This is one of a series of papers by teachers of social science research methodology in universities in Asia. It reflects their view that existing materials need to be supplemented in several important areas to be of most value to their students. The authors do not argue a need for basically different social science survey research methodology in their countries. But they are not content to say that simply adding local examples would make non-Asian materials adequate to their needs. They have tried to identify some concepts or subjects which they believe need to be treated more fully or given a different emphasis for their own students than for the usual audience of the non-Asian textbook.

Papers published so far in this series include:

Research Designs and Strategies
(Vimal Shah, Gujarat University, Ahmedabad, India)

Guidelines for Preparing Research Proposals
(Haider Ali Chaudhari, West Pakistan Agricultural University, Lyallpur, Pakistan)

Research and the Action Program
(Gelisa Castillo, University of the Philippines, Los Banos)

Reporting Research
(Vimal Shah, Gujarat University, Ahmedabad, India)

Each of these papers is a preliminary draft. This means that the authors want it tested by teachers in the classroom as a way of improving it before it appears in a more finished form. This also means that they will welcome your comments and suggestions.

The full list of teachers engaged in this effort to improve teaching materials includes:

Dr. Randolph Barker
International Rice Research Institute
Los Banos, Philippines

Dr. S.A. Qadir
Bangladesh Agr. University
Mymenung, Bangladesh

Dr. Gelisa Castillo
Department of Agr. Education
University of the Philippines
College Laguna, Philippines

Dr. Ralph Retzlaff
Dept. of Political Science
University of California
Berkeley, Calif. 94720
U.S.A.

Dr. Haider Ali Chaudhari
Faculty of Agr. Economics & Rural Sociology
West Pakistan Agr. University
Lyallpur, Pakistan

Dr. Fred Rumawas
Institut Pertanian Bogor
Bogor, Indonesia

Dr. William Collier
Agricultural Development Council
P.O. Box 62
Bogor, Indonesia

Prof. William H. Sewell
Department of Sociology
University of Wisconsin
Madison, Wisconsin 53706
U.S.A.

Prof. Mary Hollmsteiner
Inst. of Philippine Culture
Ateneo de Manila University
P.O. Box 154
Manila, Philippines

Dr. Vimal P. Shah
Department of Sociology
Gujarat University
Ahmedabad, India

Dr. Andi Hakim Nasution
Institut Pertanian Bogor
Bogor, Indonesia

Dr. Sopin Tongsan
Department of Agricultural Economics
Kasetsart University
Bangkok, Thailand
1. POPULATIONS AND SAMPLES

To get an idea whether she has put together the right amount of each of the ingredients in the chicken-curry she is preparing for her family, a housewife scoops out a spoonful to taste. But if she is a good cook, first she stirs the pot before taking a "sample". "Sampling" a pot of curry to decide whether it needs more pepper or salt is a procedure every housewife knows, and it is one of the best examples of some important principles of sampling.

The "population" in this case is the pot of chicken-curry. The "sample" is the spoonful the housewife scooped from the pot. She stirred the curry first, before she drew the sample, to make sure of an equal distribution of the ingredients. This was her way of trying to make sure that the spoonful she tasted was a good representative of the content of the pot.

Generally, how should we sample within a population to get the maximum amount of information at a cost in time and money that we can afford?

Let us consider three equally sized populations of N = 4 households each. Let us suppose that with respect to land ownership in hectares, these populations have the following make up:

<table>
<thead>
<tr>
<th>POPULATION I</th>
<th>POPULATION II</th>
<th>POPULATION III</th>
</tr>
</thead>
<tbody>
<tr>
<td>household no.: (i)</td>
<td>land owned in ha (y)</td>
<td>household no.: (i)</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
If we draw one household at random, i.e. with equal chance, from a population, the probability of obtaining a household owning $Y$ hectares of land is:

\[ P(Y = y) = \begin{cases} 
\frac{1}{4}, & y = 2, 4, 6, 8 \\
0, & \text{otherwise} 
\end{cases} \quad \text{for Population I,} \]

\[ P(Y = y) = \begin{cases} 
\frac{1}{2}, & y = 5 \\
0, & \text{otherwise} 
\end{cases} \quad \text{for Population II,} \]

and

\[ P(Y = y) = \begin{cases} 
1, & y = 5 \\
0, & \text{otherwise} 
\end{cases} \quad \text{for Population III.} \]

The true means and variances are then for:

Population I: \[ \mu = E(Y) = \frac{1}{4}(2 + 4 + 6 + 8) = 5, \]
\[ \sigma_Y^2 = \frac{1}{4}[(2 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (8 - 5)^2] = 5; \]

Population II: \[ \mu = \frac{1}{4}(4 + 6) + \frac{1}{2}(5) = 5, \]
\[ \sigma_Y^2 = \frac{1}{4}[(4 - 5)^2 + (6 - 5)^2] + \frac{1}{2}(5 - 5)^2 = .5; \]

Population III: \[ \mu = 1(5) = 5 \]
\[ \sigma_Y^2 = 1(5 - 5)^2 = 0. \]

In other words, although these three populations all have the same mean, they differ greatly on variance, Population I having the largest variance and Population III the smallest.

Now, suppose we decide to draw a sample of three households at random from Population I. We will then face four equally likely cases, i.e. the sample we have drawn will consist of households 1, 2 and 3 or households 1, 2 and 4 or households 1, 3 and 4 or households 2, 3 and 4.

From a sample of size $n$ with observed values $y_1, y_2, \ldots, y_n$, we could construct two special functions:
(a) the sample mean:
\[ \bar{y} = \frac{1}{n} (y_1 + y_2 + \ldots + y_n) \]
\[ = \frac{1}{n} \sum_{i=1}^{n} y_i \]
\[ = \bar{y}/n. \]

(b) the sample variance:
\[ s_Y^2 = \frac{1}{n-1} \left[ (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 \right] \]
\[ = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2. \]

In case the sample turns out to be made up of households 1, 2 and 3 from Population I, the respective observed values are 4, 6 and 8. Therefore for Population I, \( \bar{y} = \frac{1}{3} (4 + 6 + 8) = 6 \) and

\[ s_Y^2 = \frac{1}{2} (4 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 = 4. \] Calculating these entities for the other possible samples that might be drawn from Population I we arrive at the following results:

Sample size \( n = 3 \)

<table>
<thead>
<tr>
<th>Composition of sample (household no.)</th>
<th>Observed values</th>
<th>( \bar{y} )</th>
<th>( s_Y^2 )</th>
<th>Probability of obtaining such a sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3</td>
<td>4, 6, 8</td>
<td>6</td>
<td>4</td>
<td>1/4</td>
</tr>
<tr>
<td>1, 2, 4</td>
<td>4, 6, 2</td>
<td>4</td>
<td>4</td>
<td>1/4</td>
</tr>
<tr>
<td>1, 3, 4</td>
<td>4, 8, 2</td>
<td>4 2/3</td>
<td>9 1/3</td>
<td>1/4</td>
</tr>
<tr>
<td>2, 3, 4</td>
<td>6, 8, 2</td>
<td>5 1/3</td>
<td>9 1/3</td>
<td>1/4</td>
</tr>
</tbody>
</table>
By drawing the sample randomly, we have an equal chance of obtaining any one of these four possible samples. This also implies an equal (1/4, or one chance out of four) chance that the mean for our sample will be any particular one of the sample means shown above.

Thus \( P(\bar{y} = 6) = P(\bar{y} = 4) = P(\bar{y} = 4\frac{1}{4}) = P(\bar{y} = 5\frac{1}{2}) = \frac{1}{4} \).

Therefore \( \bar{y} \) is also a random variable, derived from the random variable \( Y \), with the probability function

\[
P(Y = y) = \begin{cases} 
1/4, & y = 2, 4, 6, 8 \\
0, & \text{otherwise}
\end{cases}
\]

by taking a random sample of 3 units out of the 4 households available in the population. The probability function of \( \bar{y} \)

\[
P(\bar{y} = m) = \begin{cases} 
1/4, & m = 4, 4\frac{1}{4}, 5\frac{1}{2}, 6 \\
0, & \text{otherwise}
\end{cases}
\]

is therefore called a derived probability function of the probability function \( P(Y = y) \).

The true mean of the sample mean is:

\[
E(\bar{y}. \cdot) = \frac{1}{4}(4 + 4\frac{1}{4} + 5\frac{1}{2} + 6) = 5 = E(Y) = \mu_Y.
\]

We notice that for a random sample, the expected value of the sample mean is equal to the true mean of the original population. We say then that the sample mean of a random sample is an unbiased estimate of the true mean of the population.

The true variance of the sample mean \( \bar{y} \) is

\[
\sigma_{\bar{y}}^2 = \frac{1}{4}(4 - 5)^2 + (4\frac{1}{4} - 5)^2 + (5\frac{1}{2} - 5)^2 + (6 - 5)^2 = \frac{1}{4} \cdot \frac{1}{9} = \frac{1}{9} \sigma_Y^2.
\]

When we are concerned with a finite population, i.e. a population with a finite number of members, like our population with \( N = 4 \) households, the true variance could be written as:

\[
\sigma_Y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu_Y)^2.
\]
It is also customary to define the true variance of the population as:

\[ S_Y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \mu_Y)^2 = \frac{N}{N-1} \sigma_Y^2. \]  

(17)

such that

\[ \sigma_Y^2 = \frac{N-1}{N} S_Y^2. \]  

(18)

in case \( N = 4, \sigma_Y^2 = \frac{3}{4} S_Y^2 \). Utilizing this formula in the relation obtained previously for the variance of a sample mean (i.e., \( \sigma_i^2 = \frac{1}{9} \sigma_Y^2 \)) we have \( \sigma_i^2 = \frac{13}{9} S_Y^2 = \frac{S_Y^2}{4} \left( \frac{1}{3} \right) = \frac{S_Y^2}{4} (1 - \frac{3}{4}) \).

For the general case where the population is of size \( N \) and the random sample is of size \( n \), it could be shown that the true variance of the sample mean is:

\[ \sigma_i^2 = \frac{S_Y^2}{n (1 - \frac{n}{N})}. \]  

(19)

The ratio \( f = \frac{n}{N} \) is called the sampling fraction. The larger the population size with respect to the sample size, the smaller the value of \( f \), and the more \( \frac{N-1}{N} \) tends to the value of 1. Therefore, for large sized populations it would not make much difference whether we use \( S_Y^2 \) or \( \sigma_Y^2 \) and \( (1 - \frac{n}{N}) \) or 1. Hence, for \( N \) infinitely large:

\[ \sigma_i^2 = \frac{1}{n} \sigma_Y^2. \]  

(20)

From (19) and (20) we conclude that the true variance of the sample mean \( \bar{Y} \) obeys the following properties:

(a) it decreases with increasing sample size,

(b) it increases with an increase of the true variance of the original population from which the sample was derived.

The first property could be demonstrated by taking samples of size 2 and 1 from Population 1. The results could be seen in the
following tables:

<table>
<thead>
<tr>
<th>Composition of sample (household number)</th>
<th>Observed values</th>
<th>$\bar{y}$</th>
<th>$s_{\bar{y}}^2$</th>
<th>Probability of obtaining sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>4, 6</td>
<td>5</td>
<td>2</td>
<td>1/6</td>
</tr>
<tr>
<td>1, 3</td>
<td>4, 8</td>
<td>6</td>
<td>2.4</td>
<td>1/6</td>
</tr>
<tr>
<td>1, 4</td>
<td>4, 2</td>
<td>3</td>
<td>1.2</td>
<td>1/6</td>
</tr>
<tr>
<td>2, 3</td>
<td>6, 8</td>
<td>7</td>
<td>2.8</td>
<td>1/6</td>
</tr>
<tr>
<td>2, 4</td>
<td>6, 2</td>
<td>4</td>
<td>4</td>
<td>1/6</td>
</tr>
<tr>
<td>3, 4</td>
<td>8, 2</td>
<td>5</td>
<td>5</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Sample size $n = 1$

<table>
<thead>
<tr>
<th>Composition of sample (household number)</th>
<th>Observed value</th>
<th>$\bar{y} = y$</th>
<th>$s_{\bar{y}}^2$</th>
<th>Probability of obtaining sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>1/4</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>6</td>
<td>-</td>
<td>1/4</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8</td>
<td>-</td>
<td>1/4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>1/4</td>
</tr>
</tbody>
</table>

If we are drawing only one single item our sample size is $n = 1$. The sample mean $\bar{y}$ reduces into $y$ and $\sigma_{\bar{y}}^2 = \sigma_y^2$. The sample variance becomes undefined because the denominator $(n - 1)$ is zero.

If the sample we took is of size $n = N = 4$, then $\bar{y} = E(Y) = 5, \sigma_{\bar{y}}^2 = 0$. The sample mean will not vary if we repeat the sampling. In other words, observing all available data will always expose the true mean.

Comparing the sample means and variances we could obtain by sampling Population I. we notice that for $n = 4$, i.e. complete enumeration, the "sample" mean is a constant value, in this case
5. If \( n = 3 \), the sample mean will vary slightly, between 4 and 6. 
If \( n = 2 \), the range of the variation will be between 3 and 7, while for \( n = 1 \), the range will be equal to that existing in the population or between 2 and 8. This shows how the variability of the sample mean is reduced when a larger sample is drawn.

To show simultaneously how the variability of the sample mean is affected by sample size and population variance, we could calculate similarly, the sample mean for various sample sizes taken from Population II and Population III. The results, together with that for Population I, could be summarized as follows:

<table>
<thead>
<tr>
<th>Sample size ((n))</th>
<th>Range of variability of sample mean</th>
<th>Variance of sample mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population I</td>
<td>Population II</td>
</tr>
<tr>
<td>1</td>
<td>2 8</td>
<td>4 6</td>
</tr>
<tr>
<td>2</td>
<td>3 7</td>
<td>4( \frac{1}{2} )</td>
</tr>
<tr>
<td>3</td>
<td>4( \frac{1}{2} )</td>
<td>5( \frac{1}{2} )</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\( S^2 \) 20/3 2/3 0
\( \sigma^2 \) 5 1/2 0

We observe that the sample means show less variance as larger samples are drawn and also less variance when the population variance itself is lower.

In any case, however, it could be shown that the expected value of the mean of a random sample is always equal to the true population mean. The randomization procedure is thus a safeguard to ascertain that the sample will represent the population as closely as possible.

How can we be sure of getting a random sample? The most simple case we could face is when we have to choose randomly one out of two available items (in other words, when the population size is \( N = 2 \) and the sample size \( n = 1 \)).

People often toss a coin to help themselves out of this problem. We tacitly assume that the coin is fair.
If we have to choose from a population of size \( N = 6 \), we can use a fair die which could be rolled to decide—a six-sided cube with each of the numbers from 1 to 6 on a side. We would need to associate each side of the die with one member of the population.

For a population with size \( N = 10 \), we would need a fair die with ten sides. Even if it does exist, the die would not work, because such a die would not have two parallel sides. However, the Japanese have invented a die with twenty sides which are pairwise parallel to each other. Every time it is rolled, a side will appear parallel to the table top on which it will come down. Each of the numerals 1, 2, 3, \ldots, 9, 0 is imprinted on two parallel sides of the die, such that the probability of occurrence of any one of the numbers 1, 2, \ldots, 9 or 0 is \( \frac{1}{20} \) or \( \frac{1}{10} \). Therefore this twenty-sided die is effectively a ten-sided die.

This twenty-sided die is called a "Random Number Generating Icosahedron Die", and is produced by the Japanese Standards Association. A set of three, packed in a special plastic case, is for sale on the market.

If we roll such an icosahedron die repeatedly and note the results in the form of a table such as given on pages \( \ldots \) and \( \ldots \), we will obtain a Table of Random Numbers. The numerals 1, 2, 3, 4, 5, 6, 7, 8, 9 and 0 are interspersed randomly in the table. This means that each of those numerals appears with a relative frequency of about 1/10.

This kind of a table of random numbers can be used to simulate the results of rolling a ten-sided die. To choose a series of numbers at random, you simply start to read numbers from it at any randomly selected point and read in any previously agreed upon direction.

The table could also be used to simulate the results of rolling a hundred-sided die, numbered from 01 to 100, by reading the numerals in the table in pairs. The symbol "00" in the table will then denote the side of the die labeled "100". Similarly the rolling of a thousand-sided die could be simulated by reading the numerals in triplets and so on.

How could we use the table to select one item at random from a population of \( N = 3 \) items? If the table is to be useful at all, we should be able to use it to simulate any \( n \)-sided die—in this case a three-sided die.
The first step in doing so is to tag the three members of the population by numbering them from 1 to 3. Now we go to the table and, starting at a randomly chosen point, read single numbers. If we read a specific number in the table which is less than 4, say 2, then this implies that item number 2 has to be chosen. However, what do we do if a number higher than 4 appears?

We could discard these numbers from our consideration, but this implies that on the average we are throwing away 70% of the tabulated numbers. This is not a very efficient way of utilizing the table. Instead, we could consider number 4 to denote that item number \((4 - 3) = 1\) is to be chosen, number 5 to denote that item number \((5 - 3) = 2\) is to be chosen, number 6 to denote that item number \((6 - 3) = 3\) is to be chosen, number 7 to denote that item number \((7 - 2 \times 3) = 1\) is to be chosen, number 8 to denote that item number \((8 - 2 \times 3) = 2\) is to be chosen and number 9 to denote that item number \((9 - 2 \times 3) = 3\) is to be chosen.

Number 0, which represents "10" might be expected to tell us that item number \((10 - 3 \times 3) = 1\) has to be chosen. However, if we do not discard this number from our consideration the condition of randomness will be violated. Item number 1 would have a larger probability of being chosen than the other items, because the probability of occurrence of any one of the ten numerals is 1/10. The appearance of either 1, 4, 7 or 0 will lead to the drawing of item number 1 with the probability of 4/10, whereas items number 2 or number 3 will respectively be drawn with the probability of 3/10, because only three out of ten numerals will lead to the drawing of one of these items. Therefore, in utilizing the table of random numbers to draw a member of a population of size \(N = 3\), we have to discard 0 wherever it appears.

As a conclusion, we now could say that to utilize the table of random numbers to select one item from a population of 3 items, we should consider all numbers not larger than 10 and not larger than the largest multiple of 3, i.e. 9. To know which item has to be chosen after a number is read from the table, we divide the observed number by 3 and determine the remainder of the division. If the remainder is 1, then we choose item number 1, if the remainder is 0, i.e. when the number read from the table was an integral multiple of 3, we choose item number 3.
This rule can be extended for populations of any size. If \( n = 29 \) say, then we have to read from the table a pair of consecutively placed numerals, which could take on the values 01, 02, \ldots, 98, 99 or 00. In reading the table we discard any number larger than \( 3 \times 29 = 87 \), the largest multiple of 29 less than 100. The item to be chosen will then be determined by the remainder of the number read from the table after division by 29, just as was the case for \( N = 3 \).

There remains the question of where to start to read a table of random numbers. Our table in this book is very limited in size. It consists of only 5000 numerals tabulated on two sheets, each of which has 2500 numerals arranged in 50 rows and 50 columns.

We have numbered each row and column from 00 to 49. This is to help us randomize the starting point of reading. Otherwise we may tend to start somewhere near the same place in the table each time we use it, and thus get a specific sequence of numbers taken from specific locations, on a specific page. The result of our drawing would then not be random any more.

Suppose that we want to choose 15 items from a population of 29 items. Then we open either of the two available pages of the table of random numbers, say the first page. With closed eyes we put the point of a pencil on the table and then open our eyes. The number closest to the pencil point and the four subsequent numerals will determine the position where we have to start reading the table, to stimulate the results of throwing a twenty-nine-sided die. Suppose that the numeral in the 28th row and 7th column was pointed. We note the result. In this case it is: 767 31.

The first numeral is odd and tells us to use page number \((7 - 3 \times 2) = 1 \). The two subsequent numerals denote the row number where we have to start our reading. In this case it is row number \((67 - 50) = 17 \), while the column number is obviously number 31. We then read the following sequence of numerals from left to right continued to the next row:

\[
5613 \ 35670 \ 10549 \ 07468 \ 38148 \ 79001 \ 02509 \ 79424 \ 39625 \ 73315 \ \ldots \text{etc.}
\]

Now we write up this sequence of numerals, pairwise, to form numbers not larger than 100. From this list of numbers we subtract
the highest possible multiple of 29 to determine the remainder.

<table>
<thead>
<tr>
<th>number from table:</th>
<th>56</th>
<th>13</th>
<th>35</th>
<th>67</th>
<th>01</th>
<th>05</th>
<th>49</th>
<th>07</th>
</tr>
</thead>
<tbody>
<tr>
<td>multiple of 29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subtracted:</td>
<td>-29</td>
<td>-0</td>
<td>-29</td>
<td>-58</td>
<td>-0</td>
<td>-0</td>
<td>-29</td>
<td>-0</td>
</tr>
<tr>
<td>remainder:</td>
<td>27</td>
<td>13</td>
<td>6</td>
<td>9</td>
<td>1</td>
<td>5</td>
<td>20</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number from table:</th>
<th>46</th>
<th>83</th>
<th>81</th>
<th>48</th>
<th>79</th>
<th>00</th>
<th>10</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>multiple of 29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subtracted:</td>
<td>-29</td>
<td>-58</td>
<td>-58</td>
<td>-29</td>
<td>-58</td>
<td>X</td>
<td>-0</td>
<td>-0</td>
</tr>
<tr>
<td>remainder:</td>
<td>17</td>
<td>25</td>
<td>23</td>
<td>19</td>
<td>21</td>
<td>10</td>
<td>25</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number from table:</th>
<th>09</th>
<th>79</th>
<th>42</th>
<th>43</th>
<th>96</th>
</tr>
</thead>
<tbody>
<tr>
<td>multiple of 29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subtracted:</td>
<td>-0</td>
<td>-58</td>
<td>-29</td>
<td>-29</td>
<td>X</td>
</tr>
<tr>
<td>remainder:</td>
<td>9</td>
<td>21</td>
<td>13</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

The circled remainders denote the items that have to be drawn from the population. A cross (X) in place of a subtractant denotes that the number to be subtracted is larger than the largest multiple of 29 less than 100. A cross below a remainder denotes that it has already appeared previously. When sampling is done without replacement, a reappearing remainder is discarded.

In the previous discussions of this chapter, we have seen that the mean of a sample drawn from a population of random variables with a specified probability distribution is itself a random variable and has a probability distribution dependent on the probability distribution of the original random variable.
Nothing has been said in general about the nature of this derived probability distribution, except about the mean and variance. As we have seen, the true mean of a random sample is equal to the true mean $\mu$ of the original random variable and the true variance of the mean of a random sample with respect to the population variance $\sigma^2$ is inversely proportional to the sample size and directly proportional to $1 - \frac{n}{N} = \frac{N-n}{N}$, the portion of the population left out of the sample. For an infinite population this ratio tends to 1 and the sample variance could safely be assumed to be equal to $\frac{\sigma^2}{n}$.

The decrease in variance of the sample mean due to the increase in sample size implies that if $n$ is large, the probability that the sample mean will assume a value near the population mean $\mu$ is high. This phenomenon is called the Law of Large Numbers.

If we know that the random variable we are concerned with is normally distributed, then it could be shown that the mean of a random sample drawn from this population is a random variable which is also normally distributed, with true mean $\mu$ and true variance $\frac{\sigma^2}{n}$. However, what would happen if the random variable we drew our sample from is not normally distributed? It could be shown that even if the original population is not normally distributed, means of random samples drawn from that population tend to act as a normal random variable, provided the sample size is sufficiently large. This property of a sample mean is called the Central Limit Theorem. Its importance for us lies in the fact that we can almost always handle any random variable as a normal random variable provided we are working with means of observations.

The sample mean could be considered as a measure of central tendency of the observed values of a sample. The sample variance on the other hand is a measure of dispersion of the observed values of the sample. The observed values of a sample will scatter around the sample mean. The smaller the sample variance, the less will be the deviation of the observed values from the sample mean.

The sample median is another measure of central tendency. It is the value of the mid-item of a series of observed values arranged
in order of size. When the number of items is odd, the median is the value of the observation situated in the middle of the ranked items. If the number of items is even, the median is arbitrarily defined as the arithmetic average of the values of the middle observations.

For example, if the following values were observed: 9, 3, 4, 2, 6, 7, then after arranging in order of size, we have the series 2, 3, 3, 4, 6, 7, 9. The middle observation is 4. Therefore the median is 4. But if another value was observed in addition to these values, say 5, then the series will look as follows: 2, 3, 3, 4, 5, 6, 7, 9. The two middle values are 4 and 5. Therefore, the median in this case $\frac{4+5}{2} = 4\frac{1}{2}$.

Another measure of central tendency, less important than the median, is the mode. The mode is the most common item of a series of observations. In the above example 3 appeared twice, while the other observations only appeared once. Hence, 3 was the mode of the above set of data.

If the data observed form a symmetric frequency table, then the mean, the mode and the median coincide. For data with a frequency distribution skewed to the left, the following relation holds:

Mean < Median < Mode,

while for a frequency distribution skewed to the right it is:

Mode < Median < Mean.

Next to the sample variance, other measures exist to measure the variability of the sample. The sample range is the most simple measure of dispersion and is the difference between the highest and lowest values in a sample. Whereas this range is a "total" range, other "partial" ranges might be of importance to an investigator.

One of the partial ranges is the quartile range. The first quartile, $Q_1$, is the observed value located one fourth of the distance from the lowest value to the highest, and the third quartile, $Q_3$, is located at three quarters of this total range. Following this definition, the second quartile, $Q_2$, is thus the sample median.

The interquartile range is the difference between the values of the third and the first quartiles ($Q_3 - Q_1$). Dividing this quantity by two, we arrive at the quartile deviation $QD = \frac{1}{2}(Q_3 - Q_1)$.

For the purpose of our future discussions, we will almost always
utilize the sample mean and the sample variance to describe the two tendencies of a sample.

2. SOME METHODS OF SAMPLING

In the previous section, most of the examples we used dealt with "known" hypothetical populations, because that made it easier to explain the properties of a random sample. Usually, however, we draw a sample from a population whose properties we do not yet know. We examine a part of this population in order to make some deductions about the whole.

The unknown properties we are trying to estimate we call the parameters of the population. The true mean \( \mu \) and the true variance \( \sigma^2 \) of the population are the parameters we most commonly try to deduce from a sample.

We also have seen that the expected value of the mean of a random sample is equal to the mean of the population \( \mu \). This implies that if we were to draw an infinite number of samples without ruining the makeup of the population, the average of the means of all these samples would be exactly the same as the true population mean \( \mu \).

This is a nice property a sample mean should have. In technical language we say that such a sample mean is an unbiased estimate of the population mean. How much variation there is among various sample means drawn from the same population depends on sample size and on the magnitudes of the true variance of the population. This is expressed by the variance of the sample mean, as is given in formula (19). Because this measure is so important in judging the quality of our estimate of the population mean, it is vital to know the magnitude of the variance of the sample mean. However, it includes an unknown factor \( S^2 \), the population variance. Because the sample variance \( s^2 \) is an unbiased estimate of \( S^2 \), we usually assume it to be an acceptable estimator of the population variance.

Sampling designs are invented primarily for the purpose of getting an unbiased estimate of the population mean with minimum variability of the estimator per unit cost.
The least complicated sampling design is what we know as *simple random sampling*. Before starting to draw a simple random sample, we first have to define what the population is that we want to sample. All of the distinct individual members of the population are our *sampling units*. The total number of sampling units $N$ defines our population size.

If we manage to produce a complete and accurate list of $N$ well defined sampling units that constitute the entire population, we have what is called a *sampling frame*.

For example, to estimate the proportion of rice fields of a village that could be irrigated in the dry monsoon, one kind of sampling frame might consist of a list of all households in that village who own a rice-field. The sampling unit is then a household.

If a list of households could not be provided but a good map is available of the village, then we could subdivide the rice-field area of the village into $N$ sub-areas that will serve as the sampling units.

The first way of devising a sampling frame uses the *household* approach, while the latter uses an *area* approach. In fact, due to scarcity of accurate maps most developing countries usually use the household approach.

Once we have defined the sampling frame, it is possible to proceed at once to draw a random sample of size $n$ to obtain the necessary data. We do not actually need to have any further information about the composition of the population. If we use truly random methods to draw it, the sample we obtain is then a simple random sample.

As an example, let us reconsider the data of Table 1 concerning the population of fish farmers at Sekardangan.

This list was obtained by complete enumeration. In practice however, we usually do not have enough time and money allocated to perform a complete census. What we manage to know is usually only a list of all sampling units, but not the value $y_i$ of the random variable $Y$ associated with each of those sampling units.

Let us suppose that we only could cover the cost of obtaining a sample of size $n = 6$ from the population of fish farmers at Sekardangan. Suppose also, that according to the outcome of a random drawing, farmers numbered 47, 05, 49, 29, 34 and 44 were included in the random sample. Let the random variable
Y be the pond area owned by a farmer. Then, the set of data collected is as given in Table 5.

Table 5 Composition of a sample random sample of size $n = 6$, drawn from the population defined in Table 1

<table>
<thead>
<tr>
<th>$i$</th>
<th>$y_i$</th>
<th>$y_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>05</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>49</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>30</td>
</tr>
<tr>
<td>Sum</td>
<td>98</td>
<td></td>
</tr>
</tbody>
</table>

The sample mean is therefore $y_\bar{a} = 98/6 = 16.33$.

The most efficient way to calculate the sample variance is to use this short-cut formula:

$$ s^2 = \frac{1}{n-1} (\Sigma y_i^2 - (\Sigma y_i)^2). $$

(21)

This formula is especially convenient if an automatic desk-top calculator is available, capable of cumulating sum of squares.

For our sample, $s^2 = \frac{1}{5}(2458 - (98)^2) = 857.33/5 = 171.47$. The estimate of the variance of the mean $s^2_y$ is obtained by substituting $s^2$ and $S^2$ in the formula for $\sigma^2$ in (18):

$$ s^2_y = \frac{s^2}{n-1}. $$

(22)

In our case, $s^2_y = \frac{171.47}{6}(1 - \frac{6}{58}) = 25.62$.

If the quantity we are interested in is not the average pond area $\mu$, but the total area of fish ponds, then what we have to estimate is the total area $N\mu$. We estimate it by $N\bar{y}$. The variance of this estimate could be shown to be:

$$ \sigma_{N\bar{y}}^2 = N^2\sigma_y^2, $$

and is estimated by

$$ s_{N\bar{y}}^2 = N^2s_y^2. $$

(24)
For our example, \( N, \gamma = 58(16.3\%) \) ha = 947.14 ha and \( s_{N, \gamma}^2 = (58)^2(25.62) \) ha = 86185.68 ha.

In many cases, the population we are interested in is a dichotomous population, where we could distinguish two classes of members, one class possessing a specified property and the other class not having that property. For example, suppose we are interested in estimating the proportion of farmers in a village who are already receptive to innovation, say the use of improved varieties. What we assume, in other words, is that the whole population of farmers of that village could be subdivided into two classes, the class of progressive and the class of conservative farmers. (We would have to write out a definition of what we mean by “progressive” and “conservative” so that every farmer would fall clearly into one of the two classes and no farmer could be included in both.)

In our calculations we will say that out of a total of \( N \) farmers in that village, \( A \) are progressive and \( N-A \) are conservative. The quantity \( A \) we usually do not know and it is therefore a parameter.

The parameter could also be expressed as the ratio \( P = \frac{A}{N} \) which is the proportion of farmers which are progressive in that village.

The formulae for the estimation of parameters by means of simple random sampling could now be adapted to estimate the value of a proportion in a population. Suppose we have drawn a simple random sample of size \( n \). We observed from this sample that \( a \) farmers are progressive. Then the proportion of progressive farmers in the sample is \( p = \frac{a}{n} \).

Let us define the random variable \( Y \) to assume the value \( y = 1 \) if the farmer we observed is a progressive farmer and \( y = 0 \) if the farmer is conservative. If in the population \( A \) progressive farmers are present, then the sum of all \( y_i \) for the population is

\[
y_1 + y_2 + \ldots + y_i + \ldots + y_N = A,
\]

because \( A \) values of \( y_i \) are equal to 1 and \((N-A)\) values assume the value 0. Therefore the true mean of \( Y \) is \( \mu = \frac{1}{N} \sum_{i=1}^{N} y_i = \)
A/N = P, while the population variance of Y is:

\[ S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2. \]

Because there are A values of \( y_i \) equal to 1 and (N-A) values equal to 0,

\[ S^2 = \frac{1}{N-1} \left[ A(1-P)^2 + (N-A)(0-P)^2 \right] \]

\[ = \frac{1}{N-1} \left[ A(1-2P + P^2) + NP^2 - AP^2 \right]. \]

Remembering that \( P = A/N \) or \( A = NP \),

\[ S^2 = \frac{N}{N-1} P(1-P). \]

and

\[ \sigma^2 = \frac{N-1}{N} S^2 = P(1-P). \]

Following the same reasoning for a sample of size \( n \) we have

\[ y_1 + y_2 + \ldots + y_i + \ldots + y_n = a, \]

and

\[ \bar{y} = a/n = p. \] (25)

We know that \( E(\bar{y}) = \mu \). Hence, \( E(p) = P \) and therefore \( p \) is an unbiased estimate of \( P \). Its variance is:

\[ \sigma_p^2 = \sigma^2 = \frac{s^2}{N} \left( 1 - \frac{n}{N} \right) \]

\[ = \frac{1}{N} \frac{P(1-P)}{N-n} \frac{N-n}{N} = \frac{N-n P(1-P)}{N-1} \frac{N}{n} \] (26)

The estimator of \( \sigma_p^2 \) is:

\[ s_p^2 = \frac{s^2}{n} \left( 1 - \frac{n}{N} \right) \]

\[ = \frac{N-n P(1-P)}{n-1} \frac{n}{N} \] (27)

We have to note carefully that the estimate of \( \sigma_p^2 \) is not obtained from \( \sigma_p^2 \) by simply substituting \( p \) for \( P \) in formula (26). The values
of \( n \) and \( N \) in the denominators of (26) have in addition to be interchanged to arrive at (27).

As an application, suppose there are 100 farmers in a village. We draw a random sample of 20 farmers and learn by interviewing them that 5 of them regularly use improved varieties, fertilizers and pesticides to obtain a good crop. Then, as an estimate we could say that \( p = a/n = 5/20 = 1/4 \) of the whole population of farmers are progressive farmers. This estimate \( p = 1/4 \) has an estimated variance \( s_p^2 = \frac{100 - 20 \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot 100}{16} = .0079 \).

Sometimes what we want to estimate is not the proportion \( p \), but the total number \( A = Np \). This is estimated by \( a = Np \) with variance \( \sigma^2 = \frac{N^2 \sigma^2}{N} \), estimated as:

\[
s_{Np}^2 = \frac{N^2 \sigma^2}{N} = \frac{N(N - n)p(1 - p)}{n - 1}
\]

For our previous example, \( A \) is estimated by \( a = 100(1/4) = 25 \) with an estimated variance of \( s_a^2 = \frac{100(80)(3)}{19(16)} = 78.95 \).

In estimating a proportion \( P \), the variance of the population is as we could see from (24) equal to \( P(1 - P) \). Therefore the population variance is determined by the value of the proportion \( P \) we want to estimate. The change in value of \( \sigma^2 \) in relation with a change in value of \( P \) is given in Table 6.

<table>
<thead>
<tr>
<th>( P )</th>
<th>0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma^2 = P(1 - P) )</td>
<td>.00</td>
<td>.09</td>
<td>.16</td>
<td>.21</td>
<td>.24</td>
<td>.25</td>
<td>.24</td>
<td>.21</td>
<td>.16</td>
<td>.09</td>
<td>.00</td>
</tr>
</tbody>
</table>

We notice that the population obtains its maximum variance for \( P = .5 \). This implies that the estimate of \( P \) is more variable if the true value of \( P \) is around .5. For such a population, we have to use a larger sample size to obtain an estimate of \( P \) with smaller variance. If \( P \) is near 0 or 1 (in other words, the population consists mostly of one kind) it is easier to estimate the true proportion \( P \) with less variance.
In some cases, what we want to estimate is the percentage $C$ of a population that has a specific property. Between $C$ and $P$ exists the relationship $C = 100P$. From this relation it could be deduced that the estimate of $C$ is $c = 100p$ with variance:

$$\sigma_c^2 = (100)^2\sigma_p^2 = \frac{N-n}{N-1} \frac{100P(100-100P)}{n}$$

$$= \frac{N-n}{N-1} \frac{C(100-C)}{n} \quad (30)$$

The estimate of $\sigma_c^2$ is:

$$s_c^2 = (100)^2s_p^2 = \frac{N-n}{n-1} \frac{100P(100-100P)}{N}$$

$$= \frac{N-n}{N-1} \frac{c(100-c)}{N} \quad (31)$$

In our previous example, if 5 out of 20 farmers in the sample are innovative, then $c = \frac{5}{20} \cdot 100 = 25$ and $s_c^2 = \frac{100 - 20(25)}{19} \cdot \frac{75}{100} = 79$.

For any particular population, the most direct way to reduce the variance of the sample mean is to increase the sample size. But a larger sample usually means a higher cost. Would there not be other means to reduce the variance of the sample mean?

An answer to this question could be sought in the direction of **stratified random sampling**. In stratified random sampling we first try to subdivide the population into smaller groupings or “strata” in which there is more variability within each grouping or stratum than there is between strata. The key to the problem of lowering the variance of the sample mean lies in fact in our ability to stratify the population into several homogeneous subpopulations. Each subpopulation or stratum can then be considered as a population on its own, having its own specific characteristics.

Let us suppose that we have subdivided a population into $g$ strata. Consider the $k$-th stratum and let:

- $y_{ki}$ = the observed value of the random variable $Y$ of the $i$-th sampling unit in the $k$-th stratum
- $N_k$ = the size of the $k$-th stratum,
- $\mu_k$ = the true mean of the $k$-th stratum,
- $S_k^2$ = the true variance of the $k$-th stratum,
- $\mu$ = the true mean of the population,
\[ S^2 = \text{the true variance of the population.} \]
\[ N = N_1 + N_2 \ldots + N_g = \text{the size of the population.} \]
\[ W_k = N_k/N = \text{the relative size of the } k\text{-th stratum with respect to the size of the population.} \]

By remembering that each stratum could be considered as a population on its own, we can then express the true mean and variance of the \( k \)-th stratum in terms of the observed values as follows:

\[ \mu_k = \frac{1}{N_k} \sum N_k y_{k,i}. \]

\[ s_k^2 = \frac{1}{N_k - 1} \sum (y_{ki} - \mu_k)^2. \]

What we are interested in is the true mean of the population:

\[ \mu = \frac{1}{N} \sum N_1 y_{1i} + \sum N_2 y_{2i} + \ldots + \sum N_g y_{gi} \]
\[ = \frac{1}{N} [N_1 \mu_1 + N_2 \mu_2 + \ldots + N_g \mu_g] \]
\[ = W_1 \mu_1 + W_2 \mu_2 + \ldots + W_g \mu_g. \]

The variance of the unstratified population is:

\[ S^2 = \frac{1}{N-1} \sum N_g \sum (y_{ki} - \mu)^2. \]

The estimation of \( \mu \) by means of simple random sampling from this population will yield a sample mean variance of \( \sigma^2 = S^2 \left(1 - \frac{n}{N}\right) \). However, if the population was stratified into \( g \) strata and from the \( k \)-th stratum we draw a simple random sample of size \( n_k \), then we will obtain a sample with mean:

\[ \bar{y}_k = \frac{1}{n_k}(y_{k1} + y_{k2} + \ldots + y_{ki} \ldots + y_{kn}) \]

and variance:
For convenience we will abbreviate the notation $\sigma_k^2$ by $\sigma^2$.

The population mean $\mu$ could now be estimated by using the mean of a stratified random sample:

$$m = W_1 \bar{y}_1 + W_2 \bar{y}_2 + \ldots + W_s \bar{y}_s. \quad (38)$$

This quantity is an unbiased estimate of $\mu$. Its variance is

$$\sigma_m^2 = W_1^2 \sigma_1^2 + W_2^2 \sigma_2^2 + \ldots + W_s^2 \sigma_s^2. \quad (39)$$

If the stratification of the population was carried out appropriately, then $\sigma_m^2$ will be less than $\sigma^2$ of a simple random sample.

As an example, let us consider a hypothetical population of size $N = 10$ with the following observed values: $18, 22, 18, 20, 4, 4, 2, 2, 2, 4$.

Then, $\mu = \frac{1}{10}(18 + 22 + 22 + 18 + 20 + 4 + 4 + 2 + 4) = 120 \times 10 = 12 \text{ and}$

$$S^2 = \frac{1}{9}[(2(18-12)^2 + 2(22-12)^2 + (20-12)^2 + 3(4-12)^2 + (6-12)^2 + (2-12)^2)]$$

$$= 73.78.$$

If $\mu$ is to be estimated by $\bar{y}$, from a simple random sample of size $n = 5$, then the variance of the sample mean would be:

$$\sigma_n^2 = \frac{73.78}{5^2} (1 - \frac{5}{10}) = 7.38.$$

However, if we could manage to subdivide the population neatly into two strata, such that the first stratum includes the more similar values $18, 20, 22, 18$ and $20$, while the second stratum comprises the values $4, 6, 4, 2$ and $4$, then:

$$\mu_1 = (18 + 22 + 22 + 18 + 20)/5 = 20, \mu_2 = (4 + 6 + 4 + 2 + 4)/5 = 4$$

and $\mu = \frac{5}{10} \mu_1 + \frac{5}{10} \mu_2 = 12.$

The variance of each stratum is respectively:
\[ S_1^2 = \frac{1}{4}[(18-20)^2 + (22-20)^2 + (22-20)^2 + (18-20)^2 + (20-20)^2] \]
\[ = 4, \text{ and} \]
\[ S_2^2 = \frac{1}{4}[(4-4)^2 + (6-4)^2 + (4-4)^2 + (2-4)^2 + (4-4)^2] = 2. \]

If the drawing of a sample was carried out by drawing \( n_1 \) units from the first stratum and \( n_2 \) units from the second stratum such that \( n_1 + n_2 = n = 5 \), then the variance of the mean of the stratified random sample is:

\[
\sigma_m^2 = \frac{1}{N^2} \left[ \frac{N_1^2 S_1^2}{n_1} (1 - \frac{n_1}{N_1}) + \frac{N_2^2 S_2^2}{n_2} (1 - \frac{n_2}{N_2}) \right]
\]
\[ = \frac{1}{n_1} (1 - n_1/5) + \frac{1}{2n_2} (1 - n_2/5). \]

Therefore the magnitude of \( \sigma_m^2 \) also depends on the allocation of the sample to the various strata. For each possible allocation of the sampling units to the various strata we will then have the following values of \( \sigma_m^2 \):

<table>
<thead>
<tr>
<th>allocation</th>
<th>( \sigma_m^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 )</td>
<td>( n_2 )</td>
</tr>
<tr>
<td>4 1</td>
<td>1(1 - 4/5)/4 + (1 - 1/5)/2 = 0.45</td>
</tr>
<tr>
<td>3 2</td>
<td>1(1 - 3/5)/3 + (1 - 2/5)/4 = 0.28</td>
</tr>
<tr>
<td>2 3</td>
<td>1(1 - 2/5)/2 + (1 - 3/5)/6 = 0.37</td>
</tr>
<tr>
<td>1 4</td>
<td>1(1 - 1/5)/1 + (1 - 4/5)/8 = 0.82</td>
</tr>
</tbody>
</table>

It seems that for this special case, the allocation \( n_1 = 3 \) and \( n_2 = 2 \) yields the most efficient estimate of \( \mu \), i.e. that estimate with the lowest variance. If sampling was performed without stratification, then the variance of the sample mean based on simple random sampling would be \( \sigma_y^2 = 7.38 \), about 26 times as large than that of the most efficient stratified random sample. It is even about 9 times as large as the least efficient stratified random sample.

Now let us suppose that we failed in trying to stratify our
population into sub-groups whose members are similar to one another, and consequently obtained the following stratification:

First Stratum: 22, 18, 4, 2; \( N_1 = 4 \).

Second Stratum: 22, 18, 20, 4, 4, 6; \( N_2 = 6 \).

Then, \( \mu_1 = \frac{(22 + 18 + 4 + 2)}{4} = 11.5 \) and \( \mu_2 = \frac{(22 + 18 + 20 + 4 + 4 + 6)}{6} = 12.33 \), while \( S_1^2 = \frac{299}{3} \) and \( S_2^2 = \frac{248}{3} \). The variance of the sample mean will then be:

\[
\sigma_m^2 = \frac{1}{n_1} \left[ \frac{4(299)}{n_1} (1 - n_1/4) + \frac{9(248)}{n_2} (1 - n_2/6) \right]
\]

For various values of \( n_1 \) and \( n_2 \), the values of \( \sigma_m^2 \) are:

<table>
<thead>
<tr>
<th>Allocation</th>
<th>( \sigma_m^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_1 n_2</td>
<td></td>
</tr>
<tr>
<td>4 1</td>
<td>24.80</td>
</tr>
<tr>
<td>3 2</td>
<td>11.25</td>
</tr>
<tr>
<td>2 3</td>
<td>8.95</td>
</tr>
<tr>
<td>1 4</td>
<td>14.40</td>
</tr>
</tbody>
</table>

We notice that none of the four allocations yields a more efficient estimator than that obtained by simple random sampling. Only one allocation, i.e. for \( n_1 = 2 \) and \( n_2 = 3 \), yields a variance of the mean close to that obtained from a simple random sample. Comparing the ratio \( n_1 : n_2 \) with that of \( N_1 : N_2 \), we see that \( n_1 : n_2 = N_1 : N_2 \), such that \( n_1/N_1 = n_2/N_2 \). An allocation satisfying this relation is called a proportional allocation, because the number of sampling units drawn from each stratum is proportional to its size. It seems that the variance of the mean of a stratified random sample based on proportional allocation will approach the variance of the sample mean of a simple random sample. This could be explained by the fact that proportional allocation tends to sample with equal weight on all sampling units of the population, which is basically the criterion for simple random sampling. Therefore, should we be forced to stratify our sampling procedure, because of any reason but efficiency, then the best approach we could do to balance this constraint is to apply proportional allocation.
From the two examples we have been discussing, we can conclude that stratification of a population gives us an estimate of the population mean with a low variance only if the variation within the strata we have formed is considerably less than the variation between strata. To accomplish this we must locate measures that will aid us in forming the strata.

The best measure for successful stratification would be the values of the observed random variable $Y$ itself. But this is of course inadmissible, because the stratification has to be done before we observe the random variable $Y$.

Past experience on the nature of the variable to be observed is, of course, helpful in forming the strata. We can work from some known properties of the sampling units which we have reason to think will influence the value $Y$ will assume. For example, if we want to estimate the manhours needed per hectare for the production of rice, we could divide the rice farmers into two strata, subsistence farmers and commercial farmers, on the presumption that the subsistence farmers will use more manhours per hectare than the commercial farmers will.

A stratified random sample is also sometimes used for the purpose of handling organizational problems. If the sampling frame consists of an administrative area subdivided into several lower level administrative units, then from administrative viewpoint it might be convenient to use these smaller administrative units as strata.

Another reason for stratification arises when the cost of drawing a sampling unit is more expensive in one location than in another. This could be caused, for instance, by a distance factor or by different local conditions.

In seeking to keep sampling cost as low as possible, the optimum allocation of a stratified random sample has to be determined by taking into consideration:

1. the relative size of each stratum,
2. the relative size of the stratum variance,
3. the relative cost of drawing one unit from each stratum.

If for the $k$-th stratum,

$n_k =$ the number of sampling units drawn from the $k$-th stratum,

$n =$ the sample size,
The stratum mean, \( \mu_k \), is the size of the \( k \)-th stratum,
\( N_k \) is the size of the population,
\( S_k = \sqrt{S_k^2} \) is the standard deviation of the \( k \)-th stratum
\( c_k \) is the cost of drawing one sampling unit from the \( k \)-th stratum, then the optimum allocation is determined by the relation

\[
n_k = \frac{nN_kS_k/\sqrt{c_k}}{\Sigma(n_kS_k/\sqrt{c_k})}.
\]

This means that more sampling units have to be drawn from those strata that have a larger size, a larger standard deviation or variance, or a lower sampling cost per unit. As an example we could consider the population of size \( N = 10 \) with observed values 18, 22, 22, 18 and 20 for the first stratum and 4, 6, 4, 2 and 4 for the second stratum. The stratum means are \( \mu_1 = 20 \) and \( \mu_2 = 4 \), while the stratum variances are \( S_1^2 = 4 \) and \( S_2^2 = 2 \).

The optimum allocation for constant cost \( c_k = c \) for all strata is reached for \( n_1 = n_1N_1S_1/(N_1S_1 + N_2S_2) = 10/(2 + 1.41) = 2.8 \) and \( n_2 = 2.2 \). This optimum allocation will be approached by taking \( n_1 = 3 \) and \( n_2 = 2 \) or \( n_1 = 2 \) and \( n_2 = 3 \).

If the cost of drawing a sampling unit is 9 times higher in the first stratum than in the second stratum, then \( c_1 = 9c_2 \) and the optimum allocation is determined by \( n_1 = (nN_1S_1/\sqrt{c_1})/(N_1S_1/\sqrt{c_1} + N_2S_2/\sqrt{c_2}) = (nN_1S_1)/(N_1S_1 + 3N_2S_2) = 1.6 \) and \( n_2 = 3.4 \). The formula tells us to economize by drawing more units from the second stratum, where sampling is less expensive.

Based on stratified random sampling we could develop more sophisticated sampling designs. Readers who are interested in this topic are invited to consult the works by Cochran\(^*\) and by Sukhatme and Sukhatme\(^*\).

---


3. PROBABILITY STATEMENTS ON ESTIMATES OF PARAMETERS

The purpose of drawing a sample from a population is ultimately to estimate the population parameters on the basis of the values in the sample.

The estimate obtained does not necessarily tell us the true parameter value. It would be nice if we could produce a statement about how dependable our estimate of the parameter is. A statement of this kind can be formed provided we know the probability distribution of our estimate as a random variable.

We will start the discussion by considering a normally distributed random variable with mean \( \mu \) and variance \( \sigma^2 \). We symbolize it in statistical shorthand by the notation:

\[
Y \sim N(\mu, \sigma^2).
\]

(41)

and read it as "Y is normally distributed with mean \( \mu \) and variance \( \sigma^2 \)." The random variable \( Y \) could be transformed into another random variable \( Z \) according to the following rule:

\[
Z = \frac{Y - \mu}{\sigma}
\]

(42)

where \( \sigma \) is the standard deviation of \( Y \). This new random variable \( Z \) has a true mean of 0 and a true variance of 1 and is normally distributed provided \( Y \) is a normal random variable.

Because we know from Figure 10 that for example

\[
P(-1.96 < Y < +1.96) = .95,
\]

we can also say that

\[
P(-1.96 < \frac{Y - \mu}{\sigma} < +1.96) = .95.
\]

(43)

We then can perform the following algebraic manipulations:

\[
P(-1.96\sigma < Y - \mu < +1.96\sigma) = .95,
\]

\[
P(Y - 1.96\sigma < \mu < Y + 1.96\sigma) = .95,
\]

and

\[
P(Y - 1.96\sigma < \mu < Y + 1.96\sigma) = .95.
\]

(44)

Therefore, if we obtained from a normal distribution with unknown mean \( \mu \) and known standard deviation \( \sigma \), say \( \sigma = 2 \), an observed
value $Y = 12$, then $P(12 - 1.96(2) < \mu < 12 + 1.96(2)) = .95$, or $P(8.08 < \mu < 15.92) = 95\%$. Stated in words, this means that with $95\%$ probability the interval between 8.08 and 15.92 will include the population mean. More plainly stated, this means that we are $95\%$ sure that the population mean is within the range of 8.08 to 15.92.

Because this interval of values will include the true mean on the average 95 times out of 100 drawings of a sample mean, it is called the $95\%$ confidence interval for the population mean. The endpoints of this interval are called the confidence limits of the interval.

A confidence interval of $99\%$ could be constructed similarly from the same observed data, starting from the probability statement $P(-2.58 < Z < +2.58) = .99$.

The confidence probability we have utilized to construct the probability statement, we call the level of confidence of the interval. In general, for any confidence level we could construct a confidence limit using the tabulated values of $P(0 < Z < z_0)$ under Figure 10.

So far we have constructed a confidence interval for a single observation. However, if a sample of size $n$ was drawn, then from the $n$ observed values of $Y$ the sample mean $\bar{y}$ could be constructed, which has the property of:

$$\bar{y} \sim N(\mu, \frac{\sigma^2}{n}).$$

Therefore, comparing (45) with (41), we now could construct from (44) a similar probability statement by substituting $y$. for $Y$ and $\sigma/\sqrt{n}$ for $\sigma$:

$$P(\bar{y} - 1.96\sigma/\sqrt{n} < \mu < \bar{y} + 1.96\sigma/\sqrt{n}) = .95.$$  (46)

This is the confidence interval statement for the population mean based on a random sample of size $n$. If $n = 1$, i.e. if we only drew a single observation, we see that (46) reduces into (44).

As an example, suppose that a random sample of size $n = 9$ yields a sample mean $\bar{y} = 12$, while we know that $\sigma = 2$. Then, the confidence interval for the mean is determined by the probability statement:

$$P(12 - 1.96(2/3) < \mu < 12 + 1.96(2/3)) = .95 \text{ or } P(10.63 < \mu < 13.63) = .95.$$

Comparing this confidence interval to that for a single observa-
tion, we see again the effect of increasing the sample size on the reliability of the sample mean as an estimate of the population mean. A narrower interval includes the population mean with the same confidence level. As \( n \) increases, we could imagine that the width of this confidence interval will decrease, leading to a more dependable estimate of \( \mu \).

The confidence limit for the population mean was constructed under the condition that the population variance, and hence also the population standard deviation \( \sigma \), is known to us. In reality however, it is more frequently unknown than known. We can only estimate \( \sigma^2 \) by the sample variance \( s^2 \). The question now arises, whether it is still valid to assume that \( \frac{Y - \mu}{s} \) is distributed as a standard normal variable. Student* observed that if this quantity is treated as a standard normal variable, then some discrepancies will arise. In honor of him this random variable

\[
t = \frac{Y - \mu}{s}
\]

is called a Student-\( t \) random variable. The shape of the probability function of this \( t \) variable deviates slightly from that of the standard normal random variable. Its peak is flatter and its tails extend

* pseudonym of W.S. Gosset, a chemist working in a beer brewery in England at the beginning of the 20th century. He signed a contract not to publish any of his findings, but his superiors allowed him to publish research of non-chemical nature under a pseudonym.
farther away from the mean, which in this case is also equal to 0 (Figure 11). The deviation from normality becomes less with increasing sample size and is practically unimportant for sample sizes larger than 30.

The longer extending tails of the \( t \)-distribution imply that a specified area covered by the normal curve between the limits \( Z = -z_0 \) and \( Z = +z_0 \) is covered by the \( t \)-curve between limits that are located farther away from 0 than \(-z_0\) and \(+z_0\) are. This implies that the probability statement of the form

\[
P(-1.96 < \frac{Y - \mu}{s} < 1.96) = .95
\]

is no longer valid because we have substituted \( s \) for \( \sigma \) in the formula. To validate this statement for the same confidence level, we have to replace the critical value 1.96 by a larger positive value. If the sample size is \( n = 2 \), such that \( r = n - 1 = 1 \), the statement becomes valid if we replace the critical value 1.96 by 12.71, leading to the valid probability statement

\[
P(-12.71 < \frac{Y - \mu}{s} < 12.71) = .95
\]

or equivalently

\[
P(Y - 12.71s < \mu < Y + 12.71s) = .95.
\]

In this case \( r = (n - 1) \) is called the number of degrees of freedom of the sample.

For larger degrees of freedom, or equivalently for larger sample sizes, the critical \( t \)-value will approach the critical z-value of the standard normal distribution. For instance, if \( r = n - 1 = 12 \), then

\[
P(-2.18 < \frac{Y - \mu}{s} < 2.18) = .95,
\]

while for \( r = 30 \), the critical \( t \)-value is 2.04. For sample sizes larger than 30, it does not matter any more whether we consider \( \frac{Y - \mu}{s} \) as a \( t \) or a \( Z \) random variable, because the critical \( t \)-values approach 1.96, the critical z-value of the standard normal \( Z \)-distribution.

In general, if \( z_{a/2} \) is the critical value of the standard normal distribution, such that the probability statement

\[
P(-z_{a/2} < \frac{Y - \mu}{\sigma} < +z_{a/2}) = 1 - \alpha
\]

is true and \( t_{a/2(r)} \) is the critical value of the Student-\( t \) distribution with \( r \) degrees of freedom, such that the probability statement

\[
P(-t_{a/2(r)} < \frac{Y - \mu}{s} < +t_{a/2(r)}) = 1 - \alpha
\]

is true.
is true, then the critical values $t_{a/2(n)}$ for various values of $r$ are given in Table 6, together with the corresponding $z_{a/2}$ values, which are the limiting values of $t_{a/2(r)}$ for $r$ tending to infinity.

If the confidence interval of $\mu$ we want to construct is to be based on the mean of a random sample of size $n$, then the confidence interval statement is of the following form:

$$P\left[ \bar{y} - t_{a/2(n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{a/2(n-1)} \frac{s}{\sqrt{n}} \right] = 1 - \alpha. \quad (49)$$

For example, an agricultural extension agent wants to estimate the amount of unstalked paddy that can be harvested by a person using an anu-anu* from an IR-22 crop. He chooses six woman harvesters at random and lets them harvest the crop at randomly chosen spots in the field for a specified time period. The amounts harvested are respectively 10, 14, 12, 16, 14 and 10 kg.

To use those results to calculate a confidence interval for the true average weight of unstalked paddy that could be harvested within a specified time period, the following data sheet could be formed:

<table>
<thead>
<tr>
<th>$y_i$</th>
<th>$y_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>196</td>
</tr>
<tr>
<td>12</td>
<td>144</td>
</tr>
<tr>
<td>16</td>
<td>256</td>
</tr>
<tr>
<td>14</td>
<td>196</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

$y_i = 76$ \quad $y_i^2 = 992$

$\bar{y} = 12.67$

The 95\% confidence statement is then

$$P(12.67 - 2.57(99) < \mu < 12.67 + 2.57(99))$$

or

$$P(10.10 < \mu < 15.24) = .95.$$ 

The confidence interval could be symbolized as $(10.10; 15.24)$.

* harvesting knife used by Javanese rice-farmers in Indonesia.
### Table 6  Critical values $t_{a, \nu}$ of the Student-$t$ distribution for various levels of $a$

<table>
<thead>
<tr>
<th>degrees of freedom ($\nu$)</th>
<th>$a = .05$</th>
<th>$a = .025$</th>
<th>$a = .01$</th>
<th>$a = .005$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.31</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
</tr>
<tr>
<td>2</td>
<td>2.92</td>
<td>4.30</td>
<td>6.97</td>
<td>9.93</td>
</tr>
<tr>
<td>3</td>
<td>2.35</td>
<td>3.18</td>
<td>4.54</td>
<td>5.84</td>
</tr>
<tr>
<td>4</td>
<td>2.13</td>
<td>2.78</td>
<td>3.75</td>
<td>4.60</td>
</tr>
<tr>
<td>5</td>
<td>2.02</td>
<td>2.57</td>
<td>3.37</td>
<td>4.03</td>
</tr>
<tr>
<td>6</td>
<td>1.94</td>
<td>2.45</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>7</td>
<td>1.90</td>
<td>2.37</td>
<td>3.00</td>
<td>3.50</td>
</tr>
<tr>
<td>8</td>
<td>1.86</td>
<td>2.31</td>
<td>2.90</td>
<td>3.36</td>
</tr>
<tr>
<td>9</td>
<td>1.83</td>
<td>2.26</td>
<td>2.82</td>
<td>3.25</td>
</tr>
<tr>
<td>10</td>
<td>1.81</td>
<td>2.23</td>
<td>2.76</td>
<td>3.17</td>
</tr>
<tr>
<td>11</td>
<td>1.80</td>
<td>2.20</td>
<td>2.72</td>
<td>3.11</td>
</tr>
<tr>
<td>12</td>
<td>1.78</td>
<td>2.18</td>
<td>2.68</td>
<td>3.06</td>
</tr>
<tr>
<td>13</td>
<td>1.77</td>
<td>2.16</td>
<td>2.65</td>
<td>3.01</td>
</tr>
<tr>
<td>14</td>
<td>1.76</td>
<td>2.15</td>
<td>2.62</td>
<td>2.98</td>
</tr>
<tr>
<td>15</td>
<td>1.75</td>
<td>2.13</td>
<td>2.60</td>
<td>2.95</td>
</tr>
<tr>
<td>16</td>
<td>1.75</td>
<td>2.12</td>
<td>2.58</td>
<td>2.92</td>
</tr>
<tr>
<td>17</td>
<td>1.74</td>
<td>2.11</td>
<td>2.57</td>
<td>2.90</td>
</tr>
<tr>
<td>18</td>
<td>1.73</td>
<td>2.10</td>
<td>2.55</td>
<td>2.88</td>
</tr>
<tr>
<td>19</td>
<td>1.73</td>
<td>2.09</td>
<td>2.54</td>
<td>2.86</td>
</tr>
<tr>
<td>20</td>
<td>1.73</td>
<td>2.09</td>
<td>2.53</td>
<td>2.85</td>
</tr>
<tr>
<td>21</td>
<td>1.72</td>
<td>2.08</td>
<td>2.52</td>
<td>2.83</td>
</tr>
<tr>
<td>22</td>
<td>1.72</td>
<td>2.07</td>
<td>2.51</td>
<td>2.82</td>
</tr>
<tr>
<td>23</td>
<td>1.71</td>
<td>2.07</td>
<td>2.50</td>
<td>2.81</td>
</tr>
<tr>
<td>24</td>
<td>1.71</td>
<td>2.06</td>
<td>2.49</td>
<td>2.80</td>
</tr>
<tr>
<td>25</td>
<td>1.71</td>
<td>2.06</td>
<td>2.48</td>
<td>2.78</td>
</tr>
<tr>
<td>26</td>
<td>1.71</td>
<td>2.06</td>
<td>2.48</td>
<td>2.78</td>
</tr>
<tr>
<td>27</td>
<td>1.70</td>
<td>2.05</td>
<td>2.47</td>
<td>2.77</td>
</tr>
<tr>
<td>28</td>
<td>1.70</td>
<td>2.05</td>
<td>2.47</td>
<td>2.76</td>
</tr>
<tr>
<td>29</td>
<td>1.70</td>
<td>2.05</td>
<td>2.46</td>
<td>2.76</td>
</tr>
<tr>
<td>30</td>
<td>1.70</td>
<td>2.04</td>
<td>2.46</td>
<td>2.75</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.65</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>

Because \( s^2 \) is an estimate of \( \sigma^2 \), we also could construct a confidence interval for the population variance. The construction of the probability statement is based on the property of the random variable \( \frac{(Y - \mu)^2}{\sigma^2} \). If \( Y \sim N(\mu, \sigma^2) \), then it could be shown that \( \frac{(Y - \mu)^2}{\sigma^2} \) is a random variable distributed according to the \( \chi^2 \) (pronounced chi square) distribution with one degree of freedom.

If we take a random sample with the observations \( y_1, y_2, \ldots, y_n \) from a normal population of \( Y \)'s with mean \( \mu \) and variance \( \sigma^2 \), then

\[
\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} = \frac{(y_1 - \mu)^2}{\sigma^2} + \frac{(y_2 - \mu)^2}{\sigma^2} + \ldots + \frac{(y_n - \mu)^2}{\sigma^2} = \chi^2_n \quad (50)
\]

is distributed according to the \( \chi^2 \) distribution with \( n \) degrees of freedom. The quantity \( \chi^2_n \) is then called a \( \chi^2 \) random variable with \( n \) degrees of freedom.

If \( \mu \) is unknown, as is usually the case, then substitution of \( \bar{y} \) for \( \mu \) in (50) results in:

\[
\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{\sigma^2} = \frac{(y_1 - \bar{y})^2}{\sigma^2} + \frac{(y_2 - \bar{y})^2}{\sigma^2} + \ldots + \frac{(y_n - \bar{y})^2}{\sigma^2} = \chi^2_{n-1} \quad (51)
\]

which is also a \( \chi^2 \) random variable, but with \( (n-1) \) degrees of freedom. It lost one degree of freedom because the parameter \( \mu \) has been replaced by its estimate \( \bar{y} \).

The probability distribution of the \( \chi^2 \) random variable in the form of \( P(0 < \chi^2 < \chi^2_{(1-\alpha)})) = 1 - \alpha \) has been tabulated for several \( \alpha \) values and degrees of freedom, such as represented by Table 7. The values tabulated in the table are the critical values \( \chi^2_{\alpha} \).

* harvesting knife used by Javanese rice-farmers in Indonesia.
the statement \( P(0 < \chi^2 < \chi^2_{a/2}) = 1 - \alpha/2 \), we could substitute \((1 - \alpha/2)\) for \(\alpha/2\) and obtain the statement \( P(0 < \chi^2 < \chi^2_{1-a/2}) = \alpha/2 \). We also could conclude that \( P(\chi^2_{a/2} < \chi^2 < \infty) = \alpha/2 \). Combining the latter two facts we know that \( P(\chi^2_{1-a/2} < \chi^2 < \chi^2_{a/2}) = 1 - \alpha/2 \). The truth of these probability statements is obvious from Figure 12.

A \( \chi^2 \) random variable with \( v \) degrees of freedom has a true mean \( E(\chi^2) = v \) and a variance equal to \( 2v \). The shape of the probability function is skewed, but tends to become symmetrical as \( v \) increases in value.

The statement \( P(\chi^2_{1-a/2} < \chi^2 < \chi^2_{a/2}) = 1 - \alpha \) is our tool to construct a confidence interval of the population variance \( \sigma^2 \) for a

Figure 12 The critical value \( \chi^2 \) divides the area under the \( \chi^2 \) curve into subareas in the proportion of \((1 - \alpha)\): \( \alpha \)
normal distribution, for in a random sample of size \( n \) for such a distribution the sample mean \( s^2 \) could be written as:

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

\[
= \frac{\sigma^2}{n-1} [(y_1 - \bar{y})^2/\sigma^2 + (y_2 - \bar{y})^2/\sigma^2 + \ldots + (y_n - \bar{y})^2/\sigma^2].
\]

Therefore,

\[
\frac{(n-1)s^2}{\sigma^2} = \frac{(y_1 - \bar{y})^2}{\sigma^2} + \frac{(y_2 - \bar{y})^2}{\sigma^2} + \ldots + \frac{(y_n - \bar{y})^2}{\sigma^2}
\]

is a \( \chi^2 \) random variable with \( n \) degrees of freedom and satisfies the probability statement:

\[
P(\chi^2_{n-2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-2}) = 1 - \alpha
\]

By algebraic manipulations we will obtain the following equivalent statements:

\[
P(\frac{1}{(n-1)s^2} < \frac{\sigma^2}{\chi^2_{n-2}} < \frac{1}{\sigma^2}) = 1 - \alpha
\]

and

\[
P\left(\frac{(n-1)s^2}{\chi^2_{n-2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-2}}\right) = 1 - \alpha.
\]

The quantity

\[
(n-1)s^2 = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2
\]

\[
= \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

is called the sum of squares of the observed values \( y_i \) of the sample.

Utilizing the statement (52) we now can calculate a confidence interval for \( \sigma^2 \) of the population which the six women rice harvesters were representing. Because the sum of squares of \( y \), abbreviated by \( \text{SSY} \) is \( 992 - \frac{(76)^2}{6} = 35.33 \) and \( \chi^2_{.975,5} = .83 \), while \( \chi^2_{.025,5} = 13, \)
36 NASOETION  SAMPLING

\[ P \left( \frac{35.33}{13} < \sigma^2 < \frac{35.33}{.83} \right) = .95, \text{ the 95\% confidence interval for } \sigma^2 \text{ is} \]

\[ \left( \frac{35.33}{13}, \frac{35.33}{.83} \right) \text{ which is equivalent to } (2.72; 42.6). \]

Table 7  Critical values of the \( \chi^2 \) variable. Tabulated values are the values of \( \chi^2_{v,\alpha} \), such that \( P(0 < \chi^2 < \chi^2_{v,\alpha}) = 1 - \alpha \) and \( P(\chi^2 < \chi^2 + x) = \alpha. \)

<table>
<thead>
<tr>
<th>( v )</th>
<th>( \alpha = .99 )</th>
<th>( \alpha = .975 )</th>
<th>( \alpha = .05 )</th>
<th>( \alpha = .025 )</th>
<th>( \alpha = .01 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.000</td>
<td>.001</td>
<td>3.8</td>
<td>5.0</td>
<td>6.6</td>
</tr>
<tr>
<td>2</td>
<td>.02</td>
<td>.05</td>
<td>6.0</td>
<td>7.4</td>
<td>9.2</td>
</tr>
<tr>
<td>3</td>
<td>.12</td>
<td>.22</td>
<td>7.8</td>
<td>9.4</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>.30</td>
<td>.55</td>
<td>9.5</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>.55</td>
<td>.83</td>
<td>11</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>.87</td>
<td>1.2</td>
<td>13</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>1.2</td>
<td>1.7</td>
<td>14</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>1.7</td>
<td>2.2</td>
<td>16</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>2.1</td>
<td>2.7</td>
<td>17</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>2.6</td>
<td>3.3</td>
<td>18</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>3.1</td>
<td>3.8</td>
<td>20</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>3.6</td>
<td>4.4</td>
<td>21</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>13</td>
<td>4.1</td>
<td>5.0</td>
<td>22</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>14</td>
<td>4.7</td>
<td>5.6</td>
<td>24</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td>15</td>
<td>5.2</td>
<td>6.3</td>
<td>25</td>
<td>28</td>
<td>31</td>
</tr>
<tr>
<td>16</td>
<td>5.8</td>
<td>6.9</td>
<td>26</td>
<td>29</td>
<td>32</td>
</tr>
<tr>
<td>17</td>
<td>6.4</td>
<td>7.6</td>
<td>28</td>
<td>30</td>
<td>33</td>
</tr>
<tr>
<td>18</td>
<td>7.0</td>
<td>8.2</td>
<td>29</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>19</td>
<td>7.6</td>
<td>8.9</td>
<td>30</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>20</td>
<td>8.3</td>
<td>9.6</td>
<td>31</td>
<td>34</td>
<td>38</td>
</tr>
<tr>
<td>21</td>
<td>8.9</td>
<td>10</td>
<td>33</td>
<td>36</td>
<td>39</td>
</tr>
<tr>
<td>22</td>
<td>9.5</td>
<td>11</td>
<td>34</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>23</td>
<td>10</td>
<td>12</td>
<td>35</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>24</td>
<td>11</td>
<td>12</td>
<td>36</td>
<td>39</td>
<td>43</td>
</tr>
<tr>
<td>25</td>
<td>12</td>
<td>13</td>
<td>38</td>
<td>41</td>
<td>44</td>
</tr>
<tr>
<td>26</td>
<td>12</td>
<td>14</td>
<td>39</td>
<td>42</td>
<td>46</td>
</tr>
<tr>
<td>27</td>
<td>13</td>
<td>15</td>
<td>40</td>
<td>43</td>
<td>47</td>
</tr>
<tr>
<td>28</td>
<td>14</td>
<td>16</td>
<td>43</td>
<td>46</td>
<td>50</td>
</tr>
<tr>
<td>29</td>
<td>15</td>
<td>17</td>
<td>44</td>
<td>47</td>
<td>51</td>
</tr>
</tbody>
</table>

We have seen in (25) that a proportion \( P \) could be estimated by the sample estimate \( \hat{p} = \frac{a}{n} \). This estimate has a variance of

\[
\sigma_{\hat{p}}^2 = \frac{N-n}{N-1} \cdot \frac{P(1-P)}{n},
\]

which for large samples could be estimated by \( s_p^2 = \frac{p(1-p)}{n} \). This deviates slightly from the finite sample estimate in (27). The central limit theorem stated that any sample mean tends to distribute normally as the sample size increases. Because the sample proportion is also a sample mean for large samples, the confidence interval for proportions could be determined by the following probability statement:

\[
P(p - z_{\alpha/2} \sqrt{p(1-p)/n} < P < p + z_{\alpha/2} \sqrt{p(1-p)/n}) = 1 - \alpha.
\]

For the special case \( 1 - \alpha = .95 \), the statement becomes:

\[
P(p = 1.96 \sqrt{p(1-p)/n} < P < p + 1.96 \sqrt{p(1-p)/n}) = .95. \tag{53}
\]

For small samples however, this probability statement is only a very rough approximation of the real situation. The confidence statement has to be expressed more exactly by means of the binomial distribution. Snedecor has tabulated the confidence limits for proportions for various sample sizes and confidence levels. In Table 8, part of his tabulation is reproduced in a simplified form.

To determine the confidence interval of a proportion \( P \), we first determine the sample value \( \hat{p} = a/n \). If it is less than or equal to \(.50\), the table can be used directly. The confidence interval can then be read in the body of the table for the specified value of \( a/n \) under the associated sample size.

For example, if 20 out of 100 seeds germinated in a germination test, then \( \hat{p} = \frac{a}{n} = \frac{20}{100} = 0.20 \). The confidence interval for the true germination capacity \( P \) is then read directly from the table under \( n = 100 \) for \( \frac{a}{n} = .20 \), i.e. 13 - 29\%.

If the value of \( \hat{p} \) observed is larger than .50, then we have to find first the sample proportion \( (1-p) \) of the complementary characteristic. For example, if 80 out of 100 persons sampled are illiterate, then \( \hat{p} = .80 > .50 \). We determine \( (1-p) = .20 \), the sample
proportion of literate people. The confidence interval is 13 - 29.9. The confidence interval for P is therefore between (100 - 29) and (100 - 13)%, or between 71 and 87.9.

Table 8  Confidence interval for the proportion P expressed in percent at the 95% confidence level

<table>
<thead>
<tr>
<th>n</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>n</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00 17</td>
<td>00 07</td>
<td>00 04</td>
<td>.26</td>
<td>15 - 41</td>
<td>18 - 36</td>
<td></td>
</tr>
<tr>
<td>01</td>
<td>00 05</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>00 11</td>
<td>00 07</td>
<td>.28</td>
<td>16 - 43</td>
<td>19 - 38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>03</td>
<td>01 - 08</td>
<td>.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>00 14</td>
<td>01 10</td>
<td>.30</td>
<td>12 - 54</td>
<td>18 - 44</td>
<td>21 - 40</td>
<td></td>
</tr>
<tr>
<td>05</td>
<td>00 25</td>
<td>02 11</td>
<td>.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06</td>
<td>01 17</td>
<td>02 12</td>
<td>.32</td>
<td>20 - 46</td>
<td>23 - 42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>03 14</td>
<td>.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>08</td>
<td>02 19</td>
<td>04 15</td>
<td>.34</td>
<td>21 - 48</td>
<td>25 - 44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>09</td>
<td>04 16</td>
<td>.35</td>
<td>15 - 59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>01 31</td>
<td>03 22</td>
<td>05 18</td>
<td>.36</td>
<td>23 - 50</td>
<td>27 - 46</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>05 19</td>
<td>.37</td>
<td></td>
<td></td>
<td></td>
<td>28 - 47</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>05 24</td>
<td>06 20</td>
<td>.38</td>
<td>25 - 53</td>
<td>28 - 48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>07 21</td>
<td>.39</td>
<td></td>
<td></td>
<td></td>
<td>29 - 49</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>06 27</td>
<td>08 22</td>
<td>.40</td>
<td>19 - 64</td>
<td>27 - 55</td>
<td>30 - 50</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>03 38</td>
<td>09 24</td>
<td>.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>07 29</td>
<td>09 25</td>
<td>.42</td>
<td>28 - 57</td>
<td>32 - 52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>10 26</td>
<td>.43</td>
<td></td>
<td></td>
<td></td>
<td>33 - 53</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>09 31</td>
<td>11 27</td>
<td>.44</td>
<td>30 - 59</td>
<td>34 - 54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>12 28</td>
<td>.45</td>
<td>23 - 68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>06 44</td>
<td>10 34</td>
<td>13 29</td>
<td>.46</td>
<td>32 - 61</td>
<td>36 - 56</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>14 30</td>
<td>.47</td>
<td></td>
<td></td>
<td></td>
<td>37 - 57</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>12 36</td>
<td>14 31</td>
<td>.48</td>
<td>34 - 63</td>
<td>38 - 58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>15 - 32</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
<td>39 - 59</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>13 38</td>
<td>16 33</td>
<td>.50</td>
<td>27 - 73</td>
<td>36 - 64</td>
<td>40 - 60</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>09 49</td>
<td>17 35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Simplified from: Snedecor, G.W. 1956. Statistical Methods Applied to Experiments in Agriculture and Biology. 5th ed., Table 1.3.1. Iowa State College Press, Ames, Iowa.
A reasonable number of copies of this publication will be made available upon request for teachers wishing to use it in their classrooms. Their only obligation will be to convey to the author their comments and suggestions for improvement. Request copies from the A/D/C Asia Office, Tanglin P.O. Box 84, Singapore 10.

Distribution of these papers on survey research methodology has been made possible by a grant to A/D/C from the International Development Research Center of Ottawa.