2.2.1 *Expectation*

Definition 2.1 If x is a random variable which assumes values $x_1, x_2, ..., x_k$ with probabilities $p_1, p_2,..., p_k$ respectively with $\Sigma\ p_i = 1$, i= 1,2, ..., k, then *expectation* of the variable x is defined as

$$E\,(x) = \sum_{i=1}^{k} p_i\, x_i \qquad (2.1|)$$

Example 2.1
A fair die is rolled once. If x denotes the number on the upper face of the die, find expected value of x.

Solution
Here, the random variable x is the number on the upper face of the die. Thus, x can take any one of the values 1, 2,...,6, each with probability 1/6. Hence using (2.1),

$$\begin{aligned} E\,(x) &= \frac{1}{6}\,(1) + \frac{1}{6}\,(2) + \frac{1}{6}\,(3) + \frac{1}{6}\,(4) + \frac{1}{6}\,(5) + \frac{1}{6}\,(6) \\ &= 3.5 \ \blacksquare \end{aligned}$$

Result 2.1 Let a and b be two constants and x a random variable, then

$$E\ (ax+b) = aE(x)+b$$

Result 2.2 If x_1, x_2,...,x_k are random variables, and a_1, a_2,..., a_k are constants, then

$$E\ (a_1x_1+a_2\ x_2+\ldots+a_kx_k) = a_1\ E\ (x_1)+a_2E(x_2)+\ldots+a_kE\ (x_k)$$

Result 2.3 If x_1,x_2,...,x_k are k mutually independent random variables, then

$$E\ (x_1.x_2\ldots x_k) = E(x_1).E(x_2)\ldots E(x_k)$$

2.2.2 *Variance and Covariance*

If x is a random variable and E(x) its expected value, then variance of x is defined as

$$\left.\begin{aligned} V(x) &= E[x\text{-}E(x)]^2 \\ &= E(x^2)\text{-}[E(x)]^2 \end{aligned}\right] \tag{2.2}$$

When the investigator is interested in linear dependence of pairs of random variables, such as income x and expenditure y, then it is indicated by a measure called *covariance* of x and y. This term is defined as

$$Cov(x,y) = E[\{x\text{-}E(x)\}\{y\text{-}E(y)\}]$$

A zero value of the covariance indicates no linear dependence between x and y.

Example 2.2

For the experiment given in example 2.1, determine the variance of random variable x.

Solution

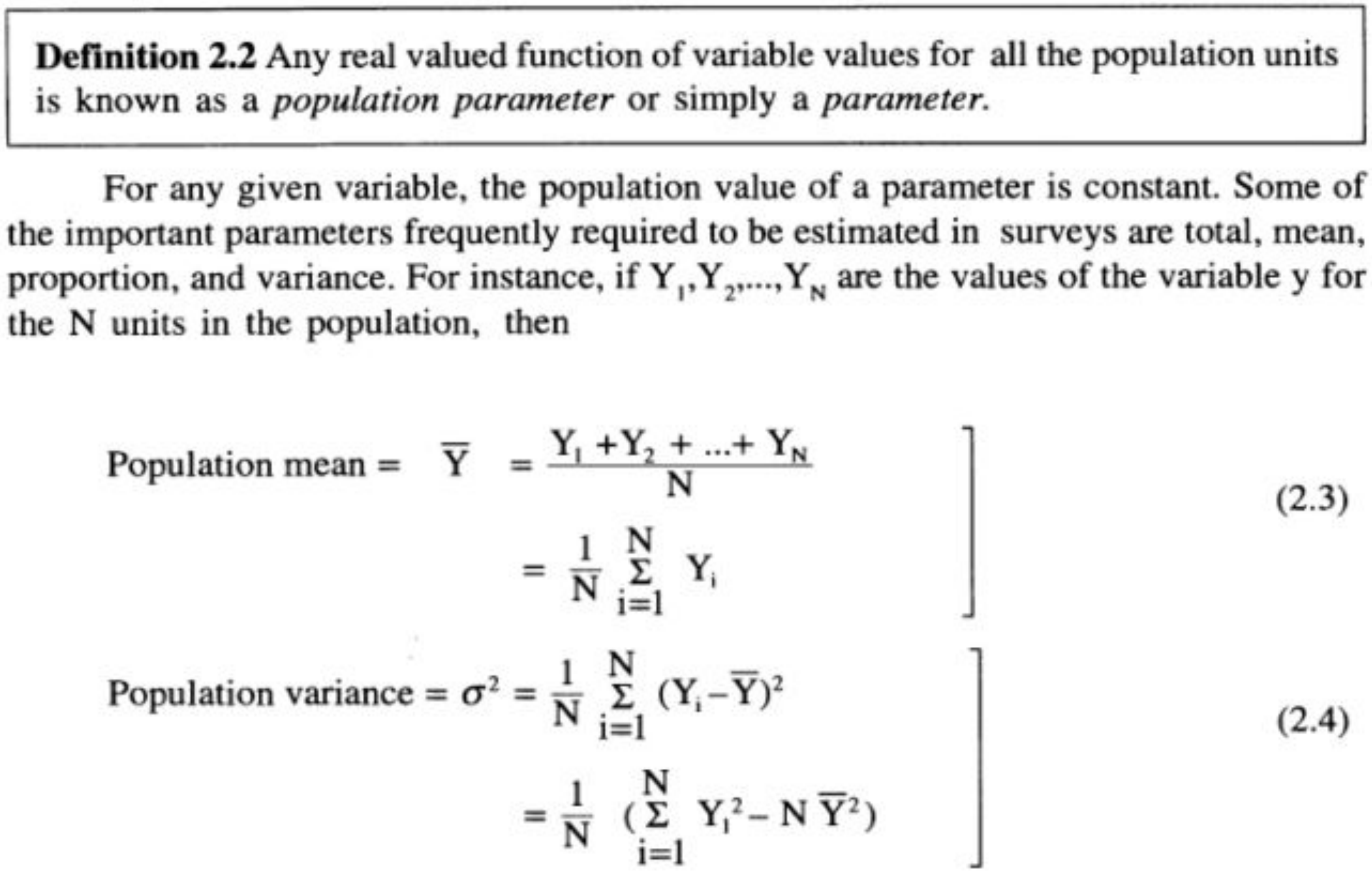
First we work out the term $E(x^2)$. This will be

$$
\begin{aligned}
E\,(x^2) &= \frac{1}{6}\,(1)^2 + \frac{1}{6}\,(2)^2 + \frac{1}{6}\,(3)^2 + \frac{1}{6}\,(4)^2 + \frac{1}{6}\,(5)^2 + \frac{1}{6}\,(6)^2 \\
&= \frac{91}{6} \\
&= 15.167
\end{aligned}
$$

On substituting in (2.2) the value of $E(x^2)$ obtained above, and of $E(x)$ from example 2.1, one gets

$$
\begin{aligned}
V(x) &= 15.167 - (3.5)^2 \\
&= 2.917 \blacksquare
\end{aligned}
$$

Population mean and population variance



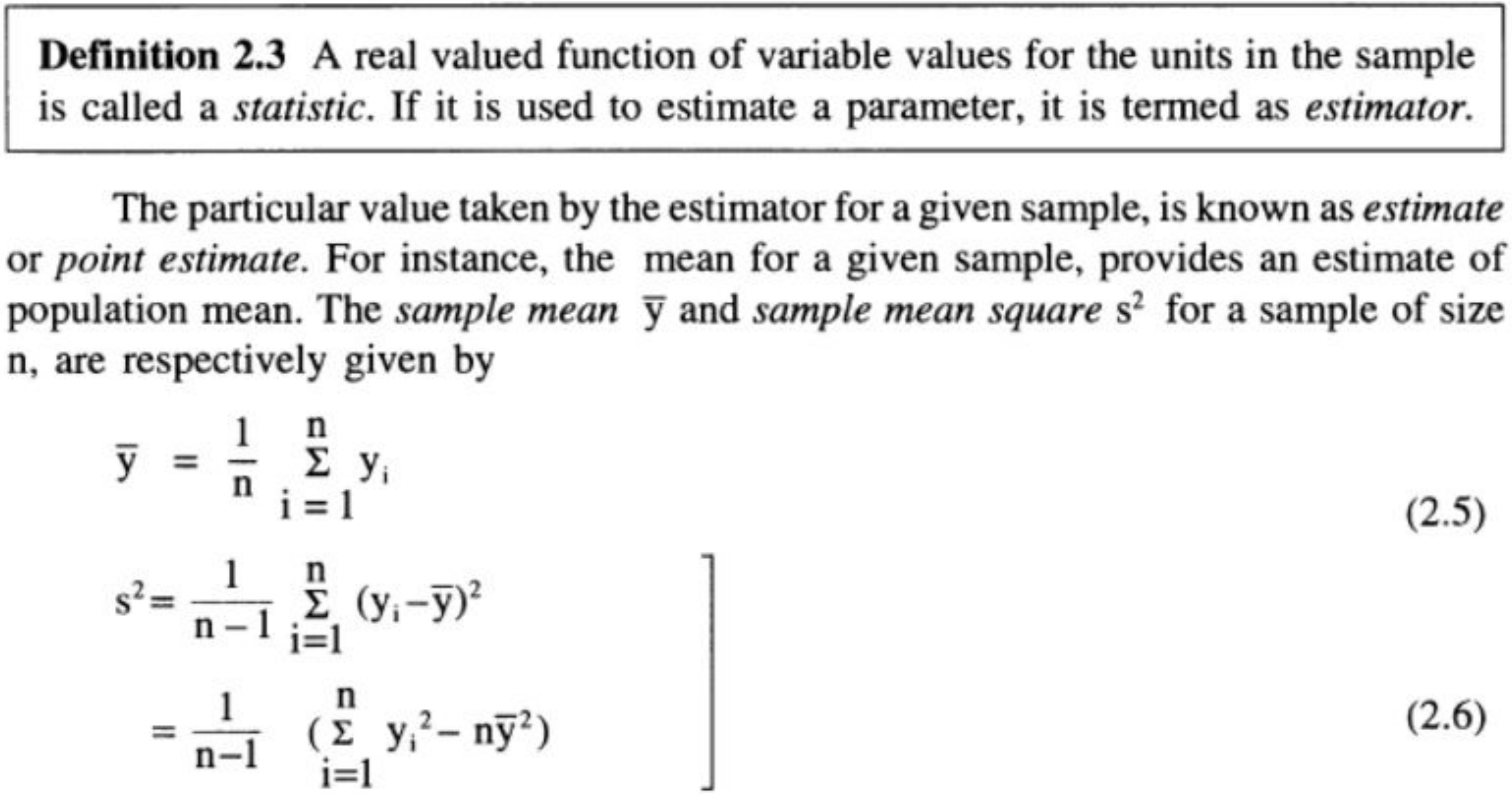
Definition 2.2 Any real valued function of variable values for all the population units is known as a *population parameter* or simply a *parameter.*

For any given variable, the population value of a parameter is constant. Some of the important parameters frequently required to be estimated in surveys are total, mean, proportion, and variance. For instance, if $Y_1, Y_2, \ldots, Y_N$ are the values of the variable y for the N units in the population, then

$$\left.\begin{aligned} \text{Population mean} = \overline{Y} &= \frac{Y_1 + Y_2 + \ldots + Y_N}{N} \\ &= \frac{1}{N} \sum_{i=1}^{N} Y_i \end{aligned}\right] \tag{2.3}$$

$$\left.\begin{aligned} \text{Population variance} = \sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \\ &= \frac{1}{N} \left(\sum_{i=1}^{N} Y_i^2 - N\,\overline{Y}^2\right) \end{aligned}\right] \tag{2.4}$$

Sample mean and sample variance(sample mean square):



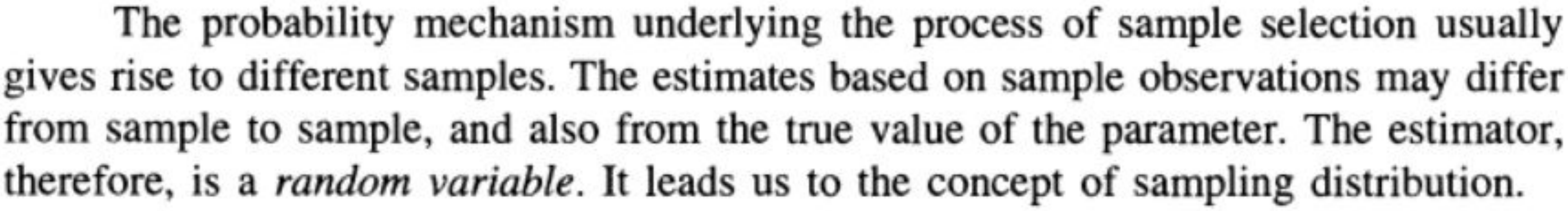
Definition 2.3 A real valued function of variable values for the units in the sample is called a *statistic*. If it is used to estimate a parameter, it is termed as *estimator*.

The particular value taken by the estimator for a given sample, is known as *estimate* or *point estimate*. For instance, the mean for a given sample, provides an estimate of population mean. The *sample mean* $\bar{y}$ and *sample mean square* s^2 for a sample of size n, are respectively given by

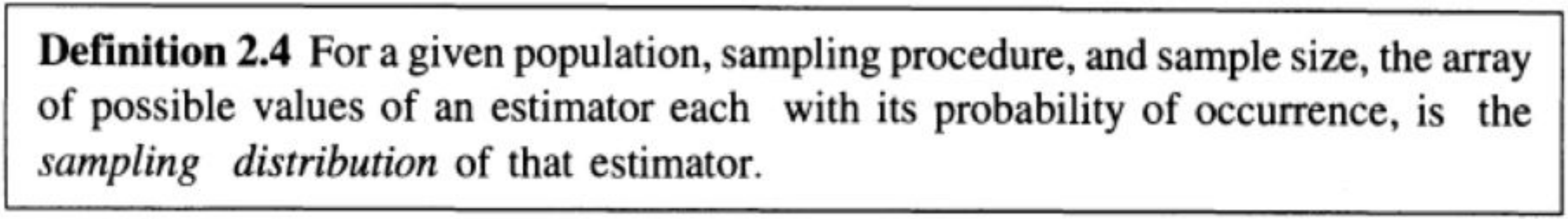
$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.5}$$

$$\left.\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= \frac{1}{n-1} \Big(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \Big) \end{aligned}\right] \tag{2.6}$$

Sampling Distribution



The probability mechanism underlying the process of sample selection usually gives rise to different samples. The estimates based on sample observations may differ from sample to sample, and also from the true value of the parameter. The estimator, therefore, is a *random variable*. It leads us to the concept of sampling distribution.



Definition 2.4 For a given population, sampling procedure, and sample size, the array of possible values of an estimator each with its probability of occurrence, is the *sampling distribution* of that estimator.

From the central limit theorem, the sample mean follows the normal distribution provided the sample size is large. As a result,

$$\frac{\bar{y}-\bar{Y}}{\text{S.E.}(\bar{y})} = \frac{\bar{y}-\bar{Y}}{\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}}}$$

approximately follows the standard normal distribution with mean zero and variance unity.

Example 2.3

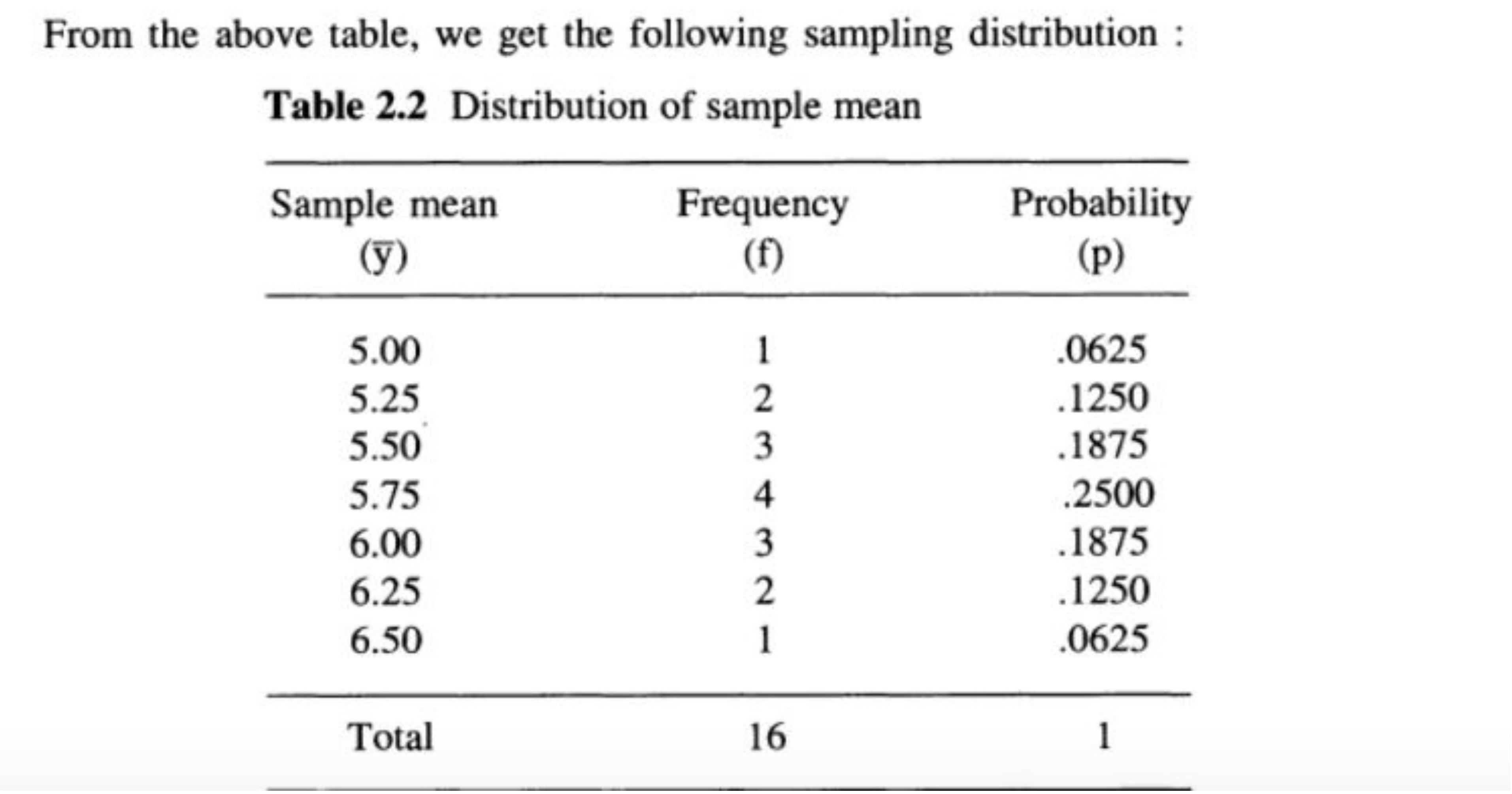
Four cows in a household marked A, B, C, and D respectively yield 5.00, 5.50, 6.00, and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield based on samples of n=2 cows, when the cows are selected with equal probabilities and WR. The procedure for drawing a sample has been explained in chapter 3.

Solution

Here N=4 and n=2. The number of possible samples in this case will be $4^2=16$. The mean of each sample, along with the cows included in the sample is given in table 2.1.

Table 2.1 Means of all possible samples

Sample	Cows in the sample	Sample mean $\bar{y}$	Sample	Cows in the sample	Sample mean $\bar{y}$
1	A, A	5.00	9	C, A	5.50
2	A, B	5.25	10	C, B	5.75
3	A, C	5.50	11	C, C	6.00
4	A, D	5.75	12	C, D	6.25
5	B, A	5.25	13	D, A	5.75
6	B, B	5.50	14	D, B	6.00
7	B, C	5.75	15	D, C	6.25
8	B, D	6.00	16	D, D	6.50

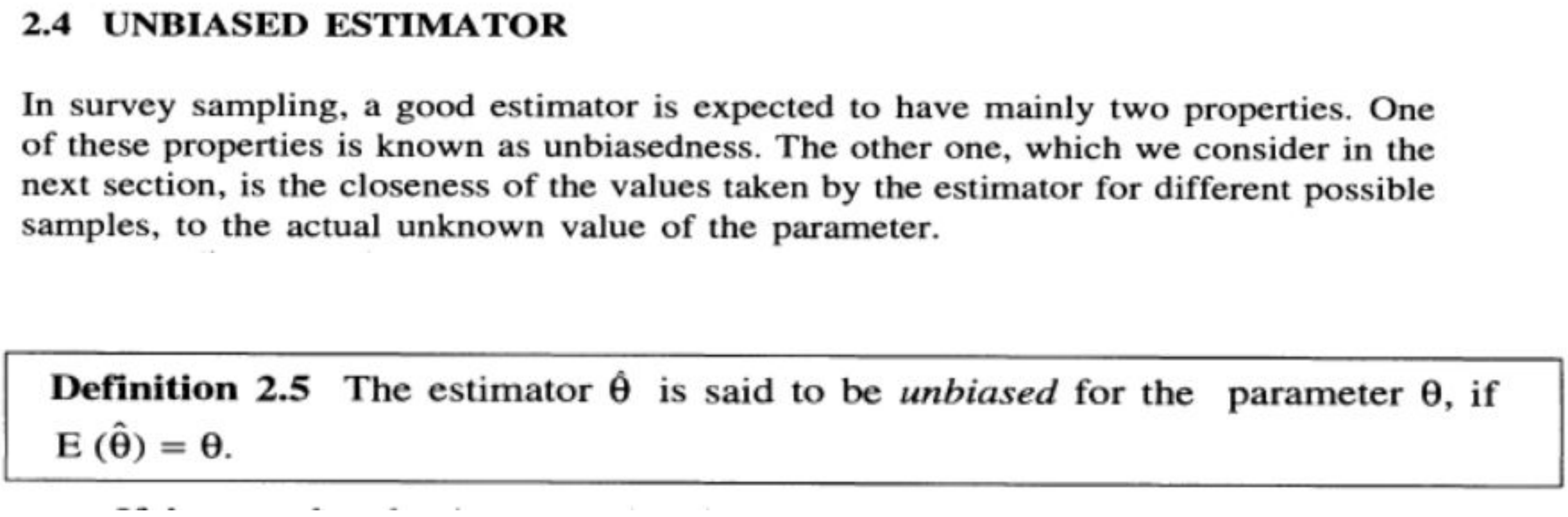


From the above table, we get the following sampling distribution :

Table 2.2 Distribution of sample mean

Sample mean ($\bar{y}$)	Frequency (f)	Probability (p)
5.00	1	.0625
5.25	2	.1250
5.50	3	.1875
5.75	4	.2500
6.00	3	.1875
6.25	2	.1250
6.50	1	.0625
Total	16	1

Properties of the estimator: 1/ Unbiased Estimator



2.4 UNBIASED ESTIMATOR

In survey sampling, a good estimator is expected to have mainly two properties. One of these properties is known as unbiasedness. The other one, which we consider in the next section, is the closeness of the values taken by the estimator for different possible samples, to the actual unknown value of the parameter.

Definition 2.5 The estimator $\hat{\theta}$ is said to be *unbiased* for the parameter θ, if $E(\hat{\theta}) = \theta$.

Example 2.4

For the data in example 2.3, verify whether the sample mean based on n=2 cows is an unbiased estimator for the average milk yield in the population ? Assume that the cows in the sample are selected with equal probabilities and with replacement.

Solution

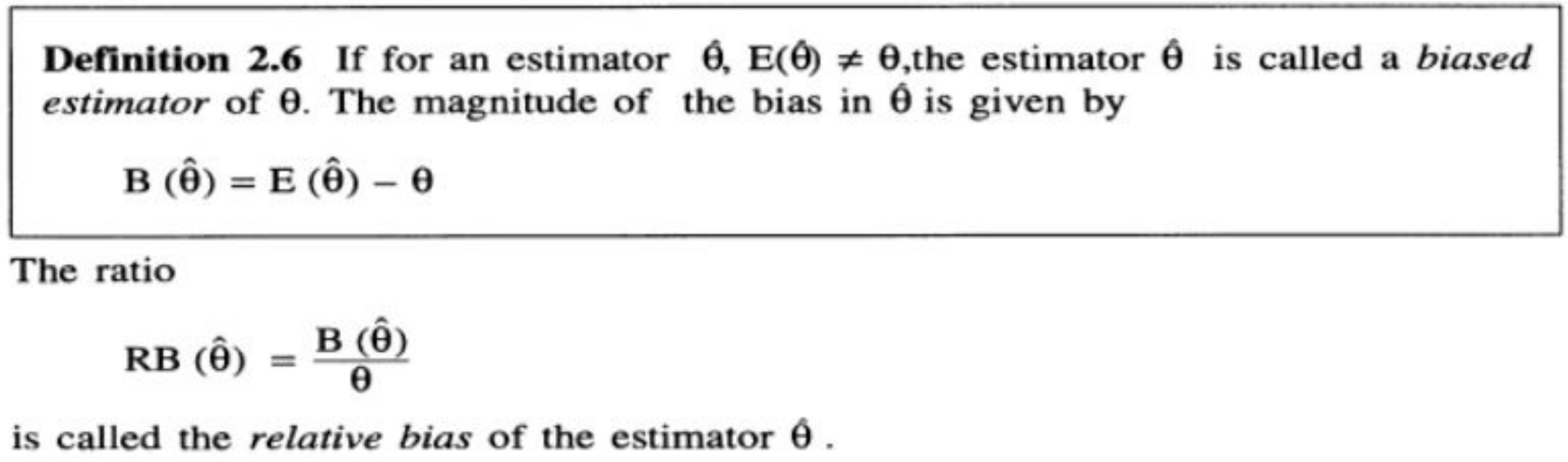
In table 2.1 are given the sample mean values for 16 possible samples of size n=2 that can be selected from a population of size N=4 with equal probabilities and with replacement. Here, sampling procedure ensures that all these samples have same chance of being selected. Hence, expected value of sample mean $\overline{y}$ will be the simple average of 16 possible sample mean values. Thus,

$$E(\overline{y}) = \frac{1}{16}(5.00+5.25+...+6.50)$$

$$= \frac{92.00}{16}$$

$$= 5.75$$

Using (2.3), it can be easily seen that the population mean $\overline{Y}$ (the parameter θ in this case) is also equal to 5.75. Hence, the sample mean $\overline{y}$ is unbiased for the population mean $\overline{Y}$. ■



Definition 2.6 If for an estimator $\hat{\theta}$, $E(\hat{\theta}) \neq \theta$,the estimator $\hat{\theta}$ is called a *biased estimator* of θ. The magnitude of the bias in $\hat{\theta}$ is given by

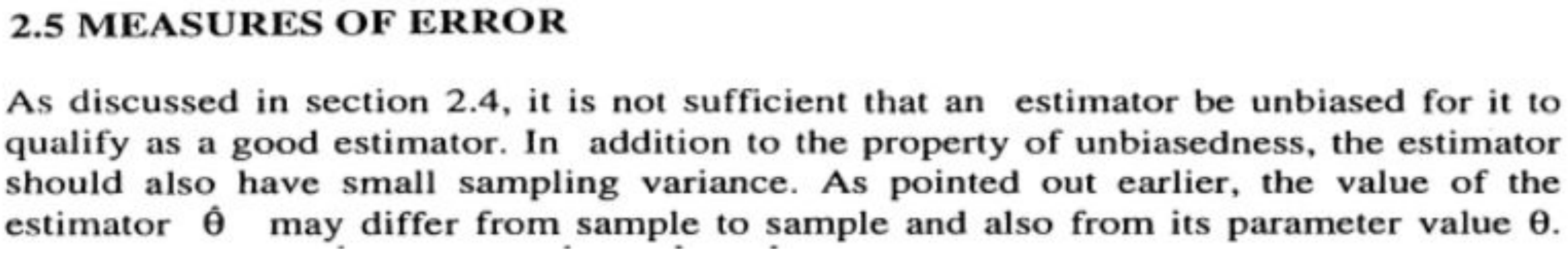
$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

The ratio

$$RB(\hat{\theta}) = \frac{B(\hat{\theta})}{\theta}$$

is called the *relative bias* of the estimator $\hat{\theta}$.

Properties of the estimator: 2/ Smaller variance



2.5 MEASURES OF ERROR

As discussed in section 2.4, it is not sufficient that an estimator be unbiased for it to qualify as a good estimator. In addition to the property of unbiasedness, the estimator should also have small sampling variance. As pointed out earlier, the value of the estimator $\hat{\theta}$ may differ from sample to sample and also from its parameter value θ.

2.5.1 *Sampling Variance*

Definition 2.7 The *sampling variance* is a measure of the divergence of the estimator values from its expected value. Alternatively, it is the variance of the sampling distribution of an estimator. In the light of (2.1) and (2.2), one gets it as

$$\begin{aligned} V(\hat{\theta}) &= E\,[\hat{\theta} - E\,(\hat{\theta})]^2 \\ &= E\,(\hat{\theta})^2 - [E\,(\hat{\theta})]^2 \end{aligned}$$

Example 2.5

For the data in example 2.3, compute the sampling variance and standard error of sample mean for the WR equal probability samples of size 2.

Solution

The sampling distribution of mean $\overline{y}$ (here the estimator $\hat{\theta}$ is the sample mean $\overline{y}$) for n=2 has been obtained in example 2.3. Also from example 2.4, $E(\overline{y}) = 5.75$. Thus from table 2.1, we have

$$V\,(\overline{y}) = \frac{1}{16}\,[(5.00)^2 + (5.25)^2 + \ldots + (6.50)^2] - (5.75)^2$$

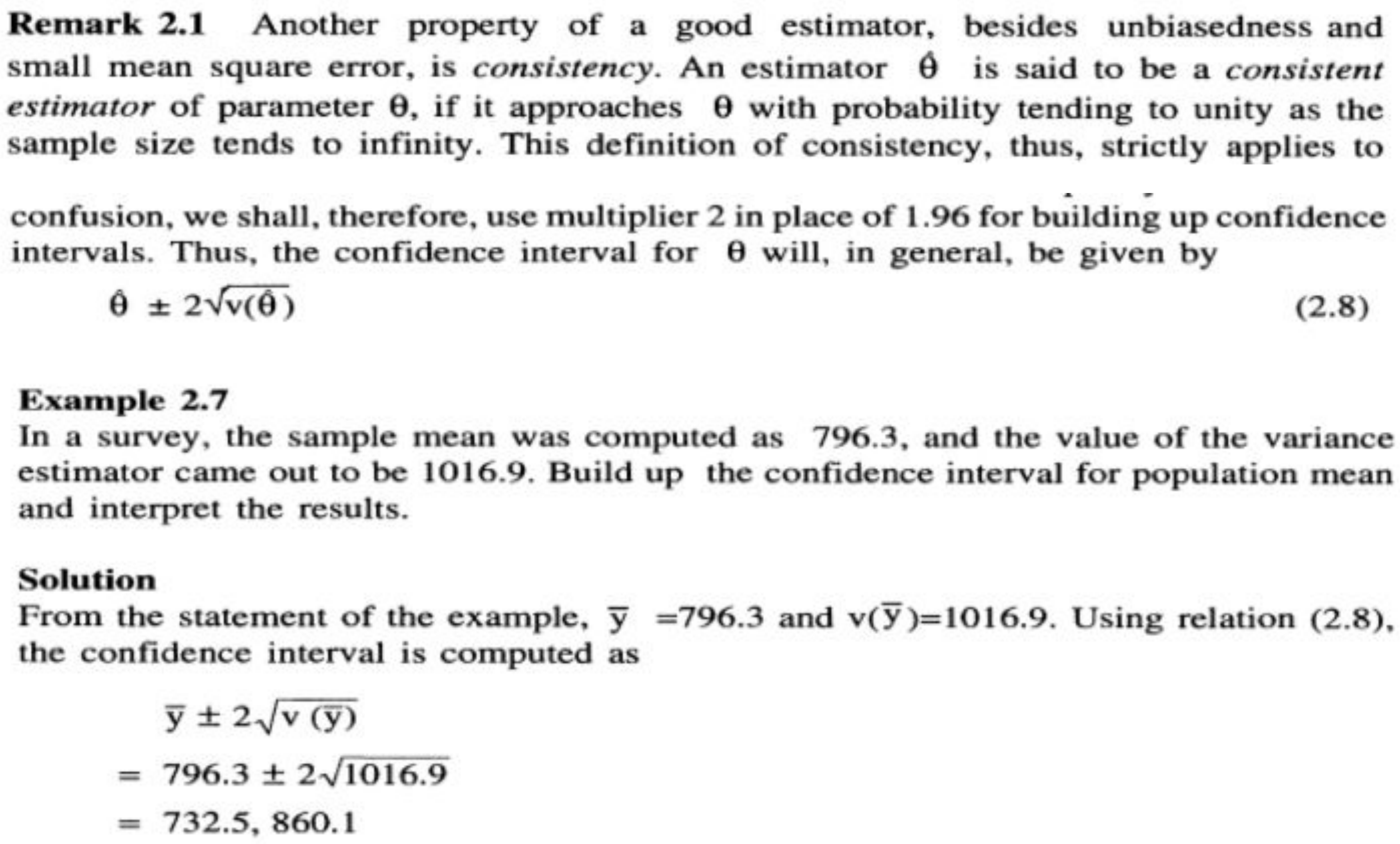
which from table 2.2, is equivalent to

$$\begin{aligned} V(\overline{y}) &= \frac{1}{16}\,[(5.00)^2\,(1) + (5.25)^2\,(2) + \ldots + (6.50)^2\,(1)] - (5.75)^2 \\ &= .15625 \end{aligned}$$

It gives

$$\begin{aligned} SE(\overline{y}) &= \sqrt{.15625} \\ &= .39528. \blacksquare \end{aligned}$$

Properties of the estimator: 3/ Consistency



Remark 2.1 Another property of a good estimator, besides unbiasedness and small mean square error, is *consistency*. An estimator $\hat{\theta}$ is said to be a *consistent estimator* of parameter θ, if it approaches θ with probability tending to unity as the sample size tends to infinity. This definition of consistency, thus, strictly applies to

confusion, we shall, therefore, use multiplier 2 in place of 1.96 for building up confidence intervals. Thus, the confidence interval for θ will, in general, be given by

$$\hat{\theta} \pm 2\sqrt{v(\hat{\theta})} \qquad (2.8)$$

Example 2.7
In a survey, the sample mean was computed as 796.3, and the value of the variance estimator came out to be 1016.9. Build up the confidence interval for population mean and interpret the results.

Solution
From the statement of the example, $\bar{y}$ =796.3 and $v(\bar{y})$=1016.9. Using relation (2.8), the confidence interval is computed as

$$\bar{y} \pm 2\sqrt{v\,(\bar{y})}$$

$$= 796.3 \pm 2\sqrt{1016.9}$$

$$= 732.5,\ 860.1$$

2.7 SAMPLE SIZE DETERMINATION

One of the important aspects in planning a sample survey is to decide about the size of the sample required for estimating the population parameter with a specified precision. The maximum difference between the estimate and the parameter value that can be tolerated on considerations of loss or gain due to policy decisions based on the sample results is termed as *permissible error, tolerable error*, or the *bound on the error of estimation*. Once the permissible error has been specified, the next objective is to determine a sample size that meets these requirements. Since the amount of error differs from sample to sample, the margin of error is specified by the probability statement

$$P[\,|\hat{\theta} - \theta\,| < B] = 1\text{-}\alpha \qquad (2.10)$$

where (1-α) may be taken as 95%, 99%, or some other desired level of confidence, and B is the permissible error. Sometimes, the permissible error is specified in terms of percentage of the value of parameter θ. Such a specification of permissible error can,

2.8 SAMPLING AND NONSAMPLING ERRORS

The probability mechanism inherent in the sampling procedure usually selects different units in different samples. The estimates based on the sample observations, as already discussed, will, therefore, differ in general from sample to sample and also from the value of the parameter under consideration. The resultant discrepancy between the sample estimate and the population parameter value is the error of the estimate. Such an error is inherent and unavoidable in any and every sampling scheme, and is termed *sampling error*. This error, however, has the favorable characteristic of being controllable through the size and design of the sample. This kind of error usually decreases with increase in sample size, and shall theoretically become nonexistent in case of complete enumeration. In many situations, the decrease is inversely proportional to the square root of sample size (figure 2.1).

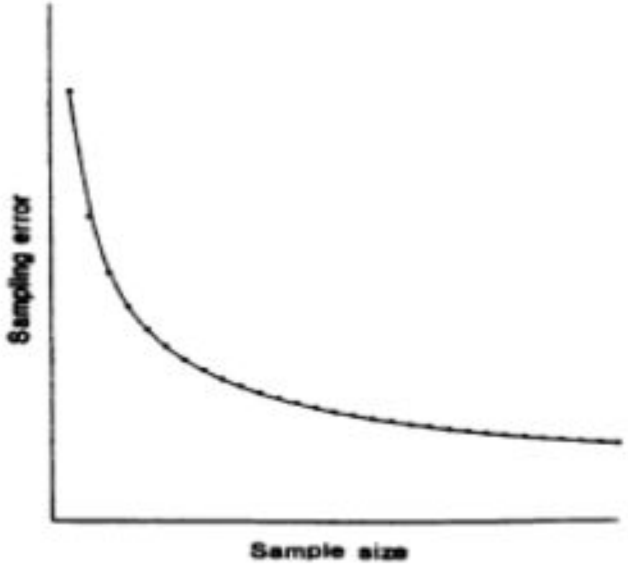
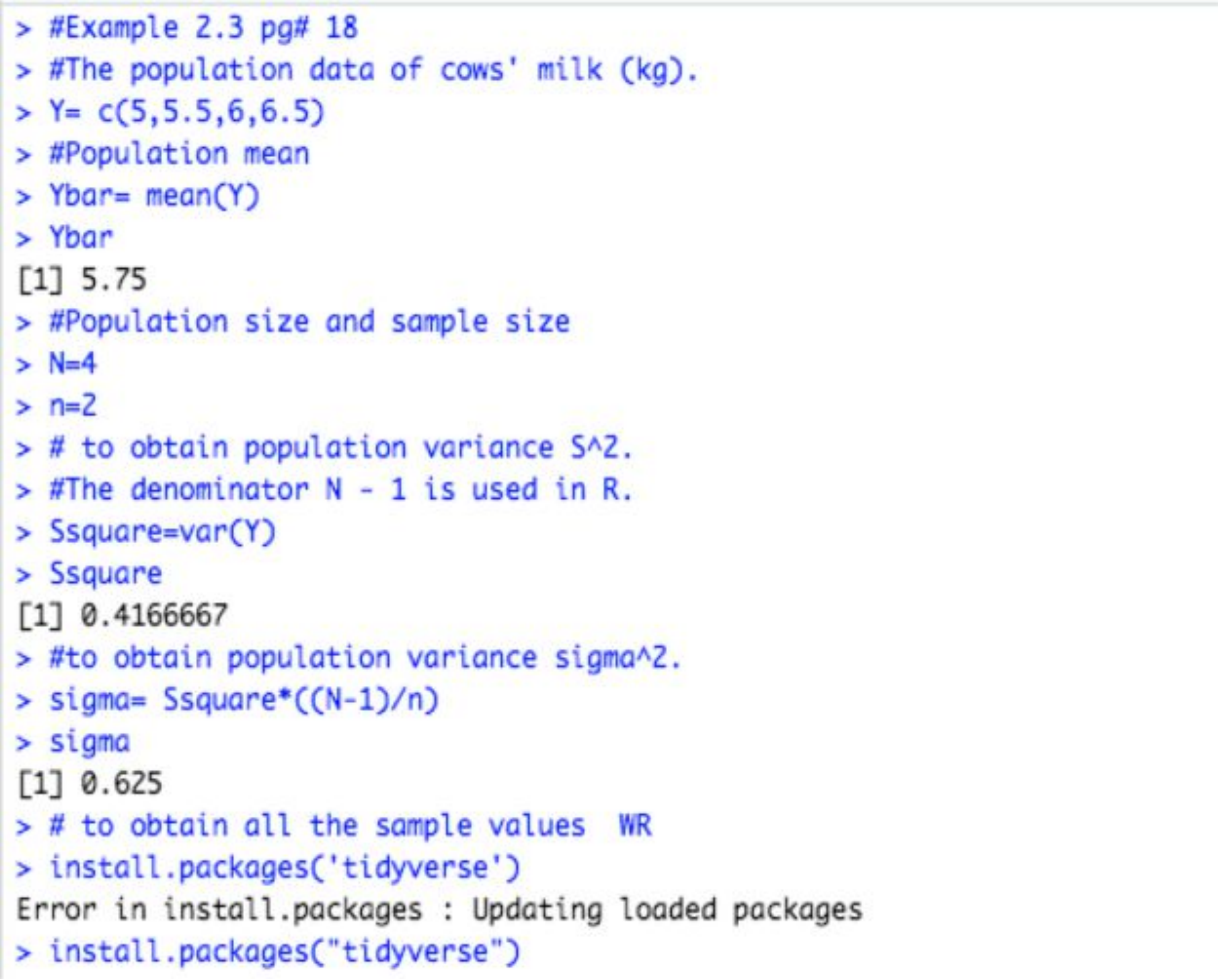
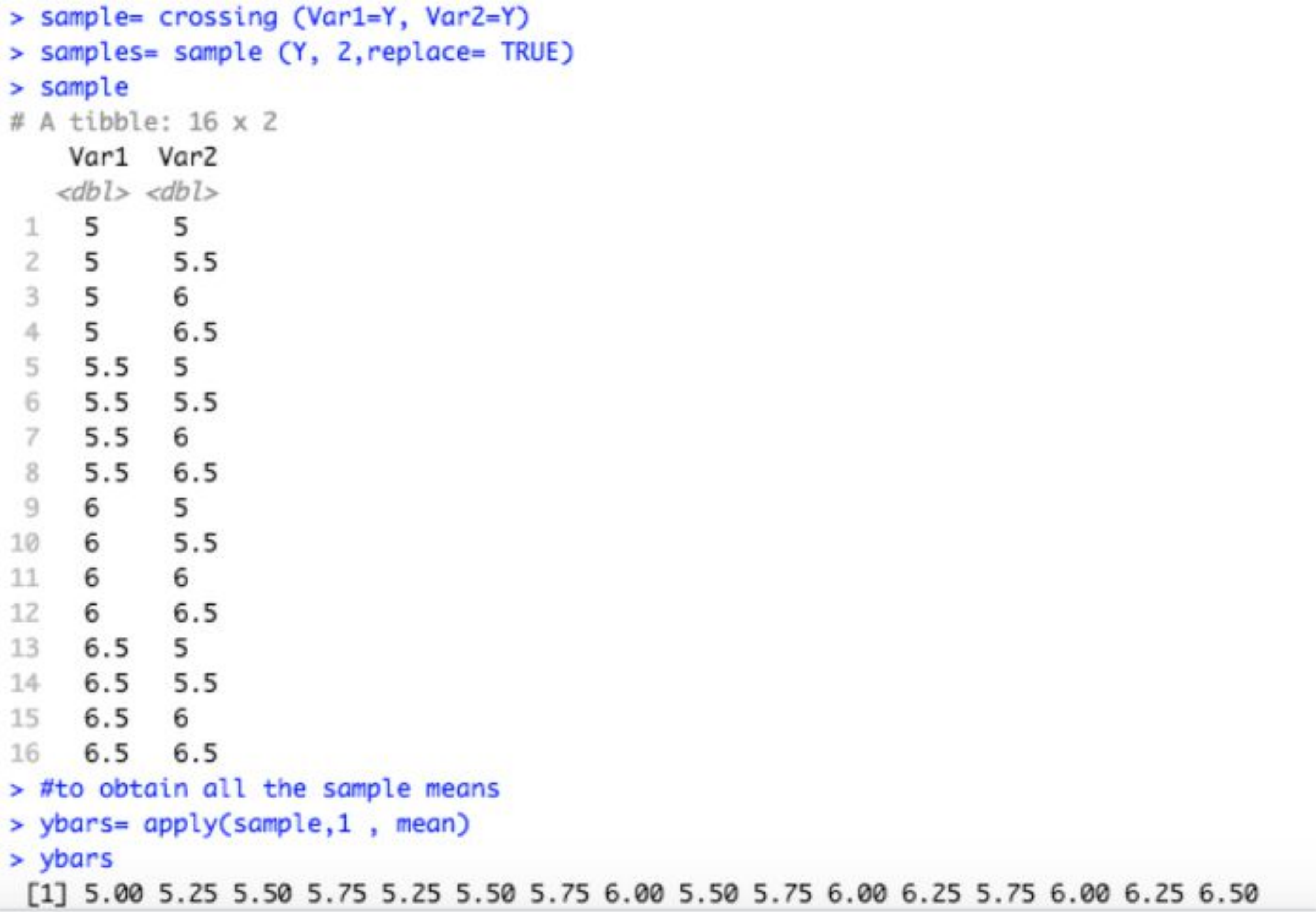


Fig 2.1 Relationship between sampling error and sample size

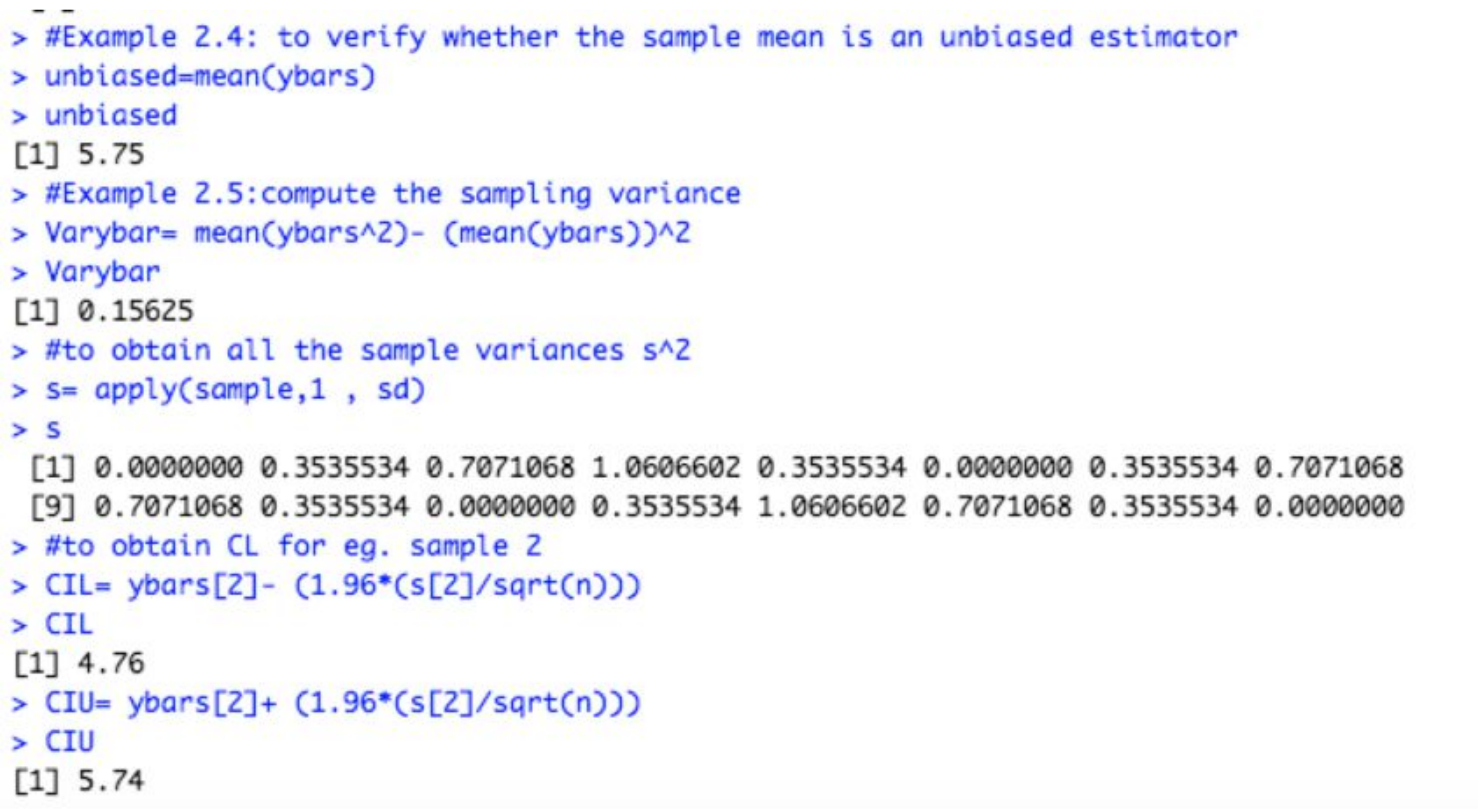
In any survey study, besides sampling error, there are also errors arising due to defective sampling procedures, ambiguity in definitions, faulty measurement techniques, mistakes in recording, errors in coding-decoding, tabulation and analysis, etc. These errors are known as *nonsampling errors*. For instance, data may be wrongly fed to the computer, or decimal places may be inadvertently changed. If the analysis of data requires transformation of per acre yield to per hectare basis, the multiplier used might be wrong.



```
> #Example 2.3 pg# 18
> #The population data of cows' milk (kg).
> Y= c(5,5.5,6,6.5)
> #Population mean
> Ybar= mean(Y)
> Ybar
[1] 5.75
> #Population size and sample size
> N=4
> n=2
> # to obtain population variance S^2.
> #The denominator N - 1 is used in R.
> Ssquare=var(Y)
> Ssquare
[1] 0.4166667
> #to obtain population variance sigma^2.
> sigma= Ssquare*((N-1)/n)
> sigma
[1] 0.625
> # to obtain all the sample values  WR
> install.packages('tidyverse')
Error in install.packages : Updating loaded packages
> install.packages("tidyverse")
```



```
> sample= crossing (Var1=Y, Var2=Y)
> samples= sample (Y, 2,replace= TRUE)
> sample
# A tibble: 16 x 2
    Var1  Var2
   <dbl> <dbl>
 1   5     5
 2   5     5.5
 3   5     6
 4   5     6.5
 5   5.5   5
 6   5.5   5.5
 7   5.5   6
 8   5.5   6.5
 9   6     5
10   6     5.5
11   6     6
12   6     6.5
13   6.5   5
14   6.5   5.5
15   6.5   6
16   6.5   6.5
> #to obtain all the sample means
> ybars= apply(sample,1 , mean)
> ybars
 [1] 5.00 5.25 5.50 5.75 5.25 5.50 5.75 6.00 5.50 5.75 6.00 6.25 5.75 6.00 6.25 6.50
```



```
> #Example 2.4: to verify whether the sample mean is an unbiased estimator
> unbiased=mean(ybars)
> unbiased
[1] 5.75
> #Example 2.5:compute the sampling variance
> Varybar= mean(ybars^2)- (mean(ybars))^2
> Varybar
[1] 0.15625
> #to obtain all the sample variances s^2
> s= apply(sample,1 , sd)
> s
 [1] 0.0000000 0.3535534 0.7071068 1.0606602 0.3535534 0.0000000 0.3535534 0.7071068
 [9] 0.7071068 0.3535534 0.0000000 0.3535534 1.0606602 0.7071068 0.3535534 0.0000000
> #to obtain CL for eg. sample 2
> CIL= ybars[2]- (1.96*(s[2]/sqrt(n)))
> CIL
[1] 4.76
> CIU= ybars[2]+ (1.96*(s[2]/sqrt(n)))
> CIU
[1] 5.74
```

HW

Page 28, Do questions 2.4, 2.9, 2.12, 2.22.