Expected Value of a Sample Estimate

BY

EARL E. HOUSEMAN

Statistical Reporting Service • U.S. Department of Agriculture • SRS No. 19, 1974

FOREWORD

The Statistical Reporting Service (SRS) has been engaged for many years in the training of agricultural statisticians from around the world. Most of these participants come under the support of the Agency for International Development (AID) training programs; however, many also come under sponsorship of the Food and Agriculture Organization into the International Statistical Programs Center of the Bureau of the Census, with which SRS is cooperating.

This treatise was developed by the SRS with the cooperation of AID and the Center, in an effort to provide improved materials for teaching and reference in the area of agricultural statistics, not only for foreign students but also for development of staff working for these agencies.

HARRY C. TRELOGAN
Administrator
Statistical Reporting Service

Washington, D. C. September 1974

PREFACE

The author has felt that applied courses in sampling should give more attention to elementary theory of expected values of a random variable. The theory pertaining to a random variable and to functions of random variables is the foundation for probability sampling. Interpretations of the accuracy of estimates from probability sample surveys are predicated on, among other things, the theory of expected values.

There are many students with career interests in surveys and the application of probability sampling who have very limited backgrounds in mathematics and statistics. Training in sampling should go beyond simply learning about sample designs in a descriptive manner. The foundations in mathematics and probability should be included. It can (1) add much to the breadth of understanding of bias, random sampling error, components of error, and other technical concepts; (2) enhance one's ability to make practical adaptations of sampling principals and correct use of formulas; and (3) make communication with mathematical statisticians easier and more meaningful.

This monograph is intended as a reference for the convenience of students in sampling. It attempts to express relevant, introductory mathematics and probability in the context of sample surveys. Although some proofs are presented, the emphasis is more on exposition of mathematical language and concepts than on the mathematics per se and rigorous proofs. Many problems are given as exercises so a student may test his interpretation or understanding of the concepts. Most of the mathematics is elementary. If a formula looks involved, it is probably because it represents a long sequence of arithmetic operations.

Each chapter begins with very simple explanations and ends at a much more advanced level. Most students with only high school algebra should have no difficulty with the first parts of each chapter. Students with a few courses in college mathematics and statistics might review the first parts of each chapter and spend considerable time studying the latter parts. In fact, some students might prefer to start with Chapter III and refer to Chapters I and II only as needed.

Discussion of expected values of random variables, as in Chapter III, was the original purpose of this monograph. Chapters I and II were added as background for Chapter III. Chapter IV focuses attention on the distribution of an estimate which is the basis for comparing the accuracy of alternative sampling plans as well as a basis for statements about the accuracy of an estimate from a sample. The content of Chapter IV is included in books on sampling, but it is important that students hear or read more than one discussion of the distribution of an estimate, especially with reference to estimates from actual sample surveys.

The author's interest and experience in training has been primarily with persons who had begun careers in agricultural surveys. I appreciate the opportunity, which the Statistical Reporting Service has provided, to prepare this monograph.

Earl E. Houseman
Statistician

CONTENTS

CONTENTS (Continued)

CHAPTER I. NOTATION AND SUMMATION

1.1 INTRODUCTION

To work with large amounts of data, an appropriate system of notation is needed. The notation must identify data by individual elements, and provide meaningful mathematical expressions for a wide variety of summaries from individual data. This chapter describes notation and introduces summation algebra, primarily with reference to data from census and sample surveys. The purpose is to acquaint students with notation and summation rather than to present statistical concepts. Initially some of the expressions might seem complex or abstract, but nothing more than sequences of operations involving addition, subtraction, multiplication, and division is involved. Exercises are included so a student may test his interpretation of different mathematical expressions. Algebraic manipulations are also discussed and some algebraic exercises are included. To a considerable degree, this chapter could be regarded as a manual of exercises for students who are interested in sampling but are not fully familiar with the summation symbol, Σ. Familiarity with the mathematical language will make the study of sampling much easier.

1.2 NOTATION AND THE SYMBOL FOR SUMMATION

"Element" will be used in this monograph as a general expression for a unit that a measurement pertains to. An element might be a farm, a person, a school, a stalk of corn. or an animal. Such units are sometimes called units of observation or reporting units. Generally, there are several characteristics or items of information about an element that one might be interested in.

"Measurement" or "value" will be used as **general terms for the** numerical value of a specified characteristic for an element. This includes assigned values. For example, the element might be a farm and the characteristic could be whether wheat is being grown or is not being grown on a farm. A value of "1" could be assigned to a farm growing wheat and a value of "0" to a farm not growing wheat. Thus, the "measurement" or "value" for a farm growing wheat would be "1" and for a farm not growing wheat the value would be "0."

Typically, a set of measurements of N elements will be expressed as follows: $X_1, X_2, \ldots, X_N$ where X refers to the characteristic that is measured and the index (subscript) to the various elements of the population (or set). For example, if there are N persons and the characteristic X is a person's height, then X_1 is the height of the first person, etc. To refer to any one of elements, not a specific element, a subscript "i" is used. Thus, X_i (read X sub i) means the value of X for any one of the N elements. A common expression would be "X_i is the value of X for the i^{th} element."

The Greek letter Σ (capital sigma) is generally used to indicate a sum. When found in an equation, it means "the sum of." For example, $\sum_{i=1}^{N} X_i$ represents the sum of all values of X from X_1 to X_N; that is,

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + \ldots + X_N.$$

The lower and upper limits of the index of summation are shown below and above the summation sign. For example, to specify the sum of X for elements 11 thru 20 one would write $\sum_{i=11}^{20} X_i$.

You might also see notation such as "ΣX_i where $i = 1, 2, \ldots, N$" which indicates there are N elements (or values) in the set indexed by serial numbers 1 thru N, or for part of a set you might see "ΣX_i where $i = 11, 12, \ldots, 20$." Generally the index of summation starts with 1; so you will often see a summation written as $\sum_i^N X_i$. That is, only the upper limit of the summation is shown and it is understood that the summation begins with $i=1$. Alternatively, when the set of values being summed is clearly understood, the lower and upper limits might not be shown. Thus, it is understood that $\sum_i X_i$ or ΣX_i is the sum of X over all values of the set under consideration. Sometimes a writer will even drop the subscript and use ΣX for the sum of all values of X. Usually the simplest notation that is adequate for the purpose is adopted. In this monograph, there will be some deliberate variation in notation to familiarize students with various representations of data.

An average is usually indicated by a "bar" over the symbol. For example, $\bar{X}$ (read "X bar," or sometimes "bar X") means the average value of X. Thus, $\bar{X} = \dfrac{\sum_{i=1}^{N} X_i}{N}$. In this case, showing the upper limit, N, of the summation makes it clear that the sum is being divided by the number of elements and $\bar{X}$ is the average of all elements. However, $\dfrac{\Sigma X_i}{N}$ would also be interpreted as the average of all values of X unless there is an indication to the contrary.

<u>Do not try to study mathematics without pencil and paper</u>. Whenever the shorthand is not clear, try writing it out in long form. This will often reduce any ambiguity and save time.

Here are some examples of mathematical shorthand:

(1) Sum of the reciprocals of X $\qquad \sum_{i=1}^{N} \frac{1}{X_i} = \frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}$

(2) Sum of the differences between X_i and a constant, C $\qquad \sum_{i=1}^{N} (X_i - C) = (X_1 - C) + (X_2 - C) + \ldots + (X_N - C)$

(3) Sum of the deviations of X_i from the average of X $\qquad \sum_{i}^{N} (X_i - \bar{X}) = (X_1 - \bar{X}) + (X_2 - \bar{X}) + \ldots + (X_N - \bar{X})$

(4) Sum of the absolute values of the differences between X_i and $\bar{X}$. (Absolute value, indicated by the vertical lines, means the positive value of the difference) $\qquad \Sigma|X_i - \bar{X}| = |X_1 - \bar{X}| + |X_2 - \bar{X}| + \ldots + |X_N - \bar{X}|$

(5) Sum of the squares of X_i $\qquad \Sigma X_i^2 = X_1^2 + X_2^2 + X_3^2 + \ldots X_N^2$

(6) Sum of squares of the deviations of X from $\bar{X}$ $\qquad \Sigma(X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + \ldots + (X_N - \bar{X})^2$

(7) Average of the squares of the deviations of X from $\bar{X}$ $\qquad \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N} = \frac{(X_1 - \bar{X})^2 + \ldots + (X_N - \bar{X})^2}{N}$

(8) Sum of products of X and Y $\qquad \sum_{i=1}^{N} X_i Y_i = X_1 Y_1 + X_2 Y_2 + \ldots + X_N Y_N$

(9) Sum of quotients of X divided by Y $\qquad \Sigma \frac{X_i}{Y_i} = \frac{X_1}{Y_1} + \frac{X_2}{Y_2} + \ldots + \frac{X_N}{Y_N}$

(10) Sum of X divided by the sum of Y $\qquad \frac{\Sigma X_i}{\Sigma Y_i} = \frac{X_1 + X_2 + \ldots + X_N}{Y_1 + Y_2 + \ldots + Y_N}$

(11) Sum of the first N digits $\qquad \sum_{i=1}^{N} i = 1+2+3+\ldots+N$

(12) $\qquad \sum_{i=1}^{N} iX_i = X_1 + 2X_2 + 3X_3 + \ldots + NX_N$

(13) $\qquad \sum_{i=1}^{6} (-1)^i X_i = -X_1 + X_2 - X_3 + X_4 - X_5 + X_6$

<u>Exercise 1.1</u>. You are given a set of four elements having the following values of X: $X_1 = 2$, $X_2 = 0$, $X_3 = 5$, $X_4 = 7$. To test your understanding of the summation notation, compute the values of the following algebraic expressions:

Expression		Answer
(1)	$\sum_{i=1}^{4} (X_i+4)$	30
(2)	$\Sigma 2(X_i-1)$	20
(3)	$2\Sigma(X_i-1)$	20
(4)	$\Sigma 2X_i-1$	27
(5)	$\bar{X} = \frac{\Sigma X_i}{N}$	3.5
(6)	ΣX_i^2	78
(7)	$\Sigma(-X_i)^2$	78
(8)	$[\Sigma X_i]^2$	196
(9)	$\Sigma(X_i^2 - X_i)$	64
(10)	$\Sigma(X_i^2) - \Sigma X_i$	64
(11)	$\Sigma i(X_i)$	45
(12)	$\Sigma(-1)^i(X_i)$	0
(13)	$\sum_{i=1}^{4} (X_i^2 - 3)$	66
(14)	$\sum_{i=1}^{4} X_i^2 - \sum_{i=1}^{4} (3)$	66

Note: $\sum_{i=1}^{4} (3)$ means find the sum of four 3's

Expression (Continued)	Answer
(15) $\Sigma(X_i - \bar{X})$	0
(16) $\frac{\Sigma(X_i - \bar{X})^2}{N-1}$	$\frac{29}{3}$
(17) $\frac{\Sigma[X_i^2 - 2X_i\bar{X} + \bar{X}^2]}{N-1}$	$\frac{29}{3}$
(18) $\frac{\Sigma X_i^2 - N\bar{X}^2}{N-1}$	$\frac{29}{3}$

Definition 1.1. The variance of X where $X = X_1, X_2, \ldots, X_N$, is defined in one of two ways:

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

or

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

The reason for the two definitions will be explained in Chapter III. The variance formulas provide measures of how much the values of X vary (deviate) from the average. The square root of the variance of X is called the standard deviation of X. The central role that the above definitions of variance and standard deviation play in sampling theory will become apparent as you study sampling. The variance of an estimate from a sample is one of the measures needed to judge the accuracy of the estimate and to evaluate alternative sampling designs. Much of the algebra and notation in this chapter is related to computation of variance. For complex sampling plans, variance formulas are complex. This chapter should help make the mathematics used in sampling more readable and more meaningful when it is encountered.

Definition 1.2. "Population" is a statistical term that refers to a set of elements from which a sample is selected ("Universe" is often used instead of "Population").

Some examples of populations are farms, retail stores, students, households, manufacturers, and hospitals. A complete definition of a population is a detailed specification of the elements that compose it. Data to be collected also need to be defined. Problems of defining populations to be surveyed should receive much attention in courses on sampling. From a defined population a sample of elements is selected, information for each element in the sample is collected, and inferences from the sample are made about the population. Nearly all populations for sample surveys are finite so the mathematics and discussion in this monograph are limited to finite populations.

In the theory of sampling, it is important to distinguish between data for elements in a sample and data for elements in the entire population. Many writers use uppercase letters when referring to the population and lowercase letters when referring to a sample. Thus $X_1,\ldots, X_N$ would represent the values of some characteristic X for the N elements of the population; and $x_1,\ldots, x_n$ would represent the values of X in a sample of n elements. The subscripts in $x_1,\ldots, x_n$ simply index the different elements in a sample and do not correspond to the subscripts in $X_1,\ldots, X_N$ which index the elements of the population. In other words, x_1 could be any one of the X_i's. Thus,

$$\frac{\sum_{i}^{N} X_i}{N} = \bar{X} \text{ represents the population mean, and}$$

$$\frac{\sum_{i}^{n} x_i}{n} = \bar{x} \text{ represents a sample mean}$$

In this chapter we will be using only uppercase letters, except for constants and subscripts, because the major emphasis is on symbolic representation of data for a set of elements and on algebra. For this purpose, it is sufficient to start with data for a set of elements and not be concerned with whether the data are for a sample of elements or for all elements in a population.

The letters X, Y, and Z are often used to represent different characteristics (variables) whereas the first letters of the alphabet are commonly used as constants. There are no fixed rules regarding notation. For example, four different variables or characteristics might be called X_1, X_2, X_3, and X_4. In that case X_{1i} might be used to represent the i^{th} value of the variable X_1. Typically, writers adopt notation that is convenient for their problems. It is not practical to completely standardize notation.

Exercise 1.2. In the list of expressions in Exercise 1.1 find the variance of X, that is, find S^2. Suppose that X_4 is 15 instead of 7. How much is the variance of X changed? Answer: From $9\frac{2}{3}$ to $44\frac{1}{3}$.

Exercise 1.3. You are given four elements having the following values of X and Y

$$X_1 = 2 \qquad X_2 = 0 \qquad X_3 = 5 \qquad X_4 = 7$$

$$Y_1 = 2 \qquad Y_2 = 3 \qquad Y_3 = 1 \qquad Y_4 = 14$$

Find the value of the following expressions:

Expression	Answer	Expression	Answer
(1) $\Sigma X_i Y_i$	107	(7) $\Sigma X_i - \Sigma Y_i$	-6
(2) $(\Sigma X_i)(\Sigma Y_i)$	280	(8) $\Sigma(X_i - Y_i)^2$	74
(3) $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$	37	(9) $\Sigma(X_i^2 - Y_i^2)$	-132
(4) $\Sigma X_i Y_i - N\bar{X}\bar{Y}$	37	(10) $\Sigma X_i^2 - \Sigma Y_i^2$	-132
(5) $\frac{1}{N} \Sigma \frac{X_i}{Y_i}$	1.625	(11) $[\Sigma(X_i - Y_i)]^2$	36
(6) $\Sigma(X_i - Y_i)$	-6	(12) $[\Sigma X_i]^2 - [\Sigma Y_i]^2$	-204

1.3 FREQUENCY DISTRIBUTIONS

Several elements in a set of N might have the same value for some characteristic X. For example, many people have the same age. Let X_j be a particular age and let N_j be the number of people in a population (set) of N people who have the age X_j. Then $\sum_{j=1}^{K} N_j = N$ where K is the number of different ages found in the population. Also $\Sigma N_j X_j$ is the sum of the ages of the N people in the population and $\frac{\Sigma N_j X_j}{\Sigma N_j}$ represents the average age of the N people. A listing of X_j and N_j is called the frequency distribution of X, since N_j is the number of times (frequency) that the age X_j is found in the population.

On the other hand, one could let X_i represent the age of the i^{th} individual in a population of N people. Notice that j was an index of age. We are now using i as an index of individuals, and the average age would be written as $\frac{\Sigma X_i}{N}$. Note that $\Sigma N_j X_j = \Sigma X_i$ and that $\frac{\Sigma N_j X_j}{\Sigma N_j} = \frac{\Sigma X_i}{N}$. The choice between these two symbolic representations of the age of people in the population is a matter of convenience and purpose.

Exercise 1.4. Suppose there are 20 elements in a set (that is, $N = 20$) and that the values of X for the 20 elements are: 4, 8, 3, 7, 8, 8, 3, 3, 7, 2, 8, 4, 8, 8, 3, 7, 8, 10, 3, 8.

(1) List the values of X_j and N_j, where j is an index of the values 2, 3, 4, 7, 8, and 10. This is the frequency distribution of X.

(2) What is K equal to?

Interpret and verify the following by making the calculations indicated:

(3) $$\sum_{i=1}^{N} X_i = \sum_{j=1}^{K} N_j X_j$$

(4) $$\frac{\Sigma X_i}{N} = \frac{\Sigma N_j X_j}{\Sigma N_j} = \bar{X}$$

(5) $$\frac{\Sigma (X_i - \bar{X})^2}{N} = \frac{\Sigma N_j (X_j - \bar{X})^2}{\Sigma N_j}$$

1.4 ALGEBRA

In arithmetic and elementary algebra, the order of the numbers when addition or multiplication is performed does not affect the results. The familiar arithmetic laws when extended to algebra involving the summation symbol lead to the following important rules or theorems:

Rule 1.1 $\Sigma(X_i - Y_i + Z_i) = \Sigma X_i - \Sigma Y_i + \Sigma Z_i$

or $\Sigma(X_{1i} + X_{2i} + \ldots + X_{Ki}) = \Sigma X_{1i} + \Sigma X_{2i} + \ldots + \Sigma X_{Ki}$

Rule 1.2 $\Sigma a X_i = a \Sigma X_i$ where _a_ is a constant

Rule 1.3 $\Sigma(X_i + b) = \Sigma X_i + Nb$ where _b_ is constant

If it is not obvious that the above equations are correct, write both sides of each equation as series and note that the difference between the two sides is a matter of the order in which the summation (arithmetic) is performed. Note that the use of parentheses in Rule 1.3 means that b is contained in the series N times. That is,

$$\sum_{i=1}^{N} (X_i+b) = (X_1+b)+(X_2+b)+\ldots+(X_N+b)$$

$$= (X_1+X_2+\ldots+X_N) + Nb$$

On the basis of Rule 1.1, we can write

$$\sum_{i=1}^{N} (X_i+b) = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} b$$

The expression $\sum_{i=1}^{N} b$ means "sum the value of b, which occurs N times." Therefore,

$$\sum_{i=1}^{N} b = Nb.$$

Notice that if the expression had been $\sum_{i}^{N} X_i+b$, then b is an amount to add to the sum, $\sum_{i}^{N} X_i$.

In many equations $\bar{X}$ will appear; for example, $\sum_{i}^{N} \bar{X}X_i$ or $\sum_{i}^{N} (X_i-\bar{X})$.

Since $\bar{X}$ is constant with regard to the summation, $\Sigma\bar{X}X_i = \bar{X}\Sigma X_i$. Thus,

$\sum_{i}(X_i-\bar{X}) = \sum_{i} X_i-\sum_{i}\bar{X} = \sum_{i}X_i - N\bar{X}$. By definition, $\bar{X} = \dfrac{\sum_{i} X_i}{N}$. Therefore,

$N\bar{X} = \sum_{i} X_i$ and $\sum_{i}(X_i-\bar{X}) = 0$.

To work with an expression like $\sum_{i}^{N}(X_i+b)^2$ we must square the quantity in parentheses before summing. Thus,

$$\sum_i (X_i + b)^2 = \Sigma(X_i^2 + 2bX_i + b^2)$$

$$= \Sigma X_i^2 + \Sigma 2bX_i + \Sigma b^2 \quad \text{Rule 1}$$

$$= \Sigma X_i^2 + 2b\Sigma X_i + Nb^2 \quad \text{Rules 2 and 3}$$

Verify this result by using series notation. Start with $(X_1+b)^2+\ldots+(X_N+b)^2$.

It is very important that the ordinary rules of algebra pertaining to the use of parentheses be observed. Students frequently make errors because inadequate attention is given to the placement of parentheses or to the interpretation of parentheses. Until you become familiar with the above rules, practice translating shorthand to series and series to shorthand. Study the following examples carefully:

(1) $\Sigma(X_i)^2 \neq (\Sigma X_i)^2$ — The left-hand side is the sum of the squares of X_i. The right-hand side is the square of the sum of X_i. On the right the parentheses are necessary. The left side could have been written ΣX_i^2.

(2) $\Sigma\left[\frac{X_i}{N}\right]^2 = \frac{\Sigma X_i^2}{N^2}$ — Rule 1.2 applies.

(3) $\Sigma(X_i+Y_i)^2 \neq \Sigma X_i^2 + \Sigma Y_i^2$ — A quantity in parentheses must be squared before taking a sum.

(4) $\Sigma(X_i^2 + Y_i^2) = \Sigma X_i^2 + \Sigma Y_i^2$ — Rule 1.1 applies

(5) $\Sigma X_i Y_i \neq (\Sigma X_i)(\Sigma Y_i)$ — The left side is the sum of products. The right side is the product of sums.

(6) $\Sigma(X_i-Y_i)^2 = \Sigma X_i^2 - 2\Sigma X_i Y_i + \Sigma Y^2$

(7) $\sum_i^N a(X_i-b) \neq a\sum_i^N X_i - ab$

(8) $\sum_{i}^{N} a(X_i - b) = a\sum_{i}^{N} X_i - Nab$

(9) $a[\sum_{i}^{N} X_i - b] = a\sum_{i}^{N} X_i - ab$

(10) $\Sigma X_i(X_i - Y_i) = \Sigma X_i^2 - \Sigma X_i Y_i$

<u>Exercise 1.5</u>. Prove the following:

In all cases, assume $i = 1, 2, \ldots, N$.

(1) $\Sigma(X_i - \bar{X}) = 0$

(2) $\Sigma\frac{X_i Y_i}{X_i^2} = \Sigma\frac{Y_i}{X_i}$

(3) $N\bar{X}^2 = \frac{(\Sigma X_i)^2}{N}$

(4) $\sum_{i=1}^{N} (aX_i + bY_i + C) = a\Sigma X_i + b\Sigma Y_i + NC$

Note: Equations (5) and (6) should be (or become) very familiar equations.

(5) $\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - N\bar{X}^2$

(6) $\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma X_i Y_i - N\bar{X}\bar{Y}$

(7) $\Sigma(\frac{X_i}{a} + Y_i)^2 = \frac{1}{a^2} \Sigma(X_i + aY_i)^2$

(8) Let $Y_i = a+bX_i$, show that $\bar{Y} = a+b\bar{X}$

and $\Sigma Y_i^2 = Na(a+2b\bar{X}) + b^2 \Sigma X_i^2$

(9) Assume that $X_i = 1$ for N_1 elements of a set and that $X_i = 0$ for N_0 of the elements. The total number of elements in the set is $N = N_1 + N_0$. Let $\frac{N_1}{N} = P$ and $\frac{N_0}{N} = Q$. Prove that $\frac{\Sigma(X_i - \bar{X})^2}{N} = PQ$.

(10) $\Sigma(X_i-d)^2 = \Sigma(X_i-\bar{X})^2 + N(\bar{X}-d)^2$. Hint: Rewrite $(X_i-d)^2$ as $[(X_i-\bar{X})+(\bar{X}-d)]^2$. Recall from elementary algebra that $(a+b)^2 = a^2+2ab+b^2$ and think of $(X_i-\bar{X})$ as a and of $(\bar{X}-d)$ as b. For what value of d is $\Sigma(X_i-d)^2$ a minimum?

1.5 DOUBLE INDEXES AND SUMMATION

When there is more than one characteristic for a set of elements, the different characteristics might be distinguished by using a different letter for each or by an index. For example, X_i and Y_i might represent the number of acres of wheat planted and the number of acres of wheat harvested on the i^{th} farm. Or, X_{ij} might be used where i is the index for the characteristics and j is the index for elements; that is, X_{ij} would be the value of characteristic X_i for the j^{th} element. However, when data on each of several characteristics for a set of elements are to be processed in the same way, it might not be necessary to use notation that distinguishes the characteristics. Thus, one might say calculate $\frac{\Sigma(X_i-\bar{X})^2}{N-1}$ for all characteristics.

More than one index is needed when the elements are classified according to more than one criterion. For example, X_{ij} might represent the value of characteristic X for the j^{th} farm in the i^{th} county; or X_{ijk} might be the value of X for the k^{th} household in the j^{th} block in the i^{th} city. As another example, suppose the processing of data for farms involves classification of farms by size and type. We might let X_{ijk} represent the value of characteristic X for the k^{th} farm in the subset of farms classified as type j and size i. If N_{ij} is the number of farms classified as type j and size i, then $\frac{\sum_{k}^{N_{ij}} X_{ijk}}{N_{ij}} = \bar{X}_{ij.}$ is the average value of X for the subset of farms classified as type j and size i.

There are two general kinds of classification--cross classification and hierarchal or nested classification. Both kinds are often involved in the same problem. However, we will discuss each separately. An example of nested classification is farms within counties, counties within States, and States within regions. Cross classification means that the data can be arranged in two or more dimensions as illustrated in the next section.

1.5.1 CROSS CLASSIFICATION

As a specific illustration of cross classification and summation with two indexes, suppose we are working with the acreages of K crops on a set of N farms. Let X_{ij} represent the acreage of the i^{th} crop on the j^{th} farm where $i = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, N$. In this case, the data could be arranged in a K by N matrix as follows:

Row (i)	Column (j)					Row total
	1		j		N	
1	X_{11}	...	X_{1j}	...	X_{1N}	$\sum_j X_{1j}$
	.					
i	X_{i1}	...	X_{ij}	...	X_{iN}	$\sum_j X_{ij}$
	.					
K	X_{K1}	...	X_{Kj}	...	X_{KN}	$\sum_j X_{Kj}$
Column total	$\sum_i X_{i1}$		$\sum_i X_{ij}$		$\sum_i X_{iN}$	$\sum_{ij}\sum X_{ij}$

The expression $\sum_{j}^{N} X_{ij}$ (or $\sum_{j} X_{ij}$) means the sum of the values of X_{ij} for a fixed value of i. Thus, with reference to the matrix, $\sum_{j} X_{ij}$ is the total of the values of X in the i^{th} row; or, with reference to the example about farms and crop acreages, $\sum_{j} X_{ij}$ would be the total acreage on all farms of whatever the i^{th} crop is. Similarly, $\sum_{i}^{K} X_{ij}$ (or $\sum_{i} X_{ij}$) is the column total for the j^{th} column, which in the example is the total for the j^{th} farm of the acreages of the K crops under consideration. The sum of all values of X could be written as $\sum_{ij}^{KN}\sum X_{ij}$ or $\sum_{ij}\sum X_{ij}$.

Double summation means the sum of sums. Breaking a double sum into parts can be an important aid to understanding it. Here are two examples:

(1) $$\sum_{ij}^{KN}\sum X_{ij} = \sum_{j}^{N} X_{1j} + \sum_{j}^{N} X_{2j} +\ldots+ \sum_{j}^{N} X_{Kj} \tag{1.1}$$

With reference to the above matrix, Equation (1.1) expresses the grand total as the sum of row totals.

(2) $$\sum_{ij}^{KN}\sum X_{ij}(Y_{ij}+a) = \underbrace{\sum_{j}^{N} X_{1j}(Y_{1j}+a)} +\ldots+ \sum_{j}^{N} X_{Kj}(Y_{Kj}+a) \tag{1.2}$$

$$\sum_{j}^{N} X_{1j}(Y_{1j}+a) = X_{11}(Y_{11}+a) +\ldots+ X_{1N}(Y_{1N}+a)$$

In Equations (1.1) and (1.2) the double sum is written as the sum of K partial sums, that is, one partial sum for each value of i.

Exercise 1.6. (a) Write an equation similar to Equation (1.1) that expresses the grand total as the sum of column totals. (b) Involved in Equation (1.2) are KN terms, $X_{ij}(Y_{ij}+a)$. Write these terms in the form of a matrix.

The rules given in Section 1.4 also apply to double summation. Thus,

$$\sum_{ij}^{KN}\sum X_{ij}(Y_{ij}+a) = \sum_{ij}^{KN}\sum X_{ij}\, Y_{ij} + a \sum_{ij}^{KN}\sum X_{ij} \tag{1.3}$$

Study Equation (1.3) with reference to the matrix called for in Exercise 1.6(b). To fully understand Equation (1.3), you might need to write out intermediate steps for getting from the left-hand side to the right-hand side of the equation.

To simplify notation, a system of dot notation is commonly used, for example:

$$\sum_j X_{ij} = X_{i\cdot}$$

$$\sum_i X_{ij} = X_{\cdot j}$$

$$\sum_{ij}\sum X_{ij} = X_{\cdot\cdot}$$

The dot in $X_{i\cdot}$ indicates that an index in addition to i is involved and $X_{i\cdot}$ is interpreted as the sum of the values of X for a fixed value of i. Similarly, $X_{\cdot j}$ is the sum of X for any fixed value of j, and $X_{\cdot\cdot}$ represents a sum over both indexes. As stated above, averages are indicated by use of a bar. Thus $\bar{X}_{i\cdot}$ is the average of X_{ij} for a fixed value of i, namely $\frac{\sum_{j=1}^{N} X_{ij}}{N} = \bar{X}_{i\cdot}$ and $\bar{X}_{\cdot\cdot}$ would represent the average of all values of X_{ij}, namely $\frac{\sum_{ij}\sum X_{ij}}{NK}$.

Here is an example of how the dot notation can simplify an algebraic expression. Suppose one wishes to refer to the sum of the squares of the row totals in the above matrix. This would be written as $\sum_i (X_{i\cdot})^2$. The sum of squares of the row means would be $\sum_{i}(\bar{X}_{i.})^2$. Without the dot notation the corresponding expressions would be $\sum_{i}^{K}(\sum_{j}^{N}X_{ij})^2$ and $\sum_{i}^{K}\left[\frac{\sum_{j}^{N}X_{ij}}{N}\right]^2$. It is very important that the parentheses be used correctly. For example, $\sum_{i}^{K}(\sum_{j}^{N}X_{ij})^2$ is not the same as $\sum_{i}^{K}\sum_{j}^{N}X_{ij}^2$. Incidentally, what is the difference between the last two expressions?

Using the dot notation, the variance of the row means could be written as follows:

$$V(\bar{X}_{i.}) = \frac{\sum_{i}^{K}(\bar{X}_{i.}-\bar{X}_{..})^2}{K-1} \tag{1.4}$$

where V stands for variance and $V(\bar{X}_{i.})$ is an expression for the variance of $\bar{X}_{i.}$. Without the dot notation, or something equivalent to it, a formula for the variance of the row means would look much more complicated.

<u>Exercise 1.7</u>. Write an equation, like Equation (1.4), for the variance of the column means.

<u>Exercise 1.8</u>. Given the following values of X_{ij}

i	j = 1	2	3	4
1	8	11	9	14
2	10	13	11	14
3	12	15	10	17

Find the value of the following algebraic expressions:

Expression	Answer	Expression	Answer
(1) $\sum_{j}^{N} X_{1j}$	42	(9) $K\sum_{j}^{N}(\bar{X}_{.j}-\bar{X}_{..})^2$	54
(2) $\frac{\sum_{j}^{N} X_{2j}}{N}$	12	(10) $\sum\sum_{ij}^{KN}(X_{ij}-\bar{X}_{.j}-\bar{X}_{i.}+\bar{X}_{..})^2$	6
(3) $\bar{X}_{3.}$	13.5	(11) $\sum\sum_{ij}^{KN} X_{ij}^2 - \frac{\left[\sum\sum_{ij}^{KN} X_{ij}\right]^2}{KN}$	78
(4) $\sum X_{i4}$	45		
(5) $\sum\sum_{ij}^{KN} X_{ij}$	144	(12) $\frac{\sum_{i}^{K} X_{i.}^2}{N} - \frac{\left[\sum\sum_{ij}^{KN} X_{ij}\right]^2}{KN}$	18
(6) $\bar{X}_{..}$	12		
(7) $\sum\sum_{ij}^{KN}(X_{ij}-\bar{X}_{..})^2$	78	(13) $\sum_{j}^{N}(X_{1j}-\bar{X}_{1.})^2$	21
(8) $N\sum_{i}^{K}(\bar{X}_{i.}-\bar{X}_{..})^2$	18	(14) $\sum\sum_{ij}^{KN}(X_{ij}-\bar{X}_{i.})^2$	60

<u>Illustration 1.1</u>. To introduce another aspect of notation, refer to the matrix on Page 15 and suppose that the values of X in row one are to be multiplied by a_1, the values of X in row two by a_2, etc. The matrix would then be

$$
\begin{matrix}
a_1X_{11} & \cdots & a_1X_{1j} & \cdots & a_1X_{1N} \\
\vdots & & \vdots & & \vdots \\
a_iX_{i1} & \cdots & a_iX_{ij} & \cdots & a_iX_{iN} \\
\vdots & & \vdots & & \vdots \\
a_KX_{K1} & \cdots & a_KX_{Kj} & \cdots & a_KX_{KN}
\end{matrix}
$$

The general term can be written as a_iX_{ij} because the index of a and the index i in X_{ij} are the same. The total of all KN values of $a_i X_{ij}$ is $\sum_{ij}^{KN}\sum a_i X_{ij}$. Since a_i is constant with respect to summation involving j, we can place a_i ahead of the summation symbol $\sum_j^N$. That is, $\sum_{ij}\sum a_i X_{ij} = \sum_i a_i \sum_j X_{ij}$.

Exercise 1.9. Refer to the matrix of values of X_{ij} in Exercise 1.8. Assume that $a_1 = -1$, $a_2 = 0$, and $a_3 = 1$.

Calculate:

(1) $\sum_{ij}\sum a_i X_{ij}$

(2) $\sum_{ij}\sum \frac{a_i X_{ij}}{N}$

(3) $\sum_{ij}\sum a_i X_{ij}^2$ Answer:-296

Show algebraically that:

(4) $\sum_{ij}\sum a_i X_{ij} = \sum_j X_{3j} - \sum_j X_{1j}$

(5) $\sum_{ij}\sum \frac{a_i X_{ij}}{N} = \bar{X}_{3.} - \bar{X}_{1.}$

(6) $\sum\sum a_i X_{ij}^2 = \sum_j X_{3j}^2 - \sum_j X_{1j}^2$

Exercise 1.10. Study the following equation and if necessary write the summations as series to be satisfied that the equation is correct:

$$\sum_{ij}^{KN}\sum (aX_{ij}+bY_{ij}) = a\sum_{ij}\sum X_{ij} + b\sum_{ij}\sum Y_{ij}$$

Illustration 1.2. Suppose

$Y_{ij} = X_{ij}+a_i+b_j+c$ where $i = 1, 2,\ldots,K$ and $j = 1, 2,\ldots,N$

The values of Y_{ij} can be arranged in matrix format as follows:

$$\begin{array}{ccc}
Y_{11} = X_{11} + a_1+b_1+c & \cdots & Y_{1N} = X_{1N} + a_1+b_N+c \\
\vdots & Y_{ij} = X_{ij} + a_i+b_j+c & \vdots \\
Y_{K1} = X_{K1} + a_k+b_1+c & \cdots & Y_{KN} = X_{KN} + a_K+b_N+c
\end{array}$$

Notice that a_i is a quantity that varies from row to row but is constant within a row and that b_j varies from column to column but is constant within a column. Applying the rules regarding the summation symbols we have

$$\sum_j Y_{ij} = \sum_j (X_{ij}+a_i+b_j+c)$$

$$= \sum_j X_{ij} + Na_i + \sum_j b + Nc$$

$$\sum_i Y_{ij} = \sum_i (X_{ij}+a_i+b_j+c)$$

$$= \sum_i X_{ij} + \sum_i a_i + Kb_j+Kc$$

$$\sum_{ij}\sum Y_{ij} = \sum_{ij}\sum (X_{ij}+a_i+b_j+c)$$

$$= \sum_{ij}\sum X_{ij} + N\sum_i a_i + K\sum_j b_j + KNc$$

<u>Illustration 1.3</u>. We have noted that $\Sigma(X_iY_i)$ <u>does</u> <u>not</u> equal $(\Sigma X_i)(\Sigma Y_i)$. (See (1) and (2) in Exercise 1.3, and (5) on Page 12). But, $\sum_{ij}\sum X_iY_j = (\sum_i X_i)(\sum_j Y_j)$ where $i = 1, 2,\ldots,K$ and $j = 1, 2,\ldots,N$. This becomes clear if we write the terms of $\sum_{ij}\sum X_iY_j$ in matrix format as follows:

$$\begin{array}{ll}
 & \underline{\text{Row Totals}} \\
X_1Y_1 + X_1Y_2 +\ldots+ X_1Y_N & X_1\Sigma Y_j \\
+ X_2Y_1 + X_2Y_2 +\ldots+ X_2Y_N & X_2\Sigma Y_j \\
\vdots & \\
+ X_KY_1 + X_KY_2 +\ldots+ X_KY_N = \sum_{ij}\sum X_iY_j & \underline{X_K\Sigma Y_j} \\
 & \Sigma X_i\Sigma Y_j
\end{array}$$

The sum of the terms in each row is shown at the right. The sum of these row totals is $X_1\Sigma Y_j + \ldots + X_K\Sigma Y_j = (X_1 + \ldots + X_K)\Sigma Y_j = \Sigma X_i \Sigma Y_j$. One could get the same final result by adding the columns first. Very often intermediate summations are of primary interest.

Exercise 1.11. Verify that $\sum\limits_{ij}\Sigma X_i Y_j = (\Sigma X_i)(\Sigma Y_j)$ using the values of X and Y in Exercise 1.3. In Exercise 1.3 the subscript of X and the subscript of Y were the same index. In the expression $\sum\limits_{ij}\Sigma X_i Y_j$ that is no longer the case.

Exercise 1.12. Prove the following:

$$(1)\quad \sum_{ij}^{KN}\Sigma(a_iX_{ij}+b_j)^2 = \sum_i^K a_i^2 \sum_j^N X_{ij}^2 + 2\sum_i^K a_i \sum_j^N b_jX_{ij} + K\sum_j^N b_j^2$$

$$(2)\quad \sum_{ij}^{KN}\Sigma a_i(X_{ij}-\bar{X}_{i.})^2 = \sum_i^K a_i \sum_j^N X_{ij}^2 - N\sum_i^K a_i\bar{X}_{i.}^2$$

$$(3)\quad \sum_{ij}^{KN}\Sigma a_i(X_{ij}-\bar{X}_{i.})(Y_{ij}-\bar{Y}_{i.}) = \sum_i^K a_i \sum_j^N X_{ij}Y_{ij} - N\sum_i^K a_i\bar{X}_{i.}\bar{Y}_{i.}$$

1.5.2 HIERARCHAL OR NESTED CLASSIFICATION

A double index does not necessarily imply that a meaningful cross classification of the data can be made. For example, X_{ij} might represent the value of X for the j^{th} farm in the i^{th} county. In this case, j simply identifies a farm within a county. There is no correspondence, for example, between farm number 5 in one county and farm number 5 in another. In fact the total number of farms varies from county to county. Suppose there are K counties and N_i farms in the i^{th} county. The total of X for the i^{th} county could be expressed as $X_{i.} = \sum\limits_j^{N_i} X_{ij}$. In the present case $\sum\limits_i^K X_{ij}$ is meaningless. The total of all values of X is $\sum\limits_{ij}^{KN_i}\Sigma X_{ij}$.

When the classification is nested, the order of the subscripts (indexes) and the order of the summation symbols from left to right should be from the highest to lowest order of classification. Thus in the above example the index for farms was on the right and the summation symbol involving this index is also on the right. In the expression $\sum_{i}^{K}\sum_{j}^{N_i} X_{ij}$, summation with respect to i cannot take place before summation with regard to j. On the other hand, when the classification is cross classification the summations can be performed in either order.

In the example of K counties and N_i farms in the i^{th} county, and in similar examples, you may think of the data as being arranged in rows (or columns):

$$
\begin{array}{l}
X_{11}, X_{12}, \dots, X_{1N_1} \\
X_{21}, X_{22}, \dots, X_{2N_2} \\
\cdot \\
\cdot \\
\cdot \\
X_{K1}, X_{K2}, \dots, X_{KN_K}
\end{array}
$$

Here are two double sums taken apart for inspection:

$$
(1) \quad \sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{..})^2 = \underbrace{\sum_{j}^{N_1}(X_{1j}-\bar{X}_{..})^2}_{} + \dots + \sum_{j}^{N_K}(X_{Kj}-\bar{X}_{..})^2 \qquad (1.5)
$$

$$
\sum_{j}^{N_1}(X_{1j}-\bar{X}_{..})^2 = (X_{11}-\bar{X}_{..})^2 + \dots + (X_{1N_1}-\bar{X}_{..})^2
$$

Equation (1.5) is the sum of squares of the deviations, $(X_{ij}-\bar{X}_{..})$, of all values of X_{ij} from the overall mean. There are $\sum_{i}^{K} N_i$ values of X_{ij}, and

$\bar{X}_{..} = \frac{\sum\limits_{ij}^{KN_i} X_{ij}}{\sum\limits_{i}^{K} N_i}$. If there was no interest in identifying the data by counties, a single index would be sufficient. Equation (1.5) would then be $\sum\limits_{i}^{N}(X_i-\bar{X})^2$.

$$(2)\quad \sum\sum_{ij}^{KN_i}(X_{ij}-\bar{X}_{i.})^2 = \underbrace{\sum_{j}^{N_1}(X_{1j}-\bar{X}_{1.})^2}_{} +\ldots+ \sum_{j}^{N_K}(X_{Kj}-\bar{X}_{K.})^2 \qquad (1.6)$$

$$\sum_{j}^{N_1}(X_{1j}-\bar{X}_{1.})^2 = (X_{11}-\bar{X}_{1.})^2 +\ldots+ (X_{1N_1}-\bar{X}_{1.})^2$$

With reference to Equation (1.6) do you recognize $\sum\limits_{j}^{N_1}(X_{1j}-\bar{X}_{1.})^2$? It involves only the subset of elements for which $i = 1$, namely $X_{11}, X_{12}, \ldots X_{1N_1}$. Note that $\bar{X}_{1.}$ is the average value of X in this subset. Hence, $\sum\limits_{j}^{N_1}(X_{1j}-\bar{X}_{1.})^2$ is the sum of the squares of the deviations of the X's in this subset from the subset mean. The double sum is the sum of K terms and each of the K terms is a sum of squares for a subset of X's, the index for the subsets being i.

Exercise 1.13. Let X_{ij} represent the value of X for the j^{th} farm in the i^{th} county. Also, let K be the number of counties and N_i be the number of farms in the i^{th} county. Suppose the values of X are as follows:

$X_{11} = 3 \quad X_{12} = 1 \quad X_{13} = 5$

$X_{21} = 4 \quad X_{22} = 6$

$X_{31} = 0 \quad X_{32} = 5 \quad X_{33} = 1 \quad X_{34} = 2$

Find the value of the following expressions:

Expression	Answer
(1) $\sum\limits_{i}^{K} N_i$	9

Expression (Continued)	Answer		
(2) $\sum_{i}^{K}\sum_{j}^{N_i} X_{ij}$	27		
(3) $X_{..}$ and $\bar{X}_{..}$	27	3	
(4) $\sum_{j}^{N_1} X_{1j} = X_{1.}$	9		
(5) $X_{2.}$ and $X_{3.}$	10	8	
(6) $\bar{X}_{1.}$, $\bar{X}_{2.}$, and $\bar{X}_{3.}$	3	5	2
(7) $\dfrac{\Sigma N_i \bar{X}_{i.}}{\Sigma N_i}$	3		
(8) $\sum_{i}^{K}(\sum_{j}^{N_i} X_{ij})^2$ or $\sum_{i}^{K} X_{i.}^2$	245		
(9) $\sum_{i}\sum_{j}(X_{ij}-\bar{X}_{..})^2$	36		
(10) $\sum_{j}^{N_1}(X_{1j}-\bar{X}_{1.})^2$	8		
(11) $\sum_{j}^{N_i}(X_{ij}-\bar{X}_{i.})^2$	8, 2, and 14 for i = 1, 2, and 3 respectively		
(12) $\sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{i.})^2$	24		
(13) $\sum_{i}^{K} N_i(\bar{X}_{i.}-\bar{X}_{..})^2$	12		
(14) $\sum_{i}^{K}\dfrac{\left[\sum_{j}^{N_i} X_{ij}\right]^2}{N_i} - \dfrac{\left[\sum_{i}^{K}\sum_{j}^{N_i} X_{ij}\right]^2}{\Sigma N_j}$	12		
(15) $\sum_{i}^{K} N_i \bar{X}_{i.}^2 - N\bar{X}_{..}^2$	12		

Expressions (14) and (15) in Exercise 1.13 are symbolic representations of the same thing. By definition

$$\sum_{j}^{N_i} X_{ij} = X_{i\cdot}\ , \quad \sum_{ij}^{KN_i}\sum X_{ij} = X_{\cdot\cdot}\ , \quad \text{and } \sum_{i}^{K} N_i = N$$

Substitution in (14) gives

$$\sum_{i}^{K} \frac{X_{i\cdot}^2}{N_i} - \frac{X_{\cdot\cdot}^2}{N} \tag{1.7}$$

Also by definition $\frac{X_{i\cdot}}{N_i} = \bar{X}_{i\cdot}$ and $\frac{X_{\cdot\cdot}}{N} = \bar{X}_{\cdot\cdot}$. Therefore $\frac{X_{i\cdot}^2}{N_i} = N_i\bar{X}_{i\cdot}^2$ and $\frac{X_{\cdot\cdot}^2}{N} = N\bar{X}^2$. Hence, by substitution, Equation (1.7) becomes $\sum_{i}^{K} N_i\bar{X}_{i\cdot}^2 - N\bar{X}_{\cdot\cdot}^2$.

<u>Exercise 1.14</u>. Prove the following:

(1) $\sum_{ij}^{KN_i}\sum X_{i\cdot}X_{ij} = \sum_{i}^{K} X_{i\cdot}^2$

(2) $\sum_{ij}^{KN_i}\sum \bar{X}_{i\cdot}(X_{ij}-\bar{X}_{i\cdot}) = 0$

(3) $\sum_{i}^{K} N_i(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2 = \sum_{i}^{K} N_i\bar{X}_{i\cdot}^2 - N\bar{X}_{\cdot\cdot}^2$

Note that this equates (13) and (15) in Exercise 1.13. The proof is similar to the proof called for in part (5) of Exercise 1.5.

(4) $\sum_{ij}^{KN_i}\sum (a_iX_{ij}-b_i)^2 = \sum_{i}^{K} a_i^2 \sum_{j}^{N_i} X_{ij}^2 - 2\sum_{i}^{K} a_ib_iX_{i\cdot} + \sum_{i}^{K} N_ib_i^2$

1.6 THE SQUARE OF A SUM

In statistics, it is often necessary to work algebraically with the square of a sum. For example,

$$(\Sigma X_i)^2 = (X_1+X_2+\ldots+X_N)^2 = X_1^2+X_1X_2+\ldots+X_2^2+X_2X_1+\ldots+X_N^2+X_NX_1+\ldots$$

The terms in the square of the sum can be written in matrix form as follows:

$$\begin{matrix} X_1X_1 & X_1X_2 & \cdots & X_1X_j & \cdots & X_1X_N \\ X_2X_1 & X_2X_2 & \cdots & X_2X_j & \cdots & X_2X_N \\ \vdots & \vdots & & \vdots & & \vdots \\ X_iX_1 & X_iX_2 & \cdots & X_iX_j & \cdots & X_iX_N \\ \vdots & \vdots & & \vdots & & \vdots \\ X_NX_1 & X_NX_2 & \cdots & X_NX_j & \cdots & X_NX_N \end{matrix}$$

The general term in this matrix is X_iX_j where X_i and X_j come from the same set of X's, namely, $X_1,\ldots,X_N$. Hence, i and j are indexes of the same set. Note that the terms along the main diagonal are the squares of the value of X and could be written as ΣX_i^2 . That is, on the main diagonal i = j and $X_iX_j = X_iX_i = X_i^2$. The remaining terms are all products of one value of X with some other value of X. For these terms the indexes are never equal. Therefore, the sum of all terms not on the main diagonal can be expressed as $\sum\limits_{i \neq j} X_iX_j$ where $i \neq j$ is used to express the fact that the summation includes all terms where i is not equal to j, that is, all terms other than those on the main diagonal. Hence, we have shown that $(\Sigma X_i)^2 = \Sigma X_i^2 + \sum\limits_{i \neq j} X_iX_j$.

Notice the symmetry of terms above and below the main diagonal: $X_1X_2 = X_2X_1, X_1X_3 = X_3X_1$, etc. When symmetry like this occurs, instead of $\sum\limits_{i \neq j} X_iX_j$ you might see an equivalent expression $2\sum\limits_{i<j} X_iX_j$. The sum of all terms above the main diagonal is $\sum\limits_{i<j} X_iX_j$. Owing to the symmetry, the sum

of the terms below the main diagonal is the same. Therefore, $\sum_{i\neq j} X_iX_j = 2\sum_{i<j} X_iX_j$.

Exercise 1.15. Express the terms of $[\sum_{i=1}^{4} X_i]^2 = [X_1+X_2+X_3+X_4]^2$ in matrix format. Let $X_1 = 2$, $X_2 = 0$, $X_3 = 5$, and $X_4 = 7$. Compute the values of ΣX_i^2 , $2\sum_{i<j} X_iX_j$, and $[\Sigma X_i]^2$. Show that $[\Sigma X_i]^2 = \Sigma X_i^2 + 2\sum_{i<j} X_iX_j$.

An important result, which we will use in Chapter 3, follows from the fact that

$$[\Sigma X_i]^2 = \Sigma X_i^2 + \sum_{i\neq j} X_iX_j \tag{1.8}$$

Let $X_i = Y_i-\bar{Y}$. Substituting $(Y_i-\bar{Y})$ for X_i in Equation 1.8 we have

$$[\Sigma(Y_i-\bar{Y})]^2 = \Sigma(Y_i-\bar{Y})^2 + \sum_{i\neq j}(Y_i-\bar{Y})(Y_j-\bar{Y})$$

We know that $[\Sigma(Y_i-\bar{Y})]^2 = 0$ because $\Sigma(Y_i-\bar{Y}) = 0$. Therefore,

$$\Sigma(Y_i-\bar{Y})^2 + \sum_{i\neq j}(Y_i-\bar{Y})(Y_j-\bar{Y}) = 0$$

It follows that $\sum_{i\neq j}(Y_i-\bar{Y})(Y_j-\bar{Y}) = -\Sigma(Y_i-\bar{Y})^2$ (1.9)

Exercise 1.16. Consider

$$\begin{aligned}\sum_{i\neq j}(Y_i-\bar{Y})(Y_j-\bar{Y}) &= \sum_{i\neq j}(Y_iY_j - \bar{Y}Y_i - \bar{Y}Y_j + \bar{Y}^2)\\ &= \sum_{i\neq j} Y_iY_j - \bar{Y}\sum_{i\neq j} Y_i - \bar{Y}\sum_{i\neq j} Y_j + \sum_{i\neq j}\bar{Y}^2\end{aligned}$$

Do you agree that $\sum_{i\neq j}\bar{Y}^2 = N(N-1)\bar{Y}^2$? With reference to the matrix layout, $\bar{Y}^2$ appears N^2 times but the specification is $i \neq j$ so we do not want to count the N times that $\bar{Y}^2$ is on the main diagonal. Try finding the values of $\sum_{i\neq j} X_i$ and $\sum_{i\neq j} X_j$ and then show that

$$\sum_{i=j} (Y_i-\bar{Y})(Y_j-\bar{Y}) = \sum_{i\neq j} Y_iY_j - N(N-1)\bar{Y}^2$$

Hint: Refer to a matrix layout. In $\sum_{i\neq j} Y_i$ how many times does Y_1 appear? Does Y_2 appear the same number of times?

1.7 SUMS OF SQUARES

For various reasons statisticians are interested in components of variation, that is, measuring the amount of variation attributable to each of more than one source. This involves computing sums of squares that correspond to the different sources of variation that are of interest. We will discuss a simple example of nested classification and a simple example of cross classification.

1.7.1 NESTED CLASSIFICATION

To be somewhat specific, reference is made to the example of K counties and N_i farms in the i^{th} county. The sum of the squares of the deviations of X_{ij} and $\bar{X}_{..}$ can be divided into two parts as shown by the following formula:

$$\sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{..})^2 = \sum_{i}^{K} N_i(\bar{X}_{i.}-\bar{X}_{..})^2 + \sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{i.})^2 \qquad (1.10)$$

The quantity on the left-hand side of Equation (1.10) is called the total sum of squares. In Exercise 1.13, Part (9), the total sum of squares was 36.

The first quantity on the right-hand side of the equation involves the squares of $(\bar{X}_{i.}-\bar{X}_{..})$, which are deviations of the class means from the overall mean. It is called the between class sum of squares or with reference to the example the between county sum of squares. In Exercise 1.13, Part (13), the between county sum of squares was computed. The answer was 12.

The last term is called the within sum of squares because it involves deviations within the classes from the class means. It was presented previously. See Equation (1.6) and the discussion pertaining to it. In Exercise 1.13, the within class sum of squares was 24, which was calculated in Part (12). Thus, from Exercise 1.13, we have the total sum of squares, 36, which equals the between, 12, plus the within, 24. This verifies Equation (1.10).

The proof of Equation 1.10 is easy if one gets started correctly. Write $X_{ij}-\bar{X}_{..} = (X_{ij}-\bar{X}_{i.}) + (\bar{X}_{i.}-\bar{X}_{..})$. This simple technique of adding and subtracting $\bar{X}_{i.}$ divides the deviation $(X_{ij}-\bar{X}_{..})$ into two parts. The proof proceeds as follows:

$$\begin{aligned}
\sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{..})^2 &= \sum_{i}\sum_{j}[(X_{ij}-\bar{X}_{i.}) + (\bar{X}_{i.}-\bar{X}_{..})]^2 \\
&= \sum_{i}\sum_{j}[(X_{ij}-\bar{X}_{i.})^2 + 2(X_{ij}-\bar{X}_{i.})(\bar{X}_{i.}-\bar{X}_{..}) + (\bar{X}_{i.}-\bar{X}_{..})^2] \\
&= \sum_{i}\sum_{j}(X_{ij}-\bar{X}_{i.})^2 + 2\sum_{i}\sum_{j}(X_{ij}-\bar{X}_{i.})(\bar{X}_{i.}-\bar{X}_{..}) + \sum_{i}\sum_{j}(\bar{X}_{i.}-\bar{X}_{..})^2
\end{aligned}$$

<u>Exercise 1.17</u>. Show that $\sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{i.})(\bar{X}_{i.}-\bar{X}_{..}) = 0$

and that $\sum_{i}^{K}\sum_{j}^{N_i}(\bar{X}_{i.}-\bar{X}_{..})^2 = \sum_{i}^{K} N_i(\bar{X}_{i.}-\bar{X}_{..})^2$

Completion of Exercise 1.17 completes the proof.

Equation (1.10) is written in a form which displays its meaning rather than in a form that is most useful for computational purposes. For computation purposes, the following relationships are commonly used:

$$\text{Total} = \sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{..})^2 = \sum_{i}\sum_{j}X_{ij}^2-N\bar{X}_{..}^2$$

$$\text{Between} = \sum_{i}^{K} N_i(\bar{X}_{i.}-\bar{X}_{..})^2 = \sum_{i} N_i\bar{X}_{i.}^2-N\bar{X}_{..}^2$$

$$\text{Within} = \sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{i.})^2 = \sum_{i}\sum_{j} X_{ij}^2-\sum_{i} N_i\bar{X}_{i.}^2$$

$$\text{where } N = \sum_{i}^{K} N_i \;, \quad \bar{X}_{i.} = \frac{\sum_{j}^{N_i} X_{ij}}{N_i}, \text{ and } \bar{X}_{..} = \frac{\sum_{i}^{K}\sum_{j}^{N_i} X_{ij}}{N}$$

Notice that the major part of arithmetic reduces to calculating $\sum_{i}^{K}\sum_{j}^{N_i} X_{ij}^2$, $\sum_{i}^{K} N_i\bar{X}_{i.}^2$, and $N\bar{X}_{..}^2$. There are variations of this that one might use. For example, one could use $\sum_{i}^{K} \frac{X_{i.}^2}{N_i}$ instead of $\sum_{i}^{K} N_i\bar{X}_{i.}^2$.

<u>Exercise 1.18</u>. Show that

$$\sum_{i}^{K}\sum_{j}^{N_i}(X_{ij}-\bar{X}_{i.})^2 = \sum_{i}\sum_{j} X_{ij}^2-\sum_{i} N_i\bar{X}_{i.}^2$$

A special case that is useful occurs when $N_i = 2$. The within sum of squares becomes

$$\sum_{i}^{K}\sum_{j}^{2}(X_{ij}-\bar{X}_{i.})^2 = \sum_{i}^{K}[(X_{i1}-\bar{X}_{i.})^2 + (X_{i2}-\bar{X}_{i.})^2]$$

Since $\bar{X}_{i.} = \frac{X_{i1}+X_{i2}}{2}$ it is easy to show that

$$(X_{i1}-\bar{X}_{i.})^2 = \frac{1}{4}(X_{i1}-X_{i2})^2$$

$$\text{and } (X_{i2}-\bar{X}_{i.})^2 = \frac{1}{4}(X_{i1}-X_{i2})^2$$

Therefore the within sum of squares is

$$\frac{1}{2}\sum_{i}^{K}(X_{i1}-X_{i2})^2$$

which is a convenient form for computation.

1.7.2 CROSS CLASSIFICATION

Reference is made to the matrix on Page 15 and to Exercise 1.8. The total sum of squares can be divided into three parts as shown by the following formula:

$$\sum_{ij}^{KN}(X_{ij}-\bar{X}_{..})^2 = N\sum_{i}^{K}(\bar{X}_{i.}-\bar{X}_{..})^2 + K\sum_{j}^{N}(\bar{X}_{.j}-\bar{X}_{..})^2 + \sum_{ij}^{KN}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..})^2 \qquad (1.11)$$

Turn to Exercise 1.8 and find the total sum of squares and the three parts. They are:

	Sum of Squares
Total	78
Rows	18
Columns	54
Remainder	6

The three parts add to the total which verifies Equation (1.11). In Exercise 1.8, the sum of squares called remainder was computed directly (see Part (10) of Exercise 1.8). In practice, the remainder sum of squares is usually obtained by subtracting the row and column sum of squares from the total.

Again, the proof of Equation (1.11) is not difficult if one makes the right start. In this case the deviation, $(X_{ij}-\bar{X}_{..})$, is divided into three parts by adding and subtracting $\bar{X}_{i.}$ and $\bar{X}_{.j}$ as follows:

$$(X_{ij}-\bar{X}_{..}) = (\bar{X}_{i.}-\bar{X}_{..}) + (\bar{X}_{.j}-\bar{X}_{..}) + (X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..}) \qquad (1.12)$$

Exercise 1.19. Prove Equation (1.11) by squaring both sides of Equation (1.12) and then doing the summation. The proof is mostly a matter of showing that the sums of the terms which are products (not squares) are zero. For example, showing that $\sum_{ij}^{KN}(\bar{X}_{i.}-\bar{X}_{..})(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..}) = 0$.

CHAPTER II. RANDOM VARIABLES AND PROBABILITY

2.1 RANDOM VARIABLES

The word "random" has a wide variety of meanings. Its use in such terms as "random events," "random variable," or "random sample," however, implies a random process such that the probability of an event occurring is known a priori. To select a random sample of elements from a population, tables of random numbers are used. There are various ways of using such tables to make a random selection so any given element will have a specified probability of being selected.

The theory of probability sampling is founded on the concept of a random variable which is a variable that, by chance, might equal any one of a defined set of values. The value of a random variable on any particular occasion is determined by a random process in such a way that the chance (probability) of its being equal to any specified value in the set is known. This is in accord with the definition of a probability sample which states that every element of the population must have a known probability (greater than zero) of being selected. A primary purpose of this chapter is to present an elementary, minimum introduction or review of probability as background for the next chapter on expected values of a random variable. This leads to a theoretical basis for sampling and for evaluating the accuracy of estimates from a probability-sample survey.

In sampling theory, we usually start with an assumed population of N elements and a measurement for each element of some characteristic X. A typical mathematical representation of the N measurements or values is $X_1,\ldots,X_i,\ldots,X_N$ where X_i is the value of the characteristic X for the i^{th} element. Associated with the i^{th} element is a probability P_i, which is the probability of obtaining it when one element is selected at random from the set of N. The P_i's will be called selection probabilities. If each element has an equal chance of selection, $P_i = \frac{1}{N}$. The P_i's need not be equal, but we will specify that each $P_i>0$. When referring to the probability of X being equal to X_i we will use $P(X_i)$ instead of P_i.

We need to be aware of a distinction between selection probability and inclusion probability, the latter being the probability of an element being included in a sample. In this chapter, much of the discussion is oriented toward selection probabilities because of its relevance to finding expected values of estimates from samples of various kinds.

Definition 2.1. A random variable is a variable that can equal any value X_i, in a defined set, with a probability $P(X_i)$.

When an element is selected at random from a population and a measurement of a characteristic of it is made, the value obtained is a random variable. As we shall see later, if a sample of elements is selected at random from a population, the sample average and other quantities calculated from the sample are random variables.

Illustration 2.1. One of the most familiar examples of a random variable is the number of dots that happen to be on the top side of a die when it comes to rest after a toss. This also illustrates the concept of probability that we are interested in; namely, the relative frequency with which a particular outcome will occur in reference to a defined set of possible outcomes. With a die there are six possible outcomes and we expect each to occur with the same frequency, 1/6, assuming the die is tossed a very large or infinite number of times. Implicit in a statement that each side of a die has a probability of 1/6 of being the top side are some assumptions about the physical structure of the die and the "randomness" of the toss.

The additive and multiplicative laws of probability can be stated in several ways depending upon the context in which they are to be used. In sampling, our interest is primarily in the outcome of one random selection or of a series of random selections that yields a probability sample. Hence, the rules or theorems for the addition or multiplication of probabilities will be stated or discussed only in the context of probability sampling.

2.2 ADDITION OF PROBABILITIES

Assume a population of N elements and a variable X which has a value X_i for the i^{th} element. That is, we have a set of values of X, namely $X_1,\ldots,X_i,\ldots,X_N$. Let $P_1,\ldots,P_i,\ldots,P_N$ be a set of selection probabilities where P_i is the probability of selecting the i^{th} element when a random selection is made. We specify that each P_i must be greater than zero and that $\sum_i^N P_i = 1$. When an element is selected at random, the probability that it is either the i^{th} element <u>or</u> the j^{th} element is $P_i + P_j$. This addition rule can be stated more generally. Let P_s be the sum of the selection probabilities for the elements in a subset of the N elements. When a random selection is made from the whole set, P_s is the probability that the element selected is from the subset and $1-P_s$ is the probability that it is not from the subset. With reference to the variable X, let $P(X_i)$ represent the probability that X equals X_i . Then $P(X_i)+P(X_j)$ represents the probability that X equals either X_i or X_j; and $P_s(X)$ could be used to represent the probability that X is equal to one of the values in the subset.

Before adding (or subtracting) probabilities one should determine whether the events are mutually exclusive and whether all possible events have been accounted for. Consider two subsets of elements, subset A and subset B, of a population of N elements. Suppose one element is selected at random. What is the probability that the selected element is a member of either subset A or subset B? Let P(A) be the probability that the selected element is from subset A; that is, P(A) is the sum of the selection probabilities for elements in subset A. P(B) is defined similarly. If the two subsets are mutually exclusive, which means that no element is in both subsets, the probability that the element selected is from either subset A or subset B is P(A) + P(B). If some elements are in both subsets, see Figure 2.1, then event A (which is the selected element being a member of subset A) and event B (which is the selected element being a member of subset B) are not mutually exclusive events. Elements included in both subsets are counted once in P(A) and once in P(B). Therefore, we must subtract P(A,B) from P(A) + P(B) where P(A,B) is the sum of the probabilities for the elements that belong to both subset A and subset B. Thus,

$$P(A \text{ or } B) = P(A) + P(B) - P(A,B)$$

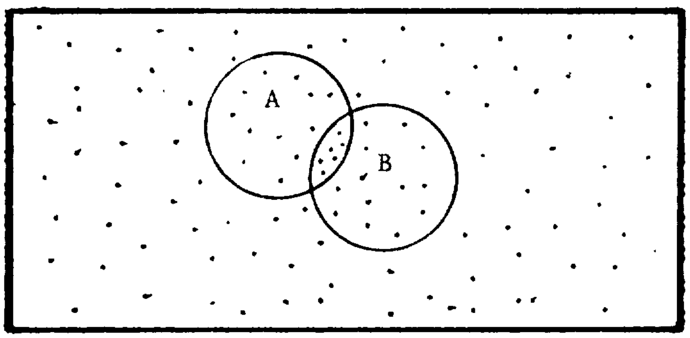

Figure 2.1

To summarize, the additive law of probability as used above could be stated as follows: If A and B are subsets of a set of all possible outcomes that could occur as a result of a random trial or selection, the probability that the outcome is in subset A or in subset B is equal to the probability that the outcome is in A plus the probability that it is in B minus the probability that it is in both A and B.

The additive law of probability extends without difficulty to three or more subsets. Draw a figure like Figure 2.1 with three subsets so that some points are common to all three subsets.. Observe that the additive law extends to three subsets as follows:

$$P(A \text{ or } B \text{ or } C)=P(A)+P(B)+P(C)-P(A,B)-P(A,C)-P(B,C)+P(A,B,C)$$

As a case for further discussion purposes, assume a population of N elements and two criteria for classification. A two-way classification of the elements could be displayed in the format of Table 2.1.

Table 2.1--A two-way classification of N elements

Y class	X class					Total
	1	...	j	...	s	
1	N_{11},P_{11}	...	N_{1j},P_{1j}	...	N_{1s},P_{1s}	$N_{1\cdot},P_{1\cdot}$
.	. .					
.	. .					
.	. .					
i	N_{i1},P_{i1}	...	N_{ij},P_{ij}	...	N_{is},P_{is}	$N_{i\cdot},P_{i\cdot}$
.	. .					
.	. .					
.	. .					
t	N_{t1},P_{t1}	...	N_{tj},P_{tj}	...	N_{ts},P_{ts}	$N_{t\cdot},P_{t\cdot}$
Total	N.1		N.j		N.s	N,P=1

The columns represent a classification of the elements in terms of criterion X; the rows represent a classification in terms of criterion Y; N_{ij} is the number of elements in X class j and Y class i; and P_{ij} is the sum of the selection probabilities for the elements in X class j and Y class i. Any one of the N elements can be classified in one and only one of the t times s cells.

Suppose one element from the population of N is selected. According to the additive law of probability we can state that

$\sum_i P_{ij} = P_{\cdot j}$ is the probability that the element selected is from X class j, and

$\sum_j P_{ij} = P_{i\cdot}$ is the probability that the element selected is from Y class i, where

P_{ij} is the probability that the element selected is from (belongs to both) X class j and Y class i.

The probabilities $P_{\cdot j}$ and $P_{i\cdot}$ are called marginal probabilities.

The probability that one randomly selected element is from X class j <u>or</u> from Y-class i is $P_{\cdot j} + P_{i\cdot} - P_{ij}$. (The answer is not $P_{\cdot j} + P_{i\cdot}$ because in $P_{\cdot j} + P_{i\cdot}$ there are N_{ij} elements in X class j and Y class i that are counted twice.)

If the probabilities of selection are equal, $P_{ij} = \frac{N_{ij}}{N}$, $P_{\cdot j} = \frac{N_{\cdot j}}{N}$, and $P_{i\cdot} = \frac{N_{i\cdot}}{N}$.

<u>Illustration 2.2</u>. Suppose there are 5,000 students in a university. Assume there are 1,600 freshmen, 1,400 sophomores, and 500 students living in dormitory A. From a list of the 5,000 students, one student is selected at random. Assuming each student had an equal chance of selection, the probability that the selected student is a freshman is $\frac{1600}{5000}$, that he is a sophomore is $\frac{1400}{5000}$, and that he is either a freshman or a sophomore is $\frac{1600}{5000} + \frac{1400}{5000}$. Also, the probability that the selected student lives in dormitory A

is $\frac{500}{5000}$. But, what is the probability that the selected student is either a freshman <u>or</u> lives in dormitory A? The question involves two classifications: one pertaining to the student's class and the other to where the student lives. The information given about the 5000 students could be arranged as follows:

Dormitory	Class			Total
	Freshmen	Sophomores	Others	
A				500
Other				4500
Total	1600	1400	2000	5000

From the above format, one can readily observe that the answer to the question depends upon how many freshmen live in dormitory A. If the problem had stated that 200 freshmen live in dormitory A, the answer would have been $\frac{1600}{5000} + \frac{500}{5000} - \frac{200}{5000}$.

Statements about probability need to be made and interpreted with great care. For example, it is not correct to say that a student has a probability of 0.1 of living in dormitory A simply because 500 students out of 5000 live in A. Unless students are assigned to dormitories by a random process with known probabilities there is no basis for stating a student's probability of living in (being assigned to) dormitory A. We are considering the outcome of a random selection.

<u>Exercise 2.1</u>. Suppose one has the following information about a population of 1000 farms:

600 produce corn

500 produce soybeans

300 produce wheat

100 produce wheat and corn

200 have one or more cows

all farms that have cows also produce corn

200 farms do not produce any crops

One farm is selected at random with equal probability from the list of 1000. What is the probability that the selected farm,

(a) produces corn? Answer: 0.6

(b) does not produce wheat?

(c) produces corn but no wheat? Answer: 0.5

(d) produces corn or wheat but not both?

(e) has no cows? Answer: 0.8

(f) produces corn or soybeans?

(g) produces corn and has no cows? Answer: 0.4

(h) produces either corn, cows, or both?

(i) does not produce corn or wheat?

One of the above questions cannot be answered.

Exercise 2.2. Assume a population of 10 elements and selection probabilities as follows:

Element	X_i	P_i	Element	X_i	P_i
1	2	.05	6	11	.15
2	7	.10	7	2	.20
3	12	.08	8	8	.05
4	0	.02	9	6	.05
5	8	.20	10	3	.10

One element is selected at random with probability P_i.

Find:

(a) $P(X=2)$, the probability that $X = 2$.

(b) $P(X>10)$, the probability that X is greater than 10.

(c) $P(X\leq 2)$, the probability that X is equal to or less than 2.

(d) $P(3<X>10)$, the probability that X is greater than 3 and less than 10

(e) $P(X\leq 3 \text{ or } X\geq 10)$, the probability that X is either equal to or less than 3 or is equal to or greater than 10.

Note: The answer to (d) and the answer to (e) should add to 1.

So far, we have been discussing the probability of an event occurring as a result of a single random selection. When more than one random selection occurs simultaneously or in succession the multiplicative law of probability is useful.

2.3 MULTIPLICATION OF PROBABILITIES

Assume a population of N elements and selection probabilities $P_1,\ldots,P_i,\ldots,P_N$. Each P_i is greater than zero and $\sum_i^N P_i = 1$. Suppose two elements are selected but before the second selection is made the first element selected is returned to the population. In this case the outcome of the first selection does not change the selection probabilities for the second selection. The two selections (events) are independent. The probability of selecting the i^{th} element first and the j^{th} element second is, P_iP_j, the product of the selection probabilities P_i and P_j. If a selected element is not returned to the population before the next selection is made, the selection probabilities for the next selection are changed. The selections are dependent.

The multiplicative law of probability, for two independent events A and B, states that the joint probability of A and B happening in the order A,B is equal to the probability that A happens times the probability that B happens. In equation form, $P(AB) = P(A)P(B)$. For the order B,A, $P(BA) = P(B)P(A)$ and we note that $P(AB) = P(BA)$. Remember, independence means that the probability of B happening is not affected by the occurrence of A and vice versa. The multiplicative law extends to any number of independent events. Thus, $P(ABC) = P(A)P(B)P(C)$.

For two dependent events A and B, the multiplicative law states that the joint probability of A and B happening in the order A,B is equal to the probability of A happening times the probability that B happens under the condition that A has already happened. In equation form $P(AB) = P(A)P(B|A)$; or for the order B,A we have $P(BA) = P(B)P(A|B)$. The vertical bar can usually be translated as "given" or "given that." The notation on the left of the bar refers to the event under consideration and the notation on the right to a condition under which the event can take place. $P(B|A)$ is called conditional probability and could be read "the probability of B, given that A has already happened," or simply "the probability of B given A." When the events are independent, $P(B|A) = P(B)$; that is, the conditional probability of B occurring is the same as the unconditional probability of B. Extending the multiplication rule to a series of three events A,B,C occurring in that order, we have $P(ABC) = P(A)P(B|A)P(C|AB)$ where $P(C|AB)$ is the probability of C occurring, given that A and B have already occurred.

2.4 SAMPLING WITH REPLACEMENT

When a sample is drawn and each selected element is returned to the population before the next selection is made, the method of sampling is called "sampling with replacement." In this case, the outcome of one selection does not change the selection probabilities for another selection.

Suppose a sample of n elements is selected with replacement. Let the values of X in the sample be $x_1, x_2, \ldots, x_n$ where x_1 is the value of X obtained on the first selection, x_2 the value obtained on the second selection, etc. Notice that x_1 is a random variable that could be equal to any value in the population set of values $X_1, X_2, \ldots, X_N$, and the probability that x_1 equals X_i is P_i. The same statement applies to x_2, etc. Since the selections are independent, the probability of getting a sample of n in a particular order is the product of the selection probabilities namely, $p(x_1)p(x_2)\ldots p(x_n)$ where $p(x_1)$ is the P_i for the element selected on the first draw, $p(x_2)$ is the P_i for the element selected on the second draw, etc.

Illustration 2.3. As an illustration, consider a sample of two elements selected with equal probability and with replacement from a population of four elements. Suppose the values of some characteristic X for the four elements are X_1, X_2, X_3, and X_4. There are 16 possibilities:

X_1,X_1	X_2,X_1	X_3,X_1	X_4,X_1
X_1,X_2	X_2,X_2	X_3,X_2	X_4,X_2
X_1,X_3	X_2,X_3	X_3,X_3	X_4,X_3
X_1,X_4	X_2,X_4	X_3,X_4	X_4,X_4

In this illustration $p(x_1)$ is always equal to $\frac{1}{4}$ and $p(x_2)$ is always $\frac{1}{4}$. Hence each of the 16 possibilities has a probability of $(\frac{1}{4})(\frac{1}{4}) = \frac{1}{16}$.

Each of the 16 possibilities is a different permutation that could be regarded as a separate sample. However, in practice (as we are not concerned about which element was selected first or second) it is more logical to disregard the order of selection. Hence, as possible samples and the probability of each occurring, we have:

Sample	Probability	Sample	Probability
X_1,X_1	1/16	X_2,X_3	1/8
X_1,X_2	1/8	X_2,X_4	1/8
X_1,X_3	1/8	X_3,X_3	1/16
X_1,X_4	1/8	X_3,X_4	1/8
X_2,X_2	1/16	X_4,X_4	1/16

Note that the sum of the probabilities is 1. That must always be the case if all possible samples have been listed with the correct probabilities. Also note that, since the probability (relative frequency of occurrence) of each sample is known, the average for each sample is a random variable. In other words, there were 10 possible samples, and any one of 10 possible sample averages could have occurred with the probability indicated. This is a simple illustration of the fact that the sample average satisfies the definition of a random variable. As the theory of sampling unfolds, we will be examining the properties of a sample average that exist as a result of its being a random variable.

Exercise 2.3. With reference to Illustration 2.3, suppose the probabilities of selection were $P_1 = \frac{1}{4}$, $P_2 = \frac{1}{8}$, $P_3 = \frac{3}{8}$, and $P_4 = \frac{1}{4}$. Find the probability of each of the ten samples. Remember the sampling is with replacement. Check your results by adding the 10 probabilities.

The sum should be 1. Partial answer: For the sample composed of elements 2 and 4 the probability is $(\frac{1}{8})(\frac{1}{4}) + (\frac{1}{4})(\frac{1}{8}) = \frac{1}{16}$.

2.5 SAMPLING WITHOUT REPLACEMENT

When a selected element is not returned to the population before the next selection is made, the sampling method is called sampling without replacement. In this case, the selection probabilities change from one draw to the next; that is, the selections (events) are dependent.

As above, assume a population of N elements with values of some characteristic X equal to $X_1, X_2, \ldots, X_N$. Let the selection probabilities for the first selection be $P_1, \ldots, P_i, \ldots P_N$ where each $P_i > 0$ and $\Sigma P_i = 1$. Suppose three elements are selected without replacement. Let x_1, x_2, and x_3 be the values of X obtained on the first, second, and third random draws, respectively. What is the probability that $x_1 = X_5$, $x_2 = X_6$, and $x_3 = X_7$? Let $P(X_5,X_6,X_7)$ represent this probability, which is the probability of selecting elements 5, 6, and 7 in that order.

According to the multiplicative probability law for dependent events,

$$P(X_5,X_6,X_7) = P(X_5)P(X_6|X_5)P(X_7|X_5,X_6)$$

It is clear that $P(X_5) = P_5$. For the second draw the selection probabilities (after element 5 is eliminated) must be adjusted so they add to 1. Hence, for the second draw the selection probabilities are $\frac{P_1}{1-P_5}, \frac{P_2}{1-P_5}, \frac{P_3}{1-P_5}, \frac{P_4}{1-P_5}, \frac{P_6}{1-P_5}, \ldots, \frac{P_N}{1-P_N}$. That is, $P(X_6|X_5) = \frac{P_6}{1-P_5}$.

Similarly, $$P(X_7|X_5,X_6) = \frac{P_7}{1-P_5-P_6}.$$

Therefore, $$P(X_5,X_6,X_7) = (P_5)(\frac{P_6}{1-P_5})(\frac{P_7}{1-P_5-P_6}) \qquad (2.1)$$

Observe that $P(X_6,X_5,X_7) = (P_6)(\frac{P_5}{1-P_6})(\frac{P_7}{1-P_6-P_5})$. Hence, $P(X_5,X_6,X_7) \neq P(X_6,X_5,X_7)$ unless $P_5 = P_6$. In general, each permutation of n elements has a different probability of occurrence unless the P_i's are all equal. To obtain the exact probability of selecting a sample composed of elements 5, 6, and 7, one would need to compute the probability for each of the six possible permutations and get the sum of the six probabilities.

Incidentally, in the actual process of selection, it is not necessary to compute a new set of selection probabilities after each selection is made. Make each selection in the same way that the first selection was made. If an element is selected which has already been drawn, ignore the random number and continue the same process of random selection until a new element is drawn.

As indicated by the very brief discussion in this section, the theory of sampling without replacement and with unequal probability of selection can be very complex. However, books on sampling present ways of circumventing the complex problems. In fact, it is practical and advantageous in many cases to use unequal probability of selection in sampling. The probability theory for sampling with equal probability of selection and without replacement is relatively simple and will be discussed in more detail.

Exercise 2.4. For a population of 4 elements there are six possible samples of two when sampling without replacement. Let $P_1 = \frac{1}{4}$, $P_2 = \frac{1}{8}$, $P_3 = \frac{3}{8}$, and $P_4 = \frac{1}{4}$. List the six possible samples and find the probability of getting each sample. Should the probabilities for the six samples add to 1? Check your results.

Exercise 2.5. Suppose two elements are selected with replacement and with equal probability from a population of 100 elements. Find the probability: (a) that element number 10 is not selected, (b) that element number 10 is selected only once, and (c) that element number 10 is selected twice? As a check, the three probabilities should add to 1. Why? Find the probability of selecting the combination of elements 10 and 20.

Exercise 2.6. Refer to Exercise 2.5 and change the specification "with replacement" to "without replacement." Answer the same questions. Why is the probability of getting the combination of elements 10 and 20 greater than it was in Exercise 2.5?

2.6 SIMPLE RANDOM SAMPLES

In practice, nearly all samples are selected without replacement. Selection of a random sample of n elements, with equal probability and without replacement, from a population of N elements is called simple random sampling (srs). One element must be selected at a time, that is, n separate random selections are required.

First, the probability of getting a particular combination of n elements will be discussed. Refer to Equation (2.1) and the discussion preceding it. The P_i's are all equal to $\frac{1}{N}$ for simple random sampling. Therefore, Equation (2.1) becomes $P(X_5,X_6,X_7) = (\frac{1}{N})(\frac{1}{N-1})(\frac{1}{N-2})$. All permutations of the three elements 5, 6, and 7 have the same probability of occurrence. There are $3! = 6$ possible permutations. Therefore, the probability that the sample is composed of the elements 5, 6, and 7 is $\frac{(1)(2)(3)}{N(N-1)(N-2)}$. Any other combination of three elements has the same probability of occurrence.

In general, all possible combinations of n elements have the same chance of selection and any particular combination of n has the following probability of being selected:

$$\frac{(1)(2)(3)\ldots(n)}{N(N-1)(N-2)\ldots(N-n+1)} = \frac{n!(N-n)!}{N!} \qquad (2.2)$$

According to a theorem on number of combinations, there are $\frac{N!}{n!(N-n)!}$ possible combinations (samples) of n elements. If each combination of n elements has the same chance of being the sample selected, the probability of selecting a specified combination must be the reciprocal of the number of combinations. This checks with Equation (2.2).

An important feature of srs that will be needed in the chapter on expected values is the fact that the j^{th} element of the population is as likely to be selected at the i^{th} random draw as any other. A general expression for the probability that the j^{th} element of the population is selected at the i^{th} drawing is

$$\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)\left(\frac{N-3}{N-2}\right)\ldots\left(\frac{N-i+1}{N-i+2}\right)\left(\frac{1}{N-i+1}\right) = \frac{1}{N} \qquad (2.3)$$

Let us check Equation 2.3 for $i = 3$. The equation becomes

$$\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)\left(\frac{1}{N-2}\right) = \frac{1}{N}$$

The probability that the j^{th} element of the population is selected at the third draw is equal to the probability that it was not selected at either the first or second draw times the conditional probability of being selected at the third draw, given that it was not selected at the first or second draw. (Remember, the sampling is without replacement). Notice that $\frac{N-1}{N}$ is the probability that the j^{th} element is not selected at the first draw and $\frac{N-2}{N-1}$ is the conditional probability that it was not selected at the second draw. Therefore, $\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)$ is the probability that the j^{th}

element has not been selected prior to the third draw. When the third draw is made, the conditional probability of selecting the j^{th} element is $\frac{1}{N-2}$. Hence the probability of selecting the j^{th} element at the third draw is $(\frac{N-1}{N})(\frac{N-2}{N-1})(\frac{1}{N-2}) = \frac{1}{N}$. This verifies Equation (2.3) for $i = 3$.

To summarize, the general result for any size of sample is that the j^{th} element in a population has a probability equal to $\frac{1}{N}$ of being selected at the i^{th} drawing. It means that x_i (the value of X obtained at the i^{th} draw) is a random variable that has a probability of $\frac{1}{N}$ of being equal to any value of the set $X_1,\ldots,X_N$.

What probability does the j^{th} element have of being included in a sample of n? We have just shown that it has a probability of $\frac{1}{N}$ of being selected at the i^{th} drawing. Therefore, any given element of the population has n chances, each equal to $\frac{1}{N}$, of being included in a sample. The element can be selected at the first draw, <u>or</u> the second draw,..., <u>or</u> the n^{th} draw and it cannot be selected twice because the sampling is without replacement. Therefore the probabilities, $\frac{1}{N}$ for each of the n draws, can be added which gives $\frac{n}{N}$ as the probability of any given element being included in the sample.

<u>Illustration 2.4</u>. Suppose one has a list of 1,000 farms which includes some farms that are out-of-scope (not eligible) for a survey. There is no way of knowing in advance whether a farm on the list is out-of-scope. A simple random sample of 200 farms is selected from the list. All 200 farms are visited but only the ones found to be in scope are included in the sample. What probability does an in-scope farm have of being in the sample? Every farm on the list of 1000 farms has a probability equal to $\frac{1}{5}$

of being in the sample of 200. All in-scope farms in the sample of 200 are included in the final sample. Therefore, the answer is $\frac{1}{5}$.

Exercise 2.7. From the following set of 12 values of X a srs of three elements is to be selected: 2, 10, 5, 8, 1, 15, 7, 8, 13, 4, 6, and 2. Find $P(\bar{x}\geq 12)$ and $P(3<\bar{x}<12)$. Remember that the total possible number of samples of 3 can readily be obtained by formula. Since every possible sample of three is equally likely, you can determine which samples will have an $\bar{x}\leq 3$ or an $\bar{x}\geq 12$ without listing all of the numerous possible samples. Answer: $P(\bar{x}\geq 12) = \frac{3}{220}$; $P(\bar{x}\leq 3) = \frac{9}{220}$; $P(3<\bar{x}<12) = \frac{208}{220}$.

2.7 SOME EXAMPLES OF RESTRICTED RANDOM SAMPLING

There are many methods other than srs that will give every element an equal chance of being in the sample, but some combinations of n elements do not have a chance of being the sample selected unless srs is used. For example, one might take every k^{th} element beginning from a random starting point between 1 and k. This is called systematic sampling. For a five percent sample k would be 20. The first element for the sample would be a random number between 1 and 20. If it is 12, then elements 12, 32, 52, etc., compose the sample. Every element has an equal chance, $\frac{1}{20}$, of being in the sample, but there are only 20 combinations of elements that have a chance of being the sample selected. Simple random sampling could have given the same sample but it is the method of sampling that characterizes a sample and determines how error due to sampling is to be estimated. One may think of sample design as a matter of choosing a method of sampling; that is, choosing restrictions to place on the process of selecting a sample so the combinations which have a chance of being the sample selected are generally "better" than many of the combinations that could occur with simple random sampling. At the same time, important properties that exist for simple random samples need to be retained. The key properties of srs will be developed in the next two chapters.

Another common method of sampling involves classification of all elements of a population into groups called strata. A sample is selected from each stratum. Suppose N_i elements of the population are in the i^{th} stratum and a simple random sample of n_i elements is selected from it. This is called stratified random sampling. It is clear that every element in the i^{th} stratum has a probability equal to $\frac{n_i}{N_i}$ of being in the sample. If the sampling fraction, $\frac{n_i}{N_i}$, is the same for all strata, every element of the population has an equal chance, namely $\frac{n_i}{N_i}$, of being in the sample. Again every element of the population has an equal chance of selection and of being in the sample selected, but some combinations that could occur when the method is srs cannot occur when stratified random sampling is used.

So far, our discussion has referred to the selection of individual elements, which are the units that data pertain to. For sampling purposes a population must be divided into parts which are called sampling units. A sample of sampling units is then selected. Sampling units and elements could be identical. But very often, it is either not possible or not practical to use individual elements as sampling units. For example, suppose a sample of households is needed. A list of households does not exist but a list of blocks covering the area to be surveyed might be available. In this case, a sample of blocks might be selected and all households within the selected blocks included in the sample. The blocks are the sampling units and the elements are households. Every element of the population should belong to one and only one sampling unit so the list of sampling units will account for all elements of the population without duplication or omission. Then, the probability of selecting any given element is the same as the probability of selecting the sampling unit that it belongs to.

Illustration 2.5. Suppose a population is composed of 1800 dwelling units located within 150 well-defined blocks. There are several possible sampling plans. A srs of 25 blocks could be selected and every dwelling unit in the selected blocks could be included in the sample. In this case, the sampling fraction is $\frac{1}{6}$ and every dwelling unit has a probability of $\frac{1}{6}$ of being in the sample. Is this a srs of dwelling units? No, but one could describe the sample as a random sample (or a probability sample) of dwelling units and state that every dwelling unit had an equal chance of being in the sample. That is, the term "simple random sample" would apply to blocks, not dwelling units. As an alternative sampling plan, if there were twelve dwelling units in each of the 150 blocks, a srs of two dwelling units could be selected from each block. This scheme, which is an example of stratified random sampling, would also give every dwelling unit a probability equal to $\frac{1}{6}$ of being in the sample.

Illustration 2.6. Suppose that a sample is desired of 100 adults living in a specified area. A list of adults does not exist, but a list of 4,000 dwelling units in the area is available. The proposed sampling plan is to select a srs of 100 dwelling units from the list. Then, the field staff is to visit the sample dwellings and list all adults living in each. Suppose there are 220 adults living in the 100 dwelling units. A simple random sample of 100 adults is selected from the list of 220. Consider the probability that an adult in the population has of being in the sample of 100 adults.

Parenthetically, we should recognize that the discussion which follows overlooks important practical problems of definition such as the definition of a dwelling unit, the definition of an adult, and the definition of living in a dwelling unit. However, assume the definitions are clear, that the list of dwelling units is complete, that no dwelling is on the list more than once, and that no ambiguity exists about whether an adult lives or does not live in a particular dwelling unit. Incomplete definitions often lead to inexact probabilities or ambiguity that gives difficulty in analyzing or interpreting results. The many practical problems should be discussed in an applied course on sampling.

It is clear that the probability of a dwelling unit being in the sample is $\frac{1}{40}$. Therefore, every person on the list of 220 had a chance of $\frac{1}{40}$ of being on the list because, under the specifications, a person lives in one and only one dwelling unit, and an adult's chance of being on the list is the same as that of the dwelling unit he lives in.

The second phase of sampling involves selecting a simple random sample of 100 adults from the list of 220. The conditional probability of an adult being in the sample of 100 is $\frac{100}{220} = \frac{5}{11}$. That is, given the fact that an adult is on the list of 220, he now has a chance of $\frac{5}{11}$ of being in the sample of 100.

Keep in mind that the probability of an event happening is its relative frequency in repeated trials. If another sample were selected following the above specifications, each dwelling unit in the population would again have a chance of $\frac{1}{40}$ of being in sample; but, the number of adults listed is not likely to be 220 so the conditional probability at the second phase depends upon the number of dwellings units in the sample blocks. Does every adult have the same chance of being in the sample? Examine the case carefully. An initial impression could be misleading. Every adult in the population has an equal chance of being listed in the first phase and every adult listed has an equal chance of being selected at the second phase. But, in terms of repetition of the whole sampling plan each person does not have exactly the same chance of being in the sample of 100. The following exercise will help clarify the situation and is a good exercise in probability.

Exercise 2.8. Assume a population of 5 d.u.'s (dwelling units) with the following numbers of adults:

Dwelling Unit	No. of Adults
1	2
2	4
3	1
4	2
5	3

A srs of two d.u.'s is selected. A srs of 2 adults is then selected from a list of all adults in the two d.u.'s. Find the probability that a specified adult in d.u. No. 1 has of being in the sample. Answer: 0.19. Find the probability that an adult in d.u. No. 2 has of being in the sample. Does the probability of an adult being in the sample appear to be related to the number of adults in his d.u.? In what way?

An alternative is to take a constant fraction of the adults listed instead of a constant number. For example, the specification might have been to select a random sample of $\frac{1}{2}$ of the adults listed in the first phase. In this case, under repeated application of the sampling specifications, the probability at the second phase does not depend on the outcome of the first phase and each adult in the population has an equal chance, $(\frac{1}{40})(\frac{1}{2}) = \frac{1}{80}$, of being selected in the sample. Notice that under this plan the number of adults in a sample will vary from sample to sample; in fact, the number of adults in the sample is a random variable.

For some surveys, interviewing more than one adult in a dwelling unit is inadvisable. Again, suppose the first phase of sampling is to select a srs of 100 dwelling units. For the second phase, consider the following: When an interviewer completes the listing of adults in a sample dwelling, he is to select one adult, from the list of those living in the dwelling, at random in accordance with a specified set of instructions. He then interviews the selected adult if available; otherwise, he returns at a time when the selected adult is available. What probability does an adult living in the area have of being in the sample? According to the multiplication theorem, the answer is $P'(D)P(A|D)$ where $P'(D)$ is the probability of the dwelling unit, in which the adult lives, being in the sample and $P(A|D)$ is the probability of the adult being selected given that his dwelling is in the sample. More specifically, $P'(D) = \frac{1}{40}$ and $P(A|D) = \frac{1}{k_i}$, where k_i is the number of adults in the i^{th} dwelling. Thus, an adult's chance, $(\frac{1}{40})(\frac{1}{k_i})$, of being in a sample is inversely proportional to the number of adults in his dwelling unit.

Exercise 2.9. Suppose there are five dwelling units and 12 persons living in the five dwelling units as follows:

Dwelling Unit	Individuals
1	1, 2
2	3, 4, 5, 6
3	7, 8
4	9
5	10, 11, 12

1. A sample of two dwelling units is selected with equal probability and without replacement. All individuals in the selected dwelling units are in the sample. What probability does individual number 4 have of being in the sample? Individual number 9?

2. Suppose from a list of the twelve individuals that one individual is selected with equal probability. From the selected individual two items of information are obtained: his age and the value of the dwelling in which he lives. Let X_1, $X_2,\ldots,X_{12}$ represent the ages of the 12 individuals and let $Y_1,\ldots,Y_5$ represent the values of the five dwelling units. Clearly, the probability of selecting the i^{th} individual is $\frac{1}{12}$ and therefore $P(X_i) = \frac{1}{12}$. Find the five probabilities $P(Y_1),\ldots,P(Y_5)$. Do you agree that $P(Y_3) = \frac{2}{12}$? As a check, $\Sigma P(Y_j)$ should equal one.

3. Suppose a sample of two individuals is selected with equal probability and <u>without replacement</u>. Let Y_{1j} be the value of Y_j obtained at the first draw and Y_{2j} be the value of Y_j obtained at the second draw. Does $P(Y_{1j}) = P(Y_{2j})$? That is, is the probability of getting Y_j on the second draw the same as it was on the first? If the answer is not evident, refer to Section 2.5.

<u>Exercise 2.10</u>. A small sample of third-grade students enrolled in public schools in a State is desired. The following plan is presented only

as an exercise and without consideration of whether it is a good one: A sample of 10 third-grade classes is to be selected. All students in the 10 classes will be included in the sample.

Step 1. Select a srs of 10 school districts.

Step 2. Within each of the 10 school districts, prepare a list of public schools having a third grade. Then select one school at random from the list.

Step 3. For each of the 10 schools resulting from Step 2, list the third-grade classes and select one class at random. (If there is only one third-grade class in the school, it is in the sample). This will give a sample of 10 classes.

Describe third-grade classes in the population which have relatively small chances of being selected. Define needed notation and write a mathematical expression representing the probability of a third-grade class being in the sample.

2.8 TWO-STAGE SAMPLING

For various reasons sampling plans often employ two or more stages of sampling. For example, a sample of counties might be selected, then within each sample county a sample of farms might be selected.

Units used at the first stage of sampling are usually called primary sampling units or psu's. The sampling units at the second stage of sampling could be called secondary sampling units. However, since there has been frequent reference earlier in this chapter to "elements of a population," the sampling units at the second stage will be called elements.

In the simple case of two-stage sampling, each element of the population is associated with one and only one primary sampling unit. Let i be the index for psu's and let j be the index for elements within a psu. Thus X_{ij} represents the value of some characteristic X for the j^{th} element in the i^{th} psu. Also, let

M = the total number of psu's,

m = the number of psu's selected for a sample,

N_i = the total number of elements in the i^{th} psu, and

n_i = the number of elements in the sample from the i^{th} psu.

Then,

$\sum_i^M N_i = N$, the total number of elements in the population, and

$\sum_i^m n_i = n$, the total number of elements in the sample.

Now consider the probability of an element being selected by a two step process: (1) Select one psu, and (2) select one element within the selected psu. Let,

P_i = the probability of selecting the i^{th} psu,

$P_{j|i}$ = the conditional probability of selecting the j^{th} element in the i^{th} psu given that the i^{th} psu has already been selected, and

P_{ij} = the overall probability of selecting the j^{th} element in the i^{th} psu.

Then,

$$P_{ij} = P_i P_{j|i}$$

If the product of the two probabilities, P_i and $P_{j|i}$, is constant for every element, then every element of the population has an equal chance of being selected. In other words, given a set of selection probabilities $P_1,\ldots,P_M$ for the psu's, one could specify that $P_{ij} = \frac{1}{N}$ and compute $P_{j|i}$, where $P_{j|i} = \frac{1}{NP_i}$, so every element of the population will have an equal chance of selection.

Exercise 2.11. Refer to Table 2.1. An element is to be selected by a three-step process as follows: (1) Select one of the Y classes (a row) with probability $\frac{N_{i.}}{N}$, (2) within the selected row select an X class (a column) with probability $\frac{N_{ij}}{N_{i.}}$, (3) within the selected cell select an element with equal probability. Does each element in the population of N elements have an equal probability of being drawn? What is the probability?

The probability of an element being included in a two-stage sample is given by

$$P'_{ij} = P'_i P'_{j|i} \tag{2.4}$$

where

P'_i = the probability that the i^{th} psu is in the sample of psu's, and

$P'_{j|i}$ = the conditional probability which the j element has of being in the sample, given that the i^{th} psu has been selected.

The inclusion probability P'_{ij} will be discussed very briefly for three important cases:

(1) Suppose a random sample of m psu's is selected with equal probability and without replacement. The probability, P'_i, of the i^{th} psu being in the sample is $f_1 = \frac{m}{M}$ where f_1 is the sampling fraction for the first-stage units. In the second stage of sampling assume that, within each of the m psu's, a constant proportion, f_2, of the elements is selected.

That is, in the i^{th} psu in the sample, a simple random sample of n_i elements out of N_i is selected, the condition being that $n_i = f_2 N_i$. Hence, the conditional probability of the j^{th} element in the i^{th} psu being in the sample is $P'_{j|i} = \frac{n_i}{N_i} = f_2$. Substituting in Equation 2.4, we have $P'_{ij} = f_1 f_2$ which shows that an element's probability of being in the sample is equal to the product of the sampling fractions at the two stages. In this case P'_{ij} is constant and is the overall sampling fraction.

Unless N_i is the same for all psu's, the size of the sample, $n_i = f_2 N_i$, varies from psu to psu. Also, since the psu's are selected at random the total size of the sample, $n = \sum_i^m n_i = f_2 \sum_i^m N_i$, is not constant with regard to repetition of the sampling plan. In practice variation in the size, n_i, of the sample from psu to psu might be very undesirable. If appropriate information is available, it is possible to select psu's with probabilities that will equalize the sample sizes n_i and also keep P'_{ij} constant.

(2) Suppose one psu is selected with probability $P_i = \frac{N_i}{N}$. This is commonly known as sampling with pps (probability proportional to size). Within the selected psu, assume that a simple random sample of k elements is selected. (If any N_i are less than k, consolidations could be made so all psu's have an N_i greater than k). Then,

$$P'_i = \frac{N_i}{N} \ , \ P'_{j|i} = \frac{k}{N_i} \ , \text{ and } P'_{ij} = \frac{N_i}{N}\frac{k}{N_i} = \frac{k}{N}$$

which means that every element of the population has an equal probability, $\frac{k}{N}$, of being included in a sample of k elements.

Extension of this sampling scheme to a sample of m psu's could encounter the complications indicated in Section 2.5. However, it was stated that means exist for circumventing those complications. Sampling books 1/ discuss this matter quite fully so we will not include it in this monograph. The point is that one can select m psu's without replacement in such a way that $m \frac{N_i}{N}$ is the probability of including the i^{th} psu in the sample. That is, $P_i' = m \frac{N_i}{N}$. If a random sample of k elements is selected with equal probability from each of the selected psu's,

$P_{j|i}' = \frac{k}{N_i}$ and

$$P_{ij}' = (m \frac{N_i}{N})(\frac{k}{N_i}) = \frac{mk}{N} = \frac{n}{N}$$

Thus, if the N_i are known exactly for all M psu's in the population, and if a list of elements in each psu is available, it is possible to select a two-stage sample of n elements so that k elements for the sample come from each of m psu's and every element of the population has an equal chance of being in the sample. In practice, however, one usually finds one of two situations: (a) there is no information on the number of elements in the psu's, or (b) the information that does exist is out-of-date. Nevertheless, out-of-date information on number of elements in the psu's can be very useful. It is also possible that a measure of size might exist which will serve, more efficiently, the purposes of sampling.

(3) Suppose that characteristic Y is used as a measure of size. Let Y_i be the value of Y for the i^{th} psu in the population and let $P_i = \frac{Y_i}{Y}$ where $Y = \sum_i^M Y_i$. A sample of m psu's is selected in such a way that $P_i' = m \frac{Y_i}{Y}$ is the probability that the i^{th} psu has of being in the sample.

1/ For example, Hansen, Hurwitz, and Madow. Sample Survey Methods and Theory. Volume I, Chapter 8. John Wiley and Sons. 1953.

With regard to the second stage of sampling, let f_{2i} be the sampling fraction for selecting a simple random sample within the i^{th} psu in the sample. That is, $P'_{j|i} = f_{2i}$. Then,

$$P'_{ij} = (m \frac{Y_i}{Y})(f_{2i}) \tag{2.5}$$

In setting sampling specifications one would decide on a fixed value for P'_{ij}. In this context P'_{ij} is the overall sampling fraction or proportion of the population that is to be included in the sample. For example, if one wanted a 5 percent sample, P'_{ij} would be .05. Or, if one knew there were approximately 50,000 elements in the population and wanted a sample of about 2,000, he would set $P'_{ij} = .04$. Hence, we will let f be the overall sampling fraction and set P'_{ij} equal to f. Decisions are also made on the measure of size to be used and on the number, m, of psu's to be selected. In Equation 2.5, this leaves f_{2i} to be determined. Thus, f_{2i} is computed as follows for each psu in the sample:

$$f_{2i} = \frac{fY}{mY_i}$$

Use of the sampling fractions f_{2i} at the second stage of sampling will give every element of the population a probability equal to f of being in the sample. A sample wherein every element of the population has an equal chance of inclusion is often called a self-weighted sample.

CHAPTER III. EXPECTED VALUES OF RANDOM VARIABLES

3.1 INTRODUCTION

The theory of expected values of random variables is used extensively in the theory of sampling; in fact, it is the foundation for sampling theory. Interpretations of the accuracy of estimates from probability samples depend heavily on the theory of expected values.

The definition of a random variable was discussed in the previous chapter. It is a variable that can take (be equal to) any one of a defined set of values with known probability. Let X_i be the value of X for the i^{th} element in a set of N elements and let P_i be the probability that the i^{th} element has of being selected by some chance operation so that P_i is known a priori. What is the expected value of X?

Definition 3.1. The expected value of a random variable X is $\sum_{i=1}^{N} P_i X_i$ where $\sum_{i=1}^{N} P_i = 1$. The mathematical notation for the expected value of X is E(X). Hence, by definition, $E(X) = \sum_{i=1}^{N} P_i X_i$.

Observe that $\Sigma P_i X_i$ is a weighted average of the values of X, the weights being the probabilities of selection. "Expected value" is a substitute expression for "average value." In other words, E means "the average value of" or "find the average value of" whatever follows E. For example, $E(X^2)$, read "the expected value of X^2," refers to the average value of the squares of the values that X can equal. That is, by definition,

$$E(X^2) = \sum_{i=1}^{N} P_i X_i^2 \quad .$$

If all of the N elements have an equal chance of being selected, all values of P_i must equal $\frac{1}{N}$ because of the requirement that $\Sigma P_i = 1$. In this case, $E(X) = \sum_{i=1}^{N} \frac{1}{N} X_i = \frac{\Sigma X_i}{N} = \bar{X}$, which is the simple average of X for all N elements.

Illustration 3.1. Assume 12 elements having values of X as follows:

$X_1 = 3$	$X_5 = 5$	$X_9 = 10$
$X_2 = 9$	$X_6 = 3$	$X_{10} = 3$
$X_3 = 3$	$X_7 = 4$	$X_{11} = 8$
$X_4 = 5$	$X_8 = 3$	$X_{12} = 4$

For this set, $E(X) = \frac{3+9+...+4}{12} = 5$, assuming each element has the same chance of selection. Or, by counting the number of times that each unique value of X occurs, a frequency distribution of X can be obtained as follows:

X_j	N_j
3	5
4	2
5	2
8	1
9	1
10	1

where X_j is a unique value of X and N_j is the number of times X_j occurs.

We noted in Chapter I that $\Sigma N_j = N$, $\Sigma N_j X_j = \Sigma X_i$, and that $\frac{\Sigma N_j X_j}{\Sigma N_j} = \frac{\Sigma X_i}{N} = \bar{X}$.

Suppose one of the X_j values is selected at random with a probability equal to P_j where $P_j = \frac{N_j}{\Sigma N_j} = \frac{N_j}{N}$. What is the expected value of X_j ? By definition $E(X_j) = \Sigma P_j X_j = \Sigma \frac{N_j}{N} X_j = \frac{\Sigma N_j X_j}{N} = \bar{X}$. The student may verify that in this illustration $E(X_j) = 5$. Note that the selection specifications were equivalent to selecting one of the 12 elements at random with equal probability.

Incidentally, a frequency distribution and a probability distribution are very similar. The probability distribution with reference to X_j would be:

X_j	P_j
3	5/12
4	2/12
5	2/12
8	1/12
9	1/12
10	1/12

The 12 values, $P_i = \frac{1}{N}$, for the 12 elements are also a probability distribution. This illustration shows two ways of treating the set of 12 elements.

When finding expected values be sure that you understand the definition of the set of values that the random variable might equal and the probabilities involved.

<u>Definition 3.2</u>. When X is a random variable, by definition the expected value of a function of X is

$$E[f(X)] = \sum_{i=1}^{N} P_i[f(X_i)]$$

Some examples of simple functions of X are: $f(X) = aX$, $f(X) = X^2$, $f(X) = a + bX + cX^2$, and $f(X) = (X-\bar{X})^2$. For each value, X_i , in a defined set there is a corresponding value of $f(X_i)$.

<u>Illustration 3.2</u>. Suppose $f(X) = 2X+3$. With reference to the set of 12 elements discussed above, there are 12 values of $f(X_i)$ as follows:

$$f(X_1) = (2)(3) + 3 = 9$$

$$f(X_2) = (2)(9) + 3 = 21$$

$$\vdots$$

$$f(X_{12}) = 2(4) + 3 = 11$$

Assuming $P_i = \frac{1}{N}$ the expected value of $f(X) = 2X+3$ would be

$$E(2X+3) = \sum_{i}^{12} \frac{1}{N}(2X_i+3) = (\frac{1}{12})(9)+(\frac{1}{12})(21)+\ldots+(\frac{1}{12})(11) = 13 \qquad (3.1)$$

In algebraic terms, for $f(X) = aX+b$, we have

$$E(aX+b) = \sum_{i=1}^{N} P_i(aX_i+b) = \Sigma P_i(aX_i) + \Sigma P_i b$$

By definition $\Sigma P_i(aX_i) = E(aX)$, and $\Sigma P_i b = E(b)$. Therefore,

$$E(aX+b) = E(aX) + E(b) \qquad (3.2)$$

Since b is constant and $\Sigma P_i = 1$, $\Sigma P_i b = b$, which leads to the first important theorem in expected values.

<u>Theorem 3.1</u>. The expected value of a constant is equal to the constant: $E(a) = a$.

By definition $E(aX) = \Sigma P_i(aX_i) = a\Sigma P_i X_i$. Since $\Sigma P_i X_i = E(X)$, we have another important theorem:

<u>Theorem 3.2</u>. The expected value of a constant times a variable equals the constant times the expected value of the variable: $E(aX) = aE(X)$.

Applying these two theorems to Equation (3.2) we have $E(aX+b) = aE(X) + b$. Therefore, with reference to Illustration 3.2, $E(2X+3) = 2E(X) + 3 = 2(5) + 3 = 13$, which is the same as the result found in Equation (3.1).

Exercise 3.1. Suppose a random variable X can take any of the following four values with the probabilities indicated:

$X_1 = 2 \quad X_2 = 5 \quad X_3 = 4 \quad X_4 = 6$

$P_1 = 2/6 \quad P_2 = 2/6 \quad P_3 = 1/6 \quad P_4 = 1/6$

(a) Find $E(X)$ Answer: 4

(b) Find $E(X^2)$ Answer: $18\frac{1}{3}$. Note that $E(X^2) \neq [E(X)]^2$

(c) Find $E(X-\bar{X})$ Answer: 0 Note: By definition

$$E(X-\bar{X}) = \sum_{i=1}^{4} P_i(X_i-\bar{X})$$

(d) Find $E(X-\bar{X})^2$ Answer: $2\frac{1}{3}$. Note: By definition

$$E(X-\bar{X})^2 = \sum_{i=1}^{4} P_i(X_i-\bar{X})^2$$

Exercise 3.2. From the following set of three values of Y_i one value is to be selected with a probability P'_i:

$Y_1 = -2 \quad Y_2 = 2 \quad Y_3 = 4$

$P'_1 = 1/4 \quad P'_2 = 2/4 \quad P'_3 = 1/4$

(a) Find $E(Y)$ Answer: $1\frac{1}{2}$

(b) Find $E(\frac{1}{Y})$ Answer: 3/16. Note: $\frac{1}{E(Y)} \neq E(\frac{1}{Y})$

(c) Find $E(Y-\bar{Y})^2$ Answer: $4\frac{3}{4}$

3.2 EXPECTED VALUE OF THE SUM OF TWO RANDOM VARIABLES

The sum of two or more random variables is also a random variable. If X and Y are two random variables, the expected value of X + Y is equal to the expected value of X plus the expected value of Y: $E(X+Y) = E(X)+E(Y)$. Two numerical illustrations will help clarify the situation.

Illustration 3.3. Consider the two random variables X and Y in Exercises 3.1 and 3.2:

$X_1 = 2$	$P_1 = \frac{2}{6}$	$Y_1 = -2$	$P_1' = \frac{1}{4}$
$X_2 = 5$	$P_2 = \frac{2}{6}$	$Y_2 = 2$	$P_2' = \frac{2}{4}$
$X_3 = 4$	$P_3 = \frac{1}{6}$	$Y_3 = 4$	$P_3' = \frac{1}{4}$
$X_4 = 6$	$P_4 = \frac{1}{6}$		

Suppose one element of the first set and one element of the second set are selected with probabilities as listed above. What is the expected value of X + Y? The joint probability of getting X_i and Y_j is P_iP_j' because the two selections are independent. Hence by definition

$$E(X + Y) = \sum_{i=1}^{4} \sum_{j=1}^{3} P_iP_j' (X_i + Y_j) \tag{3.3}$$

The possible values of X + Y and the probability of each are as follows:

X + Y	P_iP_j'	X + Y	P_iP_j'
$X_1 + Y_1 = 0$	$P_1P_1' = \frac{2}{24}$	$X_3 + Y_1 = 2$	$P_3P_1' = \frac{1}{24}$
$X_1 + Y_2 = 4$	$P_1P_2' = \frac{4}{24}$	$X_3 + Y_2 = 6$	$P_3P_2' = \frac{2}{24}$
$X_1 + Y_3 = 6$	$P_1P_3' = \frac{2}{24}$	$X_3 + Y_3 = 8$	$P_3P_3' = \frac{1}{24}$
$X_2 + Y_1 = 3$	$P_2P_1' = \frac{2}{24}$	$X_4 + Y_1 = 4$	$P_4P_1' = \frac{1}{24}$
$X_2 + Y_2 = 7$	$P_2P_2' = \frac{4}{24}$	$X_4 + Y_2 = 8$	$P_4P_2' = \frac{2}{24}$
$X_2 + Y_3 = 9$	$P_2P_3' = \frac{2}{24}$	$X_4 + Y_3 = 10$	$P_4P_3' = \frac{1}{24}$

As a check the sum of the probabilities must be 1 if all possible sums have been listed and the probability of each has been correctly determined. Substituting the values of $X_i + Y_j$ and P_iP_j' in Equation (3.3) we obtain 5.5 as follows for expected value of X + Y:

$$(\frac{2}{24})(0) + (\frac{4}{24})(4) + \ldots + (\frac{1}{24})(10) = 5.5$$

From Exercises 3.1 and 3.2 we have E(X) = 4 and E(Y) = 1.5. Therefore, E(X) + E(Y) = 4 + 1.5 = 5.5 which verifies the earlier statement that E(X + Y) = E(X) + E(Y).

Illustration 3.4. Suppose a random sample of two is selected with replacement from the population of four elements used in Exercise 3.1. Let x_1 be the first value selected and let x_2 be the second. Then x_1 and x_2 are random variables and $x_1 + x_2$ is a random variable. The possible values of $x_1 + x_2$ and the probability of each, $P(x_1,x_2)$, are listed below. Notice that each possible order of selection is treated separately.

x_1	x_2	$P(x_1,x_2)$	x_1+x_2	x_1	x_2	$P(x_1,x_2)$	x_1+x_2
X_1	X_1	4/36	4	X_3	X_1	2/36	6
X_1	X_2	4/36	7	X_3	X_2	2/36	9
X_1	X_3	2/36	6	X_3	X_3	1/36	8
X_1	X_4	2/36	8	X_3	X_4	1/36	10
X_2	X_1	4/36	7	X_4	X_1	2/36	8
X_2	X_2	4/36	10	X_4	X_2	2/36	11
X_2	X_3	2/36	9	X_4	X_3	1/36	10
X_2	X_4	2/36	11	X_4	X_4	1/36	12

By definition $E(x_1 + x_2)$ is

$$\frac{4}{36}(4) + \frac{4}{36}(7) + \frac{2}{36}(6) + \ldots + \frac{1}{36}(12) = 8$$

In Exercise 3.1 we found E(X) = 4. Since x_1 is the same random variable as X, $E(x_1) = 4$. Also, x_2 is the same random variable as X, and $E(x_2) = 4$. Therefore, $E(x_1) + E(x_2) = 8$, which verifies that $E(x_1+x_2) = E(x_1) + E(x_2)$.

In general if X and Y are two random variables, where X might equal $X_1,\ldots,X_N$ and Y might equal $Y_1,\ldots,Y_M$, then E(X + Y) = E(X)+E(Y). The proof is as follows: By definition $E(X+Y) = \sum_{ij}^{NM} P_{ij}(X_i+Y_j)$ where P_{ij} is the probability of getting the sum $X_i + Y_j$, and $\Sigma\Sigma P_{ij} = 1$. The double summation is over all possible values of $P_{ij}(X_i+Y_j)$. According to the rules for summation we may write

$$\sum_{ij}^{NM} P_{ij}(X_i+Y_j) = \sum_{ij}^{NM} P_{ij}X_i + \sum_{ij}^{NM} P_{ij}Y_j \qquad (3.4)$$

In the first term on the right, X_i is constant with regard to the summation over j; and in the second term on the right, Y_j is constant with regard to the summation over i. Therefore, the right-hand side of Equation (3.4) can be written as

$$\sum_{i}^{N} X_i \sum_{j}^{M} P_{ij} + \sum_{j}^{M} Y_j \sum_{i}^{N} P_{ij}$$

And, since $\sum_{j}^{M} P_{ij} = P_i$ and $\sum_{i}^{N} P_{ij} = P_j$, Equation (3.4) becomes

$$\sum_{ij}^{NM} P_{ij}(X_i+Y_j) = \sum_{i}^{N} X_iP_i + \sum_{j}^{M} Y_jP_j$$

By definition $\sum_{i}^{N} X_iP_i = E(X)$ and $\sum_{j}^{M} Y_jP_j = E(Y)$.

Therefore $E(X+Y) = E(X) + E(Y)$.

If the proof is not clear write the values of $P_{ij}(X_i+Y_j)$ in a matrix format. Then, follow the summation manipulations in the proof.

The above result extends to any number of random variables; that is, the expected value of a sum of random variables is the sum of the expected values of each. In fact, there is a very important theorem that applies to a linear combination of random variables.

<u>Theorem 3.3</u>. Let $u = a_1u_1 + \ldots + a_ku_k$, where $u_1,\ldots,u_k$ are random variables and $a_1,\ldots,a_k$ are constants. Then

$$E(u) = a_1E(u_1) + \ldots + a_k\ E(u_k)$$

or in summation notation

$$E(u) = E \sum_{i}^{k} a_iu_i = \sum_{i}^{k} a_iE(u_i)$$

The generality of Theorem 3.3 is impressive. For example, with reference to sampling from a population $X_1,\ldots, X_N$, u_1 might be the value of X obtained at the first draw, u_2 the value obtained at the second draw, etc. The constants could be weights. Thus, in this case, u would be a weighted average of the sample measurements. Or, suppose $\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_k$ are averages from a random sample for k different age groups. The averages are random variables and the theorem could be applied to any linear combination of the averages. In fact u_i could be any function of random variables. That is, the only condition on which the theorem is based is that u_i must be a random variable.

<u>Illustration 3.5</u>. Suppose we want to find the expected value of $(X + Y)^2$ where X and Y are random variables. Before Theorem 3.3 can be applied we must square (X + Y). Thus $E(X + Y)^2 = E(X^2 + 2XY + Y^2)$.

The application of Theorem 3.3 gives $E(X + Y)^2 = E(X)^2 + 2E(XY) + E(Y)^2$.

<u>Illustration 3.6</u>. We will now show that

$$E(X-\bar{X})(Y-\bar{Y}) = E(XY) - \bar{X}\bar{Y} \quad \text{where} \quad E(X) = \bar{X} \text{ and } E(Y) = \bar{Y}$$

Since $(X-\bar{X})(Y-\bar{Y}) = XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}$ we have

$$E(X-\bar{X})(Y-\bar{Y}) \doteq E(XY-\bar{X}Y-X\bar{Y}+\bar{X}\bar{Y})$$

and application of Theorem 3.3 gives

$$E(X-\bar{X})(Y-\bar{Y}) = E(XY) - E(\bar{X}Y) - E(Y\bar{X}) + E(\bar{X}\bar{Y})$$

Since $\bar{X}$ and $\bar{Y}$ are constant, $E(\bar{X}Y) = \bar{X}\,E(Y) = \bar{X}\bar{Y}$, $E(Y\bar{X}) = \bar{Y}\bar{X}$, and $E(\bar{X}\bar{Y}) = \bar{X}\bar{Y}$. Therefore, $E(X-\bar{X})(Y-\bar{Y}) = E(XY) - \bar{X}\bar{Y}$

Exercise 3.3. Suppose $E(X) = 6$ and $E(Y) = 4$. Find

(a) $E(2X+4Y)$ Answer: 28

(b) $[E(2X)]^2$ Answer: 144

(c) $\sqrt{E(Y)}$ Answer: 2

(d) $E(5Y-X)$ Answer: 14

Exercise 3.4. Prove the following, assuming $E(X) = \bar{X}$ and $E(Y) = \bar{Y}$:

(a) $E(X-\bar{X}) = 0$

(b) $E(aX-bY) + cE(Y) = a\bar{X} + (c-b)\bar{Y}$

(c) $E[a(X-\bar{X}) + b(Y-\bar{Y})] = 0$

(d) $E(X+a)^2 = E(X^2) + 2a\bar{X} + a^2$

(e) $E(X-\bar{X})^2 = E(X^2) - \bar{X}^2$

(f) $E(aX+bY) = 0$ for any values of a and b if $E(X) = 0$ and $E(Y) = 0$.

3.3 EXPECTED VALUE OF AN ESTIMATE

Theorem 3.3 will now be used to find the expected value of the mean of a simple random sample of n elements selected without replacement from a population of N elements. The term "simple random sample" implies equal probability of selection without replacement. The sample average is

$$\bar{x} = \frac{x_1+\ldots+x_n}{n}$$

where x_i is the value of X for the i^{th} element in the sample. Without loss of generality, we can consider the subscript of x as corresponding to the i^{th} draw; i.e., x_1 is the value of X obtained on the first draw, x_2 the value on the second, etc. As each x_i is a random variable, $\bar{x}$ is a linear combination of random variables. Therefore, Theorem 3.3 applies and

$$E(\bar{x}) = \frac{1}{n}\,[E(x_1) + \ldots + E(x_n)]$$

In the previous chapter, Section 2.6, we found that any given element of the population had a chance of $\frac{1}{N}$ of being selected on the i^{th} draw. This means that x_i is a random variable that has a probability equal to $\frac{1}{N}$ of being equal to any value of the population set $X_1,\ldots,X_N$. Therefore,

$$E(x_1) = E(x_2) = \ldots = E(x_n) = \bar{X}$$

Hence, $E(\bar{x}) = \frac{\bar{X} + \ldots + \bar{X}}{n} = \bar{X}$. The fact that $E(\bar{x}) = \bar{X}$ is one of the very important properties of an average from a simple random sample. Incidentally, $E(\bar{x}) = \bar{X}$ whether the sampling is with or without replacement.

Definition 3.3. A parameter is a quantity computed from all values in a population set. The total of X, the average of X, the proportion of elements for which $X_i<A$, or any other quantity computed from measurements including all elements of the population is a parameter. The numerical value of a parameter is usually unknown but it exists by definition.

Definition 3.4. An estimator is a mathematical formula or rule for making an estimate from a sample. The formula for a sample average, $\bar{x} = \frac{\Sigma x_i}{n}$, is a simple example of an estimator. It provides an estimate of the parameter $\bar{X} = \frac{\Sigma X_i}{N}$.

Definition 3.5. An estimate is unbiased when its expected value equals the parameter that it is an estimate of. In the above example, $\bar{x}$ is an unbiased estimate of $\bar{X}$ because $E(\bar{x}) = \bar{X}$.

Exercise 3.5. Assume a population of only four elements having values of X as follows: $X_1 = 2$, $X_2 = 5$, $X_3 = 4$, $X_4 = 6$. For simple random samples of size 2 show that the estimator $N\bar{x}$ provides an unbiased estimate of the population total, $\Sigma X_i = 17$. List all six possible samples of two and calculate $N\bar{x}$ for each. This will give the set of values that the random variable $N\bar{x}$ can be equal to. Consider the probability of each of the possible values of $N\bar{x}$ and show arithmetically that $E(N\bar{x}) = 17$.

A sample of elements from a population is not always selected by using equal probabilities of selection. Sampling with unequal probability is complicated when the sampling is without replacement, so we will limit our discussion to sampling with replacement.

Illustration 3.7. The set of four elements and the associated probabilities used in Exercise 3.1 will serve as an example of unbiased estimation when samples of two elements are selected with unequal probability and with replacement. Our estimator of the population total, $2+5+4+6 = 17$, will be $x' = \dfrac{\sum_{i=1}^{n} \frac{x_i}{p_i}}{n}$. The estimate x' is a random variable. Listed below are the set of values that x' can equal and the probability of each value occurring.

Possible Samples	x'_j	P_j
x_1 x_1	6	4/36
x_1 x_2	10.5	8/36
x_1 x_3	15	4/36
x_1 x_4	21	4/36
x_2 x_2	15	4/36
x_2 x_3	19.5	4/36
x_2 x_4	25.5	4/36
x_3 x_3	24	1/36
x_3 x_4	30	2/36
x_4 x_4	36	1/36

<u>Exercise 3.6</u>. Verify the above values of x_j' and P_j and find the expected value of x'. By definition $E(x') = \Sigma P_j x_j'$. Your answer should be 17 because x' is an unbiased estimate of the population total.

To put sampling with replacement and unequal probabilities in a general setting, assume the population is $X_1,\ldots,X_j,\ldots,X_N$ and the selection probabilities are $P_1,\ldots,P_j,\ldots,P_N$. Let x_i be the value of X for the i^{th} element in a sample of n elements and let p_i be the probability which that element had of being selected. Then $x' = \frac{\sum\limits_{i=1}^{n} \frac{x_i}{p_i}}{n}$ is an unbiased estimate of the population total. We will now show that $E(x') = \sum\limits_{j=1}^{N} X_j$.

To facilitate comparison of x' with u in Theorem 3.3, x' may be written as follows:

$$x' = \frac{1}{n}\left(\frac{x_1}{p_1}\right) +\ldots+ \frac{1}{n}\left(\frac{x_n}{p_n}\right)$$

It is now clear that $a_i = \frac{1}{n}$ and $u_i = \frac{x_i}{p_i}$. Therefore,

$$E(x') = \frac{1}{n}\left[E\left(\frac{x_1}{p_1}\right) +\ldots+ E\left(\frac{x_n}{p_n}\right)\right] \tag{3.5}$$

The quantity $\frac{x_1}{p_1}$, which is the outcome of the first random selection from the population, is a random variable that might be equal to any one of the set of values $\frac{X_1}{P_1}$,..., $\frac{X_j}{P_j}$,..., $\frac{X_N}{P_N}$. The probability that $\frac{x_1}{p_1}$ equals $\frac{X_j}{P_j}$ is P_j.

Therefore, by definition

$$E\left(\frac{x_1}{p_1}\right) = \sum_j^N P_j\left(\frac{X_j}{P_j}\right) = \sum_j^N X_j$$

Since the sampling is with replacement it is clear that any $\frac{x_i}{p_i}$ is the same random variable as $\frac{x_1}{p_1}$.

Therefore Equation (3.5) becomes

$$E(x') = \frac{1}{n}[\sum_{j}^{N} X_j + \ldots + \sum_{j}^{N} X_j]$$

Since there are n terms in the series it follows that

$$E(x') = \sum_{j}^{N} X_j \ .$$

Exercise 3.7. As a corollary show that the expected value of $\frac{x'}{n}$ is equal to the population mean.

By this time, you should be getting familiar with the idea that an estimate from a probability sample is a random variable. Persons responsible for the design and selection of samples and for making estimates from samples are concerned about the set of values, and associated probabilities, that an estimate from a sample might be equal to.

Definition 3.6. The distribution of an estimate generated by probability sampling is the sampling distribution of the estimate.

The values of x_j' and P_j in the numerical Illustration 3.7 are an example of a sampling distribution. Statisticians are primarily interested in three characteristics of a sampling distribution: (1) the mean (center) of the sampling distribution in relation to the value of the parameter being estimated, (2) a measure of the variation of possible values of an estimate from the mean of the sampling distribution, and (3) the shape of the sampling distribution. We have been discussing the first. When the expected value of an estimate equals the parameter being estimated, we know that the mean of the sampling distribution is equal to the parameter estimated. But, in practice, values of parameters are generally not known. To judge the accuracy of an estimate, we need information on all three characteristics of the sampling distribution. Let us turn now to the generally accepted measure of variation of a random variable.

3.4 VARIANCE OF A RANDOM VARIABLE

The variance of a random variable, X, is the average value of the squares of the deviation of X from its mean; that is, the average value of $(X-\bar{X})^2$. The square root of the variance is the standard deviation (error) of the variable.

Definition 3.7. In terms of expected values, the variance of a random variable, X, is $E(X-\bar{X})^2$ where $E(X) = \bar{X}$. Since X is a random variable, $(X-\bar{X})^2$ is a random variable and by definition of expected value,

$$E(X-\bar{X})^2 = \sum_i^N P_i(X_i-\bar{X})^2$$

In case $P_i = \frac{1}{N}$ we have the more familiar formula for variance, namely,

$$E(X-\bar{X})^2 = \frac{\sum_i^N (X_i-\bar{X})^2}{N} = \sigma_X^2$$

Commonly used symbols for variance include: σ^2, σ_X^2, V^2, S^2, Var(X) and V(X). Variance is often defined as $\frac{\Sigma(X_i-\bar{X})^2}{N-1}$. This will be discussed in Section 3.7.

3.4.1 VARIANCE OF THE SUM OF TWO INDEPENDENT RANDOM VARIABLES

Two random variables, X and Y, are independent if the joint probability, P_{ij}, of getting X_i and Y_j is equal to $(P_i)(P_j)$, where P_i is the probability of selecting X_i from the set of values of X, and P_j is the probability of selecting Y_j from the set of values of Y. The variance of the sum of two independent random variables is the sum of the variance of each. That is,

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

Illustration 3.8. In Illustration 3.3, X and Y were independent. We had listed all possible values of X_i+Y_j and the probability of each. From that listing we can readily compute the variance of X+Y. By definition

$$\sigma^2_{X+Y} = E[(X+Y)-(\bar{X}+\bar{Y})]^2 = \sum_{ij}\sum P_iP_j[(X_i+Y_j)-(\bar{X}+\bar{Y})]^2 \qquad (3.6)$$

Substituting in Equation (3.6) we have

$$\sigma^2_{X+Y} = \frac{2}{24}(0-5.5)^2 + \frac{4}{24}(4-5.5)^2 + \ldots + \frac{1}{24}(10-5.5)^2 = \frac{85}{12}$$

The variances of X and Y are computed as follows:

$$\sigma^2_X = E(X-\bar{X})^2 = \frac{2}{3}(2-4)^2 + \frac{2}{6}(5-4)^2 + \frac{1}{6}(4-4)^2 + \frac{1}{6}(6-4)^2 = \frac{7}{3}$$

$$\sigma^2_Y = E(Y-\bar{Y})^2 = \frac{1}{4}(-2-1.5)^2 + \frac{2}{4}(2-1.5)^2 + \frac{1}{4}(4-1.5)^2 = \frac{19}{4}$$

We now have $\sigma^2_X + \sigma^2_Y = \frac{7}{3} + \frac{19}{4} = \frac{85}{12}$ which verifies the above statement that the variance of the sum of two independent random variables is the sum of the variances.

Exercise 3.8. Prove that $E[(X+Y)-(\bar{X}+\bar{Y})]^2 = E(X+Y)^2 - (\bar{X}+\bar{Y})^2$. Then calculate the variance of X+Y in Illustration 3.3 by using the formula $\sigma^2_{X+Y} = E(X+Y)^2 - (\bar{X}+\bar{Y})^2$. The answer should agree with the result obtained in Illustration 3.8.

Exercise 3.9. Refer to Illustration 3.3 and the listing of possible values of X + Y and the probability of each. Instead of X_i+Y_j list the products $(X_i-\bar{X})(Y_j-\bar{Y})$ and show that $E(X_i-\bar{X})(Y_j-\bar{Y}) = 0$.

Exercise 3.10. Find $E(X-\bar{X})(Y-\bar{Y})$ for the numerical example used in Illustration 3.3 by the formula $E(XY) - \bar{X}\bar{Y}$ which was derived in Illustration 3.6.

3.4.2 VARIANCE OF THE SUM OF TWO DEPENDENT RANDOM VARIABLES

The variance of dependent random variables involves covariance which is defined as follows:

Definition 3.8. The covariance of two random variables, X and Y, is $E(X-\bar{X})(Y-\bar{Y})$ where $E(X) = \bar{X}$ and $E(Y) = \bar{Y}$. By definition of expected value

$$E(X-\bar{X})(Y-\bar{Y}) = \sum_{i}\sum_{j} P_{ij}(X_i-\bar{X})(Y_j-\bar{Y})$$

where the summation is over all possible values of X and Y.

Symbols commonly used for covariance are σ_{XY}, S_{XY}, and Cov(X,Y).

Since $(X+Y) - (\bar{X}+\bar{Y}) = (X-\bar{X}) + (Y-\bar{Y})$ we can derive a formula for the variance of X+Y as follows:

$$\begin{aligned}\sigma^2_{X+Y} &= E[(X+Y) - (\bar{X}+\bar{Y})]^2 \\ &= E[(X-\bar{X}) + (Y-\bar{Y})]^2 \\ &= E[(X-\bar{X})^2 + (Y-\bar{Y})^2 + 2(X-\bar{X})(Y-\bar{Y})]\end{aligned}$$

Then, according to Theorem 3.3,

$$\sigma^2_{X+Y} = E(X-\bar{X})^2 + E(Y-\bar{Y})^2 + 2E(X-\bar{X})(Y-\bar{Y})$$

and by definition we obtain,

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}$$

Sometimes σ_{XX} is used instead of σ^2_X to represent variance. Thus

$$\sigma^2_{X+Y} = \sigma_{XX} + \sigma_{YY} + 2\sigma_{XY}$$

For two independent random variables, $P_{ij} = P_iP_j$. Therefore

$$E(X-\bar{X})(Y-\bar{Y}) = \sum_{i}\sum_{j} P_iP_j\ (X_i-\bar{X})(Y_j-\bar{Y})$$

Write out in longhand, if necessary, and be satisfied that the following is correct:

$$\sum_i\sum_j P_iP_j(X_i-\bar{X})(Y_j-\bar{Y}) = \sum_i P_i(X_i-\bar{X})\sum_j P_j(Y_j-\bar{Y}) = 0 \qquad (3.7)$$

which proves that the covariance σ_{XY} is zero when X and Y are independent. Notice that in Equation (3.7) $\sum_i P_i(X_i-\bar{X}) = E(X-\bar{X})$ and $\sum_j P_j(Y_j-\bar{Y}) = E(Y-\bar{Y})$ which, for independent random variables, proves that $E(X-\bar{X})(Y-\bar{Y}) = E(X-\bar{X})\ E(Y-\bar{Y})$. When working with independent random variables the following important theorem is frequently very useful:

Theorem 3.4. The expected value of the product of independent random variables $u_1, u_2, \ldots, u_k$ is the product of their expected values:

$$E(u_1u_2\ldots u_k) = E(u_1)E(u_2)\ldots E(u_k)$$

3.5 VARIANCE OF AN ESTIMATE

The variance of an estimate from a probability sample depends upon the method of sampling. We will derive the formula for the variance of $\bar{x}$, the mean of a random sample selected with equal probability, with and without replacement. Then, the variance of an estimate of the population total will be derived for sampling with replacement and unequal probability of selection.

3.5.1 EQUAL PROBABILITY OF SELECTION

The variance of $\bar{x}$, the mean of a random sample of n elements selected with equal probabilities and with replacement from a population of N, is:

$$\mathrm{Var}(\bar{x}) = \frac{\sigma_X^2}{n}, \quad \text{where } \sigma_X^2 = \frac{\sum_i^N (X_i-\bar{X})^2}{N}$$

The proof follows:

By definition, $\mathrm{Var}(\bar{x}) = E[\bar{x}-E(\bar{x})]^2$. We have shown that $E(\bar{x}) = \bar{X}$. Therefore, $\mathrm{Var}(\bar{x}) = E(\bar{x}-\bar{X})^2$. By substitution and algebraic manipulation, we obtain

$$\mathrm{Var}(\bar{x}) = E\left[\frac{x_1+\ldots+x_n}{n} - \bar{X}\right]^2$$

$$= E\left[\frac{(x_1-\bar{X})+\ldots+(x_n-\bar{X})}{n}\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n}(x_i-\bar{X})^2 + \sum_{i\neq j}\sum(x_i-\bar{X})(x_j-\bar{X})\right].$$

Applying Theorem 3.3 we now obtain

$$\mathrm{Var}(\bar{x}) = \frac{1}{n^2}\left[\sum_{i=1}^{n}E(x_i-\bar{X})^2 + \sum_{i\neq j}\sum E(x_i-\bar{X})(x_j-\bar{X})\right] \tag{3.8}$$

In series form, Equation (3.8) can be written as

$$\mathrm{Var}(\bar{x}) = \frac{1}{n^2}\left[E(x_1-\bar{X})^2 + E(x_2-\bar{X})^2 +\ldots+ E(x_1-\bar{X})(x_2-\bar{X}) + E(x_1-\bar{X})(x_3-\bar{X})+\ldots\right]$$

Since the sampling is with replacement x_i and x_j are independent and the expected value of all of the product terms is zero. For example, $E(x_1-\bar{X})(x_2-\bar{X}) = E(x_1-\bar{X})\ E(x_2-\bar{X})$ and we know that $E(x_1-\bar{X})$ and $E(x_2-\bar{X})$ are zero. Next, consider $E(x_1-\bar{X})^2$. We have already shown that x_1 is a random variable that can be equal to any one of the population set of values $X_1,\ldots,X_N$ with equal probability. Therefore

$$E(x_1-\bar{X})^2 = \frac{\sum_{j}^{N}(X_j-\bar{X})^2}{N} = \sigma_X^2$$

The same argument applies to x_2, x_3, etc. Therefore, $\sum_{i=1}^{n} E(x_i-\bar{X})^2 = \sigma_X^2 +\ldots+ \sigma_X^2 = n\sigma_X^2$ and Equation (3.8) reduces to $\mathrm{Var}(\bar{x}) = \frac{\sigma_X^2}{n}$.

The mathematics for finding the variance of $\bar{x}$ when the sampling is without replacement is the same as sampling with replacement down to and including Equation (3.8). The expected value of a product term in Equation (3.8) is not zero because x_i and x_j are not independent. For example, on the first draw an element has a probability of $\frac{1}{N}$ of being selected, but on the second draw the probability is conditioned by the fact that the element selected on the first draw was not replaced. Consider the first product term in Equation (3.8). To find $E(x_1-\bar{X})(x_2-\bar{X})$ we need to consider the set of values that $(x_1-\bar{X})(x_2-\bar{X})$ could be equal to. Reference to the following matrix is helpful:

$$\begin{matrix} (X_1-\bar{X})^2 & (X_1-\bar{X})(X_2-\bar{X}) & \cdots & (X_1-\bar{X})(X_N-\bar{X}) \\ (X_2-\bar{X})(X_1-\bar{X}) & (X_2-\bar{X})^2 & \cdots & (X_2-\bar{X})(X_N-\bar{X}) \\ \vdots & \vdots & & \vdots \\ (X_N-\bar{X})(X_1-\bar{X}) & (X_N-\bar{X})(X_2-\bar{X}) & \cdots & (X_N-\bar{X})^2 \end{matrix}$$

The random variable $(x_1-\bar{X})(x_2-\bar{X})$ has an equal probability of being any of the products in the above matrix, except for the squared terms on the main diagonal. There are $N(N-1)$ such products. Therefore,

$$E(x_1-\bar{X})(x_2-\bar{X}) = \frac{\sum\limits_{i\neq j}^{N}\sum^{N} (X_i-\bar{X})(X_j-\bar{X})}{N(N-1)}$$

According to Equation (1.9) in Chapter 1,

$$\sum_{i\neq j}^{N}\sum^{N} (X_i-\bar{X})(X_j-\bar{X}) = -\sum_{i}^{N} (X_i-\bar{X})^2$$

Hence,

$$E(x_1-\bar{X})(x_2-\bar{X}) = -\frac{\sum\limits_{i}^{N}(X_i-\bar{X})^2}{N(N-1)} = -\frac{\sigma_X^2}{N-1}$$

The same evaluation applies to all other product terms in Equation (3.8). There are $n(n-1)$ product terms in Equation (3.8) and the expected value of each is $-\frac{\sigma_X^2}{N-1}$. Thus, Equation (3.8) becomes

$$Var(\bar{x}) = \frac{1}{n^2}\left[\sum_i^n E(x_i-\bar{X})^2 - n(n-1)\frac{\sigma_X^2}{N-1}\right]$$

Recognizing that $E(x_i-\bar{X})^2 = \sigma_X^2$ and after some easy algebraic operations the answer as follows is obtained:

$$Var(\bar{x}) = \frac{N-n}{N-1}\frac{\sigma_X^2}{n} \tag{3.9}$$

The factor $\frac{N-n}{N-1}$ is called the correction for finite population because it does not appear when infinite populations are involved or when sampling with replacement which is equivalent to sampling from an infinite population.

For two characteristics, X and Y, of elements in the same simple random sample, the covariance of $\bar{x}$ and $\bar{y}$ is given by a formula analogous to Equation (3.9); namely,

$$Cov(\bar{x},\bar{y}) = \frac{N-n}{N-1}\frac{\sigma_{XY}}{n} \tag{3.10}$$

3.5.2 UNEQUAL PROBABILITY OF SELECTION

In Section 3.3 we proved that $x' = \frac{\sum_i^n \frac{x_i}{p_i}}{n}$ is an unbiased estimate of the population total. This was for sampling with replacement and unequal probability of selection. We will now proceed to find the variance of x'.

By definition $Var(x') = E[x'- E(x')]^2$. Let $X = \sum_i^N X_i$. Then since $E(x') = X$, it follows that

$$Var(x') = E\left[\frac{\frac{x_1}{p_1}+\ldots+\frac{x_n}{p_n}}{n} - X\right]^2 = \frac{1}{n^2}E\left[\left(\frac{x_1}{p_1} - X\right)+\ldots+\left(\frac{x_n}{p_n} - X\right)\right]^2$$

$$= \frac{1}{n^2}E\left[\Sigma\left(\frac{x_i}{p_i} - X\right)^2 + \sum_{i\neq k}\sum\left(\frac{x_i}{p_i} - X\right)\left(\frac{x_k}{p_k} - X\right)\right]$$

Applying Theorem 3.3, Var(x′) becomes

$$\mathrm{Var}(x') = \frac{1}{n^2}\left[\Sigma E\left(\frac{x_i}{p_i} - X\right)^2 + \sum_{i \neq j}\Sigma E\left(\frac{x_i}{p_i} - X\right)\left(\frac{x_k}{p_k} - X\right)\right] \tag{3.11}$$

Notice the similarity of Equations (3.8) and (3.11) and that the steps leading to these two equations were the same. Again, since the sampling is with replacement, the expected value of all product terms in Equation (3.11) is zero. Therefore Equation (3.11) becomes

$$\mathrm{Var}(x') = \frac{1}{n^2}\left[\sum_i^n E\left(\frac{x_i}{p_i} - X\right)^2\right]$$

By definition $E\left(\frac{x_i}{p_i} - X\right)^2 = \sum_i^N P_i\left(\frac{X_i}{P_i} - X\right)^2$

Therefore $$\mathrm{Var}(x') = \frac{\sum_i^N P_i\left(\frac{X_i}{P_i} - X\right)^2}{n} \tag{3.12}$$

<u>Exercise 3.11</u>. (a) Refer to Exercise 3.1 and compute the variance of x′ for samples of two (that is, n = 2) using Equation (3.12). (b) Then turn to Illustration 3.7 and compute the variance of x′ from the actual values of x′. Don't overlook the fact that the values of x′ have unequal probabilities. According to Definition 3.7, the variance of x′ is $\sum_j^{10} P_j(x_j' - X)^2$ where $X = E(x')$, x_j' is one of the 10 possible values of x′, and P_j is the probability of x_j' .

3.6 VARIANCE OF A LINEAR COMBINATION

Before presenting a general theorem on the variance of a linear combination of random variables, a few key variance and covariance relationships will be given. In the following equations X and Y are random variables and a, b, c, and d are constants:

$$\mathrm{Var}(X+a) = \mathrm{Var}(X)$$

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$$

$$\mathrm{Cov}(X+a,Y+b) = \mathrm{Cov}(X,Y)$$

$$\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$$

$$\mathrm{Cov}(aX+b,cY+d) = ac\mathrm{Cov}(X,Y)$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

$$\mathrm{Var}(X+Y+a) = \mathrm{Var}(X+Y)$$

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y)$$

Illustration 3.9. The above relationships are easily verified by using the theory of expected values. For example,

$$\begin{aligned}\mathrm{Var}(aX+b) &= E[aX+b-E(aX+b)]^2 \\ &= E[aX+b-E(aX)-E(b)]^2 \\ &= E[aX-aE(X)]^2 \\ &= E[a(X-\bar{X})]^2 \\ &= a^2E(X-\bar{X})^2 = a^2\mathrm{Var}(X)\end{aligned}$$

Exercise 3.12. As in Illustration 3.9 use the theory of expected values to prove that

$$\mathrm{Cov}(aX+b,cY+d) = ac\mathrm{Cov}(X,Y)$$

As in Theorem 3.3, let $u = a_1u_1+\ldots+a_ku_k$ where $a_1,\ldots,a_k$ are constants and $u_1,\ldots,u_k$ are random variables. By definition the variance of u is

$$\mathrm{Var}(u) = E[u-E(u)]^2$$

By substitution

$$\begin{aligned}\mathrm{Var}(u) &= E[a_1u_1+\ldots+a_ku_k-E(a_1u_1+\ldots+a_ku_k)]^2 \\ &= E[a_1(u_1-\bar{u}_1)+\ldots+a_k(u_k-\bar{u}_k)]^2 \quad \text{where } E(u_i) = \bar{u}_i\end{aligned}$$

By squaring the quantity in [ ] and considering the expected values of the terms in the series, the following result is obtained.

Theorem 3.5. The variance of u, a linear combination of random variables, is given by the following equation

$$\mathrm{Var}(u) = \sum_{i}^{k} a_i^2 \sigma_i^2 + \sum_{i \neq j}\sum a_i a_j \sigma_{ij}$$

where σ_i^2 is the variance of u_i and σ_{ij} is the covariance of u_i and u_j.

Theorems 3.3 and 3.5 are very useful because many estimates from probability samples are linear combinations of random variables.

Illustration 3.10. Suppose for a srs (simple random sample) that data have been obtained for two characteristics X and Y, the sample values being $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$. What is the variance of $\bar{x}-\bar{y}$? From the theory and results that have been presented one can proceed immediately to write the answer. From Theorem 3.5 we know that $\mathrm{Var}(\bar{x}-\bar{y}) = \mathrm{Var}(\bar{x}) + \mathrm{Var}(\bar{y}) - 2\mathrm{Cov}(\bar{x},\bar{y})$. From the sampling specifications we know the variances of $\bar{x}$ and $\bar{y}$ and the covariance. See Equations (3.9) and (3.10) Thus, the following result is easily obtained:

$$\mathrm{Var}(\bar{x}-\bar{y}) = \left(\frac{N-n}{N-1}\right)\left(\frac{1}{n}\right)(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) \qquad (3.13)$$

Some readers might be curious about the relationship between covariance and correlation. By definition the correlation between X and Y is

$$r_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Therefore, one could substitute $r_{XY}\ \sigma_X\sigma_Y$ for σ_{XY} in Equation (3.13).

Exercise 3.13. In a statistical publication suppose you find 87 bushels per acre as the yield of corn in State A and 83 is the estimated yield for State B. The estimated standard errors are given as 1.5 and

2.0 bushels. You become interested in the standard error of the difference in yield between the two States and want to know how large the estimated difference is in relation to its standard error. Find the standard error of the difference. You may assume that the two yield estimates are independent because the sample selection in one State was completely independent of the other. Answer: 2.5.

Illustration 3.11. No doubt students who are familiar with sampling have already recognized the application of Theorems 3.3 and 3.5 to several sampling plans and methods of estimation. For example, for stratified random sampling, an estimator of the population total is

$$x' = N_1\bar{x}_1 + \ldots + N_k\bar{x}_k = \Sigma N_i\bar{x}_i$$

where N_i is the population number of sampling units in the i^{th} stratum and $\bar{x}_i$ is the average per sampling unit of characteristic, X, from a sample of n_i sampling units from the i^{th} stratum. According to Theorem 3.3

$$E(x') = E\Sigma N_i\bar{x}_i = \Sigma N_i E(\bar{x}_i)$$

If the sampling is such that $E(\bar{x}_i) = \bar{X}_i$ for all strata, x' is an unbiased estimate of the population total. According to Theorem 3.5

$$\text{Var}(x') = N_1^2\ \text{Var}(\bar{x}_1) + \ldots + N_k^2\ \text{Var}(\bar{x}_k) \tag{3.14}$$

There are no covariance terms in Equation (3.14) because the sample selection in one stratum is independent of another stratum. Assuming a srs from each stratum, Equation (3.14) becomes

$$\text{Var}(x') = N_1^2\left(\frac{N_1-n_1}{N_1-1}\right)\frac{\sigma_1^2}{n_1} + \ldots + N_k^2\left(\frac{N_k-n_k}{N_k-1}\right)\frac{\sigma_k^2}{n_k}$$

where σ_i^2 is the variance of X among sampling units within the i^{th} stratum.

Illustration 3.12. Suppose $x_1',\ldots,x_k'$ are independent estimates of the same quantity, T. That is, $E(x_i') = T$. Let σ_i^2 be the variance of x_i'.

Consider a weighted average of the estimates, namely

$$x' = w_1x_1' +\ldots+ w_kx_k' \tag{3.15}$$

where $\Sigma w_i = 1$. Then

$$E(x') = w_1E(x_1') +\ldots+ w_k\ E(x_k') = T \tag{3.16}$$

That is, for any set of weights where $\Sigma w_i = 1$ the expected value of x' is T. How should the weights be chosen?

The variance of x' is

$$Var(x') = w_1^2\sigma_1^2 +\ldots+ w_k^2\ \sigma_k^2$$

If we weight the estimates equally, $w_i = \frac{1}{k}$ and the variance of x' is

$$Var(x') = \frac{1}{k}\ [\frac{\Sigma\sigma_i^2}{k}] \tag{3.17}$$

which is the average variance divided by k. However, it is reasonable to give more weight to estimates having low variance. Using differential calculus we can find the weights which will minimize the variance of x'. The optimum weights are inversely proportional to the variances of the estimates. That is, $w_i \propto \frac{1}{\sigma_i^2}$

As an example, suppose one has two independent unbiased estimates of the same quantity which originate from two different samples. The optimum weighting of the two estimates would be

$$\frac{\frac{1}{\sigma_1^2}\ x_1' + \frac{1}{\sigma_2^2}\ x_2'}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

As another example, suppose $x_1',\ldots,x_k'$ are the values of X in a sample of k sampling units selected with equal probability and with replacement. In this case each x_i' is an unbiased estimate of $\bar{X}$. If we let $w_i = \frac{1}{k}$, x' is $\bar{x}$, the simple average of the sample values. Notice, as one would expect, Equation (3.16) reduces to $E(\bar{x}) = \bar{X}$. Also, since each estimate, x_i', is the same random variable that could be equal to any value in the set $X_1,\ldots X_N$, it is clear that all of the σ_i^2's must be equal to $\sigma^2 = \frac{\Sigma(X_i-\bar{X})^2}{N}$. Hence, Equation (3.17) reduces to $\frac{\sigma^2}{n}$ which agrees with the first part of Section 3.5.1.

<u>Exercise 3.14</u>. If you equate x_i' in Equation (3.15) with $\frac{x_i}{p_i}$ in Section 3.5.2 and let $w_i = \frac{1}{n}$ and k = n, then x' in Equation (3.15) is the same as $x' = \frac{\Sigma\frac{x_i}{p_i}}{n}$ in Section 3.5.2. Show that in this case Equation (3.17) becomes the same as Equation (3.12).

3.7 ESTIMATION OF VARIANCE

All of the variance formulas presented in previous sections have involved calculations from a population set of values. In practice, we have data for only a sample. Hence, we must consider means of estimating variances from sample data.

3.7.1 SIMPLE RANDOM SAMPLING

In Section 3.5.1, we found that the variance of the mean of a srs is

$$Var(\bar{x}) = \frac{N-n}{N-1}\;\frac{\sigma_X^2}{n} \tag{3.18}$$

where $$\sigma_X^2 = \frac{\overset{N}{\underset{i}{\Sigma}}(X_i-\bar{X})^2}{N}$$

As an estimator of σ_X^2, $\frac{\sum_i^n (x_i-\bar{x})^2}{n}$ seems like a natural first choice for consideration. However, when sampling finite populations, it is customary to define variance among units of the population as follows:

$$S^2 = \frac{\sum_i^N (X_i-\bar{X})^2}{N-1}$$

and to use $s^2 = \frac{\sum_i^n (x_i-\bar{x})^2}{n-1}$ as an estimator of S^2. A reason for this will become apparent when we find the expected value of s^2 as follows:

The formula for s^2 can be written in a form that is more convenient for finding $E(s^2)$. Thus,

$$s^2 = \frac{\sum_i^n (x_i-\bar{x})^2}{n-1} = \frac{\Sigma x_i^2 - n\bar{x}^2}{n-1}$$

and
$$E(s^2) = \frac{1}{n-1} [\sum_i^n E(x_i^2) - nE(\bar{x}^2)]$$

We have shown previously that x_i is a random variable that has an equal probability of being any value in the set $X_1,\ldots,X_N$. Therefore

$$E(x_i^2) = \frac{\sum_i^N X_i^2}{N} \quad \text{and} \quad \sum_i E(x_i^2) = \frac{n\Sigma X_i^2}{N}$$

Hence,
$$E(s^2) = \frac{n}{n-1} [\frac{\Sigma X_i^2}{N} - E(\bar{x}^2)] \tag{3.19}$$

We know, by definition, that $\sigma_{\bar{x}}^2 = E(\bar{x} - \bar{X})^2$ and it is easy to show that

$$E(\bar{x}-\bar{X})^2 = E(\bar{x}^2) - \bar{X}^2$$

Therefore, $E(\bar{x}^2) = \sigma_{\bar{x}}^2 + \bar{X}^2$.

By substitution in Equation (3.19) we obtain

$$E(s^2) = \frac{n}{n-1} \left[\frac{\Sigma X_i^2}{N} - \bar{X}^2 - \sigma_{\bar{x}}^2\right]$$

By definition $\sigma_X^2 = \dfrac{\Sigma(X_i-\bar{X})^2}{N} = \dfrac{\Sigma X_i^2}{N} - \bar{X}^2$ and since the specified method of sampling was srs, $\sigma_{\bar{x}}^2 = \dfrac{N-n}{N-1}\ \dfrac{\sigma_X^2}{n}$, we have $E(s^2) = \dfrac{n}{n-1} [\sigma_X^2 - \dfrac{N-n}{N-1}\ \dfrac{\sigma_X^2}{n}]$

which after simplification is

$$E(s^2) = \frac{N}{N-1}\ \sigma_X^2$$

Note from the above definitions of σ_X^2 and S^2 that

$$S^2 = \frac{N}{N-1}\ \sigma_X^2$$

Therefore $\qquad E(s^2) = S^2$

Since s^2 is an unbiased estimate of S^2, we will now substitute $\frac{N-1}{N}\ S^2$ for σ_X^2 in Equation (3.18) which gives

$$\mathrm{Var}(\bar{x}) = \frac{N-n}{N}\ \frac{S^2}{n} \tag{3.20}$$

Both Equations, (3.18) and (3.20), for the $\mathrm{Var}(\bar{x})$ give identical results and both agree with $E(\bar{x}-\bar{X})^2$ as a definition of variance. We have shown that s^2 is an unbiased estimate of S^2. Substituting s^2 for S^2 in Equation (3.20) we have

$$\mathrm{var}(\bar{x}) = \frac{N-n}{N}\ \frac{s^2}{n} \tag{3.21}$$

as an estimate of the variance of $\bar{x}$. With regard to Equation (3.18), $\frac{N-1}{N}\ s^2$ is an unbiased estimate of σ_X^2 . When $\frac{N-1}{N}\ s^2$ is substituted for σ_X^2 , Equation (3.21) is obtained.

Since in Equation (3.20), $\frac{N-n}{N}$ is exactly 1 minus the sampling fraction and s^2 is an unbiased estimate of S^2, there is some advantage to using

Equation (3.20) and $S^2 = \frac{\Sigma(X_i-\bar{X})^2}{N-1}$ as a definition of variance among sampling units in the population.

<u>Exercise 3.15</u>. For a small population of 4 elements suppose the values of X are $X_1 = 2$, $X_2 = 5$, $X_3 = 3$; and $X_4 = 6$. Consider simple random samples of size 2. There are six possible samples.

(a) For each of the six samples calculate $\bar{x}$ and s^2. That is, find the sampling distribution of $\bar{x}$ and the sampling distribution of s^2.

(b) Calculate S^2, then find $\text{Var}(\bar{x})$ using Equation (3.20).

(c) Calculate the variance among the six values of $\bar{x}$ and compare the result with $\text{Var}(\bar{x})$ obtained in (b). The results should be the same.

(d) From the sampling distribution of s^2 calculate $E(s^2)$ and verify that $E(s^2) = S^2$.

3.7.2 UNEQUAL PROBABILITY OF SELECTION

In Section 3.5.2, we derived a formula for the variance of the estimator x' where

$$x' = \frac{\Sigma\frac{x_i}{p_i}}{n} \tag{3.22}$$

The sampling was with unequal selection probabilities and with replacement. We found that the variance of x' was given by

$$\text{Var}(x') = \frac{\sum\limits_{i}^{N} P_i\left(\frac{X_i}{P_i} - X\right)^2}{n} \tag{3.23}$$

As a formula for estimating $\text{Var}(x')$ from a sample one might be inclined, as a first guess, to try a formula of the same form as Equation (3.23) but that does not work. Equation (3.23) is a weighted average of the squares of deviations $(\frac{X_i}{P_i} - X)^2$ which reflects the unequal selection probabilities. If one applied the same weighting system in a formula for estimating variance from a sample he would in effect be applying the weights twice; first, in the selection process itself and second, to the sample data. The unequal probability of selection is already incorporated into the sample itself.

As in some of the previous discussion, look at the estimator as follows:

$$x' = \frac{\frac{x_1}{p_1} + \ldots + \frac{x_n}{p_n}}{n} = \frac{x_1' + \ldots + x_n'}{n} \quad \text{where } x_i' = \frac{x_i}{p_i}$$

Each x_i' is an independent unbiased estimate of the population total. Since each value of x_i' receives an equal weight in determining x' it appears that the following formula for estimating Var(x') might work:

$$\text{var}(x') = \frac{s^2}{n} \tag{3.24}$$

where $$s^2 = \frac{\sum_i^n (x_i' - x')^2}{n-1}$$

By following an approach similar to that used in Section 3.7.1, one can prove that

$$E(s^2) = \sum_i^N P_i(\frac{X_i}{P_i} - X)^2$$

That is, Equation (3.24) does provide an unbiased estimate of Var(x') in Equation (3.23). The proof is left as an exercise.

<u>Exercise 3.16</u>. Reference is made to Exercise 3.1, Illustration 3.7, and Exercise 3.11. In Illustration 3.7 the sampling distribution of x'

(See Equation (3.22)) is given for samples of 2 from the population of 4 elements that was given in Exercise 3.1.

(a) Compute $\text{var}(x') = \frac{s^2}{n}$ (Equation (3.24)) for each of the 10 possible samples.

(b) Compute the expected value of var(x′) and compare it with the result obtained in Exercise 3.11. The results should be the same. Remember, when finding the expected value of var(x′), that the x′'s do not occur with equal frequency.

3.8 RATIO OF TWO RANDOM VARIABLES

In sampling theory and practice one frequently encounters estimates that are ratios of random variables. It was pointed out earlier that $E(\frac{u}{w}) \neq \frac{E(u)}{E(w)}$ where u and w are random variables. Formulas for the expected value of a ratio and for the variance of a ratio will now be presented without derivation. The formulas are approximations:

$$E(\frac{u}{w}) \doteq \frac{\bar{u}}{\bar{w}} + \frac{\bar{u}}{\bar{w}}\left[\frac{\sigma_w^2}{\bar{w}^2} - \frac{\rho_{uw}\sigma_u\sigma_w}{\bar{u}\bar{w}}\right] \tag{3.25}$$

$$\text{Var}(\frac{u}{w}) \doteq \left[\frac{\bar{u}}{\bar{w}}\right]^2\left[\frac{\sigma_u^2}{\bar{u}^2} + \frac{\sigma_w^2}{\bar{w}^2} - \frac{2\rho_{uw}\sigma_u\sigma_w}{\bar{u}\bar{w}}\right] \tag{3.26}$$

where

$$\bar{u} = E(u)$$

$$\bar{w} = E(w)$$

$$\sigma_u^2 = E(u-\bar{u})^2$$

$$\sigma_w^2 = E(w-\bar{w})^2$$

and

$$\rho_{uw} = \frac{\sigma_{uw}}{\sigma_u\sigma_w} \quad \text{where } \sigma_{uw} = E(u-\bar{u})(w-\bar{w})$$

For a discussion of the conditions under which Equations (3.25) and (3.26) are good approximations, reference is made to Hansen, Hurwitz, and