Chapter 8. Point Estimation

Let $\{X_1, \cdots, X_n\}$ be a random sample from a population $F(\cdot, \boldsymbol{\theta})$, where the form of F is known but the parameter $\boldsymbol{\theta}$ is unknown, and it has p components. Often we may specify $\boldsymbol{\theta} \in \Theta$, where Θ is called the parameter space.

For $N(\mu, \sigma^2)$, $p = 2$ and $\Theta = (-\infty, \infty) \times (0, \infty)$. For $Poisson(\lambda)$, $p = 1$ and $\Theta = (0, \infty)$.

Goal: to find a (point) estimator for $\boldsymbol{\theta}$.

8.1 Method of Moments Estimation

Let $\mu_k \equiv \mu_k(\boldsymbol{\theta}) = E(X_1^k)$ denote the k-th moment of the population, $k = 1, 2, \cdots$. Then μ_k depends on unknown parameter $\boldsymbol{\theta}$, as everything else on

the distribution $F(\cdot, \boldsymbol{\theta})$ are known. Denote the k-th sample moment by

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{1}{n}(X_1^k + \cdots + X_n^k).$$

The **MM estimator** $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is the solution of the p equations

$$\mu_1(\widehat{\boldsymbol{\theta}}) = M_1,\ \mu_2(\widehat{\boldsymbol{\theta}}) = M_2,\ \cdots,\ \mu_p(\widehat{\boldsymbol{\theta}}) = M_p.$$

Example 1. Let $\{X_1, \cdots, X_n\}$ be a sample from a population with mean μ and variance $\sigma^2 < \infty$. Find the MM estimator for (μ, σ^2).

There are two unknown parameters. Let

$$\mu = \mu_1 = M_1, \qquad \mu_2 = M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

This gives us $\widehat{\mu} = M_1 = \bar{X}$. Since $\sigma^2 = \mu_2 - \mu_1^2$,

$$\widehat{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Note. $E(\widehat{\sigma}^2) = E(X_1^2) - E(\bar{X}^2) = \sigma^2 + \mu^2 - (\sigma^2/n + \mu^2) = \frac{n-1}{n}\sigma^2$. We call $E(\widehat{\sigma}^2) - \sigma^2 = -\sigma^2/n$ the estimation bias. The **sample variance**

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

is a more frequently used estimator for σ^2, and it has zero-bias.

Theorem 1. Under some mild regularity conditions, the MME $\widehat{\boldsymbol{\theta}}$ is a consistent estimator in the sense that as $n \to \infty$, $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$, i.e.

$$P\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \epsilon\} \to 0 \qquad \text{for any } \epsilon > 0.$$

Further it is asymptotically normal, i.e. $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to a p-dimensional normal distribution.

8.2. Maximum likelihood estimation

8.2.1 Likelihood

Likelihood is one of the most fundamental concepts in all types of statistical inference.

Definition 1 Suppose that $\mathbf{X}$ has density function or probability function $f(\mathbf{x};\boldsymbol{\theta})$. We have observed $\mathbf{X} = \mathbf{x}$. Then the likelihood function with observation $\mathbf{x}$ is defined as

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}).$$

Density/probability function: a function of $\mathbf{x}$, specifying the distribution of random variable $\mathbf{X}$

Likelihood: a function of $\boldsymbol{\theta}$, reflecting information on $\boldsymbol{\theta}$ contained in observation $\mathbf{x}$

Note. A likelihood function represents the uncertainty on a unknown non-random constant θ, and it is **not a density or probability function**! It provides

a rational degree of belief, or
an order of preferences

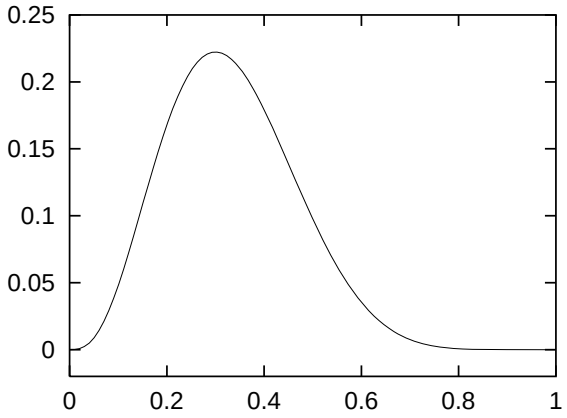
on possible values of the parameter θ. This can be seen more clearly in the simple example on next slide.

In fact, a likelihood function is often defined up to a positive constant — the constant here refers to a quantity independent of θ. But it may depending on **x**. (Note **x** is a given constant.)

Example 2. Suppose that x is the number of successes from a known number n of independent trials with unknown probability of success π. The probability function, and so the likelihood function is

$$L(\pi) = f(x; \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.$$

The likelihood function $L(\pi; x)$ can be graphed as a function of π. It changes shape for different values of x. A likelihood function for a $x = 3$ when $n = 10$ is shown in the Figure below.


0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1

Notice that the likelihood function shown above is *not* a density function. It does not have an area of 1 below it.

We use the likelihood function to compare the plausibility of different possible parameter values. For instance, the likelihood is much larger for $\pi = 0.3$ than for $\pi = 0.8$, that is the data $x = 3$ have a greater probability of being observed if $\pi = 0.3$ than if $\pi = 0.8$. This makes $\pi = 0.3$ much more **likely** as the true value for π than 0.8.

Note. In the above argument, we do not need to calculate exact probabilities under different values of $\boldsymbol{\theta}$. Only the order of those quantities matters!

Let $X_1, \cdots, X_n$ be i.i.d. with PDF $f(\cdot, \boldsymbol{\theta})$. Write $\mathbf{X} = (X_1, \cdots, X_n)'$. Then the likelihood function is

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} f(X_i, \boldsymbol{\theta}),$$

which is a product of n terms. Then the **log-likelihood function** is

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) \equiv \log\{L(\boldsymbol{\theta}; \mathbf{X})\} = \sum_{i=1}^{n} \log\{f(X_i, \boldsymbol{\theta})\},$$

which is a sum of n terms.

This explains why log-likelihood functions are often used with independent observations.

8.2.3 Maximum likelihood estimator (MLE)

The MLE is by far the most popular estimator.

Definition 2 — MLE
A *Maximum Likelihood Estimator* (MLE), $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X}) \in \Theta$, of parameter $\boldsymbol{\theta}$ is an estimator satisfying

> $L(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq L(\boldsymbol{\theta}; \mathbf{X})$ for all $\boldsymbol{\theta} \in \Theta$, or equivalently $l(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq l(\boldsymbol{\theta}; \mathbf{X})$ for all $\boldsymbol{\theta} \in \Theta$.

Obviously, a maximum likelihood estimator is the most plausible value for $\boldsymbol{\theta}$ as judged by the likelihood function. In many cases where Θ is continuous and the maximum does not occur at a boundary of Θ, $\widehat{\boldsymbol{\theta}}$ **is often the solution** of the equation

$$s(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{X}) = 0.$$

We call $s(\boldsymbol{\theta}) \equiv s(\boldsymbol{\theta}; \mathbf{X})$ **a score function**.

Example 3. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from $N(\mu, \sigma^2)$ where neither μ or σ^2 is known. Then we can find the maximum likelihood estimator from the log-likelihood

$$l(\mu, \sigma^2) = -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \mu)^2/(2\sigma^2)$$

$$= \quad -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \bar{Y})^2/(2\sigma^2) - n(\bar{Y} - \mu)^2/(2\sigma^2).$$

This is maximised by choosing $\mu = \bar{Y}$, so $\widehat{\mu} = \bar{Y}$ is the MLE for μ. It is easy to see

$$E(\widehat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu.$$

Such a estimator is called **unbiased**.

The **profile log-likelihood** remaining is

$$l(\widehat{\mu}, \sigma^2) = -n \log \sqrt{2\pi} + (n/2)(\log \sigma^{-2} - \widehat{\sigma}^2 \sigma^{-2}),$$

where $\widehat{\sigma}^2 = \sum_1^n (Y_i - \bar{Y})^2/n$. By the lemma below, the MLE for σ^2 is $\widehat{\sigma}^2$. Note that the MLE of σ^2 is *biased* since

$$E(\widehat{\sigma}^2) = (1 - 1/n)\sigma^2 \neq \sigma^2.$$

Lemma. Define $L(x) = \log(x^{-1}) - b/x$, where $b > 0$ are constants. Then $L(b) \geq L(x)$ for all $x > 0$.

Example 4. Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(π). Then

$$L(\pi) = \prod_{i=1}^{n} \pi^{X_i}(1-\pi)^{1-X_i} = \pi^{n\bar{X}}(1-\pi)^{n(1-\bar{X})}.$$

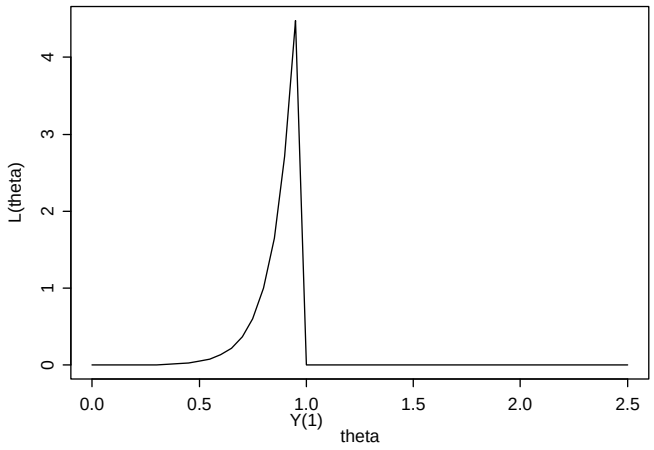
$$l(\pi) = n\bar{X} \log \pi + n(1 - \bar{X}) \log(1 - \pi).$$

Let $s(\pi) = \frac{\partial}{\partial \pi} l(\pi) = 0$, leading to $\widehat{\pi} = \bar{X}$.

Example 5. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from an exponential distribution with density function $e^{-(y-\theta)}$ for $y \geq \theta$. This is the usual exponential distribution shifted to start at θ. The Likelihood is

$$L(\theta; \mathbf{Y}) = e^{-n(\bar{Y}-\theta)} I_{\{(\theta,\infty)\}}(Y_{(1)}),$$

where $Y_{(1)}$ is the smallest observation. This likelihood is zero for $\theta > Y_{(1)}$ and increases in θ for $\theta \leq Y_{(1)}$. So the MLE $\widehat{\theta} = Y_{(1)}$, which is a boundary maximum.



Invariance property of MLEs

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$, and $\boldsymbol{\psi} = g(\boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}$ be the MLE for $\boldsymbol{\theta}$, i.e.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f(\mathbf{X}, \boldsymbol{\theta}).$$

It is obvious to see that the MLE for $\boldsymbol{\psi}$ is $\widehat{\boldsymbol{\psi}} = g(\widehat{\boldsymbol{\theta}})$.

If $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ is a `1-1` transform and $\widehat{\boldsymbol{\psi}}$ is the MLE for $\boldsymbol{\psi}$, $\widehat{\boldsymbol{\theta}} \equiv g^{-1}(\widehat{\boldsymbol{\psi}})$ is the MLE for $\boldsymbol{\theta}$.

8.2.4 Numerical computation of MLEs

In modern statistical applications, it is typically difficult to find explicit analytic forms for the maximum likelihood estimators. These estimators are found more often by iterative procedures built into computer software. An iterative scheme starts with some guess at the MLE and then steadily improves it with each iteration. The estimator is considered found when it has become numerically stable. Sometimes the iterative procedures become trapped at a local maximum which is not a global maximum. There may be a very large number of parameters in a model, which makes such local entrapment much more common.

Newton-Raphson Scheme

Suppose that the log-likelihood function function $l(\boldsymbol{\theta})$ is sufficiently smooth. Then

$$s(\widehat{\boldsymbol{\theta}}) = 0,$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE and $s(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} l(\boldsymbol{\theta})$ is the score function. Let

$$\dot{s}(\boldsymbol{\theta}) = \ddot{l}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \theta \partial \theta'} l(\boldsymbol{\theta}) = \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right).$$

Suppose $\widehat{\boldsymbol{\theta}}$ is close to the true value $\boldsymbol{\theta}^0$. By a simple Taylor expansion,

$$s(\boldsymbol{\theta}^0) = \dot{s}(\boldsymbol{\theta}^0)(\boldsymbol{\theta}^0 - \widehat{\boldsymbol{\theta}}) + o_p(||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||).$$

This leads to the approximation

$$\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^0 - \{\dot{s}(\boldsymbol{\theta}^0)\}^{-1} s(\boldsymbol{\theta}^0).$$

Since θ^0 is unknown, we use iterative estimators

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \{\dot{s}(\boldsymbol{\theta}_k)\}^{-1} s(\boldsymbol{\theta}_k) \tag{1}$$

for $k = 1, 2, \cdots$, where $\boldsymbol{\theta}_0$ is a prescribed initial value. We define $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_j$ if $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{j-1}$ differ by a small amount.

Example 6. Let $X_1, \cdots, X_n$ be a sample from Cauchy distribution with PDF

$$f(x, \theta) = \frac{1}{\pi\{1 + (x - \theta)^2\}},$$

where θ is the location parameter. The log-likelihood is

$$l(\theta) = -\sum_{i=1}^{n} \log\{1 + (X_i - \theta)^2\} - n \log \pi.$$

The MLE is the solution of $s(\widehat{\theta}) = 0$, where

$$s(\theta) = 2 \sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2}.$$

Since $s(\theta) = 0$ does not admit an explicit solution, we adopt a Newton-Raphson scheme: $\theta_{k+1} = \theta_k - s(\theta_k)/\dot{s}(\theta_k)$, where

$$\dot{s}(\theta) = 2 \sum_{i=1}^{n} \frac{(X_i - \theta)^2 - 1}{\{1 + (X_i - \theta)^2\}^2}.$$

The *R*-function below implements the above scheme.

```
cauchyMLE <- function(n, theta, init, Tiny) {
x <- rcauchy(n, theta) # x is a sample
i <- 0   # No. of iterations
theta0 <- init +10*Tiny
theta1 <- init
while(abs(theta1-theta0)>Tiny) {
    theta0 <- theta1
```

```
    x2 <- x-theta0
    x22 <- x2*x2
    t1 <- mean(x2/(x22+1))          #  s(theta0)/(2n)
    t2 <- mean((x22-1)/(x22+1)^2)  #  derivation of s(theta0)/(2n)
    theta1 <- theta0 - t1/t2
    i <- i+1
    cat(i, "iteration:", theta1, "\n")  # print out iteration values
}
cat("MLE:", theta1, "No. of iterations:", i, "\n")
}
```

By calling `cauchyMLE(100,10,11.12,0.01)`, we excute the above iterative algorithm as follows:

```
> source("cauchyMLE.r")
> cauchyMLE(100, 10, 11.12, 0.01)
1 iteration: 9.594835
2 iteration: 10.08787
3 iteration: 10.0752
4 iteration: 10.07521
```

```
MLE: 10.07521 No. of iterations: 4
```

Note the initial value is important: the iterations will not converge if $\theta_0 < 8.75$ or $\theta_0 > 11.2$ on my PC.

Choosing a good initial value is always important. For this example, the PDF is symmetric around θ, it makes sense to consider either the sample mean or sample median as an initial estimate. However $E(X_1)$ is not well-defined, so the sample mean may not be a good estimator θ. Thus we may use the sample median as the initial value for our algorithm.

The Fisher Scoring method: replace $\dot{s}(\widehat{\boldsymbol{\theta}}_k)$ in (1) by $E_{\boldsymbol{\theta}}\{\dot{s}(\boldsymbol{\theta})\}$ under $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_k$. So the algorithm is now

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - [E\{\dot{s}(\boldsymbol{\theta}_k)\}]^{-1} s(\boldsymbol{\theta}_k) = \boldsymbol{\theta}_k + \{\mathcal{I}(\boldsymbol{\theta}_k)\}^{-1} s(\boldsymbol{\theta}_k),$$

where

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{Var}\{s(\boldsymbol{\theta})\} = -E\{\dot{s}(\boldsymbol{\theta})\}$$

is the Fisher information.

Example 6 (continue). It can be shown that

$$E\{\dot{s}(\theta)\} = \frac{2n}{\pi} \int_{-\infty}^{\infty} \frac{(x-\theta)^2 - 1}{\{1 + (x-\theta)^2\}^3} dx = -n/2.$$

Hence the Fisher scoring method is

$$\theta_{k+1} = \theta_k + \frac{4}{n} \sum_{i=1}^{n} \frac{X_i - \theta_k}{1 + (X_i - \theta_k)^2}.$$

```r
cauchyMLEscoring <- function(n, theta, init, Tiny) {
   # Fisher scoring MLE for Cauchy(theta)
   # n is sample size, Tiny is the tolerance limit controls
   # iteration estimates, init is the initial value used iteration
x <- rcauchy(n, theta)
i <- 0   # No. of iterations
theta0 <- init +10*Tiny
theta1 <- init
while(abs(theta1-theta0)>Tiny) {
    theta0 <- theta1
    x2 <- x-theta0
    t1 <- mean(x2/(x2*x2+1))         # s(theta0)/(2n)
    theta1 <- theta0 + 4*t1
    i <- i+1
    cat(i, "iteration:", theta1, "\n")  # print out iteration values
}
cat("\n", "MLE:", theta1, "No. of iterations:", i, "\n")
}
```

Calling `cauchyMLEscoring(100, 10, 15, 0.1)` yields

```
MLE: 10.14528 No.  of iterations:  7
```

Note now that the range of valid initial values is much bigger.

Like most iterative algorithms, the choice of appropriate **initial values is important** to ensure the convergence to right limits. In practice multiple initial values are often used.

The differences between the Newton-Raphson and Fisher scoring methods are subtle. We make observations below

- The convergence of the Newton-Raphson algorithm is often faster when both algorithms converge

- The radius of convergence for the Fisher scoring method is often larger, making the choice of initial values less important for the scoring method.

8.3 Evaluating estimation

To measure the accuracy of an MLE or, more general, any estimation procedure, we need to define some measures for the goodness (or badness) of an estimator.

Let $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$ be an estimator of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_o$ be the (unknown) *true value* of $\boldsymbol{\theta}$. Note that

(i) exact estimation error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o$ is unknown, and
(ii) $\widehat{\boldsymbol{\theta}}$ is a random variable

we have to gauge the error

(i) in terms of a probability average, and
(ii) for all possible values of $\boldsymbol{\theta}_o \in \Theta$.

Let $P_{\boldsymbol{\theta}}$, $E_{\boldsymbol{\theta}}$ and $\text{Var}_{\boldsymbol{\theta}}$ denote the probability distribution, expectation and variance under $\boldsymbol{\theta}_o = \boldsymbol{\theta}$.

Bias: $\text{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$

Variance: $\text{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})$

Standard deviation: $\{\text{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^{1/2}$

Standard error: $\{\text{Var}_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}})\}^{1/2}$

Mean square error (MSE): $E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2$

Mean absolute error (MAE): $E_{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|$

Note that

- standard error is a meaningful measure of accuracy for (approximately) unbiased estimators only, and

- MSE (or its squared-root) should be used in general as

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \{\mathrm{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^2 + \mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}).$$

Ideally we would seek for the estimator which minimises MSE or MAE **for all $\boldsymbol{\theta} \in \Theta$** over all possible candidate estimators. Unfortunately such a global optimum rarely exists. However if we confine to some subclass of estimators, the MLE is often optimal or asymptotically optimal.

The MSE is most frequently used largely due to is technical tractability while the MAE leads to estimators which is more robust against outliers in observations.

Fisher Information

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$. The score function is

$$s(\boldsymbol{\theta}) = \dot{l}(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{X}, \boldsymbol{\theta})\}.$$

We assume certain regularity conditions so that we can take derivatives under the integral sign.

Mean of $s(\boldsymbol{\theta})$:

$$\begin{aligned} E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} &= \int s(\boldsymbol{\theta}) f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = \int \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{x}, \boldsymbol{\theta})\} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = 0. \end{aligned}$$

Variance of $s(\boldsymbol{\theta})$ — **Fisher information** matrix:

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} = E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})s(\boldsymbol{\theta})'\} = -E_{\boldsymbol{\theta}}\Big[\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\log\{f(\mathbf{X},\boldsymbol{\theta})\}\Big],$$

because

$$E_{\boldsymbol{\theta}}\Big\{\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}l(\boldsymbol{\theta})\Big\} = \int \frac{\ddot{L}(\boldsymbol{\theta})L(\boldsymbol{\theta}) - \dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x} = \int \ddot{L}(\boldsymbol{\theta})d\mathbf{x} - \int \frac{\dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x}$$
$$= -\int \frac{\dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x} = -\int s(\boldsymbol{\theta})s(\boldsymbol{\theta})'f(\mathbf{x},\boldsymbol{\theta})d\mathbf{x} = -E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})s(\boldsymbol{\theta})'\}.$$

Fisher information $\mathcal{I}(\boldsymbol{\theta})$ measures **the information on θ contained in data X**. Further if $\mathbf{X} = (X_1, \cdots, X_n)'$, and $X_1, \cdots, X_n$ are IID,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \sum_{j=1}^{n}\mathcal{I}_{X_j}(\boldsymbol{\theta}) = n\mathcal{I}_{X_1},$$

i.e. the information is additive.

For $\boldsymbol{\theta} = \theta$ is a scalar, the Fisher information is

$$\mathcal{I}(\theta) = E_\theta\{s(\theta)^2\} = -E_\theta\{\ddot{l}(\theta)\}.$$

Theorem 2. *(Cramér-Rao inequality)*
Let $\mathbf{X} \sim f(\cdot, \theta)$ which satisfying some regularity conditions. Let $T = T(\mathbf{X})$ be a statistic with $g(\theta) = E_\theta(T)$. Then for any $\theta \in \Theta$,

$$\text{Var}_\theta(T) \geq \{\dot{g}(\theta)\}^2 / \mathcal{I}(\theta).$$

The Cramér-Rao inequality specifies **a lower bound** for any *unbiased estimator* for the parameter $g(\theta)$. When the equality holds, T is the **minimum variance unbiased estimator (MVUE)** of $g(\theta)$.

Important case: For any unbiased estimator $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$,

$$\text{Var}(\widehat{\theta}) \geq 1/\mathcal{I}(\theta).$$

Multivariate case: For any unbiased estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$, $\text{Var}(\widehat{\boldsymbol{\theta}}) - \{\mathcal{I}(\boldsymbol{\theta})\}^{-1}$ is a non-negative definite matrix. Hence $\text{Var}(\widehat{\theta}_j) \geq \mathcal{I}^{jj}(\boldsymbol{\theta})$, where $\widehat{\theta}_j$ is the j-th component of $\widehat{\boldsymbol{\theta}}$, and $\mathcal{I}^{jj}(\boldsymbol{\theta})$ is the (j,j)-th element of $\{\mathcal{I}(\boldsymbol{\theta})\}^{-1}$.

Example 9. Let $X_1, \cdots, X_n$ be a sample from $N(\mu, \sigma^2)$. We consider estimators for μ, treating σ^2 as known. The score function (for one observation) is

$$\begin{aligned} s(\mu; X_1) &= \frac{\partial}{\partial \mu} \log[e^{-\frac{1}{2\sigma^2}(X_1-\mu)^2}/\sqrt{2\pi\sigma^2}] \\ &= \frac{\partial}{\partial \mu}[-\frac{1}{2\sigma^2}(X_1-\mu)^2] = (X_1-\mu)/\sigma^2. \end{aligned}$$

Note $\ddot{l}(\mu) = \dot{s}(\mu) = -\sigma^{-2}$. Hence the Fisher information based on a single observation is $\mathcal{I}_{X_1}(\mu) = \sigma^{-2}$. Therefore

$$\mathcal{I}(\mu) = \mathcal{I}_{X_1,\cdots,X_n}(\mu) = n/\sigma^2.$$

For any unbiased estimator $\widehat{\mu}$ for μ, it holds that

$$\mathrm{Var}_\mu(\widehat{\mu}) \geq \sigma^2/n,$$

which is the variance of $\bar{X}$. Hence $\bar{X}$ is the MVUE for μ.

Asymptotic properties of MLEs

Let $X_1, \cdots, X_n$ be i.i.d. with PDF $f(\cdot, \boldsymbol{\theta})$. Write

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{j=1}^{n} \log f(X_j, \boldsymbol{\theta}).$$

Let $\widehat{\boldsymbol{\theta}}$ be the MLE which maximises $l(\boldsymbol{\theta})$. Suppose f fulfils certain regularity conditions.

(a) Consistency.
The MLE is consistent in the sense that as $n \to \infty$,

$$P_{\boldsymbol{\theta}}\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \varepsilon\} \to 0$$

for any $\varepsilon > 0$.

Consistency requires that an estimator converges to the parameter to be estimated. It is a very mild and modest condition that any reasonable estimator should fulfil. The consistency condition is often used to *rule out bad estimators*.

(b) Asymptotic normality

As $n \to \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N\big(0,\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}\big).$$

For large n, it holds **approximately** that

$$\widehat{\boldsymbol{\theta}} \sim N\big(\boldsymbol{\theta},\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}/n\big).$$

Therefore *asymptotically* the MLE is unbiased and attains the Cramér-Rao lower bound. Any estimator fulfilling this condition is called **efficient**.

An approximate standard error of the j-th component of $\widehat{\boldsymbol{\theta}}$ is the square-root of the (j,j)-th element of $\{\mathcal{I}_{X_1}(\widehat{\boldsymbol{\theta}})\}^{-1}$ divided by $n^{1/2}$.