Chapter 1
Random Variables, Sampling and Estimation

1.1 Introduction

This chapter will cover the most important basic statistical theory you need in order to understand the econometric material that will be coming in the next chapters. The key topics that we will review are the following:

- *Descriptive statistics*. e.g. mean and variance.
- *Probability*. e.g. events, relative frequency, marginal and conditional probability distributions.
- *Random variables, probability distributions, and expectations*.
- *Sampling*. e.g. simple random sampling.
- *Estimation*. e.g. the distinction between and estimator and an estimate.
- *Statistical inference*. *t* and F tests.

1.2 Probabilities

1.2.1 Events

Random experiment. Process leading to two or more possible outcomes, with uncertainty as to which outcome will occur.

Flip of a coin, toss of a die, a students takes a class and either obtains an A or not.

Sample space. Set of all basic outcomes of a random experiment.

When flipping a coin, S = [head, tail].
When taking a class, S = [A, B, C, D, F, drop].
When tossing a die, S = [1, 2, 3, 4, 5, 6].
No two outcomes can occur simultaneously.

Event. Subset of basic outcomes in the sample space.

Event E_1: "Pass the class" then the subset of basic outcomes is A, B, C.

Intersection of event. When two events E_1 and E_2 have some basic outcomes in common. It is denoted by $E_1 \cap E_2$.

Event E_1: Individuals with college degree.
Event E_2: Individuals who are married.
$E_1 \cap E_2$: Individuals who have college degree and are married.

Joint probability. Probability that the intersection occurs.

Mutually exclusive events. E_1 and E_2 are mutually exclusive if $E_1 \cap E_2$ is empty.

Union of events. Denoted by $E_1 \cup E_2$. At least one of these events occurs. Either E_1, E_2, or both.

Complement. The complement of E is denoted by $\bar{E}$ and it is the set of basic outcomes of a random experiment that belongs to S, but not to E_1.

E_1 is the complement of $\bar{E}_1$
Event E_2: Individuals who are married.
E_1 and $\bar{E}$ are mutually exclusive events.

1.2.2 Probability postulates

Given a random experiment, we want to determine the probability that a particular event will occur. A probability is a measure from 0 to 1.

$0 \rightarrow$ the event will not occur.
$1 \rightarrow$ the event is certain.

When the outcomes are equally likely to occur, the probability of an event E is:

$P(E) = N_E/N$
N_E: Number of outcomes in event E.
N: Total number of outcomes in the sample space S.

Example 1: Flip of a coin, Event E is "head" then P(E) = 1/2. $N_E = 1$ and N = 2.

Example 2: Event E is "winning the lottery" then if there are 1000 lottery tickets and you bought, 2 P(E) = 2/1000 = 0.002.

Some probability rules

$P(E \cup \bar{E}) = P(E) + P(\bar{E}) = 1$.
$P(\bar{E}) = 1 - P(E)$.

Conditional probability

$P(E_1 \mid E_2)$: Probability that E_1 occurs, given that E_2 has already occurred.
$P(E_1 \mid E_2) = P(E_1 \cap E_2) / P(E_2)$ given that $P(E_2) > 0$.

Addition rule

$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

Statistically independent events

$P(E_1 \cap E_2) = P(E_1)P(E_2)$.
$P(E_1 \mid E_2) = P(E_1)P(E_2) / P(E_2) = P(E_1)$.

1.3 Discrete random variables and expectations

1.3.1 Discrete random variables

Random variable. Variable that takes numerical values determined by the outcome of a random experiment.

Examples: Hourly wage, GDP, inflation, the number when tossing a die.
Notation: Random variable X can take n possible values $x_1, x_2, \cdots x_n$.

Discrete random variable. A random variable that takes a countable number of values.

Examples: Number of years of education.

Continuous random variable. A random variable that can take any value on an internal.

Examples: Wage, GDP, exact weight.

Consider tossing two dies (green and red). This will yield 36 possible outcomes because the green can take 6 possible values and the red can take also 6 values, $6 \times 6 = 36$. The possible outcomes. Let's define the random variable X to be the sum of two dice. Therefore X can take 11 possible values, from 2 to 12. This information is summarized in the following tables.

Table 1.1 Outcomes with two dies

red / green	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Table 1.2 Frequencies and probability distributions

Value of X	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Probability (p)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

1.3.2 Expected value of random variables

Let $E(X)$ be the expected value of the random variable X. The expected value of a discrete random variable is the weighted average of all its possible values, taking the probability of each outcome as its weight. Random variable X can take n particular values $x_1, x_2, \ldots, x_n$ and the probability of x_i is given by p_i. Then we have that the expected value is given by:

$$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \sum_{i=1}^{n} x_i p_i. \tag{1.1}$$

We can also write the expected value as: $E(X) = \mu_X$. For the previous example we can calculate that the expected value is:

$$E(X) = 2 \cdot 1/36 + 3 \cdot 2/36 + \cdots + 12 \cdot 1/36 = 252/36 = 7 \tag{1.2}$$

Table 1.3 Expected value of X, two dice example

X	p	Xp
2	1/36	2/36
3	2/36	6/36
4	3/36	12/36
5	4/36	20/36
6	5/36	30/36
7	6/36	42/36
8	5/36	40/36
9	4/36	36/36
10	3/36	30/36
11	2/36	22/36
12	1/36	12/36
Total	$E(X) = \sum_{i=1}^{n} x_i p_x$	252/36 = 7

1.3.3 Expected value rules

$$E(X+Y+Z) = E(X) + E(Y) + E(Z) \tag{1.3}$$

$$E(bX) = bE(X) \quad \text{for a constant} \quad b \tag{1.4}$$

$$E(b) = b \tag{1.5}$$

For the example where $Y = b_1 + b_2X$, b_1 and b_2 are constants we want to calculate $E(X)$.

$$\begin{aligned} E(Y) &= E(b_1 + b_2X) \\ &= E(b_1) + E(b_2X) \\ &= b_1 + b_2E(X) \end{aligned} \tag{1.6}$$

1.3.4 Variance of a discrete random variable

Let $\text{var}(X)$ be the variance of the random variable X. $\text{var}(X)$ is a useful measure of the dispersion of its probability distribution. It is defined as the expected value of the square of the difference between X and its mean. That is, $(X - \mu_X)^2$, where μ_X is the population mean of X.

$$\text{var}(X) = \sigma_X^2 = E[(X - \mu_X)^2] \tag{1.7}$$

$$= (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \cdots + (x_n - \mu_X)^2 p_n \tag{1.8}$$
$$= \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i$$

Taking the square root of the variance (σ_X^2) one can obtain the standard deviation, σ_X. The standard deviation also serves as a measure of dispersion of the probability distribution. A useful way to write the variance is:

$$\sigma_X^2 = E(X^2) - \mu_X^2. \tag{1.9}$$

From the previous example of tossing two dies, we have that the population variance can be calculated as follows:

Table 1.4 Population variance, X from the two dice example

X	p	$X - \mu_X$	$(X - \mu_X)^2$	$(X - \mu_X)^2 p$
2	1/36	-5	25	0.69
3	2/36	-4	16	0.89
4	3/36	-3	9	0.75
5	4/36	-2	4	0.44
6	5/36	-1	1	0.14
7	6/36	0	0	0.00
8	5/36	1	1	0.14
9	4/36	2	4	0.44
10	3/36	3	9	0.75
11	2/36	4	16	0.89
12	1/36	5	25	0.69
Total				5.83

1.3.5 Probability density

Because discrete random variables, by definition, can only take a finite number of values, they are easy to summarize graphically. The probability distribution is the graph that links all the values that a random variable can take with its corresponding probabilities. For the two dice example above, see Figure 1.1.

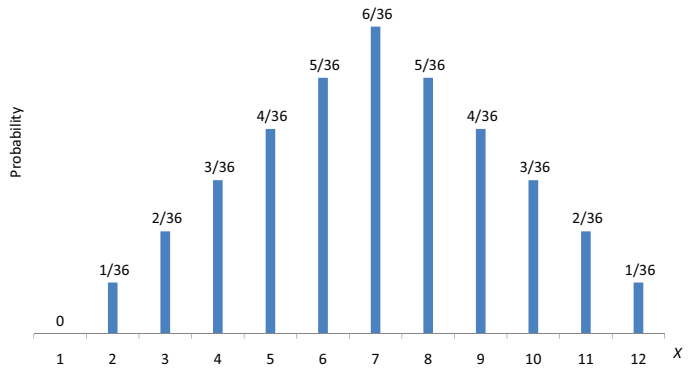

Fig. 1.1 Discrete probabilities, X from the two dice example

1.4 Continuous random variables

1.4.1 Probability density

Continuous random variables can take any value on an interval. This means that it can take an infinite number of different values, hence it is not possible to obtain a graph like the one presented in Figure 1.1 for a continuous random variable. Instead, we will define the probability of a random variable lying within a given interval. For example, the probability that the height of an individual is between 5.5 and 6 feet. This is depicted in Figure 1.2 as the shaded area below the probability density curve for the values of X between 5.5 and 6. The probability of the random variable X written as a function of the random variable is known as the probability density function. We can write this ones as $f(X)$. Then, if we use a little math we can easily find the area under the curve. Recall that the are under a curve can be obtained by taking the integral.

Probability density function. Is a function that describes the relative likelihood for a random variable to occur at a given point.

$$\int_{5.5}^{6} f(X) = 0.18 \tag{1.10}$$
$$\int_{0}^{\infty} f(X) = 1$$

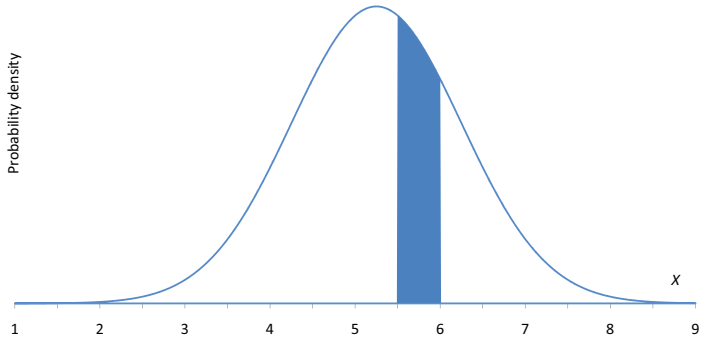

Fig. 1.2 Continuous probabilities, X from the height example

The first line in the equation above just calculates the integral under the curve $f(X)$ between the points 5.5 and 6. The second line shows that the whole area under the curve presented in Figure 1.2 is equal to one. This is for the same reason why the summation of all the bars in Figure 1.1 are also equal to one; the total probability is always equal to one.

1.4.2 Normal distribution

The normal distribution is the most widely known continuous probability distribution. The graph associated with its probability density function has a bell-shape and its is known as the Gaussian function or bell curve. Its probability density function is given by:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu^2)}{2\sigma^2}} \tag{1.11}$$

where μ is the mean and σ^2 is the variance. Figure 1.1 is an example of this distribution.

1.4.3 Expected value and variance of a continuous random variable

The basic difference between a discrete and a continuous random variable is that the second can take on infinite possible values, hence the summations signs that are used to calculate the expected value and the variance of a discrete random variable cannot be used for a continuous random variable. Instead, we use integral signs. For the expected value we have:

$$E(X) = \int X f(X) dX \tag{1.12}$$

where the integration is performed over the interval for which $f(X)$ is defined. For the variance we have:

$$\sigma_X^2 = E[(X - \mu_X)^2] = \int (X - \mu_X)^2 f(X) dX \tag{1.13}$$

1.5 Covariance and correlation

1.5.1 Covariance

When dealing with two variables, the first question you want to answer is whether these variables move together or whether they move in opposite directions. The covariance will help us answer that question. For two random variables X and Y, the covariance is defined as:

$$\text{cov}(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \tag{1.14}$$

where μ_X and μ_Y are the population means of X and Y, respectively. When to random variables are independent, their covariance is equal to zero. When $\sigma_{XY} > 0$ we say that the variables move together. When $\sigma_{XY} < 0$ they move in opposite directions.

1.5.2 Correlation

One concern when using the $\text{cov}(X,Y)$ as a measure of association is that the result is measured in the units of X times the units of Y. The correlation coefficient, that is dimensionless, overcomes this difficulty. For variables X and Y the correlation coefficient is defined as:

$$\text{corr}(X,Y) = \rho_{YX} = \frac{\sigma_{YX}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \tag{1.15}$$

The correlation coefficient is a number between -1 and 1. When it is positive, we say that there is a positive correlation between X and Y and that these two variables move in the same direction. When it is negative, we say that they move in opposite directions.

1.6 Sampling and estimators

Notice that in the two dice example we know the population characteristics, that is, the probability distribution. From this probability distribution it is easy to obtain the population mean an variance. However, what happens most of the time is that we need to rely on a data set to get estimates of the population parameters (e.g the mean and the variance). In that case the estimates of the population parameters are obtained using estimators, and the sample needs to have certain characteristics. The estimators and the sampling are the subject of this section.

1.6.1 Sampling

The most common way to obtain a sample from the population is through simple random sampling.

Simple random sampling. It is a procedure to obtain a sample from the population, where each of the observations is chosen randomly and entirely by chance. This means that each observation in the population has the same probability of being chosen.

Once the sample of the random variable X has be generated, each of the n observations can be denoted by $\{x_1, x_2, \cdots, x_n\}$.[1]

[1] The textbook Dougherty (2007) makes the distinction between the specific values of the random variable X before and after they are known, and emphasizes this distinction by using uppercase and lowercase letter. This distinction is useful only in some cases and that is why most textbooks do not make this distinction. We will follow emphasize the distinction and we will use only lowercase letters.

1.6.2 Estimators

Estimator. It is a general rule (mathematical formula) for estimating an unknown population parameter given a sample of data.

For example, an estimator for the population mean is the sample mean:

$$\bar{X} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.16}$$

An interesting feature of this estimator is that the variance of $\bar{X}$ is $1/n$ times the variance of X. The derivation is the following:

$$\sigma^2_{\bar{X}} = \text{var}(\bar{X}) \tag{1.17}$$

$$\sigma^2_{\bar{X}} = \text{var}\{\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\} \tag{1.18}$$

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}\text{var}\{\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\} \tag{1.19}$$

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}\{\text{var}(x_1) + \text{var}(x_2) + \cdots + \text{var}(x_n)\} \tag{1.20}$$

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}\{\sigma^2_X + \sigma^2_X + \cdots + \sigma^2_X\} \tag{1.21}$$

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}\{n\sigma^2_X\} = \frac{\sigma^2_X}{n} \tag{1.22}$$

Graphically, this result is shown in Figure 1.3. The distribution of X has a higher variance (it is more dispersed) than the distribution of $\bar{X}$.

1.7 Unbiasedness and efficiency

1.7.1 Unbiasedness

Because estimators are random variables, we can take expectations of the estimators. If the expectation of the estimator is equal to the true population parameter, then we say that this estimator is unbiased. Let θ be the population parameter and let $\hat{\theta}$ be a point estimator of θ. Then, $\hat{\theta}$ is unbiased if:

$$E(\hat{\theta}) = \theta \tag{1.23}$$

Example. The sample mean of X is an unbiased estimator of the population mean μ_X:

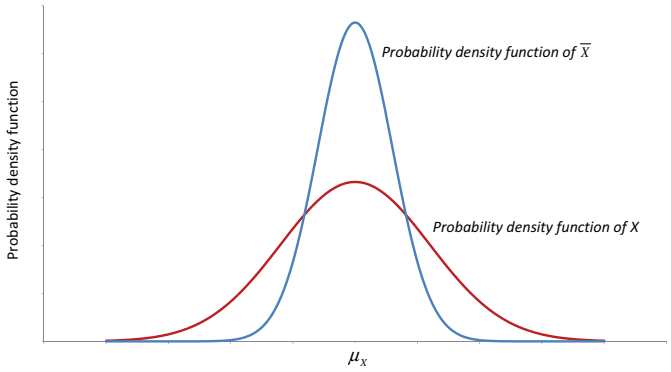

Fig. 1.3 Probability density functions of X and $\bar{X}$.

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}E(\sum_{i=1}^{n} x_i) \tag{1.24}$$
$$= \frac{1}{n}\sum_{i=1}^{n}(E(x_i)) = \frac{1}{n}\sum_{i=1}^{n}\mu_X = \frac{1}{n}n\mu_X = \mu_X$$

Unbiased estimator. An estimator is unbiased if its expected value is equal to the true population parameter.

The bias of an estimator is just the difference between its expected value and the true population parameter:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{1.25}$$

1.7.2 Efficiency

It is not only important that an estimator is on average correct (unbiased), but also that it has a high probability of being close to the true parameter. When comparing two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, we say that $\hat{\theta}_1$ is more efficient if $\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$. A comparison of the efficiency between these two estimators in presented in Figure 1.4. The estimator with higher variance, $(\hat{\theta}_2)$, is more dispersed.

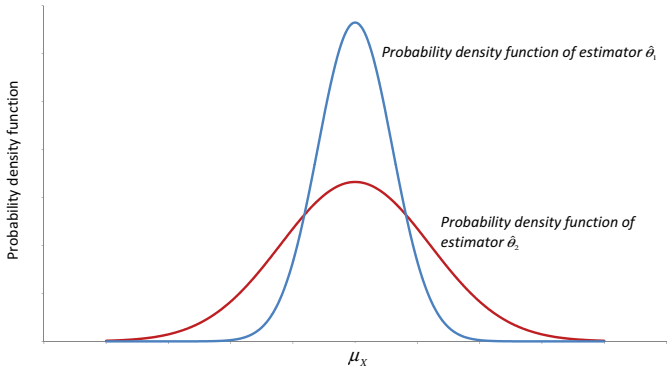

Fig. 1.4 Efficiency of estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, with $\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$.

Most efficient estimator. The estimator with the smallest variance from all unbiased estimators.

1.7.3 Unbiasedness versus efficiency

Both, unbiasedness and efficiency, are desired properties of an estimator. However, there may be conflicts in the selection between two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, if, for example, $\hat{\theta}_1$ is more efficients, but it is also biased. This case is presented in Figure 1.5.

The simplest way to select between these two estimators is to pick the one that yields the smallest mean square error (MSE):

$$\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + bias(\hat{\theta})^2 \tag{1.26}$$

1.8 Estimators for the variance, covariance, and correlation

While we have already seen the populations formulas for the variance, covariance and correlation, it is important to keep in mind that we do not have the whole population. The data sets we will be working with are just samples of the populations. The formula for the sample variance is:

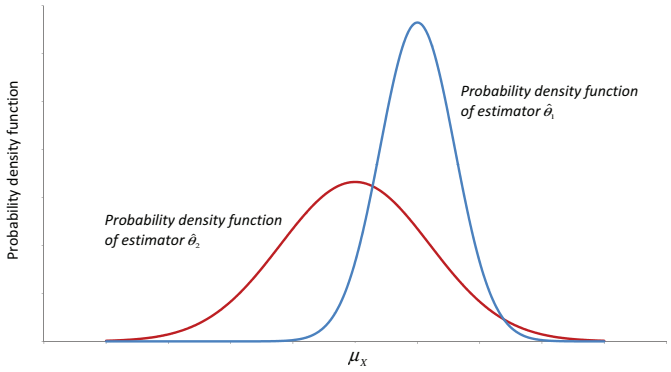

Fig. 1.5 $\hat{\theta}_2$ is unbiased, but $\hat{\theta}_1$ is more efficient.

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 \tag{1.27}$$

Notice how we changed the notation from σ^2 to s^2. The first one denotes the population variance, while the second one refers to the sample variance. An estimator for the population covariance is given by:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}). \tag{1.28}$$

Finally, the formula for the correlation coefficient, r_{XY}, is:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2 \sum_{i=1}^{n}(y_i - \bar{Y})^2}}. \tag{1.29}$$

1.9 Asymptotic properties of estimators

Asymptotic properties of estimators just refers to their properties when the number of observations in the sample grows large and approached to infinity.

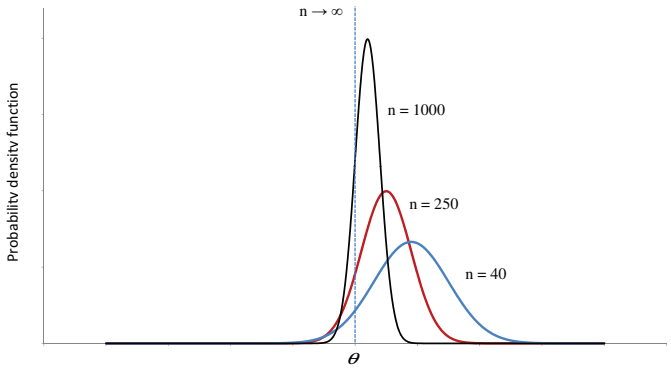

Fig. 1.6 The estimator is biased for small samples, but consistent.

1.9.1 Consistency

An estimator $\hat{\theta}$ is said to be consistent if its bias becomes smaller as the sample size grows large. Consistency is important because many of the most common estimators used in econometrics are biased, then the minimum we should expect from these estimators is that the bias becomes small as we are able to obtain larger data sets. Figure 1.6 illustrates the concept of consistency by showing how an estimator of the population parameter θ becomes unbiased as $n \to \infty$.

1.9.2 Central limit theorem

Having normally distributed random variables is important because we can then construct, for example, confidence intervals for its mean. However, what if a random variable does not follow a normal distribution? The central limit theorem gives us the answer.

Central limit theorem. States the conditions under which the mean of a sufficiently large number of independent random variables (with finite mean and variance) will be approximate a normal distribution.

Hence, even if we do not know the underlying distribution of a random variable, we will still be able to construct confidence intervals that will be approximately valid. In a numerical example, let's assume that the random variable X follows a

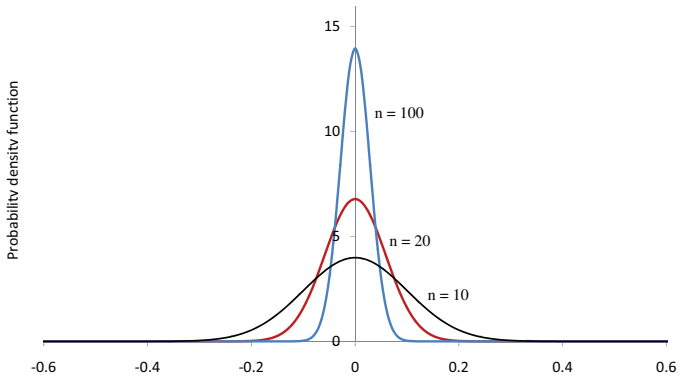

Fig. 1.7 Distribution of the sample mean of a uniform distribution.

uniform distribution [-0.5,0.5]. Hence, it is equally likely that this random variable takes any value within this range. Figure 1.7 shows the distribution of the average of this random variable for $n = 10$, 20, and 100. All of these three distributions look very close to a normal distribution.