Inferential Statistics **First semester, 2021**

Lecture 4: Point estimation

Instructor: Erlis Ruli (ruli@stat.unipd.it), Department of Statistical Sciences

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Simplifying things to the extreme, the problem of statistical inference can be stated as follows. Assume that we are given a random sample $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \dots, n$, with θ being unknown. The implied statistical model is

$$\mathcal{M} = \{F_{Y_1,\dots,Y_n}(y_1, \dots, y_n; \theta) : \theta \in \Theta\}\,.$$

Nature secretly picks $\theta = \theta_0$, i.e. it picks the model $F_{\theta_0} \in \mathcal{M}$, and uses it to generate the observed sample $y_1, \dots, y_n$. Given this observed sample, our aim is to learn (and verify hypotheses) about θ_0.

For instance, F_θ could be the Exp(λ) and Nature's choice could be Exp(1). Given $\mathcal{M}$ and the observed sample, we are asked to guess λ_0. That is, we are required to produce a guess for θ_0 which is theoretically guaranteed to be as close as possible to θ_0. We have already seen in Lecture 3 how to produce a possible guess (we denoted it by $\widehat{\theta}$) using the likelihood function.

More formally, such a guess or value for θ_0 is called *estimate* and the tool that produces such an estimate is called *estimator.* Estimation theory deals with methods for building estimators.

4.1 Statistics

Definition 4.1 *Let $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \dots, n$, and consider a function*

$$T_n = T(Y_1, \dots, Y_n)\,,$$

with $T_n : \mathbb{R}^n \to \mathbb{R}^d$ not depending on other unknown quantities. T_n is called a sample statistic or simply a statistic.

A statistic is a function applied to the random sample $Y_1, \dots, Y_n$, thus (by Lectures 0 and 1) it is an r.v. if $d = 1$, and it is an r.ve. if $d > 1$. The probability distribution of a statistic is called *sampling distribution.*

The statistic T_n evaluated at the observed sample is denoted by

$$t_n = T(y_1, \dots, y_n),$$

and is called *observed statistic.* An observed statistic is thus a real number if $d = 1$ or a vector of reals if

$d > 1$.

Examples of statistics are the sample mean $\overline{Y}$, the sample variance S^2_Y, the sample median Q_2, the sample maximum $Y_{(n)}$, etc., the empirical distribution function F_n is also a (functional) statistic. Furthermore, $T_n = (T_{1n}, T_{2n}) = (\overline{Y}, S^2_Y)$ is a bivariate statistic. Examples of observed statistics are the observed counterparts of the aforementioned statistics; the histogram is also an observed statistics.

The probability distribution of the a statistic T_n depends on F_θ but also on the function $T(\cdot)$. In some cases it can be determined analytically. Here are some notable results.

Theorem 4.1 *Let* $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$, *with* $E(Y_i) = \mu < \infty$ *and* $\mathrm{var}(Y_i) = \sigma^2 < \infty$. *Then*

(i) $E(\overline{Y}) = \mu$ *and* $\mathrm{var}(\overline{Y}) = \frac{\sigma^2}{n}$.

(ii) $E(S^2_Y) = \sigma^2$ *and* $\mathrm{var}(S^2_Y) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$, *where* $\mu_k = E(Y_1^k)$, *is the kth moment of* Y_1.

Theorem 4.1, part (i), thus tells us that whatever is the distribution of the random sample, assuming each component of the latter has finite mean μ and finite variance σ^2, the sample mean has expected value equal to μ and variance equal to σ^2 divided by n. A similar result is given also for the sample variance.

To see why (i) holds note that

$$\begin{aligned}
E(\overline{Y}) &= E\left(\frac{\sum_{i=1}^n Y_i}{n}\right) \\
&= \frac{1}{n} E(Y_1 + \ldots + Y_n) \\
&= \frac{1}{n}(E(Y_1) + \ldots + E(Y_n)) \\
&= \frac{1}{n}(\mu + \ldots + \mu) = \mu.
\end{aligned}$$

Furthermore

$$\begin{aligned}
\mathrm{var}(\overline{Y}) &= E\left((\overline{Y} - E(\overline{Y}))^2\right) = E\left(\left(\frac{\sum_{i=1}^n Y_i}{n} - \mu\right)^2\right) = \frac{1}{n^2} E\left(\sum_{i=1}^n (Y_i - \mu) \sum_{i=1}^n (Y_i - \mu)\right) \\
&= \frac{1}{n^2} E\left(\sum_{i=1}^n (Y_i - \mu)^2 + 2 \sum_{i \leq j} (Y_i - \mu)(Y_j - \mu)\right) \\
&= \frac{1}{n^2} \sum_i E((Y_i - \mu)^2) + \frac{2}{n^2} \sum_{i \leq j} E(Y_i - \mu)(Y_j - \mu)\} \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.
\end{aligned}$$

where we have used the fact that $E(Y_i - \mu)(Y_j - \mu) = \mathrm{cov}(Y_i, Y_j) = 0$ since Y_i and Y_j are independent (by Theorem 0.9(i), Lecture 0). Part (ii), and in particular the second result is more tedious to derive.

Remark 4.1

(i) *In general $\overline{Y}$ and S_Y^2 are not stochastically independent. However, as we will see in the next theorem, if the random sample is taken from the normal distribution, then these two statistics are independent.*

(ii) *Since $\overline{Y}$ is a sum of r.v. with finite mean and variance, then by the Central Limit Theorem (CLT) we have that*

$$\tfrac{\sqrt{n}(\overline{Y}-\mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1), \quad \text{as } n \to \infty.$$

The next theorem is similar except that the random sample is assumed to be normally distributed.

Theorem 4.2 *Let $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$ then*

(i) $\overline{Y} \sim N(\mu, \sigma^2/n)$.

(ii) $\frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2_{n-1}$.

(iii) *$\overline{Y}$ is independent from S_Y^2.*

(iv) $\frac{\sqrt{n}(\overline{Y}-\mu)/\sigma}{\sqrt{\frac{\frac{(n-1)S_Y^2}{\sigma^2}}{n-1}}} = \frac{\sqrt{n}(\overline{Y}-\mu)}{\sqrt{S_Y^2}} \sim t_{n-1}$.

(v) $\mathrm{var}(S_Y^2) = 2\frac{\sigma^4}{n-1}$.

Theorem 4.2(i) tells us that if the random sample has a normal distribution, then also the sample mean has exactly a normal distribution. Furthermore, the point (ii) tells us that the the "scaled" sample variance has a chi-square distribution with $n-1$ degrees of freedom and point (iii) tells us that, contrary to the general case, under the normal distribution, $\overline{Y}$ is independent from S_Y^2. Point (iv) says that the statistic $\frac{\sqrt{n}(\overline{Y}-\mu)}{\sqrt{S_Y^2}}$, follows the t-Student distribution with $n-1$ degrees of freedom. These results are the building blocks of many confidence intervals and hypothesis testing procedures as we will se in the incoming lectures. Other notable results about statistics and their distributions will be discussed in the incoming lectures.

4.2 Properties of estimators

An *estimator* is a sample statistic which aim is to provide estimates of an unknown parameter; thus if T_n is a statistics, then it is an estimator. On the other hand, the statistic evaluated at the observed sample, i.e $t_n = T(y_1, \ldots, y_n)$, is a number, thus it is an *estimate*. To simplify notation and keep the two things separately, we denote the estimator by $\widehat{\theta}_n = T(Y_1, \ldots, Y_n)$, instead of T_n. We adopt the – usual and perhaps confusing – convention in which $\widehat{\theta}_n$ denotes also the estimate. Thus, $\widehat{\theta}_n$ refers simultaneously to two different objects: estimator and estimate. This is only apparently confusing since in a given context, it will be clear to which object the notation is referring to. Sometimes we omit n and write simply $\widehat{\theta}$.

While all estimators are statistics, not all estimators are useful for a problem at hand and some estimators may be better than others. Furthermore, for a given problem there may be many estimators available and we need to have criteria for choosing the best among them. So we begin by presenting some criteria and then take up some ways of deriving estimators that are apt to be good.

4.2.1 Sufficiency

Roughly speaking, the random sample $Y_1, \ldots, Y_n$ which is assumed to be generated from the model F_θ, contains useful information about the unknown parameter θ but some information is redundant. A statistic, and thus an estimator, transforms the random sample in a random vector of lower dimension and such a transformation may lead to a loss of information about θ. Such a loss of information however will not happen if the the statistic is *sufficient*.

The statistic T_n is sufficient for θ, the parameter of the distribution F_θ, if and only if the conditional distribution of the random sample $Y_1, \ldots, Y_n$ given the value of T_n does not depend on θ, i.e. if the p.d.f.

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n | T_n) \text{ does not depend on } \theta.$$

A useful criterion for checking that a given statistic T_n is sufficient is based on the likelihood function. Indeed, T_n is sufficient if and only if the likelihood function is of the form

$$L(\theta) \propto g(T(y_1, \ldots, y_n); \theta),$$

where $g(\cdot)$ is any positive real-valued function. Here are some examples.

Example 4.1 *Let $Y_1, \ldots, Y_n$ be a random sample with $Y_i \overset{\text{iid}}{\sim}$ Geo(θ), where Geo(θ) is the geometric r.v. (see L0). The joint distribution of the random sample is then*

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} \theta(1-\theta)^{y_i} = \theta^n (1-\theta)^{\sum_{i=1}^{n} y_i}.$$

For a fixed observed sample, this is the likelihood function and it depends on $y_1, \ldots, y_n$ only through the value of their sum. So $\sum_{i=1}^{n} y_i$ is a sufficient statistic for θ.

Bijective functions of a sufficient statistics are sufficient. For instance, Example 4.1, $\overline{Y}$ is also a sufficient statistic.

Example 4.2 *Let $Y_i \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, where $\theta > 0$. The joint distribution of the random sample is then*

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} \theta^{-1} 1_{(0,\theta)}(y_i) = \theta^{-n} 1_{(0,\theta)}(y_{(n)}).$$

For a fixed observed sample, this is the likelihood function and it depends on $y_1, \ldots, y_n$ only through its maximum value $y_{(n)}$. The later is thus is a sufficient statistic for θ.

4.2.2 Unbiasedness

Let $\widehat{\theta}_n$ be an estimator for θ, based on the random sample $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$. The bias of an estimator is defined by

$$\text{bias}(\theta; \widehat{\theta}_n) = E(\widehat{\theta}_n) - \theta.$$

We say that $\widehat{\theta}_n$ is *unbiased* if $E(\widehat{\theta}_n) = \theta$ or equivalently if $\text{bias}(\theta; \widehat{\theta}_n) = 0$. Unbiasedness used to receive much attention but these days is considered less important; many interesting estimators used in practice are biased. An estimator with vanishing bias as the sample size increases, i.e. an estimator for which $\text{bias}(\theta; \widehat{\theta}_n) \to 0$ as $n \to \infty$, is called *asymptotically unbiased.*

4.2.3 Efficiency

In principle, the lower the bias, the better is the estimator. However, unbiasedness alone is not enough for judging the performance of an estimator. Indeed it is possible to build unbiased estimators which are useless; see below for an example. Another reason is that unbiasedness tells us nothing about the variability of $\widehat{\theta}_n$ with respect to the true parameter value θ. For this reason, the *mean squared error* (MSE) is a more sound criterion for judging the performance of the estimator. The MSE is defined by

$$\text{MSE}(\theta; \widehat{\theta}_n) = E(\widehat{\theta}_n - \theta)^2 = \text{var}(\widehat{\theta}_n) + (\text{bias}(\theta; \widehat{\theta}_n))^2.$$

Thus if the estimator $\widehat{\theta}_n$ is unbiased, $\text{MSE}(\theta; \widehat{\theta}_n)$ is just the variance of the estimator.

If $\widehat{\theta}_{n1}$ and $\widehat{\theta}_{n2}$ are two competing estimators for θ and $\text{MSE}(\theta; \widehat{\theta}_{n1}) < \text{MSE}(\theta; \widehat{\theta}_{n2})$ then $\widehat{\theta}_{n1}$ is said to be *relatively more efficient* than $\widehat{\theta}_{n2}$. The *relative efficiency* of $\widehat{\theta}_{n1}$ with respect to $\widehat{\theta}_{n2}$ is defined by the ratio

$$\text{eff}(\theta; \widehat{\theta}_{n1}, \widehat{\theta}_{n2}) = \tfrac{\text{MSE}(\theta; \widehat{\theta}_{n2})}{\text{MSE}(\theta; \widehat{\theta}_{n1})}.$$

Thus, if $\text{eff}(\theta; \widehat{\theta}_{n1}, \widehat{\theta}_{n2}) > 1$ ($\text{eff}(\theta; \widehat{\theta}_{n1}, \widehat{\theta}_{n2}) < 1$), θ_1 is relatively more (less) efficient than θ_2 and instead of choosing estimators with lowest or zero bias, we prefer to select estimators with lowest MSE or with highest relative efficiency. Here is an example.

Example 4.3 *By Theorem 4.1(i) we know that $E(\overline{Y}) = \mu$, thus the sample mean is an unbiased estimator of the true mean. Let us denote by $\theta = (\mu, \sigma^2)$. It follows that $\text{MSE}(\theta; \overline{Y}) = \text{var}(\overline{Y}) = \sigma^2/n$. Consider the alternative statistic $T_n = Y_1$ the first observation in the sample is also unbiased, because $E(Y_1) = \mu$. However, its mean squared error is $\text{var}(Y_1) = \sigma^2$, which is larger than that of $\overline{Y}$ when $n > 1$.*

Remark 4.2 *Just like the bias also the MSE, and thus also the efficiency, may depend on θ. Indeed, an estimator that has lowest MSE for a certain unknown parameter value, may not have lowest MSE for all possible parameter values. For instance, Figure 4.1 shows the MSE as a function of θ for two hypothetical estimators. From this figure we can conclude that for $\theta < 0$, Estimator 1 is better, i.e. it is more efficient, than Estimator 2; for $\theta > 0$ Estimator 2 is better than Estimator 1, and for $\theta = 0$ the two estimators are equally good, at least in terms of efficiency.*

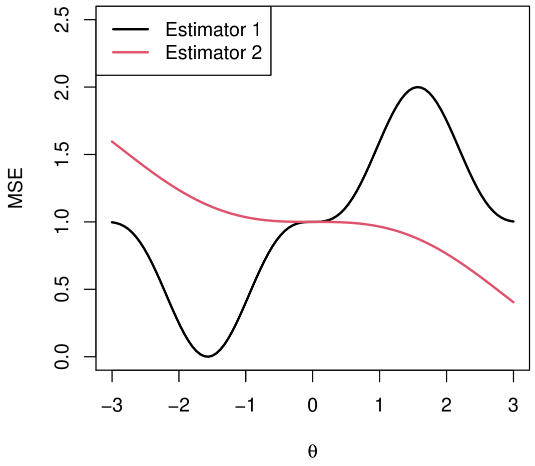

Figure 4.1: MSE as a function of θ for two hypothetical estimators.

Given a class of estimators we can always choose the one with highest efficiency, but there may be estimators outside the given class which could be more efficient. The question thus is:

is there a lower bound for the variance of an estimator ? That is, is there a *most efficient* estimator among all estimators?

In this generality the answer is negative; but in the rather wide (but still restricted) class of unbiased estimators, there is a lower bound of variances – a "best" variance – which can sometimes be achieved. This is stated formally in the next theorem.

Theorem 4.3 (Cramér-Rao's inequality) *Consider the random sample $Y_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$, and let $f(y;\theta)$ be the p.d.f. of Y_i for all i. Furthermore, let $\widehat{\theta}_n = T(Y_1, \ldots, Y_n)$ be an unbiased estimator for θ. Then*

$$\mathrm{var}(\widehat{\theta}_n) \geq I_n(\theta)^{-1},$$

where $I_n(\theta) = nE\left[\left(\frac{\mathrm{d}\log f(Y_1;\theta)}{\mathrm{d}\theta}\right)^2\right]$.

If there exists an estimator $\widehat{\theta}_n$ such that $\text{var}(\widehat{\theta}_n) = I_n(\theta)^{-1}$, then $\widehat{\theta}_n$ is called *efficient*. For such an unbiased estimator we define the efficiency by

$$\text{eff}(\theta; \widehat{\theta}_n) = (\text{var}(\widehat{\theta}_n) I_n(\theta))^{-1}.$$

This measure of efficiency is used even when $\widehat{\theta}_n$ is asymptotically unbiased.

4.2.4 Consistency

The essence of the notion of consistency of an estimator is that when the estimator is applied to the whole population as the sample, it produces the true value of the parameter being estimated. More formally, if $\widehat{\theta}_n$ is an estimator for θ defined for every sample size n, then $\widehat{\theta}_n$ is *consistent* if for any $\epsilon > 0$

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| < \epsilon) = 1,$$

or more compactly if $\widehat{\theta}_n \xrightarrow{P} \theta$. Here is an example.

Example 4.4 *Let $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, 1)$, $i = 1, \ldots, n$. Let $\widehat{\mu}_n = \overline{Y}$ be an estimator for μ. Since $\sigma^2 = 1$, $\overline{Y} \sim \text{N}(\mu, 1/n)$, so for any $\epsilon > 0$,*

$$\begin{aligned} \lim_{n\to\infty} P(|\widehat{\mu}_n - \mu| < \epsilon) &= \lim_{n\to\infty} (\Phi(\sqrt{n}\epsilon) - \Phi(-\sqrt{n}\epsilon)) \\ &= \Phi(\infty) - \Phi(-\infty) = 1 - 0 = 1. \end{aligned}$$

Thus $\widehat{\mu}_n = \overline{Y}$ is consistent. Here we used the fact that $\lim_{n\to\infty} \Phi(\sqrt{n}\epsilon) = 1$ since $\Phi(\cdot)$ is a d.f. (see Theorem 0.5(iii), Lecture 0).

The following result connects bias and efficiency with consistency and provides an easier way for checking the consistency of an estimator.

Theorem 4.4 *For an estimator $\widehat{\theta}_n$, if $\lim_{n\to\infty} \text{bias}(\theta; \widehat{\theta}_n) = 0$ and $\lim_{n\to\infty} \text{var}(\widehat{\theta}_n) = 0$, then $\widehat{\theta}_n$ is consistent.*

Before closing this section we give some further examples.

Example 4.5 *Let $Y_1, \ldots, Y_n$ be an i.i.d random sample with $E(Y_i) = \mu$ and $\text{var}(Y_i) = \sigma^2$, $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$, $i = 1, \ldots, n$. We wish to estimate μ, while σ^2 is known.*

Since $E(\overline{Y}) = \mu$ and $\text{var}(\overline{Y}) = \sigma^2/n$, the estimator $\widehat{\mu}_n = \overline{Y}$ for μ is consistent, for any value of σ^2.

By the CLT we know that for large n, $\overline{Y} \xrightarrow{d} \mathrm{N}(\mu, \sigma^2/n)$. Thus for large n we can use $\mathrm{N}(\mu, \sigma^2/n)$ as an approximation to the sampling distribution of $\overline{Y}$. Nevertheless, If $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then $\overline{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$ and $\overline{Y}$ is efficient for μ (check!).

Example 4.6 *Let $Y_1, \ldots, Y_n$ be an i.i.d random sample with $E(Y_i) = \mu$ and $\mathrm{var}(Y_i) = \sigma^2$, $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$, $i = 1, \ldots, n$. We wish to estimate σ^2, while μ is known.*

Since we know the mean of Y_i is μ, then a reasonable estimator for σ^2 is $\widehat{\sigma^2}_\mu = \frac{\sum_{i=1}^n (Y_i - \mu)^2}{n}$. It can be proved that $E\left(\widehat{\sigma^2}_\mu\right) = \sigma^2$ and that $\widehat{\sigma^2}_\mu$ is consistent for σ^2.

If $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then $\frac{n\widehat{\sigma^2}_\mu}{\sigma^2} \sim \chi^2_n$.

Example 4.7 *Let $Y_1, \ldots, Y_n$ be an i.i.d random sample with $E(Y_i) = \mu$ and $\mathrm{var}(Y_i) = \sigma^2$, $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$, $i = 1, \ldots, n$. We wish to estimate σ^2 but this time μ is unknown. In this case, μ is also called nuisance parameter, while σ^2 is the parameter of interest.*

Since μ is unknown, we cannot use $\widehat{\sigma^2}_\mu$ as an estimator for σ^2. However, we can use the sample variance S^2. We have already proved that $E(S^2) = \sigma^2$ (see Theorem 4.1) so S^2 is an unbiased estimator for σ^2. Furthermore, from Theorem 4.1 (ii) we have that $\lim_{n\to\infty} \mathrm{var}(S^2) = 0$, so S^2 is consistent.

Recall from Theorem 4.2, that if $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Example 4.8 *Let $Y_i \overset{\text{iid}}{\sim} \mathrm{Ber}(\theta)$, $i = 1, \ldots, n$ where the success probability θ is unknown. Note that $\theta = E(Y_i)$, thus for estimating θ we can use the same estimator as in Example 4.5. Let $\widehat{\theta}_n = \overline{Y}$, where*

$$\overline{Y} = \tfrac{1}{n} \sum_{i=1}^n Y_i = \tfrac{1}{n} \sum_{i=1}^n 1_{Y_i = 1} \,,$$

with $1_{Y_i=1}$ being 1 if $Y_i = 1$ and 0 otherwise. The estimator $\widehat{\theta}_n$ can also be interpreted as the proportion of Y_i's that are equal to 1, i.e. the sample proportion of successes or simply the sample proportion. By the same results of Example 4.5 we can conclude that $\widehat{\theta}_n$ is consistent. Furthermore, in this case $\mathrm{var}(\widehat{\theta}_n) = I_n(\theta)^{-1}$, so $\widehat{\theta}$ is efficient.

The sampling distribution of $\widehat{\theta}_n$ can be determined exactly. Indeed if we let $\tilde{Y} = n\widehat{\theta}_n$,

$$\tilde{Y} \sim \mathrm{Bin}(n, \theta),$$

since $\tilde{Y} = \sum_{i=1}^n Y_i$. Otherwise, if n is high enough, we can appeal to the CLT to get the approximate sampling distribution

$$\widehat{\theta} \mathrel{\dot\sim} \mathrm{N}\left(\theta, \tfrac{\theta(1-\theta)}{n}\right).$$

Example 4.9 *Let $Y_i \overset{\text{iid}}{\sim} \mathrm{Unif}(0, \theta)$, $i = 1, \ldots, n$. Since $EY_i = \frac{\theta}{2}$, it makes sense to consider as an estimator* *for θ the statistic $2\overline{Y}$. Indeed $2\overline{Y}$ is unbiased because, for all $\theta > 0$,*

$$2\overline{Y} = 2E\left(\tfrac{1}{n}\sum_{i=1}^{n} Y_i\right) = \tfrac{2}{n}\sum_{i=1}^{n} E(Y_i) = \tfrac{2}{n}\sum_{i=1}^{n} \tfrac{\theta}{2} = \theta\,.$$

The MSE for this estimator is

$$\mathrm{MSE}(\theta;\widehat{\theta}_n) = \tfrac{\theta^2}{3n},$$

which converges to 0 *as $n \to \infty$.*

Compare this estimator with the (biased estimator) $Y_{(n)}$ for which

$$\mathrm{MSE}(\theta;Y_{(n)}) = \mathrm{var}(Y_{(n)}) + \mathrm{bias}(\theta;Y_{(n)}) = \tfrac{2\theta^2}{(n+1)(n+1)}.$$

4.3 Methods for building estimators

So far we studied the performance of estimators that have suggested themselves on intuitive grounds. If intuition does not suggest an estimator (and even when it does), it helps to have methods for deriving estimators. We illustrate three of them.

4.3.1 Method of moments

To estimate a single function of parameters, the method of moments is to use that same function of the corresponding sample moments. In particular, the mean of a distribution is estimated as the sample mean and a population variance is estimated as $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$. To estimate k parameters, the method of moments is first to express those parameters in terms of the first k population moments and then to replace those population moments with the corresponding sample moments.

Example 4.10 *The variance of an r.v. Y, assuming it exists, is $\sigma^2 = E(Y^2) - (E(Y))^2$. If $Y_1, \ldots, Y_n$ is an i.i.d. random sample, the method of moments is to replace $E(Y^2)$ by its sample version $\overline{Y^2}$ and $(E(Y))^2$ by its sample version $(\overline{Y})^2$. With this replacements we get $(n-1)S^2/n = \overline{Y^2} - (\overline{Y})^2$, which is the method of moments estimator for σ^2.*

Example 4.11 *Given the random sample $Y_i \overset{\text{iid}}{\sim} \mathrm{Ga}(\alpha,\lambda)$, $i = 1, \ldots, n$ let's estimate the parameters α and λ. For this, recall that for $Y \sim \mathrm{Ga}(\alpha,\lambda)$, the expectation is $E(Y) = \alpha/\lambda$, and the variance is $\mathrm{var}(Y) = \alpha/\lambda^2 = E(Y^2) - [E(Y)]^2$. To get the method of moments estimators we replace $E(Y)$ by $\overline{Y}$ and $E(Y^2)$ by $\overline{Y^2} = \sum_{i=1}^{n} Y_i^2$ to get*

$$\widehat{\lambda}_{\mathrm{MM}} = \tfrac{n\overline{Y}}{(n-1)S^2}, \quad \widehat{\alpha}_{\mathrm{MM}} = \tfrac{n(\overline{Y})^2}{(n-1)S^2}.$$

The method of moments gives estimators that are consistent and asymptotically normal (thanks to the CLT) but typically not efficient. However, the usefulness of this method stems from the fact that often the estimators are easy to calculate and can thus be used as starting points for better estimators such as those obtained by the method of maximum likelihood.

4.3.2 Method of least squares

In some situations the random sample $Y_1, \ldots, Y_n$ may be expressed as a "signal plus noise" relation, such as

$$Y_i = g_i(\theta) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the signal $g_i(\theta)$ is a known deterministic function up the the unknown parameter θ and ϵ_i is the noise, which is a random component. For instance, Y_i could be the measurement of some physical quantity which is believed to be equal to $g_i(\theta)$ plus some noise due to accidental errors or because of pure random fluctuations of the phenomenon under study; this noise is represented by the r.v. ϵ_i. It is reasonable to assume that for a large number of measurements, the negative fluctuations compensates the positive ones, thus $E(\epsilon_i) = 0$. Furthermore, it is reasonable to assume that the random fluctuations possess a variance, e.g. $\text{var}(\epsilon_i) = \sigma^2$ and the fluctuations are independent, all other things held equal, i.e. $\text{cov}(\epsilon_i, \epsilon_j) = 0$, $i, j = 1, \ldots, n$, $i \neq j$.

The method of *least squares* (LS) consists in estimating θ through the estimator

$$\widehat{\theta}_{\text{LS}} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - g_i(\theta))^2.$$

Here it is an example.

Example 4.12 *Suppose $Y_1, \ldots, Y_n$ are bacterial counts measured at time points $t_1, \ldots, t_n$ in a culture of cells and we aim at studying their growth rate in time. A possible model for this problem is the following linear multiple regression model*

$$\begin{aligned} Y_i &= g(t_i; \theta) + \epsilon_i, \\ g(t_i; \theta) &= \theta_1 + \theta_2 t_i + \theta_3 t_i^2, \quad i = 1, \ldots, n, \end{aligned}$$

where $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ is unknown. The solution in θ to the system of equations

$$\frac{d}{d\theta} \sum_{i=1}^{n} (Y_i - g_i(\theta))^2 = 0,$$

is the LS estimator for θ. If we let $Y = (Y_1, \ldots, Y_n)$ and $X = [1_n \,|\, T \,|\, T^2]$ be an $n \times 3$ matrix, where $1_n = (1, \ldots, 1)$ is the all-ones vector of dimension n, $T = (t_1, \ldots, t_n)$ and $T^2 = (t_1^2, \ldots, t_n^2)$, then the LS *estimator can be defined compactly by*

$$\widehat{\theta}_{\text{LS}} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y.$$

Under suitable conditions the LS estimator is unbiased, asymptotically efficient and asymptotically normally distributed. However, it use is limited only to those problems which can be stated as a signal plus noise relation.

4.3.3 Method of maximum likelihood

In Lecture 3, we defined $\widehat{\theta}_n \in \Theta$ as the point at which the likelihood function achieves its highest value. More formally, the point

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} L(\theta),$$

is called a *maximum likelihood estimate* (MLE) of θ. Note that for a random sample $Y_1, \ldots, Y_n$, $\widehat{\theta}_n$ is a r.v. or a r.ve., depending on the dimension of Θ. Furthermore, $\widehat{\theta}_n$ does not depend on θ, i.e. it is a statistic, so in random samples it is an estimator. In the random sample case, $\widehat{\theta}_n$ is called a *maximum likelihood estimator* (MLE). Use of the symbol $\widehat{\theta}_n$ for both observed sample and random sample cases is unfortunate but widespread, so we adhere to this use convention.

In practical applications it is often easier to work with the natural logarithm of the likelihood function $L(\theta)$, i.e. the log-likelihood function $\ell(\theta)$, if it is applicable. We illustrate the method of maximum likelihood by means of examples.

Example 4.13 *Consider Example 3.2 in Lecture 3. The log-likelihood function is*

$$\ell(\lambda) = -n\lambda + \sum_{i=1}^{n} y_i \log(\lambda) + \text{const.}.$$

From the first-order condition $d\ell(\lambda)/d\lambda = 0$ we get solution $\widehat{\lambda} = \overline{y}$. Furthermore, we have that $d^2\ell(\lambda)/d\lambda^2 < 0$ for all $\lambda > 0$, when at least one $y_i > 0$, thus $\widehat{\lambda}$ is actually a global maximum thus $\widehat{\lambda} = \overline{y}$ is the MLE of λ.

Example 4.14 *Consider Example 3.4 in Lecture 3. The log-likelihood function is*

$$\ell(\mu, \sigma^2) = -\tfrac{n}{2}\log(\sigma^2) + \tfrac{n\mu^2}{2\sigma^2} - \tfrac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \overline{y}\tfrac{n\mu}{\sigma^2} + \text{const.}$$

By solving the first order conditions we find $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma^2})$, where $\widehat{\mu} = \overline{y}$ and $\widehat{\sigma^2} = \sum_{i=1}^{n}(y_i - \overline{y})^2/n$. Furthermore, after some algebra, it can be shown that the observed information matrix $J(\mu, \sigma^2)$ at $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma}^2)$ is positive definite, thus $\widehat{\theta}$ is at least a local maximum.

Actually, in this case, $\widehat{\theta}$ is a global maximum. In general multi-parameter problems, showing that $\widehat{\theta}$ is a global maximum could be a formidable task, if possible at all, but in this case it is immediate. First note that for any $\sigma^2 > 0$,

$$\tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (y_i-\overline{y})^2/\sigma^2} \geq \tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (y_i-\mu)^2/\sigma^2}, \quad \textit{for any } \mu \in \mathbb{R}.$$

Thus to check that $\widehat{\theta}$ is a global maximum it suffices to check that $\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (y_i-\overline{y})^2/\sigma^2}$ achieves its global maximum at $\sigma^2 = n^{-1}\sum_{i=1}^n (y_i - \overline{y})^2$. This can be done by straightforward univariate calculus.

The estimators $\widehat{\mu} = \overline{Y}$ and $\widehat{\sigma^2} = n^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$ are the MLE's of μ and σ^2, respectively; alternatively we can say that $\widehat{\theta} = \left(\overline{Y}, n^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2\right)$ is the MLE of $\theta = (\mu, \sigma^2)$.

Example 4.15 *Suppose that $Y_1, \ldots, Y_m$, the counts of the m possibilities or cells $b_1, b_2, \ldots, b_m$ follow a multinomial distribution with total count $n = \sum_{i=1}^m Y_i$ and cell probabilities $\theta_1, \ldots, \theta_m$. Our aim is to estimate the parameter vector $\theta = (\theta_1, \ldots, \theta_m)$ from the observed cell counts $y_1, \ldots, y_m$. The joint p.d.f is*

$$f_{Y_1,\ldots,Y_m}(y_1, \ldots, y_m; \theta_1, \ldots, \theta_m) = \tfrac{n!}{\prod_{i=1}^m y_i!} \prod_{j=1}^n \theta_j^{y_j}, \quad \theta_j > 0, \quad \sum_{j=1}^m \theta_j = 1.$$

The log-likelihood function is

$$\ell(\theta_1, \ldots, \theta_m) = \log n! - \sum_{i=1}^m \log y_i! + \sum_{i=1}^m y_i \log \theta_i.$$

To maximise this function subject to the constraint $\sum_{j=1}^m \theta_j = 1$, we introduce a Lagrange multiplier and maximise the extended log-likelihood function

$$\Lambda(\theta_1, \ldots, \theta_m; \lambda) = \ell(\theta_1, \ldots, \theta_m) + \lambda\left(\sum_{i=1}^n \theta_i - 1\right).$$

Solving the partial derivatives equal to zero in θ_j, leads to the system

$$\theta_j = -\tfrac{y_j}{\lambda}, \quad j = 1, \ldots, m.$$

Summing both sides of this equation we get $\sum_j \theta_j = 1 = -\frac{n}{\lambda}$, thus $\lambda = -n$. Therefore, $\widehat{\theta_i} = y_i/n$, $i = 1, \ldots, n$. The log-likelihood function is strictly concave thus $\widehat{\theta} = (y_1/n, \ldots, y_m/n)$ is a global maximum, so it is the MLE of θ.

Suppose now that the multinomial cell probabilities $\theta_1, \ldots, \theta_m$ are functions of other unknown parameters τ;

that is $\theta_i = \theta_i(\tau)$. The log-likelihood of τ is

$$\ell(\tau) = \log n! - \sum_{i=1}^{m} \log y_i! + \sum_{i=1}^{m} y_i \log \theta_i(\tau).$$

Here is a concrete example.

If gene frequencies are in equilibrium, according to the Hardy-Winberg law, the genotypes AA, Aa *and* aa *occur in a population with frequencies $(1-\tau)^2$, $2\tau(1-\tau)$ and τ^2. In a sample of Chinese population of Hong Kong in 1937, blood types occur with the following frequencies, where M and N are red cell antigens*

	Blood Type			
	M	MN	N	Total
Frequency	342	500	187	1029

If we denote by y_1, y_2, y_3 the observed counts and let $n = 1029$, then the log-likelihood of τ is

$$\begin{aligned}\ell(\tau) &= \log n! - \sum_{i=1}^{3} \log y_i! + y_1 \log(1-\tau)^2 + y_2 \log[2\tau(1-\tau)] + y_3 \log \tau^2 \\ &= \log n! - \sum_{i=1}^{3} \log y_i! + (2y_1 + y_2)\log(1-\tau) + (2y_3 + y_2)\log\tau + y_2 \log 2.\end{aligned}$$

This time there is no need to incorporate the constraint that the cell probabilities sum to 1, since the functional form of $\theta_i(\tau)$ is such that $\sum_i \theta(\tau) = 1$. Setting solving $\mathrm{d}\ell(\tau)/\mathrm{d}\tau = 0$ in τ, we have

$$-\tfrac{2y_1+y_2}{1-\tau} + \tfrac{2y_3+y_2}{\tau} = 0.$$

Which gives the solution

$$\widehat{\tau} = \tfrac{2y_3+y_2}{2n}.$$

The method of maximum likelihood so widely used that is considered the gold standard method of estimation in parametric statistical models. One of the reason of this widespread usage is that, under certain rather broad conditions, the MLE $\widehat{\theta}_n$ possesses many of the above properties of estimators, as n diverges. We study more closely the properties of the MLE in the next section

4.4 Properties of the maximum likelihood estimator

So far we were mostly concerned about two features of the sampling distribution of an estimator: location and scale. Sometimes however these two features are not enough and it is useful to know the whole d.f. of an estimator. Unfortunately, the exact sampling distribution of many practical estimators is difficult or even impossible to calculate analytically. Nevertheless, there are two viable alternatives to this:

- simulation, e.g. the method of *bootstrap*;
- asymptotic approximations, e.g. Central Limit Theorem, etc.

We will touch upon the bootstrap near the end of this course. As far as asymptotic approximations are concerned, we saw in some of the examples of Section 4.2 that sometimes the sampling distribution of $\widehat{\theta}_n$ can be determined exactly, for each n and in other cases it can be approximated by the CLT for large n.

In the following we concentrate on the MLE and study some of its most relevant properties and we spend some time explaining what these properties mean and why they are good things.

Theorem 4.5 *Under suitable condition on F_θ, the following results hold.*

(i) If T_n is a sufficient statistic for θ, then the MLE of θ is a function of T_n.

(ii) The MLE is equivariant, i.e. if $\widehat{\theta}_n$ is the MLE of θ then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$.

(iii) The MLE is consistent, i.e. $\widehat{\theta}_n \xrightarrow{P} \theta_0$, where θ_0 is the true parameter value.

(iv) The MLE is asymptotically efficient: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large n.

(v) The MLE is asymptotically normal: $(\widehat{\theta}_n - \theta_0)/\mathrm{se}(\widehat{\theta}_n) \xrightarrow{d} N(0,1)$ as $n \to \infty$, where $\mathrm{se}(\widehat{\theta}) = \sqrt{\mathrm{var}(\widehat{\theta}_n)}$ and can be often computed or approximated analytically.

Property (i) tells us that the MLE is a sufficient statistic, whenever there exists one. To see why this is the case, let T_n be sufficient. By the likelihood factorisation criterion we have

$$L(\theta) = q(\theta; t_n)q(y_1, \ldots, y_n)$$

Since the maximum of $L(\theta)$ is also the maximum of $q(\theta; t_n)$ then, it is clear that $\widehat{\theta}_n$ will depend on the observed sample $y_1, \ldots, y_n$ through $t_n = T(y_1, \ldots, y_n)$ so, the MLE is sufficient.

To see why property (ii) is true, for simplicity we restrict ourselves to a function $g : \Theta \to \mathcal{T}$ being bijective. Let $h = g^{-1} : \mathcal{T} \to \Theta$ denote the inverse of g. Then $\theta = h(\tau)$ and $\tau = g(\theta)$, so we have that

$$L(\theta) = L(h(\tau)),$$

This means that the likelihood seen as function of θ is identical to the likelihood seen as function of τ. But then $L(\widehat{\theta}) = L(h(\widehat{\tau}))$, so $\widehat{\theta} = h(\widehat{\tau})$ and thus $\widehat{\tau} = g(\widehat{\theta})$. Here is an example on this point.

Example 4.16 *Let $Y_i \overset{\text{iid}}{\sim} \mathrm{Poi}(\theta)$, $i = 1, \ldots, n$ be a random sample and consider the observed sample $y_1, \ldots, y_n$. Our aim is to estimate $\tau = e^\theta$ from this observed sample. Given $L(\theta)$ the likelihood function of θ, the MLE* *of θ is found to be $\widehat{\theta} = \overline{y}$. Our aim is to estimate τ and by the equivariance principle thus we readily have that the MLE of τ is $\widehat{\tau} = e^{\widehat{\theta}} = e^{\overline{y}}$.*

Alternatively, we can ignore the aforementioned principle and try to get the MLE of τ by maximising the log-likelihood function with respect to τ. For this we have first to reparametrise the likelihood function in terms of τ.

To reparametrise in terms of τ note first that $\tau = e^{\theta}$, so $\tau : \mathbb{R}_{>0} \to (1, +\infty)$. Thus the inverse of g is $\theta = \log\tau$. Thus $L(\theta) = L(\log\tau)$. The log-likelihood function of τ is thus

$$\ell(\log\tau) = -n\log\tau + \sum_{i=1}^{n} y_i \log(\log\tau) + const..$$

Maximising $\ell(\log\tau)$ with respect to τ we get the maximum $\widehat{\tau} = e^{\overline{y}}$. Thus the MLE of τ is again $\widehat{\tau} = e^{\overline{y}}$.

To see why property (iii) is true, let $Y_i \overset{\text{iid}}{\sim} F_{\theta_0}$, thus the sample is generated from F with parameter θ_0 $(i = 1, \ldots, n)$. The idea is that the MLE will be closer and closer to θ_0 as $n \to \infty$. Let $L(\theta; Y_1, \ldots Y_n)$ be the likelihood function evaluated at the random sample and consider maximising its average

$$\tfrac{1}{n}\ell(\theta; Y_1, \ldots, Y_n) = \tfrac{1}{n}\sum_{i=1}^{n} \log f(Y_i; \theta).$$

Let $W_i = \log f(Y_i; \theta)$ be the r.v. obtained by transforming Y_i by the log-density function. Then $\frac{1}{n}\ell(\theta; Y_1, \ldots, Y_n) = \overline{W}$, the sample average of $W_1, \ldots, W_n$. As n tends to infinity, the Law of Large Numbers implies that

$$\begin{aligned} \overline{W} \xrightarrow{P} E_{\theta_0}\left(\overline{W}\right) &= E_{\theta_0}(W_1) \\ &= \int \log f(t;\theta) f(t;\theta_0)\, dt. \end{aligned}$$

This result tells us that the sample average log-likelihood function converges point-wise to the function $E_{\theta_0}(W_1)$. It is thus plausible that, for large n, the maximum of $\ell(\theta; Y_1, \ldots, Y_n)$ $\widehat{\theta}$ should be close to the maximum of $E(\log f(Y_1; \theta))^1$. (An involved argument is necessary to establish this.) But the maximum of the latter is achieved in θ_0 To see this, consider its derivative

$$\tfrac{\partial}{\partial\theta} \int \log f(t;\theta) f(t;\theta_0)\, dt = \int \frac{\frac{\partial}{\partial\theta} f(t;\theta)}{f(t;\theta)} f(t;\theta_0) dt.$$

If $\theta = \theta_0$ this equation becomes

$$\int \tfrac{\partial}{\partial\theta} f(t;\theta_0) dt = \tfrac{\partial}{\partial\theta} \int f(t;\theta_0) dt = \tfrac{\partial}{\partial\theta}(1) = 0,$$

[1] Note that $E(\log f(Y_i;\theta))$ is the same for all i so by convention we use $i = 1$ and thus $E(\log f(Y_1;\theta))$.

which shows that θ_0 is a stationary point and hopefully a maximum. Note that we have interchanged differentiation with integration and that the assumption of smoothness of f must be strong enough to justify this.

Lastly, property *(iv)* is related to *(v)*. To see why these properties are true we need some further definitions.

Definition 4.2 *For $Y_i \overset{\text{iid}}{\sim} F_\theta$ a random sample of size n, with p.d.f. $f(y;\theta)$, the* <u>*score function*</u> *is defined by*

$$\begin{aligned} s(\theta; Y_1, \ldots, Y_n) &= \tfrac{\mathrm{d} \log L(\theta; Y_1, \ldots, Y_n)}{\mathrm{d}\theta} \\ &= \tfrac{\mathrm{d}\ell(\theta; Y_1, \ldots, Y_n)}{\mathrm{d}\theta} \\ &= \sum_{i=1}^{n} \tfrac{\mathrm{d} \log f(Y_i;\theta)}{\mathrm{d}\theta} \\ &= \sum_{i=1}^{n} s(\theta; Y_i). \end{aligned}$$

The <u>*Fisher information*</u>, *also called* <u>*expected information*</u>, *is defined to be*

$$\begin{aligned} I_n(\theta) &= \operatorname{var}\left(s(\theta; Y_1, \ldots, Y_n)\right) \\ &= \operatorname{var}\left(\sum_{i=1}^{n} s(\theta; Y_i)\right) \\ &= \sum_{i=1}^{n} \operatorname{var}(s(\theta; Y_i)). \end{aligned}$$

For $n = 1$ we also write $I(\theta)$ instead of $I_1(\theta)$. Note that $I_n(\theta)$ here is exactly the same as that given earlier in Theorem 4.3, Here we only give to $I_n(\theta)$ a proper name, and show how it is related to the score function. Indeed, it can be shown that $E(s(Y_i;\theta)) = 0$ for any i. It then follows that $\operatorname{var}(s(\theta;Y_1)) = E\left(s(\theta;Y_1)^2\right)$ (recall Footnote 2 above about the convention!). In fact we have the following result.

Theorem 4.6 *For $Y_i \overset{\text{iid}}{\sim} F_\theta$ a random sample of size n, with p.d.f. $f(y;\theta)$*

$$I_n(\theta) = nI(\theta).$$

Furthermore,

$$\begin{aligned} I(\theta) &= -E\left(\tfrac{\mathrm{d}^2 \log f(Y_1;\theta)}{\mathrm{d}\theta^2}\right) \\ &= -\int \left(\tfrac{\mathrm{d}^2 \log f(y;\theta)}{\mathrm{d}\theta^2}\right) f(y;\theta)\mathrm{d}y. \end{aligned}$$

Here is then what we wanted to see.

Theorem 4.7 (Asymptotic Normality of the MLE) *Under appropriate regularity conditions, the following hold:*

1. $(\widehat{\theta}_n - \theta_0)\sqrt{I_n(\theta_0)} \xrightarrow{d} \mathrm{N}(0,1)$.
2. $(\widehat{\theta}_n - \theta_0)\sqrt{I_n(\widehat{\theta})} \xrightarrow{d} \mathrm{N}(0,1)$.

The first statement says that $\widehat{\theta}_n$ is asymptotically distributed as $\mathrm{N}(\theta_0, I_n(\theta_0)^{-1})$. Thus, $\mathrm{var}(\widehat{\theta}_n)$ the asymptotical variance of the MLE is equal to $I_n(\theta_0)^{-1}$, so the MLE is *asymptotically* efficient. The second statement says that the MLE still has the same normal distribution even though the unknown θ_0 is replaced by the MLE. It can be shown that the the MLE has still a normal distribution even if we replace $I_n(\theta)$ by $J_n(\theta)$ or $I_n(\widehat{\theta})$ by $J_n(\widehat{\theta})$; $J_n(\theta)$ is the observed information (see Lecture 3).

Example 4.17 *Consider again Example 4.13. To determine the distribution of* $\widehat{\lambda}_n = \overline{Y}$ *note that* $s(\lambda; Y_1) = -1 + Y_i/\lambda$, *thus* $\frac{ds(\lambda;Y_i)}{d\lambda} = \frac{d^2 \log f(\lambda;Y_i)}{d\lambda^2} = -Y_i/\lambda^2$. *It follows that*

$$J_n(\lambda) = -\sum_{i=1}^{n} \tfrac{d^2 \log f(\lambda;Y_i)}{d\lambda^2} = \sum_{i=1}^{n} Y_i/\lambda^2$$

$$\begin{aligned} I_n(\lambda) &= nI(\lambda) \\ &= nE\left(s(\lambda;Y_1)^2\right) \\ &= -nE\left((Y_1/\lambda - 1)^2\right) = n/\lambda. \end{aligned}$$

Thus by Theorem 4.7 and letting λ_0 *be the true parameter value, we have that*

$$\widehat{\lambda}_n \xrightarrow{d} \mathrm{N}(\lambda_0, \lambda_0/n), \quad \text{and} \quad \widehat{\lambda}_n \xrightarrow{d} \mathrm{N}(\lambda_0, \overline{Y}/n),$$

as $n \to \infty$. *Clearly, in this example the distribution of the MLE can be determined exactly. Indeed,*

$$n\widehat{\lambda}_n = \sum_{i=1}^{n} Y_i \sim \mathrm{Poi}(n\lambda).$$

Remark 4.3

(i) *In the above example we see that* $E(J_n(\lambda)) = I_n(\lambda)$, *thus since we assume* $I_n(\lambda) < \infty$, *by the WLLN we have that* $J_n(\lambda) \xrightarrow{P} I_n(\lambda)$; *this result is not limited to the present example but holds more generally. Indeed, in all regular models for which the MLE is consistent, the Fisher information is equal to the expectation of the observed information.*

(ii) *The result* $\widehat{\lambda}_n \xrightarrow{d} \mathrm{N}(\lambda_0, \lambda_0/n)$, *where* $\widehat{\lambda}_n = \overline{Y}$ *can be obtained also by appealing to the CLT (do you see why?)*

(iii) In many practical applications the Fisher information is difficult to compute, if possible at all. Furthermore, θ_0 is unknown. However, thanks to Theorem 4.7, using $J_n(\widehat{\theta})$ in place of $I_n(\theta_0)$ we still have the sample limiting distribution for the MLE. Thus, a practical advise is: if the Fisher information is computable, then use the standard error of the MLE given by $\widehat{\mathrm{se}}(\widehat{\theta}) = \sqrt{1/I_n(\widehat{\theta})}$; if not, use the standard error $\widehat{\mathrm{se}}(\widehat{\theta}) = \sqrt{1/J_n(\widehat{\theta})}$.

The normality of the MLE is not limited to the scalar parameter case. Indeed, we have the following result.

Theorem 4.8 *Under appropriate regularity conditions on the model F_θ, $\theta \in \Theta \subseteq \mathbb{R}^p$, then:*

(i) $\widehat{\theta}_n \xrightarrow{d} \mathrm{N}_p\left(\theta, I_n(\widehat{\theta}_n)^{-1}\right)$;

(ii) $\widehat{\theta}_n \xrightarrow{d} \mathrm{N}_p\left(\theta, J_n(\widehat{\theta}_n)^{-1}\right)$.

This result is extremely useful in practice since, it also says that each component of $\widehat{\theta}_n$ is approximately normally distributed. In particular, if we let θ_i be ith component of the vector θ, $\widehat{\theta}_{n,i}$ the ith component of $\widehat{\theta}_n$ and if we let $I_n(\theta)^{ii}$ denote the cell (i,i) of the matrix $I_n(\theta)^{-1}$, then we have that

$$(\widehat{\theta}_{n,i} - \theta_i)/\sqrt{I_n(\theta)^{ii}} \xrightarrow{d} \mathrm{N}\left(0,1\right).$$

A similar result holds with I_n replaced by J_n.

References

[HMC20] Hogg, R. V., McKeen, J. W. and Craig, A. T. (2018) *Introduction to Mathematical Statistics*, 8th edition, global ed., Pearson Education, Chapp. 4, 6 and 7.