Chapter 4. Statistical Inference

Populations are characterized by parameters. The goal of Inferential Statistics is to make inferences (estimates) about one or more population parameters on the basis of a sample.

1 Estimation; Basic Notions

We will refer to the parameter to be estimated as the **target parameter** and denote it by θ.
Two types of estimation will be considered: **point estimate**, when the result of the estimation is one single value and **interval estimate**, when the estimate is an interval enclosing the value of the target parameter. In either case, the actual estimation is accomplished by an **estimator**, a rule, a formula, or a procedure that leads us to the value of an estimate, based on the data from a sample.

Throughout this chapter, we consider a characteristic X (relative to a population), whose pdf $f(x;\theta)$ depends on the parameter θ, which is to be estimated. If X is discrete, then f represents the probability distribution function, while if X is continuous, f is the probability density function.

As before, we consider a random sample of size n, i.e. sample variables $X_1, \dots, X_n$, which are **independent and identically distributed (iid)**, having the same pdf as X. The notations introduced in the previous chapter for some sample functions still stand.

A **point estimator** for (the estimation of) the target parameter θ is a sample function (statistic)

$$\overline{\theta} = \overline{\theta}(X_1, X_2, \dots, X_n).$$

Other notations may be used, such as $\hat{\theta}$ or $\tilde{\theta}$.

Each statistic is a random variable because it is computed from random data. It has a so-called *sampling distribution*. Each statistic estimates the corresponding population parameter and adds certain information about the distribution of X, the variable of interest. The value of the point estimator, the **point estimate**, is the actual approximation of the unknown parameter.

Many different point estimators may be obtained for the same target parameter. Some are considered “good”, others “bad”, some “better” than others. We need some criteria to decide on one estimator versus another.

For one thing, it is highly desirable that the sampling distribution of an estimator $\overline{\theta}$ to be “clustered” around the target parameter. In simple terms, we *expect* that the value the point estimator provides to be the actual value of the parameter it estimates. This justifies the following notion.

Definition 1.1. *A point estimator $\overline{\theta}$ is called an* ***unbiased*** *estimator for θ if*

$$E(\overline{\theta}) \quad = \quad \theta. \tag{1.1}$$

The ***bias*** *of $\overline{\theta}$ is the value $B = E(\overline{\theta}) - \theta$.*

Unbiasedness means that in the long-run, collecting a large number of samples and computing $\overline{\theta}$ from each of them, on the average we hit the unknown parameter θ exactly. In other words, in a long run, unbiased estimators neither underestimate nor overestimate the parameter.

Example 1.2.
1. Recall from Proposition 2.2. (Chapter 3, Lecture 4) that for the sample mean, as a random variable, we have $E(\overline{X}) = \mu$. Thus the sample mean is an unbiased estimator for the population mean.
2. By Proposition Proposition 2.8. (Chapter 3, Lecture 4), the sample central moment of order 2 *is not* an unbiased estimator for the population central moment of order 2 (or it is a *biased* estimator), since

$$E(\overline{\mu}_2) = \frac{n-2}{n}\mu_2 \neq \mu_2 = \sigma^2.$$

3. However, the sample variance

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

is an unbiased estimator for the population variance, since $E(s^2) = \sigma^2$ (see Remark 2.11. in Chapter 3, Lecture 4). That was the main reason for the way the sample variance was defined.

Another desirable trait for a point estimator is that its values do not vary too much from the value of the target parameter. So we need to evaluate variability of computed statistics and especially parameter estimators. That can be accomplished by computing the following statistic.

Definition 1.3. *The* ***standard error*** *of an estimator $\overline{\theta}$, denoted by $\sigma_{\overline{\theta}}$, is its standard deviation*

$$\sigma_{\overline{\theta}} = \sigma(\overline{\theta}) = \mathrm{Std}(\overline{\theta}) = \sqrt{V(\overline{\theta})}.$$

Both population and sample variances are measured in squared units. Therefore, it is convenient to have standard deviations that are comparable with our variable of interest, X. As a measure of variability, standard errors show precision and reliability of estimators. They show how much estimators of the same target parameter θ can vary if they are computed from different samples. Ideally, we would like to deal with unbiased or nearly unbiased estimators that have *low* standard error.

Definition 1.4. *An unbiased estimator* $\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n)$ *for* θ *is called a* ***minimum-variance unbiased estimator (MVUE)****, if it has lower variance than any other unbiased estimator for* θ*,*

$$V(\overline{\theta}) \leq V(\hat{\theta}),\ \forall \hat{\theta} \ \text{with} \ E(\hat{\theta}) = \theta.$$

Remark 1.5. It can be shown that if an unbiased estimator exists for a parameter, then an MVUE also exists and it is unique. However, they are not easy to produce!

In Table 1, we present some common unbiased estimators, their means and their standard errors.

Target Param. θ	Sample Size	Pt. Estimator $\overline{\theta}$	Mean $E(\overline{\theta})$	St. Error $\sigma_{\overline{\theta}}$
μ	n	$\overline{X}$	μ	$\dfrac{\sigma}{\sqrt{n}}$
p	n	$\overline{p}$	p	$\sqrt{\dfrac{pq}{n}}$
$\mu_1 - \mu_2$	n_1, n_2	$\overline{X}_1 - \overline{X}_2$	$\mu_1 - \mu_2$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1, n_2	$\overline{p}_1 - \overline{p}_2$	$p_1 - p_2$	$\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$

Table 1: Common Unbiased Estimators

Remark 1.6.

1. The expected values and the standard errors in Table 1 are valid regardless of the form of the density function of the underlying population. Proposition 2.2. (Chapter 3, Lecture 4) states that for the sample mean. Similar results hold for the other three point estimators.

2. For large samples (as $n, n_1, n_2 \to \infty$), all four estimators have probability densities that are approximately Normal. The Central Limit Theorem and similar theorems justify these statements. Recall, in practice, it was determined that "large" means $n > 30$ for one sample and $n_1 + n_2 > 40$ for two samples.

2 Properties of Point Estimators

2.1 Fisher's Information and Efficient Estimators

Definition 2.1. *The* ***likelihood function*** *of a sample $X_1, \ldots, X_n$ is the joint probability function of the sample (seen as a vector), i.e. the sample function*

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta), \tag{2.1}$$

whose value $L(x_1, \ldots, x_n; \theta) = \prod\limits_{i=1}^{n} f(x_i; \theta)$ represents the joint probability distribution (in the discrete case) or the joint density (in the continuous case) of the random vector $(X_1, \ldots, X_n)$.

Definition 2.2. *For a sample of size n,* ***Fisher's (quantity of) information*** *relative to θ, is the quantity*

$$I_n(\theta) = E\left[\left(\frac{\partial \ln L(X_1, \ldots, X_n; \theta)}{\partial \theta}\right)^2\right], \tag{2.2}$$

if the likelihood function L is differentiable with respect to θ.

Remark 2.3. Fisher's information is a way of measuring the amount of information that a random sample $X_1, \ldots, X_n$ carries about an unknown parameter θ, upon which the likelihood function depends. Formally, it is the expected value of the *observed information* (or the variance of the *score*).

An easier computational formula than (2.2) is given below.

Proposition 2.4. *If the range of X does not depend on θ and the likelihood function L is twice differentiable with respect to θ, then*

$$I_n(\theta) = -E\left[\frac{\partial^2 \ln L(X_1, \ldots, X_n; \theta)}{\partial \theta^2}\right], \tag{2.3}$$

Corollary 2.5. *If the range of X does not depend on θ, then*

$$I_n(\theta) = nI_1(\theta). \tag{2.4}$$

Proof. By (2.1), we have

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \ln f(X_i; \theta), \\ \frac{\partial^2 \ln L}{\partial \theta^2} &= \sum_{i=1}^{n} \frac{\partial^2 \ln f(X_i; \theta)}{\partial \theta^2}. \end{aligned}$$

By Proposition 2.4,

$$I_n(\theta) = -\sum_{i=1}^{n} E\left[\frac{\partial^2 \ln f(X_i; \theta)}{\partial \theta^2}\right] = \sum_{i=1}^{n} I_1(\theta) = nI_1(\theta).$$

□

Recall that we seek unbiased estimators with *small* variance. At the very least, we hope that the variance gets smaller as the sample size increases. This is the idea in the next definition.

Definition 2.6. *An estimator $\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n)$ is called an* ***absolutely correct*** *estimator for θ, if it satisfies the conditions*

(i) $E(\overline{\theta}) = \theta$,

(ii) $\lim\limits_{n\to\infty} V(\overline{\theta}) = 0.$

Remark 2.7. The sample mean $\overline{X}$ is an absolutely correct estimator for the theoretical mean $\mu = E(X)$. More generally, the sample moment of order k, $\overline{\nu}_k$, is an absolutely correct estimator for the population moment of order k, $\nu_k = E\left(X^k\right)$.

Recall that we seek unbiased estimators with *small* variance. A MVUE has the *lowest* variance that an unbiased estimator can possibly have. The next result tells us exactly how low that can be, under certain conditions.

Theorem 2.8 (Cramér-Rao Inequality)**.** *Let X be a characteristic whose probability function $f(x;\theta)$ is differentiable with respect to θ and let $\overline{\theta} = \overline{\theta}(X_1, \dots, X_n)$ be an absolutely correct estimator for θ. Then*

$$V(\overline{\theta}) \geq \frac{1}{I_n(\theta)}. \tag{2.5}$$

Definition 2.9. *Let $\overline{\theta} = \overline{\theta}(X_1, \dots, X_n)$ be an absolutely correct estimator for θ. The* ***efficiency*** *of $\overline{\theta}$ is the quantity*

$$e(\overline{\theta}) = \frac{I_n^{-1}(\theta)}{V(\overline{\theta})} = \frac{1}{I_n(\theta)V(\overline{\theta})}. \tag{2.6}$$

The estimator $\overline{\theta}$ is said to be ***efficient*** *for θ, if $e(\overline{\theta}) = 1$.*

Remark 2.10.

1. So, by Theorem 2.8, the efficiency $e(\overline{\theta})$ is the minimum possible variance for an unbiased estimator $\overline{\theta}$ divided by its actual variance. Its value is always $e(\overline{\theta}) \leq 1$.
2. An efficient estimator may not exist, but if it does, it is also the MVUE. This is because an efficient estimator maintains equality on the Cramér-Rao inequality for all parameter values, which means it attains the minimum variance for all parameters. The MVUE, even if it exists, is not necessarily efficient.

Example 2.11. Let X be a characteristic with pdf

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

for $x > 0$ and 0, otherwise, where $\theta > 0$ is unknown. For a random sample $X_1, \dots, X_n$, consider the estimator $\overline{\theta} = \frac{1}{2}\overline{X}$. Show that it is absolutely correct and find its efficiency.

Solution. First, let us check that $f(x;\theta)$ is indeed a density function.

$$\int\limits_{\mathbb{R}} f(x)\, dx = \frac{1}{\theta^2} \int\limits_0^\infty x e^{-\frac{x}{\theta}}\, dx,$$

which, with the change of variables $u = \frac{x}{\theta}$, is equal to

$$\int\limits_0^\infty u e^{-u}\, du = \Gamma(2) = 1,$$

where $\Gamma(a) = \int\limits_0^\infty x^{a-1}e^{-x}\,dx$ is Euler's Gamma function. Recall that $\Gamma(n+1) = n!$.

With the same change of variables, we compute

$$\begin{aligned} E(X) &= \int\limits_{\mathbb{R}} xf(x)\,dx = \frac{1}{\theta^2}\int\limits_0^\infty x^2e^{-\frac{x}{\theta}}\,dx = \theta\int\limits_0^\infty u^2e^{-u}\,du = \theta\,\Gamma(3) = 2\theta, \\ E(X^2) &= \int\limits_{\mathbb{R}} x^2f(x)\,dx = \frac{1}{\theta^2}\int\limits_0^\infty x^3e^{-\frac{x}{\theta}}\,dx = \theta^2\int\limits_0^\infty u^3e^{-u}\,du = \theta^2\,\Gamma(4) = 6\theta^2, \\ V(X) &= E(X^2) - (E(X))^2 = 6\theta^2 - 4\theta^2 = 2\theta^2. \end{aligned}$$

Then for $\overline{\theta}$ we have

$$E(\overline{\theta}) = \frac{1}{2}E(\overline{X}) = \frac{1}{2}E(X) = \theta,$$

which means $\overline{\theta}$ is unbiased and

$$V(\overline{\theta}) = \frac{1}{4}V(\overline{X}) = \frac{1}{4}\frac{V(X)}{n} = \frac{\theta^2}{2n} \to 0, \text{ as } n \to \infty,$$

so $\overline{\theta}$ is absolutely correct.

To compute Fisher's information, since the range of X does not depend on θ, we use (2.4). We have

$$L(X_1;\theta) = \frac{1}{\theta^2}X_1e^{-\frac{1}{\theta}X_1}, \;\; \ln L = -2\ln\theta + \ln X_1 - \frac{1}{\theta}X_1,$$

so

$$\frac{\partial \ln L}{\partial\theta} = -\frac{2}{\theta} + \frac{1}{\theta^2}X_1, \;\; \frac{\partial^2 \ln L}{\partial\theta^2} = \frac{2}{\theta^2} - \frac{2}{\theta^3}X_1.$$

Then

$$I_1(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial\theta^2}\right) = -\frac{2}{\theta^2} + \frac{2}{\theta^3}E(X_1) = -\frac{2}{\theta^2} + \frac{4}{\theta^2} = \frac{2}{\theta^2}.$$

Thus

$$I_n(\theta) = \frac{2n}{\theta^2} \;\text{ and }\; e(\overline{\theta}) = 1,$$

so $\overline{\theta} = \frac{1}{2}\overline{X}$ is an efficient estimator and, by Remark 2.10, also the MVUE for θ.

■

2.2 Consistent Estimators

Recall that we seek estimators $\overline{\theta}$ that are unbiased ($E(\overline{\theta}) = \theta$), or, at least, *nearly* unbiased. We would expect that as the sample size n increases, $\overline{\theta}$ gets "closer" to θ, at least in a probabilistic sense. That is the idea behind consistent estimators.

Definition 2.12. *An estimator $\overline{\theta} = \overline{\theta}_n$, found from a sample of size n, is said to be a* ***consistent*** *estimator for θ, if $\overline{\theta}_n \xrightarrow{p} \theta$, i.e. if for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\Big(|\overline{\theta}_n - \theta| < \varepsilon\Big) = 1.$$

The property of consistency of a point estimator ensures the fact that the larger the sample size, the better the estimate. The estimate "improves consistently" with increasing the sample size. The notions of *unbiasedness* and *consistency* seem to be very close, however they are not equivalent: Unbiasedness is a statement about the expected value of the sampling distribution of the estimator. Consistency is a statement about "where the sampling distribution of the estimator is going" as the sample size increases. Let us consider a few examples.

Example 2.13. Let $X_1, \dots, X_n$ be a random sample drawn from a $N(\mu, \sigma)$ population, with both parameters $\mu \in \mathbb{R}, \sigma > 0$ unknown.

For estimating the mean μ, consider the estimator $\overline{\mu} = X_1$. Obviously it is an unbiased estimator for μ, since

$$E(X_1) \;=\; E(X) \;= \mu.$$

But, $\overline{\mu}$ is *not* consistent, since its distribution does *not* become more concentrated around μ as the sample size increases, it stays $N(\mu, \sigma)$, no matter how large the sample size gets.

To estimate the variance σ^2, let $\overline{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. We know (from Proposition 3.5, Lecture 5) that

$$E\left(\overline{\sigma}^2\right) \;=\; \frac{n-1}{n}\,\sigma^2 \;\neq\; \sigma^2,$$

so $\overline{\sigma}^2$ is *not* unbiased. On the other hand, we have

$$V\left(\overline{\sigma}^2\right) \;=\; \frac{2(n-1)}{n^2}\sigma^4.$$

So, we see that the distribution of $\overline{\sigma}^2$ is becoming more and more concentrated at σ^2 as the sample size increases, since its mean is converging to σ^2 and its variance is converging to 0, as $n \to \infty$. Thus, $\overline{\sigma}^2$ is a consistent estimator for σ^2.

Example 2.14. Let $X_1, \dots, X_n$ be a random sample drawn from a population with pdf

$$\begin{pmatrix} -a & a \\ 0.5 & 0.5 \end{pmatrix},$$

with $a > 0$ unknown.
Consider the estimators $\hat{\theta}_1 = \max\{X_1, \dots, X_n\}$ and $\hat{\theta}_2 = \overline{X}$, for the estimation of a. Let us study their unbiasedness and their consistency.

First, we compute the population mean and variance

$$\begin{aligned} E(X) &= -a \cdot 0.5 + a \cdot 0.5 = 0, \\ V(X) &= E\left(X^2\right) - (E(X))^2 = a^2. \end{aligned}$$

Let us find the pdf of $\hat{\theta}_1$. Obviously, $\hat{\theta}_1$ can only take the values a or $-a$. The only way that the maximum of the X_i's is $-a$ is if *all* variables X_i take the value $-a$. That means that

$$\begin{aligned} P(\hat{\theta}_1 = -a) &= P(X_1 = -a) \dots P(X_n = -a) = \frac{1}{2^n} \text{ and, consequently,} \\ P(\hat{\theta}_1 = a) &= 1 - \frac{1}{2^n}. \end{aligned}$$

Thus, the pdf of $\hat{\theta}_1$ is

$$\hat{\theta}_1 \begin{pmatrix} -a & a \\ \dfrac{1}{2^n} & 1 - \dfrac{1}{2^n} \end{pmatrix},$$

and its mean is

$$E(\hat{\theta}_1) = -\frac{a}{2^n} + a\left(1 - \frac{1}{2^n}\right) = a\left(1 - \frac{1}{2^{n-1}}\right) < a.$$

So $\hat{\theta}_1$ is *biased*. However, it is a consistent estimator of a because the error probability $\frac{1}{2^n}$converges to 0 as the sample size increases, so the limit of the pdf of $\hat{\theta}_1$ as $n \to \infty$ is the constant random variable $\begin{pmatrix} a \\ 1 \end{pmatrix}$.

For the second estimator, we have

$$E(\hat{\theta}_2) = E(\overline{X}) = E(X) = 0 \neq a,$$

so $\hat{\theta}_2$ is also *biased* for the estimation of a.

By the WLLN,

$$\overline{X} \overset{p}{\to} E(X),, \text{ i.e.}$$

$$\hat{\theta}_2 \overset{p}{\to} 0 \neq a,$$

so $\hat{\theta}_2$ is neither unbiased, nor consistent.

Proposition 2.15. *An absolutely correct estimator is consistent.*

Proof. Let $\overline{\theta}$ be an absolutely correct estimator. By Chebyshev's inequality, for every $\varepsilon > 0$,

$$P(|\overline{\theta} - E(\overline{\theta})| \geq \varepsilon) \;\leq\; \frac{V(\overline{\theta})}{\varepsilon^2}.$$

Since $\overline{\theta}$ is absolutely correct, it is unbiased, $E(\overline{\theta}) = \theta$, so we have

$$0 \;\leq\; P(|\overline{\theta} - \theta| \geq \varepsilon) \;\leq\; \frac{V(\overline{\theta})}{\varepsilon^2}.$$

Let $n \to \infty$ to get

$$\lim_{n\to\infty} P(|\overline{\theta} - \theta| \geq \varepsilon) \;=\; 0.$$

Taking the probability of the contrary event,

$$\lim_{n\to\infty} P(|\overline{\theta} - \theta| < \varepsilon) \;=\; 1.$$

Thus, $\overline{\theta}$ is a consistent estimator. □

Remark 2.16. The sample moment of order k, $\overline{\nu}_k$, is a consistent estimator for the population moment of order k, $\nu_k = E\left(X^k\right)$, since it is absolutely correct. In particular, the sample mean $\overline{X}$ is a consistent estimator for the theoretical mean $\mu = E(X)$.