Chapter 2

Frequentist Parameter Estimation

2.1 Overview

Parameter estimation can be used to infer unknowns about the real world (e.g, the frequency of a given disease among individuals with a certain genetic mutation) and to estimate the distribution of the data in machine learning problems.

There are two main frameworks for parameter estimation:

- Frequentist methods: In the frequentists' perspective, the true parameter value θ^* is unknown and fixed. The estimate $\hat{\theta}$ is a function of the data, which provides a single "best" estimate of θ^*. Frequentists have different methods for estimation including *maximum likelihood*, which we will discuss in detail, and the *moment method*, which finds the parameters by solving equations obtained by equating empirical moments and theoretical moments.
- 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 estimation

Suppose data x is collected. We model this data as a realization of a random variable X with distribution p_X, which has an unknown parameter θ^*. The probability of observing x, assuming θ, is $p_X(x;\theta)$. To estimate θ^*, **Maximum likelihood estimation (MLE)** chooses the parameter that assigns the highest probability to the data:

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta} p_X(x;\theta).$$

The expression $p(x;\theta)$, viewed as a function of θ, is called the **likelihood**; hence the name maximum likelihood estimation. As shorthand, we use $L(\theta) = p_X(x;\theta)$ 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 $\ell(\theta)$:

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \ln p_X(x;\theta)$$

Example 2.1. In this example, we attempt to show the intuition behind maximum likelihood. Suppose that a given road has heavy traffic or light traffic. We denote the probability of light traffic by θ^*. To estimate

data, we count the number of times X that the road has light traffic in a period of 100 days. After collecting this data, we observe that $X = 65$. We have

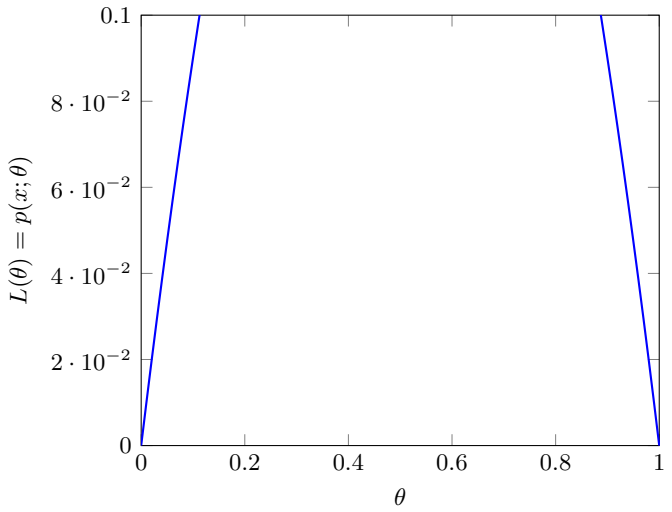
$$p_X(65;\theta) = \binom{100}{65}\theta^{65}(1-\theta)^{35}$$

Let's try a few different choices for θ, e.g., $\theta \in \{0.2, 0.4, 0.6, 0.8\}$, and see which one makes more sense:

$$\begin{aligned} p(65, \theta = 0.2) &= 1.6 \times 10^{-22}, \\ p(65, \theta = 0.4) &= 0.00000026, \\ p(65, \theta = 0.6) &= 0.0491, \\ p(65, \theta = 0.8) &= 0.00019, \end{aligned}$$

If $\theta = 0.2$, the probability of 65 days with light traffic is extremely small. So observing $x = 65$ 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. In the figure below, $L(\theta) = p(x;\theta)$ is plotted as a function of θ. This is the likelihood function.



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

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

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 $l(\theta)$.

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

Note that this result is intuitive as it agrees with our observation that 65% of the days had light traffic. △

A note on notation: In general, our data is a vector, which we denote by bold symbols such as $\boldsymbol{x}$. The corresponding random variable is $\boldsymbol{X}$.

Example 2.2 (**Parameters of the normal distribution**)**.** A device for measuring an unknown quantity μ^* (e.g., the mass of an electron) is used n times producing values $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. 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 $Y_i \sim \mathcal{N}(\mu^*, (\sigma^*)^2)$.

Suppose we have collected data $\boldsymbol{y} = (y_1, \ldots, y_n)$. We consider the problem in two cases: μ^* is unknown but σ^* is known; and both μ^* and σ^* are unknown.

- Known σ^*, unknown μ^*: We have

$$p_{Y_i}(y_i; \mu) = \frac{1}{\sigma^* \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma^*} \right)^2 \right)$$

$$L(\mu) = p_{\boldsymbol{Y}}(\boldsymbol{y}; \mu) = \prod_{i=1}^{n} p_{Y_i}(y_i; \mu)$$

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

and so

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

- Unknown σ^*, 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

$$\hat{\mu}_{\text{mle}} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y},$$

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

△

2.2.1 Maximum likelihood and the closest distribution

We have described maximum likelihood as aiming to find a distribution that gives a high probability to the observed data. An alternative view relates it to the empirical distribution of the data, denoted $p_{\boldsymbol{x}}$. Given $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, let $\#_x$ denote the number of times x appears in $\boldsymbol{x}$. The empirical distribution is given as

$$p_{\boldsymbol{x}}(x) = \frac{\#_x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\boldsymbol{x} = \boldsymbol{x}_i)$$

where $\mathbb{1}(\cdot)$ equals 1 if the enclosed condition is true and 0 otherwise.

Now consider a parameterized family of distributions p_θ. It makes sense to choose θ such that p_θ is close to $p_{\boldsymbol{x}}$. In other words, we would like θ to be chosen such that p_θ describes the observed data well. A standard way of measuring the "closeness" of p_θ to the empirical distribution $p_{\boldsymbol{x}}$ is relative entropy, $D_{KL}(p_{\boldsymbol{x}}||p_\theta)$.

It turns out the closest distribution is in fact given by maximum likelihood, i.e.,

$$\hat{\theta}_{\text{mle}} = \arg\min_\theta D_{KL}(p_{\boldsymbol{x}}||p_\theta). \tag{2.3}$$

This fact provide further evidence for the soundness of MLE strategy. Note in particular that if there exists θ such that $p_\theta = p_{\boldsymbol{x}}$, it will be chosen by MLE. This is because relative entropy is always non-negative and equals to 0 if and only if the two distributions are the same. So choosing $p_\theta = p_{\boldsymbol{x}}$, if possible, provides the smallest value for the relative entropy, i.e., 0.

Exercise 2.3. Prove (2.3). △

Exercise 2.4. (†) Note that relative entropy is not symmetric. Instead of $D_{KL}(p_{\boldsymbol{x}}||p_\theta)$, we could minimize $D_{KL}(p_\theta||p_{\boldsymbol{x}})$. What are the differences between the two formulations and which one is more suitable for parameter estimation? △

2.3 Properties of Estimators

Maximum likelihood is just one way of estimating parameters. We can choose any function of the data as the estimate. For instance, in Example 2.2, we could choose the middle (median) 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?

Clearly, we would like the estimate to be close to the true value. But stating this condition in a rigorous probabilistic way is a bit challenging in the frequentist framework. We are specifically interested in the error:

$$\hat{\theta}(\boldsymbol{x}) - \theta^*,$$

where $\hat{\theta}(\boldsymbol{x})$ is the estimate based on data $\boldsymbol{x}$ and θ^* is the true value[1]. Evaluating $\hat{\theta}(\boldsymbol{x})$ is difficult because, obviously, the true value is unknown.

So instead of finding the specific error, we may try to find the probability that the true value θ^* is within say 10% of the estimate $\hat{\theta}$. But after the estimate is produced based on a given data set, the estimate is a deterministic value. For instance, in Example 2.1, the MLE is given as $\hat{\theta}_{\text{mle}} = 0.65$. So questions such as "What is the probability that the difference between θ^* and $\hat{\theta}(\boldsymbol{x})$ is larger than 0.05?" are not meaningful because, while θ^* is unknown, both θ^* and $\hat{\theta}(\boldsymbol{x})$ are deterministic after data is collected and the estimation task is performed.

The solution to these difficulties is to study the properties of the estimator not based on a specific realization $\boldsymbol{x}$ of the data but in general, over all possible data sets that could be produced and all the resulting estimated

[1]Note the slight abuse of notation: sometimes θ is used as the generic parameter, e.g., as the argument of the likelihood function, and sometimes as the true value of the parameter. The distinction should be clear from the context

values. We can think of the thought experiment in which many, many, data sets are collected and the estimation task is performed based on each. The estimate itself is a random variable because each time we perform the estimation task, new data samples are obtained and these are random, following a certain distribution. In other words, instead of considering a single estimate $\hat{\theta}(\boldsymbol{x})$ for a specific realization $\boldsymbol{x}$, we study the estimator $\hat{\theta}(\boldsymbol{X})$, i.e., a random variable. Then it makes sense to ask "What is the probability that the difference between θ^* and $\hat{\theta}(\boldsymbol{X})$ is larger than 0.05?" since $\hat{\theta}(\boldsymbol{X})$ is a random variable with some distribution. It may be difficult to find the distribution of $\hat{\theta}(\boldsymbol{x})$ and it may depend on the unknown parameter θ^* but at least the question is meaningful. In this section, we will see some of the evaluation criteria based on this view.

A note on notation: Typically, we use θ as the generic parameter, with θ^* denoting its true value, according to which $\boldsymbol{X}$ is distributed. For a given data $\boldsymbol{x}$, the estimate is shown by $\hat{\theta}(\boldsymbol{x})$ or $\hat{\theta}$. So, $\hat{\theta}$ denotes both the estimator, i.e., a function that produces the estimate given the data, and the estimate; the intent should be clear from the context. Finally, we may use $\hat{\Theta} = \hat{\theta}(\boldsymbol{X})$ to denote the estimate as a random variable.

2.3.1 Bias

Bias is the expected estimation error,

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

As discussed, the expected value is taken over the randomness in $\boldsymbol{X}$. Bias of the estimator tells us 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 2.5 (Example 2.1 continued)**.** Previously, we obtained the maximum likelihood estimate for the probability θ of having light traffic. Let us find its bias. Again we collect data over 100 days and let X denote the number of days when there is light traffic. We know that $\hat{\theta}_{\text{mle}}$, as a function of data, is given by

$$\hat{\theta}_{\text{mle}}(X) = \frac{X}{100},$$

Note that instead of using a specific value for the number of days with light traffic, such as 65, we use a random variable X representing this quantity. Dropping the dependence on X for simplicity, the expected value of $\hat{\theta}_{\text{mle}}$ is given by

$$\mathbb{E}\Big[\hat{\theta}_{\text{mle}}(\boldsymbol{X})\Big] = \frac{\mathbb{E}[X]}{100}.$$

Assuming θ^* to be the true value, the number X of days when there is light traffic follows $\text{Bin}(100, \theta^*)$, and so $\mathbb{E}[X] = 100\theta^*$. It follows that

$$\mathbb{E}\Big[\hat{\theta}_{\text{mle}}(\boldsymbol{X})\Big] = \frac{100\theta^*}{100} = \theta^*.$$

Hence, the maximum likelihood estimate is an unbiased estimator. △

Example 2.6. Given iid data $\boldsymbol{y} = (y_1, \ldots, y_n), n \geq 3$, with mean θ^*, let us find the bias of each of the following estimators,

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

Let Y_i be the random variable corresponding to observation y_i and $\bar{Y} = \sum_{i=1}^{n} Y$. We have

$$\begin{aligned}
\mathbb{E}\,\hat{\theta}_1(\boldsymbol{Y}) &= \mathbb{E}\,\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,Y_i = \frac{1}{n}\sum_{i=1}^{n}\theta^* = \theta^*,\\
\mathbb{E}\,\hat{\theta}_2(\boldsymbol{Y}) &= \mathbb{E}\,Y_1 = \theta^*,\\
\mathbb{E}\,\hat{\theta}_3(\boldsymbol{Y}) &= \mathbb{E}\left[\frac{2Y_2 + Y_3}{3}\right] = \frac{2\,\mathbb{E}\,Y_2 + \mathbb{E}\,Y_3}{3} = \theta^*.
\end{aligned}$$

So all of these estimators are unbiased. △

Example 2.7. Given n samples $\boldsymbol{y} = (y_1, \ldots, y_n)$ from a distribution with mean μ^* and variance $(\sigma^*)^2$, 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

$$\mathbb{E}[\hat{\mu}(\boldsymbol{Y})] = \mathbb{E}[\bar{Y}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{1}{n}n\,\mathbb{E}[Y_1] = \mu^*$$

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

$$\mathbb{E}\big[\hat{\sigma}^2(\boldsymbol{Y})\big] = \frac{n-1}{n}(\sigma^*)^2$$

and the bias of estimating $(\sigma^*)^2$ is

$$\mathbb{E}\big[\hat{\sigma}^2(\boldsymbol{Y})\big] - (\sigma^*)^2 = -\frac{1}{n}(\sigma^*)^2.$$

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

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

△

Example 2.8. [1, Example 2.8.2] An urn has m^* balls, numbered $1, 2, ..., m^*$. Suppose however that m^* is unknown to us. We pick one random ball from the urn and the number on the ball is y. We estimate m^* using maximum likelihood. First, let Y be the random variable corresponding to observation y, with distribution $p_Y(y; m^*)$. We have

$$p_Y(y; m) = \begin{cases} \frac{1}{m} & y \le m,\\ 0 & y > m. \end{cases}$$

and thus

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

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

$$\begin{aligned}
\mathbb{E}[\hat{m}_{\text{mle}}(Y)] = \mathbb{E}[Y] &= \sum_{i=1}^{m^*} i \cdot \frac{1}{m^*} = \frac{m^* + 1}{2},\\
\text{Bias}(\hat{m}_{\text{mle}}) &= \frac{m^* + 1}{2} - m^* = -\frac{m^* - 1}{2},
\end{aligned}$$

which means that the ML estimator tends to underestimates m^* by almost a factor of 2. △

Example 2.9 (**Linear unbiased estimator**)**.** Can we design an unbiased estimator for Example 2.8? There are many options, but for simplicity we may choose an estimator that is linear in the data, in particular, one of the form

$$\hat{m}_L(y) = ay + b.$$

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

$$\mathbb{E}[\hat{m}_L(Y)] = a\,\mathbb{E}\,Y + b = a\frac{m^* + 1}{2} + b.$$

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

$$\hat{m}_L(y) = 2y - 1.$$

△

Example 2.10 (**Survival of Humanity (!)**)**.** The human species will eventually die out. We use two methods to estimate the total number of humans m who will ever live. Let humans be enumerated by birth order as $h_1, h_2, ..., h_y, ..., h_m$, 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 y is random, the problem is similar to estimating the number of balls in an urn in Example 2.8.

Assuming that 100 billion humans have been born so far, we have $\hat{m}_{\text{mle}} = 100$ billion and $\hat{m}_L = 200$ billion. The ML estimate 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. △

2.3.2 Mean squared error and variance

Example 2.11. 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." △

For an estimator $\hat{\theta}$, where the random variable describing data is denoted by $\boldsymbol{X}$, the mean squared error (MSE) is defined as

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}(\boldsymbol{X}) - \theta^*\right)^2\right].$$

The smaller the MSE, the more accurate the estimator.

Let $\hat{\Theta} = \hat{\theta}(\boldsymbol{X})$. Note that

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

where, the third equality uses the fact that $\mathbb{E}\,\hat{\Theta} - \theta^*$ is a deterministic constant and the fourth equality the fact that $\mathbb{E}\left[\left(\hat{\Theta} - \mathbb{E}\,\hat{\Theta}\right)\right] = 0$. Hence,

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

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

Example 2.12 (Example 2.1 re-revisit)**.** We saw in Example 2.5 that the maximum likelihood estimate for the probability of traffic θ^* is unbiased. Now, let us find its variance. Again, we write $\hat{\theta}_{\text{mle}}(\boldsymbol{X}) = \frac{X}{100}$ and

$$\text{Var}(\hat{\theta}_{\text{mle}}) = \frac{\text{Var}(X)}{100^2} = \frac{\theta^*(1-\theta^*)}{100},$$

where X is the number of days without traffic, which follows $\text{Bin}(100, \theta^*)$ with variance $100\theta^*(1-\theta^*)$. As we can see, the variance (hence, MSE) increases as the true value of θ^* approaches $1/2$, i.e., every data point contains more uncertainty. Furthermore, we can extend this result to the more general case where we collect data for n days. By the same argument, we get

$$\text{MSE}(\hat{\theta}_{\text{mle}}) = \text{Var}(\hat{\theta}_{\text{mle}}) = \frac{\theta^*(1-\theta^*)}{n}.$$

△

Example 2.13. Consider data $\boldsymbol{y} = (y_1, ..., y_n)$, where the corresponding random variables Y_i are iid with distribution $\mathcal{N}(\mu, \sigma^2)$. The ML estimator for the mean μ is $\hat{\theta}_{\text{mle}}(\boldsymbol{y}) = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is unbiased. We have

$$\text{MSE}(\hat{\theta}_{\text{mle}}) = \text{Var}(\bar{Y}) = \frac{\sigma^2}{n}.$$

△

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

Exercise 2.14. For the estimators in Example 2.6, find the MSE, assuming the variance is $(\sigma^*)^2$. △

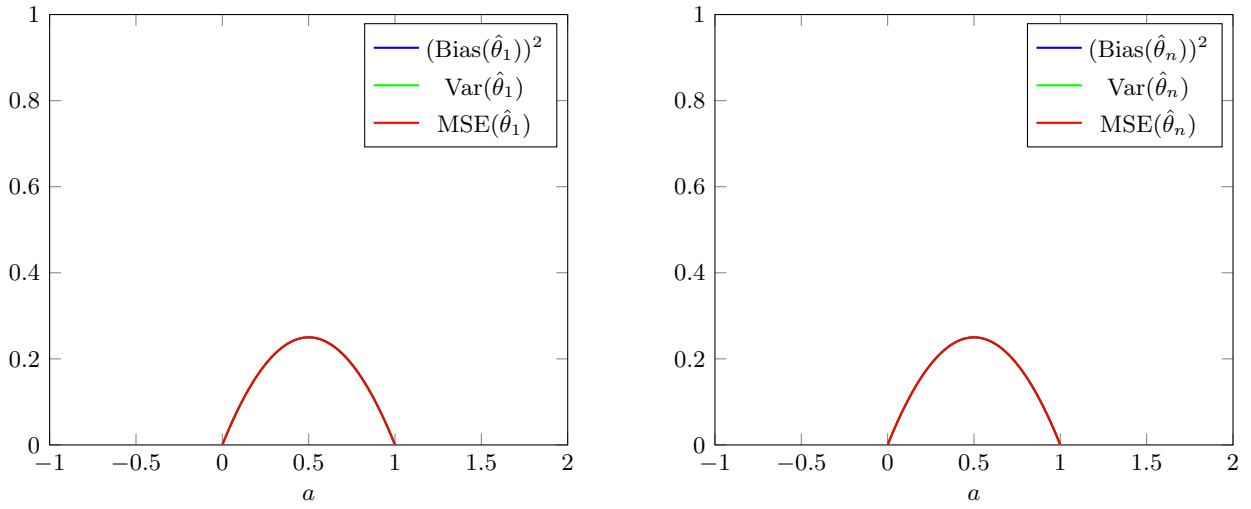
Exercise 2.15 (Bias-variance trade-off)**.** Given iid data $\boldsymbol{y} = (y_1, \ldots, y_n), n \geq 3$, with mean θ^* and variance σ^2, show that the MSE of

$$\hat{\theta}_1 = ay_1,$$
$$\hat{\theta}_n = a\bar{y} = \frac{a}{n}\sum_{i=1}^n y_i,$$

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

$$\text{MSE}(\hat{\theta}_1) = (a-1)^2(\theta^*)^2 + a^2\sigma^2,$$
$$\text{MSE}(\hat{\theta}_n) = (a-1)^2(\theta^*)^2 + a^2\sigma^2/n.$$

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 $n = 10$, for $\theta^* = 0.5, \sigma^2 = 0.1$. 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

Consider an estimator $\hat{\theta}_n(\boldsymbol{x})$ based on n samples $\boldsymbol{x} = (x_1, \ldots, x_n)$. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be the random variables that describe the n data samples and let $\hat{\Theta}_n = \hat{\theta}_n(\boldsymbol{X})$ be the random variable that corresponds to the estimate. The estimator $\hat{\theta}_n$ 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 2.16. The ML and linear estimators described in Examples 2.8 and 2.9 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 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. Define

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

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

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

For the ML estimator, we have (why?)

$$\hat{m}_{\text{mle},n} = \max_i y_i.$$

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

- As $n \to \infty$, by LLN, $\bar{Y}$ converges to the mean of the distribution, i.e., $\mathbb{E}[Y_1] = \frac{m^*+1}{2}$. Hence, $\hat{m}_{L,n} \to 2 \cdot \frac{m^*+1}{2} - 1 = m^*$.
- For the ML estimator, as $n \to \infty$, at some point, we will pick the ball numbered m^* and so we will eventually have $\hat{m}_{\text{mle}} = m^*$.

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

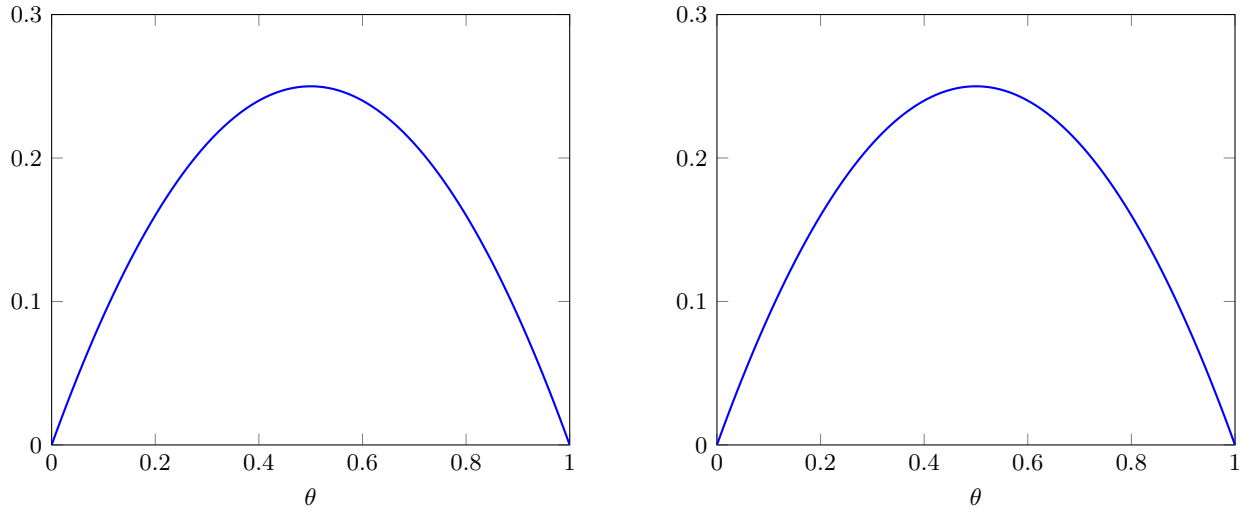


Figure 2.1: The likelihood function on the left demonstrates strong dependence on θ compared to the one on the right.

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 $\boldsymbol{X}$ depends on θ. If the dependence is strong, i.e., for values of θ other than the true value θ^*, 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 $\boldsymbol{X}$, whose distribution is given by p with parameter θ^*. Assuming $\boldsymbol{X} = \boldsymbol{x}$, the log-likelihood is $p(\boldsymbol{x};\theta)$. The sharpness of the log-likelihood $\ell(\theta)$ at the true value θ^* can be quantified as

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

Given the randomness of the data $\boldsymbol{X}$, the above quantity is random,

$$-\frac{\partial^2 \ln p(\boldsymbol{X};\theta)}{\partial \theta^2}\bigg|_{\theta=\theta^*}$$

So to average over the data, we define

$$I(\theta^*) = -\,\mathbb{E}\left[\frac{\partial^2 \ln p(\boldsymbol{X};\theta)}{\partial \theta^2}\bigg|_{\theta=\theta^*}\right] = -\int \frac{\partial^2 \ln p(\boldsymbol{x};\theta)}{\partial \theta^2}\bigg|_{\theta=\theta^*} p(\boldsymbol{x};\theta^*)d\boldsymbol{x},$$

which is called the **Fisher Information**.

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

Theorem 2.17 (CRLB). *Given that the log-likelihood $\ell(\theta)$ satisfies certain regularity conditions, 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., $\text{Var}(\hat{\theta}) = 1/I(\theta^*)$, 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

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

Example 2.18. In Example 2.2, where we estimated the mean μ^* 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,

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

Observe that

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

Based on the CRLB, the variance of the estimator satisfies

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

The variance of the estimator is $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$. 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.

References

[1] Bruce Hajek. *Random Processes for Engineers*. Illinois, 2014. URL: http://hajek.ece.illinois.edu/Papers/randomprocJuly14.pdf (visited on 01/30/2017).