Back to Chapter 4 (Inferential Statistics)

Short Review of Estimation Theory

- we refer to the parameter to be estimated as the **target parameter** and denote it by θ;
- 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;
- we consider a random sample of size n, i.e. sample variables $X_1, \ldots, X_n$, which are **independent and identically distributed (iid)**, having the same pdf as X.

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

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

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

Definition. The **sample mean** is the sample function defined by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Proposition. Let X be a population characteristic with mean $E(X) = \mu$ and variance $V(X) = \sigma^2$. Then

$$E\left(\overline{X}\right) = \mu \text{ and } V\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Definition. The statistic

$$\overline{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

is called the **sample moment of order k**.

Proposition. Let X be a characteristic with the property that for $k \in \mathbb{N}$, the theoretical moment $\nu_{2k} = \nu_{2k}(X) = E\left(X^{2k}\right)$ exists. Then

$$E\left(\overline{\nu}_k\right) = \nu_k \text{ and } V\left(\overline{\nu}_k\right) = \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right).$$

Definition. The statistic

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

is called the **sample variance**.
The statistic $s = \sqrt{s^2}$ is called the **sample standard deviation**.

Proposition. Let X be a characteristic with variance $V(X) = \mu_2 = \sigma^2$ and for which the theoretical moment $\nu_4 = E\left(X^4\right)$ exists. Then

$$E\left(s^2\right) = \sigma^2 \text{ and } V\left(s^2\right) = \frac{1}{n(n-1)}\Big[(n-1)\mu_4 - (n-3)\sigma^4\Big].$$

5 Properties of Point Estimators

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.

5.1 Unbiased Estimators

For one thing, it is highly desirable that the sampling distribution of an estimator $\overline{\theta}$ is "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 5.1. *A point estimator $\overline{\theta}$ is called an **unbiased** estimator for θ if*

$$E(\overline{\theta}) \;=\; \theta. \tag{5.1}$$

*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 5.2.
1. Recall 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. More generally, the sample moment of order $k, \overline{\nu}_k$, is an unbiased estimator for the population

moment of order k, $\nu_k = E(X^k)$, since $E(\overline{\nu}_k) = \nu_k$.
3. 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

$$E(\overline{\mu}_2) = \frac{n-1}{n}\mu_2 \neq \mu_2 = \sigma^2.$$

4. 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 $E(s^2) = \sigma^2$. 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 5.3. *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.

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

Remark 5.4.
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.
2. For large samples (as $n, n_1, n_2 \to \infty$), all these estimators have probability densities that are approximately Normal. The Central Limit Theorem and similar theorems justify these statements. In practice, it was determined that "large" means $n > 30$ for one sample and $n_1 + n_2 > 40$ for two samples.

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

Table 1: Common Unbiased Estimators

5.2 Absolutely Correct and Consistent Estimators

Recall that we seek unbiased (or nearly unbiased) estimators that have *low* standard error or at least, "decreasing" standard error. There are several ways to interpret that.

One way is for the variance to decrease as the sample size increases.

Definition 5.5. *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) $\lim_{n \to \infty} V(\overline{\theta}) = 0$.

Remark 5.6. 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, $\nu_k = E\left(X^k\right)$. In fact, *all* the unbiased estimators in Table 1 are absolutely correct.

Also, 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 5.7. *An estimator $\overline{\theta} = \overline{\theta}_n$, found from a sample of size n, is said to be a **consistent** estimator for θ, if $\overline{\theta}_n \xrightarrow{p} \theta$ ($\overline{\theta}_n$ converges in probability to θ), 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 following assertion is a direct consequence of Chebyshev's inequality.

Proposition 5.8. *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 5.9. 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 $\mu = E(X)$.

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 5.10. Let $X_1, \ldots, X_n$ 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.

Example 5.11. Let $X_1, \ldots, X_n$ be a random sample drawn from a population with pdf

$$X \left(\begin{array}{cc} -a & a \\ 0.5 & 0.5 \end{array} \right),$$

with $a > 0$ unknown.
Consider the estimator $\hat{\theta} = \max\{X_1, \ldots, X_n\}$ for the estimation of a.

First, we compute the population mean and variance

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

the last assertion following from the fact that $X^2 \equiv a^2$ (X^2 takes a single value, namely a^2, with probability 1).

Let us find the pdf of $\hat{\theta}$. Obviously, $\hat{\theta}$ 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

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

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

$$\hat{\theta} \left(\begin{array}{cc} -a & a \\ \dfrac{1}{2^n} & 1 - \dfrac{1}{2^n} \end{array} \right),$$

and its mean is

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

So $\hat{\theta}$ 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 $\hat{\theta}$ as $n \to \infty$ is the constant random variable $\begin{pmatrix} a \\ 1 \end{pmatrix}$.

5.3 Method of Moments

So far, we have discussed desirable properties of point estimators, how to distinguish "good" from "bad" or "better" estimators, based on how reliable they are in approximating the value of a population parameter. In all the procedures we analyzed and all the examples we discussed, the value of a point estimator $\overline{\theta}$ was given for a target parameter θ, based on sample variables $X_1, X_2, \ldots, X_n$, i.e. $\overline{\theta} = \overline{\theta}(X_1, X_2, \ldots, X_n)$. But *how* to *actually* find an estimator, an approximating value? Sometimes, such a value may be "guessed" from past experience or from observing many samples over time. But, most of the time, we need mathematical ways of producing a point estimator, which can then be analyzed from the various points of view discussed in the previous section.

There are several popular methods for pointwise estimation. In what follows, we present one of the oldest and easiest methods for obtaining point estimators, first formalized by K. Pearson in the late 1800's, the *method of moments*.

Let us recall, for a population characteristic X, we define the *moments of order* k as

$$\nu_k \;\; = \;\; E\left(X^k\right) \;\; = \;\; \begin{cases} \displaystyle\sum_{i \in I} x_i^k p_i, & \text{if } X \text{ is discrete with pdf } X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I} \\ \\ \displaystyle\int_{\mathbb{R}} x^k f(x)\, dx, & \text{if } X \text{ is continuous with pdf } f : \mathbb{R} \to \mathbb{R}. \end{cases} \tag{5.2}$$

For a sample drawn from the distribution of X, i.e. sample variables $X_1, \ldots, X_n$ (iid), the *sample moments of order* k are defined by

$$\overline{\nu}_k \;\; = \;\; \frac{1}{n} \sum_{i=1}^{n} X_i^k. \tag{5.3}$$

Let us recall that

$$\begin{aligned} E(\overline{\nu}_k) &= \nu_k, \\ V(\overline{\nu}_k) &= \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right) \to 0, \text{ as } n \to \infty, \end{aligned} \tag{5.4}$$

so the sample moment of order k is an *absolutely correct* (and, hence, a *consistent*) estimator for the population moment of the same order.

That is precisely the idea of this method. Since our sample comes from a family of distributions $\{f(\theta)\}$, we choose such a member of this family whose properties are close to properties of our data. Namely, we shall match the moments. As the theoretical (population) moments in (5.2) contain the target parameters that are to be estimated, while the sample moments in (5.3) are all known, computable from the sample data, simply set the two to be equal and solve the resulting system. To estimate k parameters, equate the first k population and sample moments:

$$\left\{ \begin{array}{ccc} \nu_1 & = & \overline{\nu}_1 \\ \dots & \dots & \dots \\ \nu_k & = & \overline{\nu}_k \end{array} \right. \tag{5.5}$$

The left-hand sides of these equations depend on the distribution parameters. The right-hand sides can be computed from data. The **method of moments estimator** is the solution of this $k \times k$ system of equations.

Remark 5.12. We state, without proof, the fact that an estimator $\overline{\theta}_n$ obtained by the method of moments is a *consistent* estimator.

Example 5.13. Consider a $Poisson$ distribution of parameter $\lambda > 0$, unknown. Its pdf is

$$X \left(\begin{array}{c} k \\ \dfrac{\lambda^k}{k!} e^{-\lambda} \end{array} \right)_{k=0,1,\dots}.$$

Let us estimate the parameter λ by the method of moments, based on a sample $\{X_1, \dots, X_n\}$.
Solution. There is only one unknown parameter, hence we write one equation:

$$\nu_1 = \overline{\nu}_1,$$

where

$$\begin{aligned} \nu_1 &= \mu = \lambda \text{ is the mean of the Poisson distribution and} \\ \overline{\nu}_1 &= \overline{X} = \frac{X_1 + \ldots + X_n}{n} \text{ is the sample mean.} \end{aligned}$$

So, we are solving the simple equation

$$\lambda = \overline{X}.$$

"Solving" it for λ, we obtain

$$\overline{\lambda} = \overline{X},$$

the method of moments estimator of λ. So, for instance, if we have the sample

$$\{7, 7, 11, 6, 5, 6, 7, 4\},$$

based on that, we find the estimator

$$\overline{\lambda} = \frac{53}{8} = 6.625.$$

■

Example 5.14. The following sample

$$\{-1, 1, 1, 2, -1, 2, 1, 1, 1, 2\}$$

was drawn from a distribution with pdf

$$X \begin{pmatrix} -1 & 1 & 2 \\ \frac{1}{4}\theta & 1 - \frac{1}{2}\theta & \frac{1}{4}\theta \end{pmatrix},$$

with $0 < \theta < 2$, unknown. What is the method of moments estimator of θ?

Solution. Again, we have one unknown, so one equation

$$\nu_1 = \overline{\nu}_1,$$

where

$$\nu_1 \;=\; \mu \;=\; -1 \cdot \frac{1}{4}\theta + 1 \cdot \left(1 - \frac{1}{2}\theta\right) + 2 \cdot \frac{1}{4}\theta \;=\; 1 - \frac{\theta}{4}$$

is the population mean and

$$\overline{\nu}_1 \;=\; \overline{X} \;=\; \frac{X_1 + \ldots + X_{10}}{10} \;=\; \frac{9}{10}$$

is the sample mean. So, we have

$$1 - \frac{\theta}{4} \;=\; \overline{X},$$

which yields the estimator

$$\hat{\theta} \;=\; 4(1 - \overline{X}) \;=\; \frac{2}{5} \;=\; 0.4.$$

■

Example 5.15. Let us recall the example we used before (Example 4.4, Lecture 2), where to evaluate the effectiveness of a processor, a sample of CPU times for $n = 30$ randomly chosen jobs (in seconds) was considered:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

The histogram we did suggested that the CPU times have a *Gamma* distribution with some unknown parameters $\alpha > 0$ and $\lambda > 0$. Let us use this sample to estimate them by the method of moments.

Solution. For the Gamma distribution with parameters $\alpha, \lambda > 0$, it is known that the population mean and variance are given by

$$\begin{aligned} \mu &= \alpha\lambda, \\ \sigma^2 &= \alpha\lambda^2. \end{aligned}$$

There are two unknown parameters, so we will need *two* equations to estimate them:

$$\begin{cases} \nu_1 &= \overline{\nu}_1 \\ \nu_2 &= \overline{\nu}_2 \end{cases}$$

We have

$$\begin{aligned} \nu_1 &= \mu = \alpha\lambda, \\ \overline{\nu}_1 &= \overline{X}, \end{aligned}$$

so, the first equation will be

$$\alpha\lambda = \overline{X}.$$

For the second equation, we need $\nu_2 = E(X^2)$. Recall the (more efficient) computational formula for the variance:

$$V(X) = E(X^2) - \Big(E(X)\Big)^2.$$

From here, we find

$$E(X^2) = V(X) + \Big(E(X)\Big)^2 = \alpha\lambda^2 + (\alpha\lambda)^2 = \alpha\lambda^2(1+\alpha).$$

For this sample, we found (in Lecture 3) that the sample mean and variance are

$$\begin{aligned} \overline{\nu}_1 &= \overline{X} = 48.2333 \text{ and} \\ s^2 &= 703.1506 \end{aligned}$$

The sample moment of order 2 is computed (from the sample data):

$$\overline{\nu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{90185}{30} = 3006.167.$$

So, we solve the system

$$\begin{cases} \alpha\lambda &= 48.2333 \\ \alpha\lambda^2(1+\alpha) &= 3006.167, \end{cases}$$

with solution

$$\begin{cases} \overline{\alpha} &= 3.4227, \\ \overline{\lambda} &= 14.0922. \end{cases}$$

Alternatively, since we already had the variance (the population *central* moment of order 2) computed, we could use *that* for our second equation. In other words, we consider the (population and sample) moments of order 1 (for the first equation) and the (population and sample) *central* moments of order 2 (for the second equation). We cam also compute the sample central moment of

order 2, as

$$\overline{\mu}_2 \;=\; \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \;=\; \frac{70^2+\ldots+19^2-30\cdot 48.2333^2}{30} \;=\; \frac{90185-69794}{30} \;=\; 679.7.$$

So, now we solve the system

$$\begin{cases} \alpha\lambda &= 48.2333 \\ \alpha\lambda^2 &= 679.7, \end{cases}$$

which has the solution

$$\begin{cases} \hat{\alpha} &= 3.4228, \\ \hat{\lambda} &= 14.0919. \end{cases}$$

The two estimates are very close and we can use either. ■

Remark 5.16. Method of moments estimates are typically easy to compute. However, on rare occasions, when k equations are not enough to estimate k parameters, higher moments (i.e. more equations) can be considered. Also, as we have seen in Example 5.15, *central* (population and sample) moments can be used, to make computations easier.

5.4 Estimation of Standard Errors

An important question when estimating parameters: How good are the estimators that we learned in previous sections? Standard errors can serve as measures of their accuracy. To estimate them, we derive an expression for the standard error and estimate all the unknown parameters in it.

Example 5.17. In Example 5.13, we estimated the parameter λ of a $Poisson$ distribution by

$$\overline{\lambda} \;=\; \overline{X},$$

using the method of moments. Let us estimate its standard error, based on the sample

$$\{7, 7, 11, 6, 5, 6, 7, 4\},$$

for which $\overline{X} = 6.625$ and $s = 2.0659$.

Solution. Recall that for a $Poisson(\lambda)$ distribution, the mean and the variance are

$$\mu \;=\; \sigma^2 \;=\; \lambda.$$

Also, we know that $V(\overline{X}) = \dfrac{V(X)}{n}$, hence,

$$\mathrm{Std}(\overline{X}) \;=\; \sqrt{V(\overline{X})} \;=\; \sqrt{\frac{V(X)}{n}} \;=\; \frac{\sigma}{\sqrt{n}}.$$

So, there are (at least) two ways to estimate the standard error of $\overline{\lambda}$.

On one hand, $\sigma = \sqrt{\lambda}$ for the $Poisson(\lambda)$ distribution, so we can estimate

$$\sigma_{\overline{\lambda},1} \;=\; \frac{\sigma}{\sqrt{n}} \;=\; \sqrt{\frac{\lambda}{n}} \;\approx\; \sqrt{\frac{\overline{\lambda}}{n}} \;=\; \sqrt{\frac{\overline{X}}{n}} \;=\; 0.91.$$

On the other hand, we can use the sample standard deviation as an estimate for the population one and get the estimate

$$\sigma_{\overline{\lambda},2} \;=\; \frac{\sigma}{\sqrt{n}} \;\approx\; \frac{s}{\sqrt{n}} \;=\; 0.7304.$$

Both estimates of the standard error $\sigma_{\overline{\lambda}}$ are rather small, so the estimator $\overline{\lambda}$ seems good.

■

Remark 5.18. Estimation of standard errors can become much harder for just slightly more complex estimators. In some cases, a nice analytic formula for $\sigma_{\overline{\theta}}$ may not exist. Then, other, more modern methods must be employed, such as *bootstrapping*, a method based on computer simulations.