

3.4 Sample Proportions

Definition 3.1. Assume a subpopulation A of a population consists of items that have a certain attribute. The **population proportion** is then the probability

$$p = P(i \in A), \quad (3.1)$$

i.e. the probability for a randomly selected item i to have this attribute.

The **sample proportion** is

$$\bar{p} = \frac{\text{number of sampled items from } A}{n}, \quad (3.2)$$

where n is the sample size.

Proposition 3.2. Let p be a population proportion. Then

$$E(\bar{p}) = p, \quad V(\bar{p}) = \frac{p(1-p)}{n} = \frac{pq}{n} \quad \text{and} \quad \sigma(\bar{p}) = \sqrt{\frac{pq}{n}}, \quad (3.3)$$

where $q = 1 - p$.

Proof. We use the indicator random variable

$$X_i = \begin{cases} 1, & i \in A \\ 0, & i \notin A \end{cases}.$$

Then $X_i \in \text{Bern}(p)$ and, so, we know that $E(X_i) = p$ and $V(X_i) = pq$, for every $i = 1, \dots, n$.

But notice that $\bar{p} = \frac{1}{n} \sum_{i=1}^n X_i$, i.e. the sample mean of the sample X_1, \dots, X_n . Thus, by Proposition 3.5 (Lecture 4), we have

$$\begin{aligned} E(\bar{p}) &= p, \\ V(\bar{p}) &= \frac{pq}{n}, \\ \sigma(\bar{p}) &= \sqrt{\frac{pq}{n}}. \end{aligned}$$

□

3.5 Sample Functions for Comparing Two Populations

It will be necessary sometimes to compare characteristics of two populations. For that, we will need results on sample functions referring to both collections. Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, relative to two populations. We draw from both populations independent random samples of sizes n_1 and n_2 , respectively. Denote the two sets of random variables by

$$X_{11}, \dots, X_{1n_1} \text{ and } X_{21}, \dots, X_{2n_2}.$$

Then we have two sample means and two sample variances, given by

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}, \quad \bar{X}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j}$$

and

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2, \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2,$$

respectively. In addition, denote by

$$s_p^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the **pooled variance** of the two samples, i.e. a variance that considers (pools) the data from both samples.

In inferential Statistics, when comparing the means of two populations, we will look at their difference and try to estimate it. Regarding that, we have the following result.

Proposition 3.3. *Let $X_{(1)}, X_{(2)}$ be two population characteristics with means $E(X_{(i)}) = \mu_i$ and variances $V(X_{(i)}) = \sigma_i^2, i = 1, 2$. Then*

$$\begin{aligned} E(\bar{X}_1 - \bar{X}_2) &= \mu_1 - \mu_2, \\ V(\bar{X}_1 - \bar{X}_2) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned} \tag{3.4}$$

In a similar fashion, we can compare two population proportions. Again, the random variable of interest is their difference.

Proposition 3.4. Assume we have two population proportions p_1 and p_2 . From each population we draw independent samples of size n_1 and n_2 , respectively, which yield the population proportions \bar{p}_1 and \bar{p}_2 . Then

$$\begin{aligned} E(\bar{p}_1 - \bar{p}_2) &= p_1 - p_2, \\ V(\bar{p}_1 - \bar{p}_2) &= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}, \end{aligned} \tag{3.5}$$

with $q_i = 1 - p_i, i = 1, 2$.

Notations of the sample functions and their corresponding population characteristics.

Function	Population (theoretical)	Sample
Mean	$\mu = E(X)$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
Variance	$\sigma^2 = V(X)$	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
Standard deviation	$\sigma = \sqrt{V(X)}$	$s = \sqrt{s^2}$
Moment of order k	$\nu_k = E(X^k)$	$\bar{\nu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$
Central moment of order k	$\mu_k = E[(X - E(X))^k]$	$\bar{\mu}_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$
Proportion	$p = P(i \in A)$	$\bar{p} = \frac{\text{number of } X_i \text{ from } A}{n}$

Table 1: Notations

Chapter 4. Inferential Statistics

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.

In this chapter, we discuss how

- *to estimate parameters of the distribution.* The methods in the previous chapter mostly concern measure of location (mean, median, quantiles) and variability (variance, standard deviation, interquartile range). As we know, this does not cover all possible parameters, and thus, we still lack a general methodology of estimation.
- *to construct confidence intervals.* Any estimator, computed from a collected random sample instead of the whole population, is understood as only an approximation of the corresponding parameter. Instead of one estimator that is subject to a sampling error, it is often more reasonable to produce an interval that will contain the true population parameter with a certain known high probability.
- *to test hypotheses.* That is, we shall use the collected sample to verify statements and claims about the population. As a result of each test, a statement is either rejected on basis of the observed data or accepted (not rejected). Sampling error in this analysis results in a possibility of wrongfully accepting or rejecting the hypothesis; however, we can design tests to control the probability of such errors.

Results of such statistical analysis are used for making decisions under uncertainty, developing optimal strategies, forecasting, evaluating and controlling performance and so on.

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.

Definition 1.1. A *point estimator* for (the estimation of) the target parameter θ is a sample function (statistic)

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

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.

2 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.

2.1 Unbiased Estimators

For one thing, it is highly desirable that the sampling distribution of an estimator $\bar{\theta}$ 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 2.1. A point estimator $\bar{\theta}$ is called an **unbiased** estimator for θ if

$$E(\bar{\theta}) = \theta. \tag{2.1}$$

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

Unbiasedness means that in the long-run, collecting a large number of samples and computing $\bar{\theta}$ 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 2.2.

1. Recall from Proposition 3.5 (Lecture 4) that for the sample mean, as a random variable, we have $E(\bar{X}) = \mu$. Thus the sample mean is an *unbiased* estimator for the population mean.
2. More generally, the sample moment of order k , $\bar{\nu}_k$, is an unbiased estimator for the population moment of order k , $\nu_k = E(X^k)$, since $E(\bar{\nu}_k) = \nu_k$ (Proposition 3.9 in Lecture 4).
3. By Proposition Proposition 3.10 (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

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

4. However, the sample variance

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

is an unbiased estimator for the population variance, since $E(s^2) = \sigma^2$ (see Remark 3.13 in 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 2.3. The *standard error* of an estimator $\bar{\theta}$, denoted by $\sigma_{\bar{\theta}}$, is its standard deviation

$$\sigma_{\bar{\theta}} = \sigma(\bar{\theta}) = \text{Std}(\bar{\theta}) = \sqrt{V(\bar{\theta})}.$$

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 2 we present some common unbiased estimators, their means and their standard errors.

Remark 2.4.

1. The expected values and the standard errors in Table 2 are valid regardless of the form of the density function of the underlying population.

Target Param. θ	Sample Size	Pt. Estimator $\bar{\theta}$	Mean $E(\bar{\theta})$	St. Error $\sigma_{\bar{\theta}}$
μ	n	\bar{X}	μ	$\frac{\sigma}{\sqrt{n}}$
ν_k	n	$\bar{\nu}_k$	ν_k	$\sqrt{\frac{\nu_{2k} - \nu_k^2}{n}}$
p	n	\bar{p}	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1, n_2	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1, n_2	$\bar{p}_1 - \bar{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

Table 2: Common Unbiased Estimators

2. For large samples (as $n, n_1, n_2 \rightarrow \infty$), 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.

2.2 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 2.5. An estimator $\bar{\theta} = \bar{\theta}(X_1, \dots, X_n)$ is called an **absolutely correct** estimator for θ , if it satisfies the conditions

- (i) $E(\bar{\theta}) = \theta$,
- (ii) $\lim_{n \rightarrow \infty} V(\bar{\theta}) = 0$.

Remark 2.6. The sample mean \bar{X} is an absolutely correct estimator for the theoretical mean $\mu = E(X)$. More generally, the sample moment of order k , $\bar{\nu}_k$, is an absolutely correct estimator for the population moment of order k , $\nu_k = E(X^k)$. In fact, *all* the unbiased estimators in Table 2 are absolutely correct.

Also, we would expect that as the sample size n increases, $\bar{\theta}$ gets “closer” to θ , at least in a probabilistic sense. That is the idea behind consistent estimators.

Definition 2.7. An estimator $\bar{\theta} = \bar{\theta}_n$, found from a sample of size n , is said to be a **consistent** estimator for θ , if $\bar{\theta}_n \xrightarrow{p} \theta$ ($\bar{\theta}_n$ converges in probability to θ), i.e. if for every $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P\left(|\bar{\theta}_n - \theta| < \varepsilon\right) = 1.$$

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 2.8. An absolutely correct estimator is consistent.

Remark 2.9. The sample moment of order k , $\bar{\nu}_k$, is a consistent estimator for the population moment of order k , $\nu_k = E(X^k)$, since it is absolutely correct. In particular, the sample mean \bar{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 2.10. 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 $\bar{\mu} = X_1$. Obviously it is an unbiased estimator for μ , since

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

But, $\bar{\mu}$ 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 2.11. Let X_1, \dots, X_n be a random sample drawn from a population with pdf

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

with $a > 0$ unknown.

Consider the estimator $\hat{\theta} = \max\{X_1, \dots, X_n\}$ for the estimation of a .

First, we compute the population mean and variance

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

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

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

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

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

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 $n \rightarrow \infty$ is the constant random variable $\begin{pmatrix} a \\ 1 \end{pmatrix}$.

2.3 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 $\bar{\theta}$ was given for a target parameter θ , based on sample variables X_1, X_2, \dots, X_n , i.e. $\bar{\theta} = \bar{\theta}(X_1, X_2, \dots, X_n)$. 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

$$\nu_k = E(X^k) = \begin{cases} \sum_{i \in I} x_i^k p_i, & \text{if } X \text{ is discrete with pdf } X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I} \\ \int_{\mathbb{R}} x^k f(x) dx, & \text{if } X \text{ is continuous with pdf } f : \mathbb{R} \rightarrow \mathbb{R}. \end{cases} \quad (2.2)$$

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

$$\bar{\nu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k. \quad (2.3)$$

Also, let us recall (from Proposition 3.9, Lecture 4) that

$$\begin{aligned} E(\bar{\nu}_k) &= \nu_k, \\ V(\bar{\nu}_k) &= \frac{1}{n} (\nu_{2k} - \nu_k^2) \rightarrow 0, \text{ as } n \rightarrow \infty, \end{aligned} \quad (2.4)$$

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 (2.2) contain the target parameters that are to be estimated, while the sample moments in (2.3) 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:

$$\begin{cases} \nu_1 &= \bar{\nu}_1 \\ \dots &\dots \dots \\ \nu_k &= \bar{\nu}_k \end{cases} \quad (2.5)$$

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

Example 2.13. Consider a *Poisson* distribution of parameter $\lambda > 0$, unknown. Its pdf is

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

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:

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

where $\nu_1 = \mu = \lambda$ is the mean of the Poisson distribution and $\bar{\nu}_1 = \bar{X} = \frac{X_1 + \dots + X_n}{n}$ is the sample mean. So, we are solving the simple equation

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

“Solving” it for λ , we obtain

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

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

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

■

Example 2.14. The following sample

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

was drawn from a distribution with pdf

$$X \left(\begin{array}{ccc} -1 & 1 & 2 \\ \frac{1}{4}\theta & 1 - \frac{1}{2}\theta & \frac{1}{4}\theta \end{array} \right),$$

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

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

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

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

$$\bar{\nu}_1 = \bar{X} = \frac{X_1 + \dots + X_{10}}{10} = \frac{9}{10}$$

is the sample mean. So, we have

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

which yields the estimator

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

■

Example 2.15. Let us recall the example we used before (Example 4.4, Lecture 3), 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. We have $\nu_1 = \mu$. Since we also have the variance (the population *central* moment of order 2), let us 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). Recall that for this sample, we found (in Lectures 3 and 4) that the sample mean and variance are

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

We can also compute the sample central moment of order 2, as

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

So, we solve the system

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

which has the solution

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

■

Remark 2.16. 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 2.15, *central* (population and sample) moments can be used, to make computations easier.