define one other quantity to obtain $V_n(\theta_0)$:

Fisher's information in θ based on the observation X is defined as

$$I(\theta) = E_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right]$$

For computing $I(\theta)$, we often use the result that

$$I(\theta) = -E_\theta \left[\left(\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right) \right]$$

This quantity can be estimated in several ways:

$$I(\hat{\theta}_n) \quad - \quad \text{Plug in}$$

$$\hat{I}(\hat{\theta}_n) = -\frac{1}{n} \frac{\partial^2 \ell(\hat{\theta}_n)}{\partial \theta^2} \quad - \quad \text{Hessian or observed information}$$

Fisher's Information in a Random Sample

Suppose now that we have a random sample: $X_1, \ldots, X_n$ are independent with pmf or pdf $f_\theta(x)$. The likelihood is

$$L(\theta) = f_\theta(x_1) \times \cdots \times f_\theta(x_n).$$

Then

$$\log(L(\theta)) \quad = \quad \log[f_\theta(x_1) \times \cdots \times f_\theta(x_n)] \; = \; \sum_{i=1}^{n} \log(f_\theta(x_i))$$

and

$$\frac{\partial^2 \log(L(\theta))}{\partial \theta^2} = \sum_{i=1}^{n} \frac{\partial^2 \log(f_\theta(x_i))}{\partial \theta^2}$$

Then the Fisher information $I_n(\theta)$ in $X_1, \ldots, X_n$ is given by

$$I_n(\theta) = -E_\theta \left[\frac{\partial^2 \log(L(\theta))}{\partial \theta^2} \right] = \sum_{i=1}^{n} -E_\theta \left[\frac{\partial^2 \log(f_\theta(X_i))}{\partial \theta^2} \right] = nI(\theta)$$

One can show that

$$V_n(\theta_0) = \frac{1}{nI(\theta_0)}.$$

Thus, we can say that

$$\sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$

Remarks:

- We say that $\hat{\theta}_n$ is asympotically unbiased, which means that in large samples the MLE has approximately the desired mean.
- $\hat{\theta}_n$ is asymptotically efficient. This means that in large sample, it has the smallest variance among all asymptotically unbiased estimators.
- The third result says that we can use a relatively simple distribution to provide confidence intervals for θ. In general, the actual sampling distribution of $\hat{\theta}_n$ is very messy.

- $\sqrt{\hat{V}_n} = \left[\sqrt{nI(\hat{\theta})}\right]^{-1}$ provides the estimated *asymptotic standard error* (ASE) for $\hat{\theta}$.

- $-\dfrac{\partial^2 \ell(\hat{\theta})}{\partial \theta^2}$ measures the *curvature* of the log-likelihood function.

- The greater the curvature, the greater the information about θ and the smaller the ASE.

Example 44 again The log-likelihood for a random sample from the Bernoulli distribution is

$$\ell(\theta) = \log(L(\theta|y)) = y\log(\theta) + (n-y)\log(1-\theta)$$

and the log likelihood for a single Bernoulli rv is

$$\log(f_\theta(y_i)) = y_i\log(\theta) + (1-y_i)\log(1-\theta)$$

The first and second derivatives are

$$\frac{\partial \log(f_\theta(y_i))}{\partial\theta} = \frac{y_i}{\theta} - \frac{1-y_i}{1-\theta}$$

$$\frac{\partial^2 \log(f_\theta(y_i))}{\partial\theta^2} = -\frac{y_i}{\theta^2} - \frac{1-y_i}{(1-\theta)^2}$$

Fisher's information in θ for a single Bernoulli observation is

$$I(\theta) = -E_\theta\left[\frac{\partial^2 \log(f_\theta(Y_i))}{\partial\theta^2}\right] = \frac{\theta}{\theta^2} + \frac{(1-\theta)}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

Then Fisher's information in θ for Bernoulli sample is $nI(\theta) = \frac{n}{\theta(1-\theta)}$,

$$V_n(\theta_0) = \frac{1}{nI(\theta_0)} = \frac{\theta_0(1-\theta_0)}{n},$$

and the estimated asymptotic standard error is

$$ASE = \sqrt{V_n(\hat{\theta})} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

The asymptotic properties of the MLE imply that

$$\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$

and that

$$\hat{\theta} \overset{\text{approx}}{\sim} N(\theta_0, \frac{\theta_0(1-\theta_0)}{n})$$

This equivalent to our earlier normal approximation to the binomial distribution.

Example 45 again Suppose that $X \sim N(\mu, \sigma^2)$ where σ^2 is known. Find the Fisher information in X.

$$
\begin{aligned}
f_\mu(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \\
\log(f\mu(x))) &= -\log(\sqrt{2\pi}\sigma) - (x-\mu)^2/2\sigma^2 \\
\frac{\partial \log(f_\mu(x))}{\partial \mu} &= 2(x-\mu)/2\sigma^2 = \frac{x-\mu}{\sigma^2} \\
I(\mu) &= E\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{E[(X-\mu)^2]}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}
\end{aligned}
$$

Example 45 continued Suppose now that $X_1, \ldots, X_n$ form a random sample from a $N(\mu, \sigma^2)$ population where σ^2 is known. Find the Fisher information in $X_1, \ldots, X_n$.

From the preceding slide,

$$I(\mu) = \frac{1}{\sigma^2}.$$

Then the Fisher information $I_n(\mu)$ in $X_1, \ldots, X_n$ is given by

$$I_n(\mu) = nI(\mu) = \frac{n}{\sigma^2}.$$

The large sample results for the MLE imply that

$$\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{\sigma^2}} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$

We know that this result holds exactly for any n.

Example 46, again Find the asymptotic distribution of the mle.

We need to obtain Fisher's information, $I(\lambda)$. The likelihood and log-likelihood of a single X are

$$f_\lambda(x) = \lambda e^{-\lambda x} \quad \text{and} \quad \log(f_\lambda(x)) = \log(\lambda) - \lambda x.$$

Then

$$\frac{\partial \log(f_\lambda(x))}{\partial \lambda} = \frac{1}{\lambda} - x,$$

and

$$\frac{\partial^2 \log(f_\lambda(x))}{\partial \lambda^2} = -\frac{1}{\lambda^2}.$$

Hence, $I(\lambda) = 1/\lambda^2$ and

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda_0)}{\lambda_0} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$