Chapter 2

Frequentist Parameter Estimation

2.1 Parameter Estimation

In order to find the distribution of the data, we need to estimate the parameters of the distribution. We have two frameworks for doing so:

- Frequentist methods: frequentists have different methods for estimation including:
  - *Maximum likelihood*
  - least squares
  - moment method
- Bayesian methods: Parameters are considered to be random and are treated as such. The Bayesian method provides a unified approach consisting of the following steps:
  1. Start with the prior distribution for the parameter
  2. Collect data
  3. Obtain posterior distribution by updating the prior distribution using data and Bayes' theorem

2.2 Maximum likelihood: Introduction and Examples

Suppose data $\mathcal{D}$ is collected and is assumed to be derived from a distribution p with unknown parameter θ. Let the probability of observing $\mathcal{D}$, assuming θ, be denoted by $p(\mathcal{D};\theta)$. **Maximum likelihood** estimation finds θ that maximizes $p(\mathcal{D},\theta)$:

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(\mathcal{D};\theta)$$

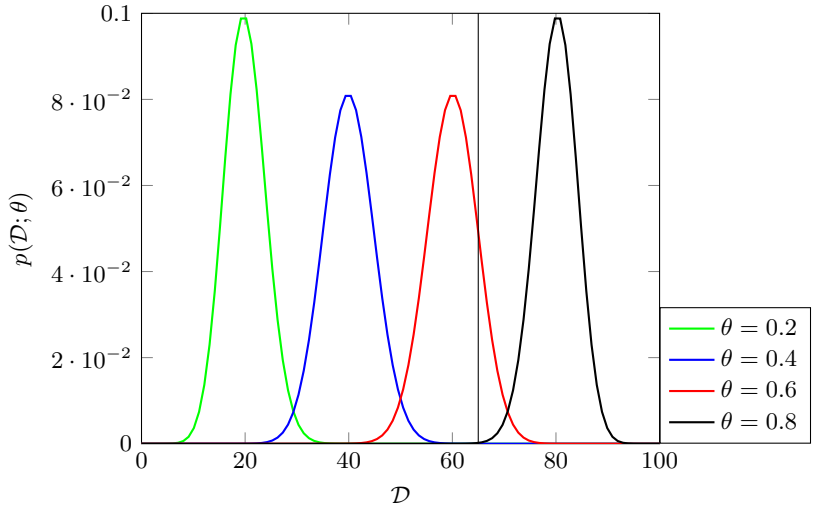
The expression $p(\mathcal{D};\theta)$, viewed as a function of θ, is called the **likelihood**; hence the name maximum likelihood estimation. As shorthand, we use $L(\theta) = p(\mathcal{D};\theta)$ and $\ell(\theta) = \ln L(\theta)$, where $\ell(\theta)$ is the **log-likelihood**. Clearly, the value of θ that maximizes $L(\theta)$ is the same as the one that maximizes $\ell(\theta)$:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \ln p(\mathcal{D};\theta)$$

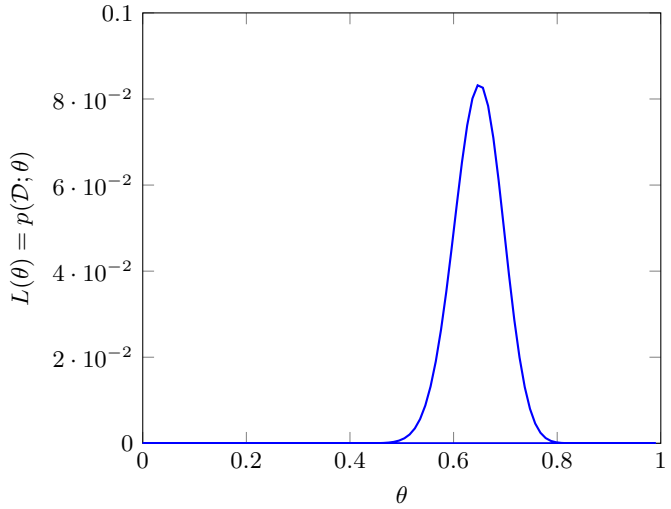
Example 39. In this example, we attempt to show the intuition behind maximum likelihood. Let T be a binary random variable such that $T = 1$ if there is traffic and $T = 0$ if there is no traffic. Suppose that data $\mathcal{D}$ collected over 100 days indicates that 65 days had no traffic. We have

$$\Pr(T = 0) = \theta$$
$$p(\mathcal{D}; \theta) = \binom{100}{65} \theta^{65} (1 - \theta)^{35}$$

Let's try a few different choices for θ and see which one makes more sense. In the figure below, $p(\mathcal{D}; \theta)$ is plotted for $\theta \in \{0.2, 0.4, 0.6, 0.8\}$. The vertical line indicates the observation, i.e., 65 days with no traffic. Which is a more appropriate value for θ?



If $\theta = 0.2$, the probability of 65 days with no traffic is very small. So observing $\mathcal{D} = 65$ would be very unlikely, which in turn would make $\theta = 0.2$ an unreasonable guess. Among the presented choices, $\theta = 0.6$ appears the most reasonable. This reasoning suggests the following: *The value of the parameter that assigns a higher probability to the observation is a better choice.* Since we are not limited to a specific set of choices, we can find the parameter that **maximizes** the probability of the observation, i.e., the maximum-likelihood estimate. In the figure below, $L(\theta) = p(\mathcal{D}, \theta)$ is plotted as a function of θ. This is the likelihood.



We can see that $\theta = 0.65$ maximizes the likelihood and hence is the maximum-likelihood estimate. We can show this also analytically. First, the likelihood is given as

$$L(\theta) = p(\mathcal{D}; \theta) = \binom{100}{65} \theta^{65}(1-\theta)^{35}.$$

We usually use the log-likelihood as the function to optimize:

$$\ell(\theta) = \log L(\theta) = \log\left(\binom{100}{65} \theta^{65}(1-\theta)^{35}\right) \doteq 65 \log\theta + 35\log(1-\theta), \tag{2.1}$$

where $\doteq$ denotes equality but with ignoring additive terms that are constant in θ (and thus do not alter the value of θ that maximize the log-likelihood). We differentiate $\ell(\theta)$ to find the value of θ that maximizes $l(\theta)$.

$$\frac{d\ell(\theta)}{d\theta} = \frac{65}{\theta} - \frac{35}{1-\theta} = 0 \implies 65 - 65\theta = 35\theta \implies \hat{\theta}_{ML} = \frac{65}{100}. \tag{2.2}$$

Note that this result is intuitive as it agrees with our observation that 65% of the days had no traffic.

Example 40 (Parameters of the normal distribution). A device for measuring an unknown quantity μ is used n times producing values $\mathcal{D} = \boldsymbol{y} = (y_1, \ldots, y_n)$. Each measurement is independent and for each i we have $y_i = \mu + z_i$, where z_i is the measurement noise satisfying $z_i \sim \mathcal{N}(0, \sigma^2)$. Note that this implies $y_i \sim \mathcal{N}(\mu, \sigma^2)$. We consider the problem in two cases: mu is unknown but σ^2 is known; and both μ and σ are unknown.

- Known σ^2, unknown μ: We have

$$p(y_i;\mu) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right)$$
$$p(\boldsymbol{y};\mu) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right)$$
$$L(\mu) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right)$$
$$\ell(\mu) = \sum_{i=1}^{n}\left(-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right) \doteq -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2$$

  and so

$$\frac{d\ell}{d\mu} = \sum_{i=1}^{n}\frac{y_i-\mu}{\sigma} = 0 \implies \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}.$$

- Unknown σ^2, μ: We have

$$\ell(\mu,\sigma) = \sum_{i=1}^{n}\left(-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right) \doteq -n\ln\sigma - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2$$

  and so

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

  Solving this system of equations for μ and σ yields

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y},$$
$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

2.3 Properties of Estimators

Maximum likelihood is just one way of estimating parameters. For example, in Example 40, we could choose the middle value among $y_1,\ldots,y_n$ as the estimate for μ. Given the fact that there are many estimators, how do we evaluate them and select one? In this section, we will see some of the evaluation criteria.

2.3.1 Estimation error and bias

For an estimator $\hat{\theta}$ of θ, assume $\mathcal{D}$ is collected. Then the error is given is

$$\hat{\theta}(\mathcal{D}) - \theta,$$

where $\hat{\theta}(\mathcal{D})$ is the estimate based on data $\mathcal{D}$.

For a given estimation task that is performed once, since we do not know the true value, we cannot find $\hat{\theta}(\mathcal{D}) - \theta$. Even if we know the true value, the error is the result of only one experiment and does not tell us much about the general behavior of the estimator.

However, we can think of the thought experiment in which estimation is performed many times and consider the behavior of the estimator and its error. For example, we may consider whether the result would be generally an overestimate or an underestimate? The key point in answering such questions is that *the estimate itself is a random value because each time we perform the estimation task, new data samples are obtained and these are random, following a certain distribution.* So for example, we can talk about the expected error. In other words, since $\mathcal{D}$ is random (although its distribution is the same in each experiment), so is $\hat{\theta}(\mathcal{D})$.

So we can consider the expected error, known as **bias**,

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}(\mathcal{D}) - \theta] \tag{2.3}$$

The expected value is taken over $\mathcal{D}$. However, the dependence on data is often implicit and we write

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta \tag{2.4}$$

Bias of the estimator tells us that whether in general the estimator over- or under-estimates the true value. If bias is equal to 0, then the estimator is called **unbiased**.

Example 41. Given n samples $y_1, \ldots, y_n$ from a distribution with mean μ and variance σ^2, are the estimators

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

for the mean and variance, respectively, unbiased? For $\hat{\mu}$, we have

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[\bar{y}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[y_i] = \frac{1}{n}\cdot n \cdot \mathbb{E}[y_1] = \mu$$

and so the ML estimator for the mean is unbiased. We can show (how?) that

$$\mathbb{E}\big[\hat{\sigma}^2\big] = \frac{n-1}{n}\sigma^2$$

and the bias of estimating σ^2 is

$$\mathbb{E}\big[\hat{\sigma}^2\big] - \sigma^2 = -\frac{1}{n}\sigma^2.$$

Based on this, we can create an unbiased estimator for the variance as

$$\hat{\sigma}_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Example 42. [1, Example 2.8.2] An urn has N balls, numbered $1, 2, ..., N$. Suppose however that N is unknown to us. We pick one random ball from the urn and the number on the ball is y. We Estimate N using maximum likelihood. First, for $p(y; N)$ we have

$$p(y; N) = \begin{cases} \frac{1}{N} & y \leq N, \\ 0 & y > N. \end{cases}$$

and thus

$$L(N) = \begin{cases} \frac{1}{N} & N \geq y, \\ 0 & N < y. \end{cases}$$

Hence, $L(N)$ is maximized by choosing $N = y$ and so $\hat{N}_{ML} = y$. To find the bias of $\hat{N}_{ML}$,

$$\mathbb{E}[\hat{N}_{ML}] = \mathbb{E}[y] = \sum_{i=1}^{N} i \cdot \frac{1}{N} = \frac{N+1}{2},$$
$$\text{Bias}(\hat{N}_{ML}) = \frac{N+1}{2} - N = -\frac{N-1}{2},$$

which means that the ML estimator tends to underestimates N by almost a factor of 2.

Example 43 (Linear unbiased estimator). Can we design an unbiased estimator for Example 42? There are many options but for simplicity we may choose an estimator that is linear in the data, in particular, one of the form

$$\hat{N}_L = ay + b.$$

We find a and b such that $\hat{N}_L$ is unbiased. We have

$$\mathbb{E}[\hat{N}_L] = a\,\mathbb{E}\,y + b = a\frac{N+1}{2} + b.$$

Setting this equal to N (equality should hold for any N) yields $a = 2$ and $b = -1$, i.e.,

$$\hat{N}_L = 2y - 1.$$

Example 44 (Survival of Humanity (!)). The human species will eventually die out. We use the two methods to estimate the total number of humans N who will ever live. Let humans be enumerated as $h_1, h_2, ..., h_y, ..., h_N$, where h_1 represents Adam, h_2 represents Eve, h_y represents you, and h_N represents the last human to live. Assuming that your birth order is random, this is similar to the urn in Example 42.

Assuming that 100 billion have been born so far, we have $\hat{N}_{ML} = 100$ billion and $\hat{N}_L = 200$ billion. The ML estimates predicts that the end is here. Further, assuming that there will be 140 million births each year, the unbiased estimator predicts the end of humanity to occur in around 700 years.

Exercise 45. Given iid data $\mathcal{D} = (y_1, \ldots, y_n), n \geq 3$, with mean θ, find the bias of each of the following estimators,

$$\hat{\theta}_1 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$
$$\hat{\theta}_2 = y_1,$$
$$\hat{\theta}_3 = \frac{2y_2 + y_3}{3}.$$

2.3.2 Mean squared error and variance

Example 46. Consider an unbiased estimator $\hat{\theta}$ and define $\hat{\theta}' = \hat{\theta} + W$, where W is a zero-mean random variable with a large variance. Now, $\hat{\theta}'$ is unbiased, similar to $\hat{\theta}$, but it is not a good estimator (regardless of how good $\hat{\theta}$ is). So clearly, being unbiased alone is not sufficient to ensure that an estimator is "good."

The mean squared error (MSE) is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] &= \mathbb{E}\left[\left(\left(\hat{\theta} - \mathbb{E}\,\hat{\theta}\right) - \left(\theta - \mathbb{E}\,\hat{\theta}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\,\hat{\theta}\right)^2\right] - 2\,\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\,\hat{\theta}\right)\right]\left(\theta - \mathbb{E}\,\hat{\theta}\right) + \left(\theta - \mathbb{E}\,\hat{\theta}\right)^2 \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\,\hat{\theta}\right)^2\right] + \left(\theta - \mathbb{E}\,\hat{\theta}\right)^2
\end{aligned}$$

and hence

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

For unbiased estimators, the variance of the estimator becomes an important quantity since it is equal to the MSE.

Example 47. Consider data $\mathcal{D} = \{y_1, ..., y_n\}$, where y_i are iid with distribution $\mathcal{N}(\mu, \sigma^2)$. The ML estimator for the mean $\hat{\mu}_{ML} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is unbiased. We have

$$\mathrm{MSE}(\hat{\mu}_{ML}) = \mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n}.$$

Note that as n increases, the MSE decreases and the estimate becomes more accurate, as would be expected. This property is studied next.

Exercise 48. For the estimators in Exercise 45, find the MSE, assuming the variance is σ^2.

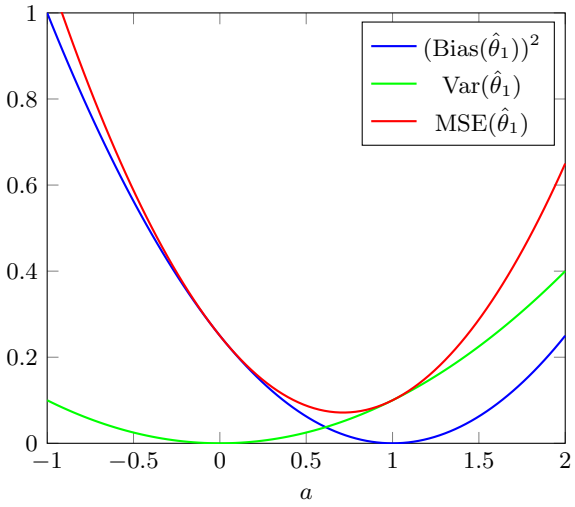
Exercise 49 (Bias-variance trade-off)**.** Given iid data $\mathcal{D} = (y_1, \ldots, y_n), n \geq 3$, with mean θ and variance σ^2, the MSE of

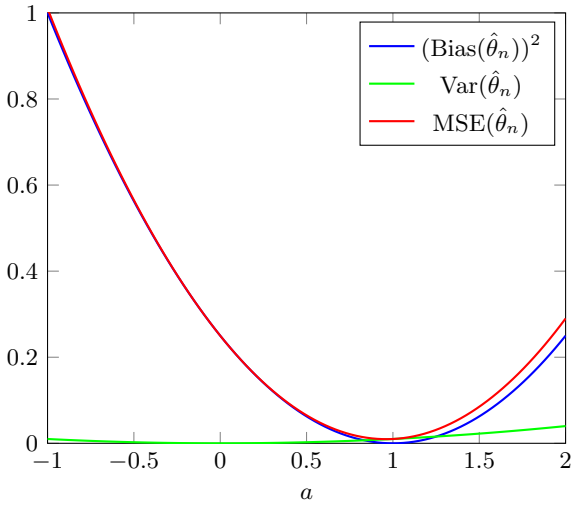
$$\begin{aligned}
\hat{\theta}_1 &= a y_1, \\
\hat{\theta}_n &= a\bar{y} = \frac{a}{n}\sum_{i=1}^{n} y_i,
\end{aligned}$$

for some constant $a \in \mathbb{R}$ is given as

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}_1) &= (a-1)^2\theta^2 + a^2\sigma^2, \\
\mathrm{MSE}(\hat{\theta}_n) &= (a-1)^2\theta^2 + a^2\sigma^2/n.
\end{aligned}$$

What is a good value for a? Does anything other than $a = 1$ make sense? The components of the MSE are given in the plots below for $\hat{\theta}_1$ and $\hat{\theta}_n$ with $n = 10$. A trade-off between the bias and variance is evident. Why is it not feasible to design an estimator by optimizing for a? What is the difference between estimation based on little data ($\hat{\theta}_1$) and a lot of data ($\hat{\theta}_n, n = 10$)?





2.3.3 Consistency

An estimator $\hat{\theta}_n$ based on n samples is said to be **consistent** if $\hat{\theta}_n \to \theta$ as $n \to \infty$. More precisely, for all $\epsilon > 0$, we need

$$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| \geq \epsilon) = 0.$$

In other words, the estimator is accurate if the size of the data is large.

Example 50. The ML and linear estimators described in Examples 42 and 43 are very different for a single data point. But how do they behave if we have a lot of data. First we need to define these for n data samples. Suppose that we take n samples from the urn with replacement, resulting in $\mathcal{D} = (y_1, y_2, \ldots, y_n)$. Define

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

To extend the linear estimator to n data points, we can choose

$$\hat{N}_{L,n} = 2\bar{y} - 1.$$

For the ML estimator, we have (why?)

$$\hat{N}_{ML,n} = \max_i y_i.$$

Both of these, although they look very different, are consistent and converge to N as $n \to \infty$.

- As $n \to \infty$, by LLN, $\bar{y}$ converges to the mean of the distribution, i.e., $\mathbb{E}\, y_1 = \frac{N+1}{2}$. Hence, $\hat{N}_{L,n} \to 2\frac{N+1}{2} - 1 = N$.
- For the ML estimator, as $n \to \infty$, at some point, we will pick the ball numbered N and so we will eventually have $\hat{N}_{ML} = N$.

Given the two estimators, the bad news is that the estimators disagree significantly for small data. However, as the size of sample data increases, the two estimators agree.

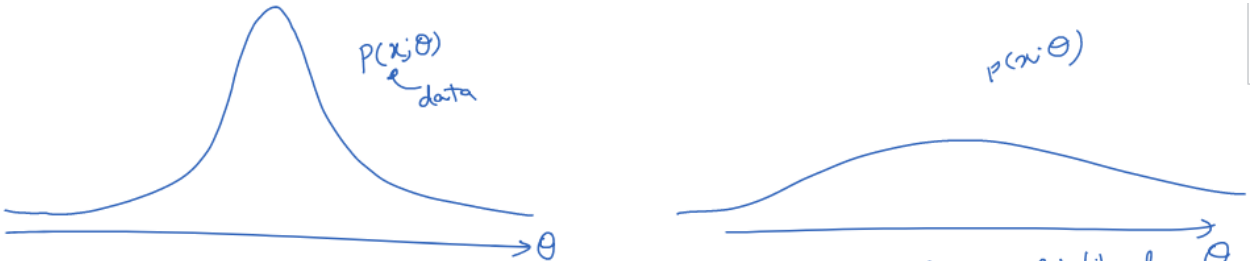


Figure 2.1: The log-likelihood on the left demonstrates strong dependence on θ compared to the one on the right.

2.4 The Cramer-Rao lower bound*

For an unbiased estimator, the MSE is equal to the variance, and thus the variance represents the accuracy of the estimator. This leads to the following question: *For a given distribution of data, what is the smallest possible variance of an unbiased estimator?*

The accuracy of estimating a parameter θ depends on how strongly the distribution of the data depends on θ. If the dependence is strong, i.e., for values of θ other than the true value, the probability of the observed data falls sharply, then we may expect to find θ with accuracy. On the other hand, if the dependence is week, then it will be difficult to find θ with precision. These two cases are shown in Figure 2.1.

Let the data be encoded as a vector x, i.e., $\mathcal{D} = x$. The sharpness of the log-likelihood $\ell(\theta)$ can be quantified as

$$-\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}. \tag{2.5}$$

Given the randomness of data the above quantity is random. So to average over the data, we define

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right] = -\int \frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2} p(x;\theta) dx,$$

which is called the **Fisher Information**.

The following theorem provides a lower bound on the variance as is referred to as the Cramer-Rao lower bound (CRLB).

Theorem 51 (CRLB). *Given that the log-likelihood $\ell(\theta)$ satisfies a certain regularity condition*[1]*, the variance of any unbiased estimator $\hat{\theta}$ of θ satisfies*

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

If an estimator achieves the CRLB, i.e., $\mathrm{Var}(\hat{\theta}) = 1/I(\theta)$, then it is called **efficient**.

As a special case, consider when we have n iid data points, and denote the estimator based on this data as $\hat{\theta}_n$. Denote the Fisher information based on n data points as $I_n(\theta)$ and based on one data point as $I_1(\theta) = I(\theta)$. Since the Fisher information is additive (Why? Hint: definition), we have $I_n(\theta) = nI(\theta)$. Thus, the variance of an unbiased estimator $\hat{\theta}_n$ based on n independent observations satisfies

$$\mathrm{Var}(\hat{\theta}_n) \geq \frac{1}{nI(\theta)}. \tag{2.6}$$

[1]The regularity condition is $\mathbb{E}\left[\frac{\partial \ell(\theta)}{\partial \theta}\right] = 0$, for all θ.

Example 52. In Example 40, where we estimated the mean of a Gaussian distribution with known σ^2 based on n iid samples $y_1, \ldots, y_n$, the log-likelihood, ignoring constant terms, was given as

$$\ell(\mu) \doteq -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}.$$

And,

$$\frac{\partial \ell(\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu). \tag{2.7}$$

Then regularity condition is satisfied since

$$\mathbb{E}\left[\frac{\partial \ell(\mu)}{\partial \mu}\right] = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbb{E}[y_i - \mu] = 0,$$

for all μ. Furthermore,

$$\frac{\partial^2 \ell(\mu)}{\partial \mu^2} = -\frac{n}{\sigma^2} \implies I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\mu)}{\partial \mu^2}\right] = \frac{n}{\sigma^2}.$$

Based on the CRLB, the variance of the estimator satisfies

$$\mathrm{Var}(\hat{\mu}) \geq \frac{\sigma^2}{n}.$$

The the variance of the estimator is $\mathrm{Var}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$. Hence, the ML estimator is efficient in this case.

2.5 Asymptotic normality of the MLE

As shown before, the maximum likelihood estimator is not necessarily unbiased. However, if we have a large amount of data, under some regularity conditions, the ML estimator $\hat{\theta}_n$ based on n iid data points satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, I^{-1}(\theta)).$$

So for large data, $\hat{\theta}_n$ is nearly normally distributed with mean θ (hence unbiased) and variance $I^{-1}(\theta)/n$ (efficient).

While we stated the CRLB and the asymptotic normality of the MLE for scalar parameters, almost identical results also hold for a vector of parameters.