2.3 Methods of Estimation

2.3.1 Method of Moments

The Method of Moments is a simple technique based on the idea that the sample moments are "natural" estimators of population moments.

The k*-th population moment* of a random variable Y is

$$\mu_k' = E(Y^k), \quad k = 1, 2, \ldots$$

and the k*-th sample moment* of a sample $Y_1, \ldots, Y_n$ is

$$m_k' = \frac{1}{n}\sum_{i=1}^{n} Y_i^k, \quad k = 1, 2, \ldots.$$

If $Y_1, \ldots, Y_n$ are assumed to be independent and identically distributed then the Method of Moments estimators of the distribution parameters $\vartheta_1, \ldots, \vartheta_p$ are obtained by solving the set of p equations:

$$\mu_k' = m_k', \quad k = 1, 2, \ldots, p.$$

Under fairly general conditions, Method of Moments estimators are asymptotically normal and asymptotically unbiased. However, they are not, in general, efficient.

Example 2.17. Let $Y_i \underset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. We will find the Method of Moments estimators of μ and σ^2.

We have $\mu_1' = E(Y) = \mu$, $\mu_2' = E(Y^2) = \sigma^2 + \mu^2$, $m_1' = \overline{Y}$ and $m_2' = \sum_{i=1}^{n} Y_i^2/n$. So, the Method of Moments estimators of μ and σ^2 satisfy the equations

$$\widehat{\mu} = \overline{Y}$$

$$\widehat{\sigma}^2 + \widehat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

Thus, we obtain

$$\widehat{\mu} = \overline{Y}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

□

Estimators obtained by the Method of Moments are not always unique.

Example 2.18. Let $Y_i \underset{iid}{\sim} \text{Poisson}(\lambda)$. We will find the Method of Moments estimator of λ. We know that for this distribution $\text{E}(Y_i) = \text{var}(Y_i) = \lambda$. Hence By comparing the first and second population and sample moments we get two different estimators of the same parameter,

$$\widehat{\lambda}_1 = \overline{Y}$$
$$\widehat{\lambda}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2.$$

□

Exercise 2.11. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\text{T}}$ be a random sample from the distribution with the pdf given by

$$f(y;\vartheta) = \begin{cases} \frac{2}{\vartheta^2}(\vartheta - y), & y \in [0, \vartheta], \\ 0, & \text{elsewhere.} \end{cases}$$

Find an estimator of ϑ using the Method of Moments.

2.3.2 Method of Maximum Likelihood

This method was introduced by R.A.Fisher and it is the most common method of constructing estimators. We will illustrate the method by the following simple example.

Example 2.19. Assume that $Y_i \underset{iid}{\sim} \text{Bernoulli}(p)$, $i = 1, 2, 3, 4$, with probability of success equal to p, where $p \in \Theta = \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\}$ i.e., p belongs to the parameter space of only three elements. We want to estimate parameter p based on observations of the random sample $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^{\text{T}}$.

The joint pmf is

$$P(\boldsymbol{Y} = \boldsymbol{y}; p) = \prod_{i=1}^{4} P(Y_i = y_i; p) = p^{\sum_{i=1}^{4} y_i}(1-p)^{4-\sum_{i=1}^{4} y_i}.$$

The different values of the joint pmf for all $p \in \Theta$ are given in the table below

p	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	$\sum_{i=1}^{4} y_i$
$P(\boldsymbol{Y}=\boldsymbol{y};p)$	$\frac{81}{256}$	$\frac{16}{256}$	$\frac{1}{256}$	0
	$\frac{27}{256}$	$\frac{16}{256}$	$\frac{3}{256}$	1
	$\frac{9}{256}$	$\frac{16}{256}$	$\frac{9}{256}$	2
	$\frac{3}{256}$	$\frac{16}{256}$	$\frac{27}{256}$	3
	$\frac{1}{256}$	$\frac{16}{256}$	$\frac{81}{256}$	4

We see that $P(\sum_{i=1}^{4} Y_i = 0)$ is largest when $p = \frac{1}{4}$. It can be interpreted that when the observed value of the random sample is $(0,0,0,0)^{\mathrm{T}}$ the most likely value of the parameter p is $\widehat{p} = \frac{1}{4}$. Then, this value can be considered as an estimate of p. Similarly, we can conclude that when the observed value of the random sample is, for example, $(0,1,1,0)^{\mathrm{T}}$, then the most likely value of the parameter is $\widehat{p} = \frac{1}{2}$. Altogether, we have

$\widehat{p} = \frac{1}{4}$ if we observe all failures or just one success;
$\widehat{p} = \frac{1}{2}$ if we observe two failures and two successes;
$\widehat{p} = \frac{3}{4}$ if we observe three successes and one failure or four successes.

Note that, for each point $(y_1, y_2, y_3, y_4)^{\mathrm{T}}$, the estimate $\widehat{p}$ is the value of parameter p for which the joint mass function, treated as a function of p, attains maximum (or its largest value).

Here, we treat the joint pmf as a function of parameter p for a given $\boldsymbol{y}$. Such a function is called the *likelihood function* and it is denoted by $L(p|\boldsymbol{y})$. □

Now we introduce a formal definition of the *Maximum Likelihood Estimator* (MLE).

Definition 2.11. *The* $\mathrm{MLE}(\boldsymbol{\vartheta})$ *is the statistic* $T(\boldsymbol{Y}) = \widehat{\boldsymbol{\vartheta}}$ *whose value for a given* $\boldsymbol{y}$ *satisfies the condition*

$$L(\widehat{\boldsymbol{\vartheta}}|\boldsymbol{y}) = \sup_{\boldsymbol{\vartheta}\in\Theta} L(\boldsymbol{\vartheta}|\boldsymbol{y}),$$

where $L(\boldsymbol{\vartheta}|\boldsymbol{y})$ *is the likelihood function for* $\boldsymbol{\vartheta}$.

Properties of MLE

The MLEs are invariant, that is

$$\mathrm{MLE}(g(\boldsymbol{\vartheta})) = g(\mathrm{MLE}(\boldsymbol{\vartheta})) = g(\widehat{\boldsymbol{\vartheta}}).$$

MLEs are asymptotically normal and asymptptically unbiased. Also, they are efficient, that is

$$\text{eff}(g(\widehat{\boldsymbol{\vartheta}})) = \lim_{n\to\infty} \frac{\text{CRLB}(g(\boldsymbol{\vartheta}))}{\text{var}\, g(\widehat{\boldsymbol{\vartheta}})} = 1.$$

In this case, for large n, $\text{var}\, g(\widehat{\boldsymbol{\vartheta}})$ is approximately equal to the CRLB. Therefore, for large n,

$$g(\widehat{\boldsymbol{\vartheta}}) \sim \mathcal{N}\left(g(\boldsymbol{\vartheta}), \text{CRLB}(g(\boldsymbol{\vartheta}))\right)$$

approximately. This is called the **asymptotic distribution** of $g(\widehat{\boldsymbol{\vartheta}})$.

Example 2.20. Suppose that $Y_1, \ldots, Y_n$ are independent $\text{Poisson}(\lambda)$ random variables. Then the likelihood is

$$L(\lambda|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}}{\prod_{i=1}^n y_i!}.$$

We need to find the value of λ which maximizes the likelihood. This value will also maximize $\ell(\lambda|\boldsymbol{y}) = \log L(\lambda|\boldsymbol{y})$, which is easier to work with. Now, we have

$$\ell(\lambda|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log\lambda - n\lambda - \sum_{i=1}^{n} \log(y_i!).$$

The value of λ which maximizes $\ell(\lambda|\boldsymbol{y})$ is the solution of $d\ell/d\lambda = 0$. Thus, solving the equation

$$\frac{d\ell}{d\lambda} = \frac{\sum_{i=1}^n y_i}{\lambda} - n = 0$$

yields the estimator $\hat{\lambda} = T(\boldsymbol{Y}) = \sum_{i=1}^n Y_i/n = \overline{Y}$, which is the same as the Method of Moments estimator. The second derivative is negative for all λ hence, $\widehat{\lambda}$ indeed maximizes the log-likelihood. □

Example 2.21. Suppose that $Y_1, \ldots, Y_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables. Then the likelihood is

$$\begin{aligned} L(\mu, \sigma^2|\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \end{aligned}$$

and so the log-likelihood is

$$\ell(\mu, \sigma^2|\boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Thus, we have

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

and

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{2\pi\sigma^2}2\pi + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Setting these equations to zero, we obtain

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0 \Rightarrow \sum_{i=1}^{n} y_i = n\hat{\mu},$$

so that $\hat{\mu} = \overline{Y}$ is the maximum likelihood estimator of μ, and

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(y_i - \overline{y})^2 = 0 \Rightarrow n\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

so that $\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2/n$ is the maximum likelihood estimator of σ^2, which are the same as the Method of Moments estimators. □

Exercise 2.12. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a random sample from a Gamma distribution, $\mathrm{Gamma}(\lambda, \alpha)$, with the following pdf

$$f(y; \lambda, \alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}, \quad \text{for } y > 0.$$

Assume that the parameter α is known.

(a) Identify the complete sufficient statistic for λ.

(b) Find the $\mathrm{MLE}[g(\lambda)]$, where $g(\lambda) = \frac{1}{\lambda}$. Is the estimator a function of the complete sufficient statistic?

(c) Knowing that $\mathrm{E}(Y_i) = \alpha\frac{1}{\lambda}$ for all $i = 1, \ldots, n$, check that the $\mathrm{MLE}[g(\lambda)]$ is an unbiased estimator of $g(\lambda)$. What can you conclude about the properties of the estimator?

2.3.3 Method of Least Squares

If $Y_1, \ldots, Y_n$ are independent random variables, which have the same variance and higher-order moments, and, if each $E(Y_i)$ is a linear function of $\vartheta_1, \ldots, \vartheta_p$, then the Least Squares estimates of $\vartheta_1, \ldots, \vartheta_p$ are obtained by minimizing

$$S(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \{Y_i - E(Y_i)\}^2.$$

The Least Squares estimator of ϑ_j has minimum variance amongst all **linear** unbiased estimators of ϑ_j and is known as the **best linear unbiased estimator** (BLUE). If the Y_is have a normal distribution, then the Least Squares estimator of ϑ_j is the Maximum Likelihood estimator, has a normal distribution and is the MVUE.

Example 2.22. Suppose that $Y_1, \ldots, Y_{n_1}$ are independent $\mathcal{N}(\mu_1, \sigma^2)$ random variables and that $Y_{n_1+1}, \ldots, Y_n$ are independent $\mathcal{N}(\mu_2, \sigma^2)$ random variables. Find the least squares estimators of μ_1 and μ_2.

Since

$$E(Y_i) = \begin{cases} \mu_1, & i = 1, \ldots, n_1, \\ \mu_2, & i = n_1 + 1, \ldots, n, \end{cases}$$

it is a linear function of μ_1 and μ_2. The Least Squares estimators are obtained by minimizing

$$S = \sum_{i=1}^{n} \{Y_i - E(Y_i)\}^2 = \sum_{i=1}^{n_1} (Y_i - \mu_1)^2 + \sum_{i=n_1+1}^{n} (Y_i - \mu_2)^2.$$

Now,

$$\frac{\partial S}{\partial \mu_1} = -2 \sum_{i=1}^{n_1} (Y_i - \mu_1) = 0 \Rightarrow \widehat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i = \overline{Y}_1$$

and

$$\frac{\partial S}{\partial \mu_2} = -2 \sum_{i=n_1+1}^{n} (Y_i - \mu_2) = 0 \Rightarrow \widehat{\mu}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n} Y_i = \overline{Y}_2,$$

where $n_2 = n - n_1$. So, we estimate the mean of each group in the population by the mean of the corresponding sample. □

Example 2.23. Suppose that $Y_i \sim \mathcal{N}(\beta_0+\beta_1 x_i, \sigma^2)$ independently for $i = 1, 2, \ldots, n$, where x_i is some **explanatory variable**. This is called the simple linear regression model. Find the least squares estimators of β_0 and β_1.

Since $E(Y_i) = \beta_0 + \beta_1 x_i$, it is a linear function of β_0 and β_1. So we can obtain the least squares estimates by minimizing

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Now,

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} Y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \Rightarrow \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

and

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(Y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i Y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Substituting the first equation into the second one, we have

$$\sum_{i=1}^{n} x_i Y_i - (\overline{y} - \widehat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \Rightarrow \left(n\overline{x}^2 - \sum_{i=1}^{n} x_i^2 \right) \widehat{\beta}_1 = n\overline{xy} - \sum_{i=1}^{n} x_i Y_i.$$

Hence, we have the estimators

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x} \quad \text{and} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

These are the Least Squares estimators of the **regression coefficients** β_0 and β_1.

□

Exercise 2.13. Given data, $(x_1, y_1), \ldots, (x_n, y_n)$, assume that $Y_i \sim \mathcal{N}(\beta_0+\beta_1 x_i, \sigma^2)$ independently for $i = 1, 2, \ldots, n$ and that σ^2 is known.

(a) Show that the Maximum Likelihood Estimators of β_0 and β_1 must be the same as the Least Squares Estimators of these parameters.

(b) The quench bath temperature in a heat treatment operation was thought to affect the Rockwell hardness of a certain coil spring. An experiment was run in which several springs were treated under four different temperatures. The table gives values of the set temperatures (x) and the observed hardness (y, coded). Assuming that hardness depends on temperature linearly and the variance of the r.vs is constant we may write the following model:

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i, \quad \mathrm{var}(Y_i) = \sigma^2.$$

Calculate the LS estimates of β_0 and of β_1. What is the estimate of the expected hardness given the temperature $x = 40$?

Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x	30	30	30	30	40	40	40	50	50	50	60	60	60	60
y	55.8	59.1	54.8	54.6	43.1	42.2	45.2	31.6	30.9	30.8	17.5	20.5	17.2	16.9

In this section we were considering so called point estimators. We used various methods, such as the Method of Moments, Maximum Likelihood or Least Squares, to derive them. We may also construct the estimators using the Rao-Blackwell Theorem. The estimators are functions

$$T(Y_1, \ldots, Y_n) \to \Theta,$$

that is, their values belong to the parameter space. However, the values vary with the observed sample. If the estimator is MVUE we may expect that "on average" the calculated estimates are very close to the true parameter and also that the variability of the estimates is the smallest possible.

Sometimes it is more appropriate to construct an interval which covers the unknown parameter with high probability and whose limits depend on the sample. We introduce such intervals in the next section. The point estimators are used in constructing the intervals.