

This dissertation has been
microfilmed exactly as received 68-17,494

WHORTON, Jr., Elbert Benjamin, 1938-
THE DEVELOPMENT AND INVESTIGATION OF SOME
EXTENSIONS TO THE EDERER-MYERS-MANTEL
PROCEDURE AND TEST FOR CLUSTERING.

The University of Oklahoma, Ph.D., 1968
Statistics

University Microfilms, Inc., Ann Arbor, Michigan

THE UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

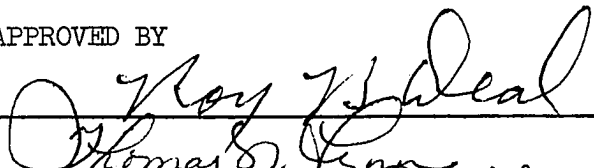
THE DEVELOPMENT AND INVESTIGATION OF SOME EXTENSIONS
TO THE EDERER-MYERS-MANTEL PROCEDURE AND TEST
FOR CLUSTERING

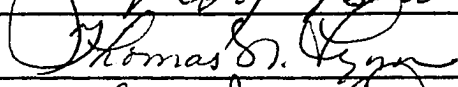
A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
degree of
DOCTOR OF PHILOSOPHY

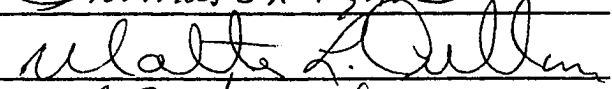
BY
ELBERT B. WHORTON, JR.
Oklahoma City, Oklahoma
1968

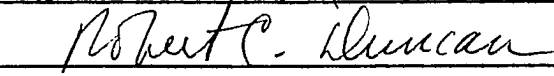
THE DEVELOPMENT AND INVESTIGATION OF SOME EXTENSIONS
TO THE EDERER-MYERS-MANTEL PROCEDURE AND TEST
FOR CLUSTERING

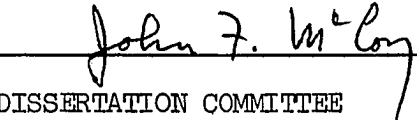
APPROVED BY











DISSERTATION COMMITTEE

ACKNOWLEDGMENT

I wish to express my sincerest appreciation to Dr. Roy B. Deal, Jr. for all his constructive suggestions and efforts throughout the research and developmental aspects of this paper. My gratitude is extended to Dr. Edward N. Brandt, Jr. for his guidance in initiating this study. I am also deeply indebted to Dr. Robert C. Duncan and Dr. John F. McCoy for their comments and evaluations and to Dr. Walter Cullinan and Dr. Thomas N. Lynn who so kindly served on the advisory and reading committee.

I am extremely grateful to Mrs. Rose Titsworth for her willing care and exactness in typing this manuscript. Thanks are also extended to those graduate students in the Department of Biostatistics and Epidemiology and the personnel in the Computer Facility of the University of Oklahoma Medical Center who helped make this study possible.

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	v
LIST OF ILLUSTRATIONS.....	vi
Chapter	
I. INTRODUCTION AND PURPOSE FOR THE STUDY.....	1
II. REVIEW OF PREVIOUS WORK.....	5
III. DEFINITIONS AND POPULATION NOTATIONS.....	13
IV. UNEQUAL SIZES OF THE STRATA POPULATIONS AT RISK WITHIN REGIONS.....	24
V. THE SAMPLE AND SAMPLE POPULATION.....	34
VI. TEST STATISTIC RECOMMENDED FOR USE WHEN $k < K$ REGIONS ARE SAMPLED AT RANDOM.....	44
VII. SUMMARY.....	53
LIST OF REFERENCES.....	55
APPENDIX.....	56

LIST OF TABLES

Table	Page
1. Number of Cases Observed per Stratum by Regions (Hypothetical).....	7
2. Expectation and Variance of X_{gh_m} per Region for the Data in Table 1.....	7
3. Possible Distinct Distributions of 3 Units in 5 Cells.....	9
4. Observed Cases X_{gh} per Stratum U_{gh} in Region U_g (Hypothetical).....	23
5. Expectation and Variance of X_{gh_m} per Region U_g for the Data in Table 4.....	23
6. Number of Cases Observed X_{gh} per Regional Stratum U_{gh} (Hypothetical).....	29
7. Transformed Values X_{gh}^* per Regional Stratum U_{gh} for the Data in Table 6.....	29
8. Expectations and Variances of Transformed Values $X_{gh_m}^*$	29
9. Randomly Generated Cases X_{gh} per Stratum U_{gh} , Given: $L_g = 5$, $X_g = 805$	31
10. Test Statistic Values and Results for 20 of the 100 Independent Samples of Size $k = 5$ Regions.....	52

LIST OF ILLUSTRATIONS

Figure	Page
1. Regions of Interest U_g in U	14
2. Strata U_{gh} Within Region of Interest U_g	14
3. Primary Sampling Units C_{ghi} in Stratum U_{gh}	15
4. A Possible Clustering Situation.....	20
5. A Possible Clustering Situation.....	21
6. A Hypothetical Clustering Situation.....	22

THE DEVELOPMENT AND INVESTIGATION OF SOME EXTENSIONS
TO THE EDERER-MYERS-MANTEL PROCEDURE AND TEST
FOR CLUSTERING

CHAPTER I

INTRODUCTION AND PURPOSE FOR THE STUDY

Cluster analysis has become a subject of major importance in a concentrated attempt to develop techniques for grouping a set of points into disjoint sets of points (Bonner 1964, Gower 1967).

Interest has also arisen however concerning still another type of clustering problem (Ederer, Myers, and Mantel 1964, Stark and Mantel 1967) which is the subject of this investigation. In this type of clustering study one attempts to ascertain if for a defined set of points there exists a tendency for a "large" number of these points to occur (cluster) into any one of a set of predetermined spaces (called strata). Moreover, in application several independent groups of strata (called regions of interest) are defined, and each bit of stratum clustering information per region is cumulated over all regions for a more reliable ascertainment. As an example, consider strata to be the four seasons in a year, and let regions of interest be years. One may like to know if there is any real tendency for persons (points) possessing a certain characteristic to cluster seasonally over all the

years included in the study.

Potential applications of a general technique to investigate this type of clustering problem might be numerous. Epidemiologists in particular might benefit, since many of their studies are fundamentally concerned with the identification of a factor or factors which cause different geographic groupings of persons with some particular condition. Therefore, it is of some importance that a general and flexible procedure be available to determine if non-random geographic clustering actually exists as a preliminary basis preceding a more extensive search for the factor or factors.

There are a number of assumptions and requirements which must be met in order to use the Ederer, Myers and Mantel (EMM) procedure and Chi-square test statistic. While the current procedure is directly applicable in those situations where these assumptions are justifiable, there exist important potential investigations where some of these assumptions are frequently unattainable. The purpose for this paper is to investigate the more commonly encountered situations in which these requirements cannot be met and to generalize the existing techniques so that many of these problems can be analyzed.

One such requirement of the current procedure is that several population values be known in advance (see CHAPTER II). These values are frequently unknown, especially in many prospective studies, often because a total enumeration of the defined population is too costly for consideration. Hence, some method for effectively obtaining these required values is necessary. To this end the rapid development of the theory and practice of sampling has made it possible to obtain

probability estimates for many characteristics and attributes which could only be "estimated" on an intuitive basis by experts or otherwise determined by relying on existing data which is possibly unsuitable. By using the efficiency and flexibility of sampling, procedures for estimating each necessary value have been determined and included in this study.

The basic EMM procedure and test is given and discussed in CHAPTER II using a portion of the notation system which is later introduced in CHAPTER III. The assumptions, limitations, and the capabilities of the procedure are discussed together with hypothetical illustrations for clarification.

The new system of notations, definitions concerning the population partition structure, required partition values, and examples are introduced in CHAPTER III.

CHAPTER IV deals with the frequently encountered difficulty associated with unequal sizes of population at risk among the strata per region of interest. As previously mentioned this situation, if it exists, violates a major assumption which is necessary for valid application of the EMM procedure. A suitable transformation is determined so that an appropriate analysis can be performed using the transformed values. The appropriateness of this transformation is evaluated empirically through the use of an electronic computer to generate populations having known characteristics.

A flexible hierarchal sample population is defined in CHAPTER V. It corresponds to the population partition structure which is given in CHAPTER III. Methods for estimating each of the parameter values as

required in CHAPTERS II and IV are also included. Alternate sample designs are introduced which must be used in certain situations that depend upon the amount of pertinent information initially available. In addition, the concept of sample sizes is considered in CHAPTER V. A method for deriving the size of sample necessary to select from each stratum is included, and an example is given.

In the event that only a sample of regions is selected from the total number of regions in the population for analysis, the current EMM Chi-square test statistic is no longer applicable. Yet, as the costs and time required to include additional regions of interest in the study become large, a sampling procedure becomes necessary and the problem emerges. To overcome this difficulty, a test statistic is developed in CHAPTER VI suitable for testing the null hypothesis when a random sample of k regions is selected from the total population of K defined regions. The test statistic is experimentally evaluated, again through the use of an electronic computer.

Finally results and conclusions are summarized in CHAPTER VII.

CHAPTER II

REVIEW OF PREVIOUS WORK

This section will describe the basic Ederer, Myers and Mantel procedure and test statistic for clustering and serves as a basis for its later extension and supplementation. The EMM technique was developed in part to ascertain whether or not there existed any significant yearly clustering of children diagnosed as having leukemia over a large number of town units. In the study there was a total of $K = 73$ five-year town units (regions of interest: U_g , $g=1,2,\dots,K$) with each region, U_g , containing $L_g = 5$ single-year time periods (strata: U_{gh} , $h=1,2,\dots,L_g$) of data. The total number of cases diagnosed per region, X_g , is the sum of all the yearly cases, X_{gh} , within that region.

The EMM procedure is based on the underlying principle that: in the absence of any real stratum clustering, the total number of regional units, X_g , observed as possessing a certain characteristic, should be distributed independently and at random among the L_g strata with probability $1/L_g$ that any unit occurs in a particular stratum if the strata are all the same size. With this formulation it is possible to determine for suitably given¹ L_g and X_g the expected value and

¹Ederer, Myers, and Mantel (1964) and Stark and Mantel (1967) give tabulated values for $E(X_{gh_m})$ and $\text{Var}(X_{gh_m})$ when $L_g = 2, 3, 4, 5$ and $X_g \geq 2$. Refer to the appendix in this paper for these values when $L_g = 2, 3, \dots, 10$ and $X_g \geq 100$.

variance of X_{gh_m} which denotes the largest (maximum) number of units with the characteristic arising in any single stratum. The word "any" should be noted carefully. Thus for every region which is included in the study, one is able to determine the "conditional" expectation and variance of X_{gh_m} , denoted by $E(X_{gh_m})$ and $\text{Var}(X_{gh_m})$ respectively, where $g = 1, 2, \dots, K$. All the observed values of the X_{gh_m} deviations from their expected value are then cumulated over the K regions of interest and these cumulated deviations are tested by a single degree of freedom Chi-square. That is to say, if the null hypothesis of no stratum clustering is true, the variate

$$\chi^2_{(1)} = \frac{\left[\frac{\sum_g^K X_{gh_m} - \sum_g^K E(X_{gh_m})}{K} \right]^2}{\sum_g^K \text{Var}(X_{gh_m})} \quad (1)$$

has approximately a one degree of freedom Chi-square distribution.

In equation (1) and in other equations involving summations, the sums will always extend from a lower limit of one to the indicated upper limit.

Consider now the following hypothetical set of data to illustrate some details of this technique. Table 1 shows the data values for:

- (1) the number of units possessing the characteristic per stratum, X_{gh} ,
- (2) the number of such units contained in each region, $X_g = \sum_h^{L_g} X_{gh}$,
and
- (3) the number of strata per region, L_g .

Furthermore, in Table 2 each regional expectation and variance of X_{gh_m}

TABLE 1
 NUMBER OF CASES OBSERVED PER STRATUM
 BY REGIONS (HYPOTHETICAL)

Region U_g	Number of Strata L_g	Cases X_g	Cases, X_{gh} , in Stratum U_{gh}				
			X_{g1}	X_{g2}	X_{g3}	X_{g4}	X_{g5}
U_1	5	6	2	0	3	0	1
U_2	5	3	2	0	0	1	0
U_3	5	10	0	3	2	4	1
U_4	5	4	1	0	0	1	2

TABLE 2
 EXPECTATION AND VARIANCE OF X_{gh_m} PER REGION
 FOR THE DATA IN TABLE 1

Region U_g	X_{gh_m}	$E(X_{gh_m})$	$Var(X_{gh_m})$
U_1	3	2.57	0.45
U_2	2	1.56	0.33
U_3	4	3.76	0.69
U_4	2	1.95	0.35
Totals	11.0	9.84	1.82

is given. If the values L_g and X_g are known per region U_g , then one can derive every regional value for $E(X_{gh_m})$ and $\text{Var}(X_{gh_m})$. The following example briefly illustrates this derivation procedure.

Consider only the single region $U_g = U_2$ in Table 1 where $L_2 = 5$ and $X_g = 3$. Referring now to Table 3, notice that these $X_2 = 3$ units can distribute themselves over the $L_2 = 5$ strata in three distinct ways:

- (1) all three units can occur in the same stratum,
- (2) two units can occur together in a single stratum and the third can occur alone, or
- (3) they can each occur in a different stratum.

In terms of occupancy numbers (Feller, 1957), the $X_2 = 3$ units can only have the three possible types of distributions shown in Table 3. It can be shown that under the null hypothesis, the probability of observing the distribution of type $(X_{2h}) = (X_{21}, X_{22}, X_{23}, X_{24}, X_{25})$ is given by

$$\Pr(X_{21}, X_{22}, \dots, X_{25}) = \frac{X_2!}{X_{21}!X_{22}!X_{23}!X_{24}!X_{25}!} \frac{L_2!}{n_0!n_1!\dots n_{X_2}!} \left(\frac{1}{L_2}\right)^{X_2}$$

where: $L_2 = 5$ denotes the number of strata in region U_2 ,

X_{2h} denotes the number of units in stratum h of region U_2 ,

and n_0 denotes the number of strata containing exactly zero units.

Therefore it follows that

$$\begin{aligned} \Pr(3,0,0,0,0) &= \left(\frac{3!}{3!}\right) \left(\frac{5!}{4!1!}\right) \left(\frac{1}{5}\right)^3 \\ &= .04 \end{aligned}$$

is the probability of observing the distribution $(3,0,0,0,0)$.

Similarly

$$\Pr(2,1,0,0,0) = .48 ,$$

and

$$\Pr(1,1,1,0,0) = .48 .$$

TABLE 3

POSSIBLE DISTINCT DISTRIBUTIONS OF 3 UNITS IN 5 CELLS

Distribution Number	Cell Number				
	1	2	3	4	5
1	3	0	0	0	0
2	2	1	0	0	0
3	1	1	1	0	0

It should be noted that, if several regions actually presented the first distribution in Table 3, it would be an indication that the $X_2 = 3$ units have a tendency to cluster.

Now the expected value of X_{gh_m} is given by

$$E(X_{gh_m}) = \sum X_{gh_m} \cdot \Pr(X_{gh_m}) \quad (2)$$

where the summation extends over all possible values that X_{gh_m} can assume, when L_g and X_g are both fixed. Thus, in the example where $L_2 = 5$ and $X_2 = 3$,

$$\begin{aligned} E(X_{2h_m}) &= 3(.04) + 2(.48) + 1(.48) \\ &= 1.56 . \end{aligned}$$

Furthermore, the variance of X_{gh_m} is given by

$$\begin{aligned} \text{Var}(X_{gh_m}) &= E[X_{gh_m} - E(X_{gh_m})]^2 \\ &= E(X_{gh_m}^2) - [E(X_{gh_m})]^2, \end{aligned} \quad (3)$$

where $E(X_{gh_m}^2) = \sum X_{gh_m}^2 \cdot \Pr(X_{gh_m})$, and the summation extends

over the same values for X_{gh_m} as in equation (2) above.

In the example, since

$$\begin{aligned} E(X_{2h_m}^2) &= 9(.04) + 4(.48) + 1(.48) \\ &= 2.76, \end{aligned}$$

it follows that the variance given by equation (3) is

$$\begin{aligned} \text{Var}(X_{2h_m}) &= 2.76 - (1.56)^2 \\ &= .3264 . \end{aligned}$$

Each region is treated in similar manner, and the results are listed in Table 2. Finally, the EMM Chi-square value is computed from equation (1) as

$$\begin{aligned} \chi_{(1)}^2 &= \frac{(11.0 - 9.84)^2}{1.82} \\ &= 0.74 \end{aligned}$$

The null hypothesis is not rejected at level of significance $\alpha = .05$, since $\chi_{(1)}^2 < \chi_{(1)}^2, 1 - \alpha$.

There are some important conditions which must be met to insure the validity of the Ederer, Myers and Mantel procedure. They are summarized at this point for special emphasis, since the purpose of this paper is to determine methods for coping with them.

1. A major requirement can be stated as follows: all of the L_g strata within each region of interest must have an equal chance, $1/L_g$, of containing any unit with the characteristic. This frequently imposes the requirement that the size of each regional stratum population at risk, Y_{gh} , must be equal. It can be seen that this condition eliminates a number of potential cluster investigations. Especially it effects those studies developed so that the L_g strata represent

different geographic or socio-economic areas, because often these areas have unequal population sizes.

2. A second consideration refers to the method of determining both the regions of interest and the various strata within each of these regions once the study hypothesis has been defined. The information which evolves from any cluster analysis is closely related to how effectively these partitions are determined. This is important, since the EMM procedure only provides a method to test for clustering given that the strata and regions of interest have already been determined.
3. Third, the procedure assumes that each population value required (CHAPTER III) for analysis is known. These values may possibly be available in retrospective studies; however, in prospective studies situations frequently arise where these values are not known.
4. A knowledge of the various expectations and variances of X_{ghm} for varying values of L_g and X_g is essential for use of this procedure. The authors originally gave these values only for $L_g = 5$ and $X_g \leq 15$. Stark and Mantel (1967) later derived and made available these expectations and variances when $L_g = 3, 4, 5$ and $X_g \geq 2$. The availability of these values when $L_g = 2, 3, \dots, 10$ and $X_g \geq 100$ would be necessary for use in studies concerning non-rare population characteristics.
5. Finally it was stated that the Chi-square test statistic is valid and achieves power when the number of regions of interest included is "large". Often, however, increased costs and time

prevent this number from being large. Thus in such a situation one may prefer to select a random sample from the total population of regions. Unfortunately the EMM test statistic is no longer appropriate, because of the introduction of sampling variation, and hence a different test statistic is required.

CHAPTER III

DEFINITIONS AND POPULATION NOTATIONS

The original article presenting the Ederer-Myers-Mantel (1964) procedure and test contains a notation system which is entirely adequate in many situations. However, in order to integrate effectively the different topics to be discussed in subsequent chapters (and in CHAPTER II) of this paper, it is necessary that a more flexible system of notation be introduced at this point. The application of various methodologies and statistical concepts plays a major role in attaining the objectives of this investigation. Sampling theory and estimation procedures are among those aspects involved.

Population Partition Structure

In what follows, the study population shall be defined according to a specific hierarchal classification system. That is, groups may be composed of sub-groups, sub-groups may be composed of sub-sub-groups, etc. This "nesting" gives rise to "hierarchal classifications".

Consider first a population denoted by U of size $c(U) = N$ units, and the first partition (U_g) of size $c(U_g) = N_g$ units such that $N_g \neq 0$, ($g = 1, 2, \dots, K$), and $\sum_{g=1}^K N_g = N$. Call U_g the g^{th} region of interest in U . Hence, K is the total number of regions of interest into which the population has been first divided (see Figure 1).

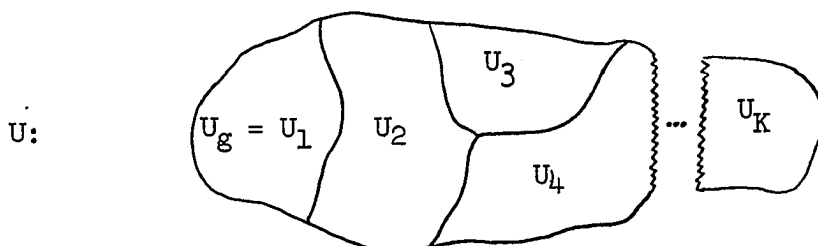


Fig. 1—Regions of interest U_g in U .

Next, considering only a single region of interest, U_g , define another partition (U_{gh}) of U_g such that the size of U_{gh} is $c(U_{gh}) = N_{gh}$ units, where $N_{gh} \neq 0$, ($h = 1, 2, \dots, L_g$) and $\sum_h^{L_g} N_{gh} = N_g$. Call U_{gh} the h^{th} stratum of region U_g . Hence L_g is the number of strata into which each region has been divided (see Figure 2).

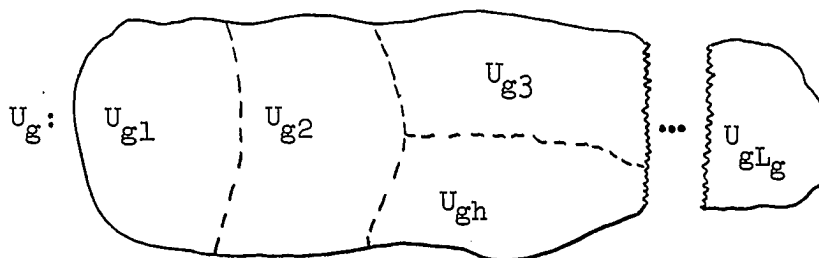


Fig. 2—Strata U_{gh} within region of interest U_g .

Both the regions of interest and the strata per region are determined according to definite objectives. The principle underlying these objectives is discussed in the section Principles Underlying Strata and Region Determinations of this chapter. It is sufficient to mention here that the null hypothesis (discussed in CHAPTER II) is

based upon the determination of these strata within each region. The regions are in a sense desirable replications over the defined population.

Finally consider a single stratum, U_{gh} , and define a partition (C_{ghi}) of U_{gh} such that the size of C_{ghi} is $c(C_{ghi}) = N_{ghi}$ where $N_{ghi} \neq 0$, ($i = 1, 2, \dots, M_{gh}$), and $\sum_{i=1}^{M_{gh}} N_{ghi} = N_{gh}$. Call C_{ghi} the i^{th} primary sampling unit (psu) within the h^{th} stratum of region U_g (see Figure 3).

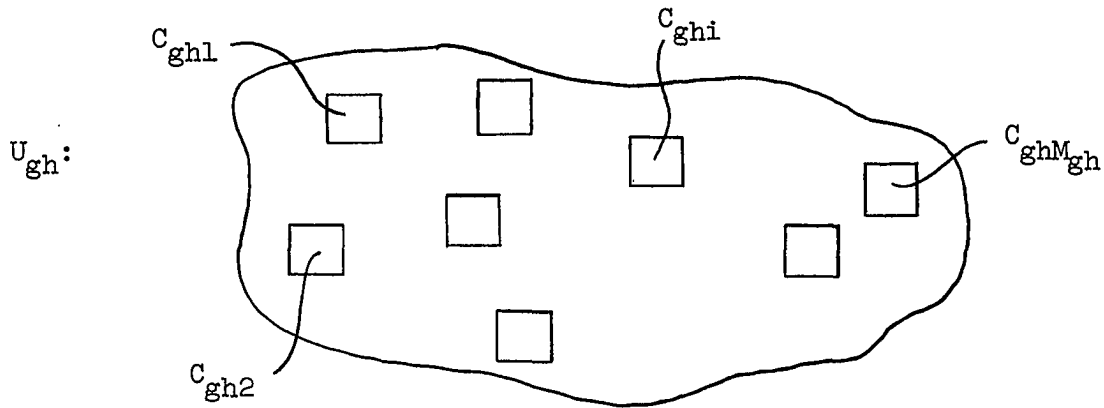


Fig. 3—Primary sampling units C_{ghi} in stratum U_{gh} .

The purpose for defining the above primary sampling units is to introduce a system whereby sample estimators can be obtained for each required stratum parameter for the cases in which a simple random sample of the elementary risk units cannot be selected. This is a frequently encountered problem, and the change in notation is to retain this distinction. The pertinent parameters are defined in another section of this chapter, while the corresponding sample estimators are given in CHAPTER V.

In summary the above structure defines a three level

partitioning of the total population of N units such that

$$\sum_i^{M_{gh}} N_{ghi} = N_{gh} = c(U_{gh}) ,$$

$$\sum_h^{L_g} \sum_i^{M_{gh}} N_{ghi} = \sum_h^{L_g} N_{gh} = N_g = c(U_g) ,$$

and

$$\sum_g^K \sum_h^{L_g} \sum_i^{M_{gh}} N_{ghi} = \sum_g^K N_g = N = c(U)$$

are true. This hierarchy can be extended as more classifications become necessary.

Partition Parameters

The following values, as defined, relate to various partition characteristics or attributes, all or part of which are required in the appropriate cluster analysis. Too, the conditions for valid use of the EMM procedure involves placing certain restrictions on some of these parameters (see CHAPTER II).

Define

$$X_{ghij} = \begin{cases} 1 & \text{if the } ghij^{\text{th}} \text{ risk unit possesses the} \\ & \text{characteristic, and} \\ 0 & \text{otherwise} \end{cases}$$

and $Y_{ghij} = 1$ always.

Let $Y_{ghi} = \sum_i^{N_{ghi}} Y_{ghij}$ denote the total number of units at risk in primary sampling unit (psu) C_{ghi} ,

$X_{ghi} = \sum_j^{N_{ghi}} X_{ghij}$ denote the number of units in psu C_{ghi} possessing the

and $R_{ghi} = X_{ghi}/Y_{ghi}$

attribute,

denote the proportion of Y_{ghi} units in C_{ghi} possessing the attribute.

Then let $Y_{gh} = \sum_i^{M_{gh}} Y_{ghi}$

denote the total number of units at risk in stratum U_{gh} ,

$$X_{gh} = \sum_i^{M_{gh}} X_{ghi}$$

denote the number of units in U_{gh} possessing the attribute,

and $R_{gh} = X_{gh}/Y_{gh}$

denote the corresponding proportion of units possessing the attribute.

Furthermore, let

$$Y_g = \sum_h^{L_g} Y_{gh}$$

denote the total number of units at risk in region U_g ,

$$X_g = \sum_h^{L_g} X_{gh}$$

denote the number of units possessing the attribute in region U_g ,

and $R_g = X_g/Y_g$

denote the proportion of units possessing the attribute in region U_g .

Finally, let

$$Y = \sum_g^K Y_g$$

denote the total number of units at risk in the population U ,

$$X = \sum_g^K X_g$$

denote the number of units in U possessing the attribute,

and $R = X/Y$ denote the corresponding proportion in U.

Now define the following

X_{gh_m} = the largest number of units with the attribute arising in any single stratum of region U_g ,

R_{gh_m} = the corresponding stratum proportion,

and $\sum_g^K X_{gh_m}$ = the sum of all K regional values of X_{gh_m} in population U.

All stratum values X_{gh} may be transformed as necessary according to the conditions and methodology discussed in CHAPTERS II and IV. These values are defined below for completeness.

Let

X_{gh}^* denote the number of units possessing the attribute which would be expected in stratum U_{gh} if every stratum in region U_g had contained \bar{Y}_g units at risk,

and $X_g^* = \sum_h^{L_g} X_{gh}^*$ denote the corresponding regional total.

Therefore $X_{gh_m}^*$ denotes the largest value corresponding to the definition above for X_{gh_m} .

Principles Underlying Strata and Region Determinations

Consider a study in which a researcher has already defined a population totaling Y units, and suspects that in those X ($X \leq Y$) units which possess a particular characteristic there exists non-random

clustering. Non-random clustering implies that these X units consist of groups of units (clusters), and that the occurrence of these clusters is not explainable entirely by random grouping alone. That is, the existence of some factor (or factors) causes non-random clustering. A researcher's fundamental interest might be in further investigations regarding possible factors related to the clustering, once a valid analysis indicates non-randomness.

Consider the following two situations:

(a) Assume that a researcher has some notion concerning a reason for clustering and would like to ascertain the validity of this notion. In this case he might be able to define K regions of interest, as well as L_g sub-divisions (strata) within each of the regions in the following manner.

(1) Determination of the strata per region should be based on that suspected cause of clustering. The strata should be defined so that they differ among themselves and are homogenous within themselves regarding the suspected cause. In this way the null hypothesis of no strata clustering of units possessing the characteristic would be formulated.

(2) The K regions of interest (ideally) should represent K homogenous images of the true cluster pattern; that is, tendencies to cluster because of some factor, should remain consistent from region to region in the population.

Now if the suspected cause is correct, by subsequently

observing the number of units X_{gh} per stratum (as in Figure 4) a subsequent analysis should indicate this within-region clustering and reject the null hypothesis of randomly distributed points.

It is to be noted that Figure 4 (and similar figures) is merely a diagram for illustrating the occurrence of those units possessing the characteristic in a corresponding partition and is not meant to specify either the density or the exact location of units in a particular partition.

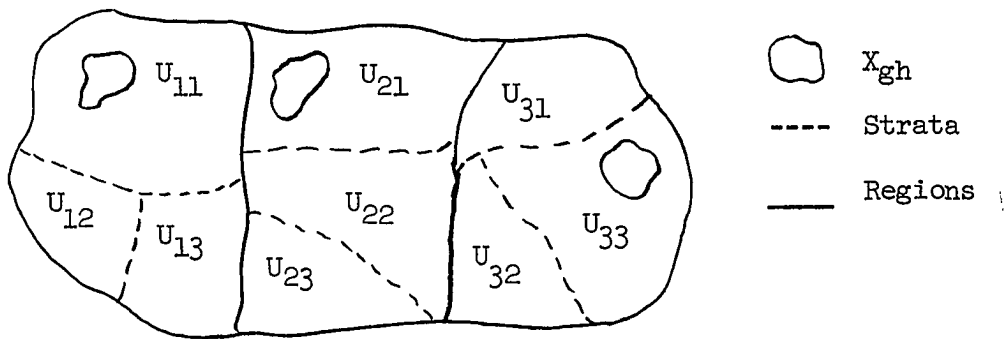


Fig. 4—A possible clustering situation.

- (b) Assume that a researcher has no prior notion about the factor responsible for clustering, yet suspects that clustering does exist. There is no fixed basis for determining the strata per region. One might arbitrarily define both the strata and the regions of interest in some attempt to demonstrate strata clustering; however, depending on the true but unknown cluster pattern, this arbitrary choice could yield very misleading conclusions.

For example the partition structure shown in Figure 5 indicates that the clustering tendencies remain consistent from region to region. But, since now the strata boundaries divide each regional cluster into three parts, strata clustering could easily go undetected even though it does exist.

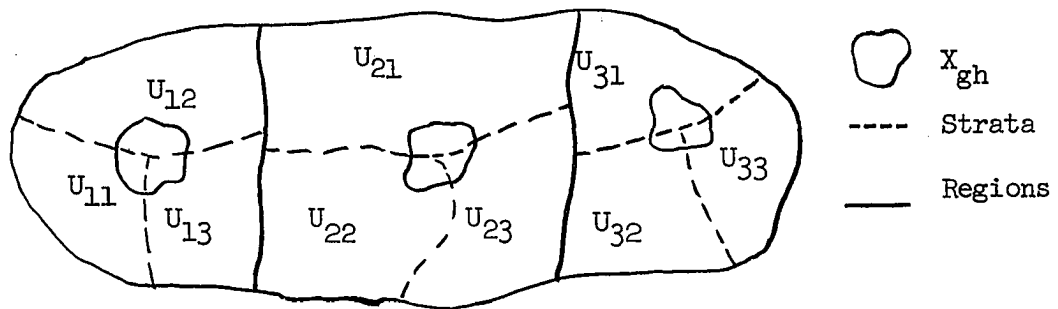


Fig. 5—A possible clustering situation.

The two situations were included to illustrate a very important point. Any one of several clustering patterns might exist, and for each there is an efficient partition structure for determining the pattern's existence. However, as mentioned above, there are also many ways to select both the regions and the strata per region, and each alternative can yield a different conclusion. One should concentrate much initial effort in logically determining the most appropriate partition structure. The purpose of the EMM procedure and all proposed extensions given in this paper is to detect non-random clustering of units within regions given that the regions, strata, and population have already been defined. Further, it is not intended to identify any specific cause for non-random clustering because of possible confounding with unsuspected factors.

A Sample Problem

Consider the notations given in the section Partition Parameters and the hypothetical situation indicated by Figure 6. This population U is composed of $K = 2$ regions (U_1, U_2) with $L_g = 3$ strata per region (U_{11}, U_{12}, U_{13}) and (U_{21}, U_{22}, U_{23}).

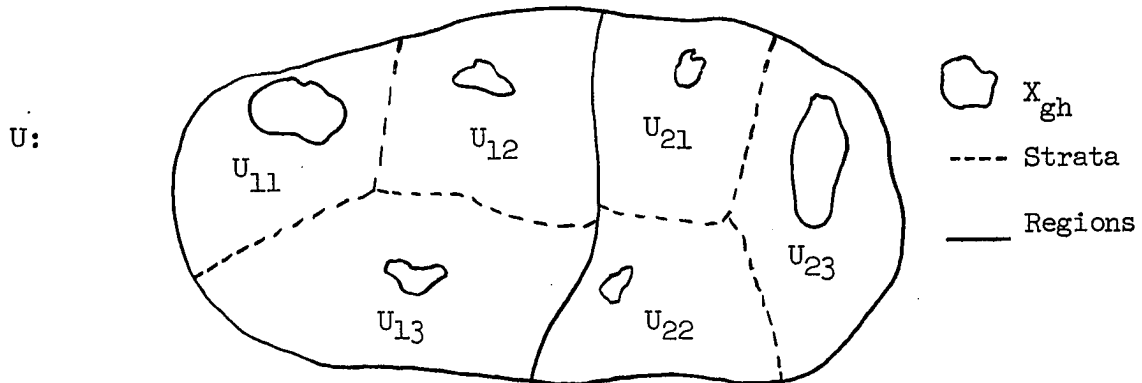


Fig. 6—A hypothetical clustering situation.

The corresponding values for Figure 6 (hypothetical) are listed in Tables 4 and 5. Notice that each stratum value Y_{gh} is equal to \bar{Y}_g , therefore $X_{gh}^* = X_{gh}$.

The EMM test statistic can be calculated from equation (1) as

$$\begin{aligned} \chi_o^2 &= \frac{(109.0 - 64.05)^2}{13.12} \\ &= 154.3 \end{aligned}$$

Since $\chi_o^2 > \chi_{(1)}^2$, 1-.05 the null hypothesis is rejected at level of significance $\alpha < .05$. Notice in Table 4 that the values $X_{11} = 35$ and $X_{33} = 74$ are both quite large as compared to the other regional strata values. One might suspect that they were abnormally large even before an analysis. In Table 5 it is clear that these values were both large

as compared to their corresponding expected values, and consequently the null hypothesis of randomly distributed units was rejected.

TABLE 4

OBSERVED CASES X_{gh} PER STRATUM U_{gh}
IN REGION U_g (HYPOTHETICAL)

Regions U_g	No. Strata L_g	$X_g =$ L_g $\sum_h X_{gh}$	$Y_g =$ L_g $\sum_h Y_{gh}$	STRATA U_{gh}					
				U_{g1}		U_{g2}		U_{g3}	
				X_{g1}	Y_{g1}	X_{g2}	Y_{g2}	X_{g3}	Y_{g3}
U_1	3	47	300	35	100	5	100	7	100
U_2	3	120	600	20	200	26	200	74	200

TABLE 5

EXPECTATION AND VARIANCE OF X_{gh_m} PER REGION U_g
FOR THE DATA IN TABLE 4

Region U_g	Observed X_{gh_m}	$E(X_{gh_m})$	$Var(X_{gh_m})$
U_1	35	19.06	3.80
U_2	74	45.00	9.32
Totals	109	64.06	13.12

CHAPTER IV

UNEQUAL SIZES OF THE STRATA POPULATIONS

AT RISK WITHIN REGIONS

In the discussion in CHAPTER II, regarding the conditions and assumptions under which the EMM procedure and test might validly be applied in testing the hypothesis of randomly distributed units possessing a certain characteristic, it was emphasized that each regional stratum must have an equal chance of containing any single unit with this characteristic. The requirement was necessary in order that each of the several regional expectations and variances involved in the Chi-square test statistic could feasibly be derived. Often this requirement imposes the restriction that the number of units (say, persons at risk) within each stratum be equal for all strata contained in that region. Clearly if this were not required, an observed number of persons per stratum having the characteristic would vary as the population at risk per stratum changed and not necessarily as a result of clustering. Consequently any significant stratum clustering of such units would be confused (or confounded) with risk population size differences.

The condition is justifiable in certain instances. The leukemia study (Ederer, Myers and Mantel 1964), for example, was developed such that the strata represented single year time periods, while the

regions of interest were defined as geographic areas (towns). It was necessary to assume therefore, that each geographic area's population at risk remained stable over the $L_g = 5$ yearly periods, and that any change which might occur would not seriously alter the conclusions of the analysis. However, Ederer, Myers and Mantel were not willing to assume that this same stability would remain over a 15 year period ($L_g = 15$).

There exists a number of potential clustering problems where the strata might represent different geographic or socio-economic areas. These area types generally do not contain equal population sizes at risk, and therefore the corresponding strata probabilities are not equal.

For these reasons it seemed desirable to develop a technique to cope with this situation. In what follows a method for transforming each observed value, X_{gh} , to a new value, X_{gh}^* , is developed. This transformation is determined in such a way that the resulting values, X_{gh}^* , are effectively based upon equally populated strata per region, and therefore suitable for analysis.

Derivation of the Transformation on X_{gh}

There are two intuitively desirable conditions which are involved in developing this transformation. The initial condition refers to the strata proportions $R_{gh} = X_{gh}/Y_{gh}$. It will be required that each of these proportions is preserved under the transformations on X_{gh} . That is to say $R_{gh}^* = R_{gh}$. The second requirement concerns an appropriate value of N_0 which is a regional constant defined in such a way that

$$R_{gh}^* = X_{gh}^*/N_0.$$

Consider now the first condition

$$R_{gh}^* = R_{gh} \quad (h = 1, 2, \dots, L_g), (g = 1, 2, \dots, K). \quad (4)$$

It follows that since

$$R_{gh}^* = R_{gh} = \frac{X_{gh}}{Y_{gh}},$$

then $X_{gh} = R_{gh}^* (Y_{gh})$. (5)

But also from above, R_{gh}^* is expressed in terms of X_{gh}^* as

$$R_{gh}^* = \frac{X_{gh}^*}{N_0},$$

and therefore equation (5) becomes

$$X_{gh} = \frac{X_{gh}^*}{N_0} (Y_{gh}).$$

Hence

$$\begin{aligned} X_{gh}^* &= \frac{X_{gh}}{Y_{gh}} (N_0) \\ &= R_{gh} (N_0) \end{aligned} \quad (6)$$

is the required value under the transformation, where N_0 is the regional constant. To this point the transformation is developed so that the first requirement, given by equation (4), is maintained for any arbitrary value N_0 .

The second condition concerns an appropriate value for N_0 per region. This value of N_0 should represent a value which is near the actual sizes of the strata populations at risk, Y_{gh} in region U_g . It seems appropriate to let $N_0 = \bar{Y}_g$, where \bar{Y}_g denotes the average risk population size per stratum in region U_g . When this is done, and when all strata values for Y_{gh} are exactly equal within a given region, the transformation given by equation (6) becomes

$$\begin{aligned} X_{gh}^* &= \frac{X_{gh}}{Y_{gh}} (\bar{Y}_g) \\ &= X_{gh}, \end{aligned}$$

since by construction $Y_{gh} = Y_{gh'}$, when $h \neq h'$, and therefore the regional mean

$$\bar{Y}_g = \frac{1}{L_g} \sum_h^{L_g} Y_{gh} \quad (7)$$

becomes

$$\begin{aligned} \bar{Y}_g &= \frac{1}{L_g} (L_g)(Y_{gh}) \\ &= Y_{gh} \end{aligned} \quad (8)$$

Hence, if in addition to the first condition, the second condition is also true, this transformation does not alter the original values of X_{gh} when equation (8) is satisfied. This is a desirable feature of the transformation. Consequently, subject to the two previous conditions, the recommended transformation can be expressed as

$$X_{gh}^* = R_{gh}(\bar{Y}_g), \quad (9)$$

where

$$R_{gh} = X_{gh}/Y_{gh}$$

and

$$\bar{Y}_g \text{ is given in equation (7).}$$

X_{gh}^* may now be defined as that number of risk units possessing the characteristic in stratum U_{gh} which would have been expected to occur if every stratum U_{gh} ($h = 1, 2, \dots, L_g$) had contained \bar{Y}_g units at risk.

The usefulness of the form of the transformation expressed in equation (9) is amplified in CHAPTER V where the procedures for estimating each stratum value for X_{gh}^* from a probability sample are considered. Since each of the transformations requires only the values R_{gh} and \bar{Y}_g , the corresponding estimators will be used to estimate

independently each stratum value for R_{gh} at a level of precision which is pre-assigned by the researcher. Furthermore, in this situation one would have a stratified sample design available with a combined sample of size, $n_g = \sum_h^{L_g} n_{gh}$, large enough to estimate the regional parameter \bar{Y}_g by \bar{y}_g with a high degree of precision. It is noticed that one can effectively estimate X_{gh}^* without requiring that the individual values X_{gh} and Y_{gh} be known.

An Example of the Use of the Transformation

Referring to the hypothetical set of data which is given in Table 6 it is seen that $Y_{gh} \neq Y_{gh'}$, ($h \neq h'$). Applying the transformation given by equation (9), the values for X_{gh}^* listed in Table 7 follow. In addition, once the values for $X_g^* = \sum_h^{L_g} X_{gh}^*$ and L_g are known, each regional expectation and variance of $X_{gh_m}^*$ can be found and listed as in Table 8. Inserting the values given in Table 8 in equation (1), one calculates the corresponding Chi-square test statistic as

$$\begin{aligned} \chi_{(1)}^2 &= \frac{\left[\begin{array}{c} K \\ \sum_g X_{gh_m}^* - \sum_g E(X_{gh_m}^*) \end{array} \right]^2}{\begin{array}{c} K \\ \sum_g \text{Var}(X_{gh_m}^*) \end{array}} \\ &= \frac{(240.3 - 132.5)^2}{26.82} \\ &\doteq 434.0 \end{aligned}$$

Since $\chi_{(1)}^2 > \chi_{(1)}^2$, 1-.05 the null hypothesis is rejected at level of significance, $\alpha < .05$.

Evaluation of the Transformation

For the case where the risk population sizes Y_{gh} are equal

TABLE 6

NUMBER OF CASES OBSERVED X_{gh} PER REGIONAL STRATUM U_{gh} (HYPOTHETICAL)

Region U_g	L_g	Cases X_g	Y_g	STRATUM U_{gh}								
				U_{g1}			U_{g2}			U_{g3}		
				X_{g1}	Y_{g1}	R_{g1}	X_{g2}	Y_{g2}	R_{g2}	X_{g3}	Y_{g3}	R_{g3}
U_1	3	161	540	15	100	.15	11	140	.08	135	300	.45
U_2	3	158	540	8	100	.08	18	140	.13	132	300	.44
U_3	3	154	540	7	100	.07	14	140	.10	133	300	.443

TABLE 7

TRANSFORMED VALUES X_{gh}^* PER REGIONAL STRATUM U_{gh} FOR THE DATA IN TABLE 6

Region U_g	L_g	Cases X_g^*	\bar{Y}_g	STRATUM U_{gh}					
				U_{g1}		U_{g2}		U_{g3}	
				X_{g1}^*	R_{g1}^*	X_{g2}^*	R_{g2}^*	X_{g3}^*	R_{g3}^*
U_1	3	122.21	180	27.0	.15	14.21	.08	81.0	.45
U_2	3	116.9	180	14.4	.08	23.20	.13	79.3	.44
U_3	3	110.6	180	12.6	.07	18.00	.10	80.0	.443

TABLE 8

EXPECTATIONS AND VARIANCES OF TRANSFORMED VALUES $X_{gh_m}^*$

Region U_g	$X_{gh_m}^*$	$E(X_{gh_m}^*)$	$Var(X_{gh_m}^*)$
U_1	81.0	45.0	9.32
U_2	79.3	44.5	9.00
U_3	80.0	43.0	8.50
Totals	240.3	132.5	26.82

per region the transformation is clearly applicable. However for unequal Y_{gh} it was felt that a theoretical treatment of the effect of the transformation upon the appropriateness of the EMM procedure and test statistic should be preceded by an empirical evaluation. To this end, an experimental investigation was carried out making extensive use of an electronic computer. This investigation was conducted in the following manner.

Thirty regions of interest, $K = 30$, were determined each possessing $L_g = 5$ strata containing unequal risk population sizes, Y_{gh} (see Table 9 for a sample region). Consider only a single region, U_g . A number, X_g , which represents the number of risk units possessing a certain characteristic in region U_g , was randomly selected. The values X_g and Y_g were chosen so that each proportion $R_g = X_g/Y_g$, was approximately equal to .20. Using this information as input to the computer, the X_g units were then allocated to the $L_g = 5$ strata by a random process which was so developed that the number of units, X_{gh} , generated into stratum U_{gh} was in proportion to population at risk, Y_{gh} , contained in that stratum, except for random variations. The transformation was then performed by using equation (9) for each generated value, X_{gh} . The process was repeated in each of the $K = 30$ regions. From the resulting transformed values, X_{gh}^* , each X_g^* was found, each $E(X_{gh_m}^*)$ and $Var(X_{gh_m}^*)$ was determined, and the corresponding EMM Chi-square test statistic was computed.

The process described above was repeated yielding a total of 228 independent experiments. In these experiments, given that the null hypothesis is true, the hypothesis was rejected thirteen times.

TABLE 9

RANDOMLY GENERATED CASES X_{gh} PER STRATUM U_{gh} ,
 GIVEN: $L_g = 5$, $X_g = 805$

Stratum U_{gh}	Risk Population Y_{gh}	Generated Cases X_{gh}	Rate R_{gh}	Transformed Values X_{gh}^*
U_{g1}	940	175	.1861	148.9362
U_{g2}	700	134	.1914	153.1429
U_{g3}	475	94	.1978	158.3158
U_{g4}	310	61	.1967	157.4194
U_{g5}	1575	341	.2165	173.2063
Totals	4000	$X_g = 805$.2012	$X_g^* = 791.0179$

Thus an estimate of the experiment-wise Type I error rate is $\hat{\alpha} = 13/228 \cong .06$. Since this value is not significantly different from the theoretical value of $\alpha = .05$, it is concluded that the transformation may validly be used. In addition, the process was repeated 175 times in a manner similar to the above, except that all values of X_g were selected such that $R_g \cong .003$. Of these 175 experiments, only 11 were rejected.

It was then decided to investigate the power of the test statistic under this transformation, and to this end a particular alternate hypothesis was defined. Since power may be thought of as the probability of rejecting the null hypothesis when in fact the alternate is true, it is this probability which one must estimate. In order to investigate this probability experimentally a technique

for randomly generating data sets in such a way that the alternate hypothesis is true had to be determined. This resulted in modifying the computer input values for Y_{gh} to achieve the proper random generation process. Once this was accomplished, the data sets repeatedly generated, transformations made using the original values of Y_{gh} , and the numerous independent test of the null hypothesis performed; then the experimental estimate of the power would simply be the null hypothesis rejection rate.

To this end the alternate hypothesis was determined as follows. It was decided to investigate how well the test statistic under the transformation could detect an absolute 6% increase in an arbitrarily selected stratum base rate of R_{gh} . Under the null hypothesis, each stratum base rate is equal to the regional rate $R_g = X_g/Y_g$. Recall that each of the random base rates, R_g , ($g = 1, 2, \dots, 30$), were previously determined by randomly selecting values for X_g when each regional risk population size, Y_g , was fixed. Now, the computer input value, Y_{gh} , corresponding to the randomly selected regional stratum where the 6% increase was to occur, had to be modified in order to randomly generate the data sets, X_{gh} , ($g = 1, 2, \dots, 30$), ($h = 1, 2, \dots, L_g$), under the alternate hypothesis. A method was therefore found which made it possible to compute each required input value, and consequently the appropriate sets of data were randomly generated. The complete process was repeated until 150 experiments had been performed. Each experiment was analyzed using the transformation and the original values, Y_{gh} , and the null hypothesis was rejected 147 times. Therefore, the experimental estimate of the power

when this particular alternate hypothesis is true, is $147/150 \cong .98$.

In summary it was found that the transformation given by equation (9) is appropriate and does not seriously alter the Type I error rate. Too, power was maintained, at least under the particular alternate hypothesis selected. It should be noted here that this study of power considered only one of many possible alternatives, a fixed number ($K = 30$) of regions per experiment, and $\alpha = .05$. A more extensive investigation of the power concept is reserved for later consideration.

CHAPTER V

THE SAMPLE AND SAMPLE POPULATION

The Necessity for Sampling

The rapid development of the theory and practice of sampling has led to highly efficient methodology which has made it possible to obtain probability estimates of desired population characteristics or attributes which previously could only be "estimated" on an intuitive basis by experts, or otherwise indirectly approximated from existing information. Realizing this efficiency, professional people in many fields have used various sampling techniques for precisely estimating the desired population parameters.

Such is the case in various segments of this particular type of clustering investigation. Recall that in CHAPTERS II and III the values necessarily required for use in the EMM test statistic were discussed. It was further shown in CHAPTER IV that one must ultimately obtain each stratum value X_{gh}^* where ($g = 1, 2, \dots, K$) and ($h = 1, 2, \dots, L_g$). The value X_{gh}^* , a transformed X_{gh} , is that expected number of risk units possessing a certain characteristic in stratum U_{gh} , determined in such a way that all regional strata effectively contain risk populations of equal size \bar{Y}_g . From equation (9) in CHAPTER IV each stratum value X_{gh}^* is computed as

$$X_{gh}^* = R_{gh}(\bar{Y}_g)$$

where

$$R_{gh} = X_{gh}/Y_{gh}, \quad (g = 1, 2, \dots, K), \quad (h = 1, 2, \dots, L_g),$$

and

$$\bar{Y}_g = \frac{1}{L_g} \sum_h^{L_g} Y_{gh} .$$

Hence it is clear that one must either possess or in some way obtain the values, R_{gh} and \bar{Y}_g . The following sections of this chapter involve practical methodologies for estimating X_{gh}^* , where the method chosen depends upon what pertinent information is available in the initial phase of the study. Furthermore, corresponding sample sizes are derived which insure that each eventual sample estimator is precise, where the measure of precision is predetermined by the researcher.

In the first section that follows, an effective and flexible sample population is defined which corresponds to the partition structure as given in CHAPTER III. The following sections introduce procedures for estimating each required parameter. Since the type of estimator required depends upon the information already available, this gives rise to multiple sampling situations. Some of these commonly encountered situations together with the corresponding estimators are included.

The Sample Population Defined

For every strata parameter required for use in the EMM test statistic as discussed in CHAPTERS II, III, and IV there exists an effective sample estimator that is computed from a sample of risk (elementary) units selected by the method of simple random sampling from the total population of risk units. However, as is frequently the case, a random selection of these sample units within each stratum

cannot conveniently or accurately be made. This situation can arise, for example, when complete lists of all risk units are not available from which to select the sample. Therefore, in order to obtain a required sample size, one must rely on either a single or multi-stage sampling design. It is because of this necessity that the partitions per stratum, $C_{ghi}(i=1,2,\dots,M_{gh})$, were defined. The optimum choice of design depends upon many factors among which are those relating to cost, time available, the precision required, and data already available.

The sample population defined as follows corresponds to the population partition structure given in CHAPTER III.

Assume that $k < K$ regions of interest, (U_1, U_2, \dots, U_k) , are randomly sampled from the total number of regions (U_1, U_2, \dots, U_K) in the population, that m_{gh} primary sampling units are sampled from the total number of primary units, M_{gh} , in each stratum and that n_{ghi} risk units are selected at random from the N_{ghi} units per included primary sampling unit C_{ghi} . This hierarchal sampling scheme results in samples of size

$$n = \sum_g^k \sum_h^{L_g} \sum_i^{m_{gh}} n_{ghi} ,$$

$$n_g = \sum_h^{L_g} \sum_i^{m_{gh}} n_{ghi} ,$$

and

$$n_{gh} = \sum_i^{m_{gh}} n_{ghi}$$

which can be used for estimating parameters in U , U_g , and U_{gh} respectively.

It should be noted that the above system is sufficient to

define a stratified two-stage hierarchal sample per region, and a two-stage hierarchal sample within each stratum.

Sample Estimators

This section includes procedures for estimating

$$X_{gh}^* = R_{gh}(\bar{Y}_g)$$

under various situations which depend upon the amount of pertinent information already available.

Assume first that the size of the risk population, Y_{gh} , is known in stratum U_{gh} , and that a complete list of this risk population is available from which to draw a sample. The estimator of X_{gh}^* is then given by

$$\hat{X}_{gh}^* = r_{gh}(\bar{Y}_g) ,$$

where

$$r_{gh} = \frac{\frac{N_{gh}}{n_{gh}} \sum_j^{n_{gh}} x_{ghj}}{\frac{N_{gh}}{n_{gh}} \sum_j^{n_{gh}} y_{ghj}} ,$$

$$x_{ghj} = \begin{cases} 1, & \text{if the } ghj^{\text{th}} \text{ risk unit possesses the characteristic} \\ 0, & \text{otherwise} \end{cases}$$

$$y_{ghj} = 1 \text{ for each risk unit in the sample,}$$

$$\text{and } \bar{Y}_g = \frac{1}{L_g} \sum_h^{L_g} Y_{gh} .$$

It is noted that no sample of primary sampling units is necessary since a list of all risk units exists.

Assume now that the value Y_{gh} is not known, and hence no list of the risk units is available. In this case, assuming that suitable

primary sampling units can be defined, the estimator of X_{gh}^* is

$$\hat{X}_{gh}^* = r_{gh}(\bar{y}_g) ,$$

where now

$$r_{gh} = \frac{\frac{M_{gh}}{m_{gh}} \sum_i^{m_{gh}} \frac{N_{ghi}}{n_{ghi}} \sum_j^{n_{ghi}} x_{ghij}}{\frac{M_{gh}}{m_{gh}} \sum_i^{m_{gh}} \frac{N_{ghi}}{n_{ghi}} \sum_j^{n_{ghi}} y_{ghij}} \quad (10)$$

$$\bar{y}_g = \frac{1}{L_g} \sum_i^{M_{gh}} \frac{m_{gh}}{m_{gh}} \sum_j^{N_{ghi}} \frac{n_{ghi}}{n_{ghi}} y_{ghij} \quad (11)$$

The sample variances which correspond to the above sample estimators are given in a number of standard textbooks on probability sampling. Moreover, techniques for optimizing the number of primary sampling units to include per stratum, as well as the optimum number of risk units to select per included primary sampling unit, are given. This optimization process involves determination of the best sampling procedure in consideration of costs, time available, and precision required.

Derived Sample Size Per Stratum

In planning any study in which probability sampling techniques are employed, a point is always reached at which some decision must be made about the size of the sample or samples. This decision is an important one, since oversampling causes a waste in resources. On the other hand undersampling in critical areas will diminish the validity and hence the utility of an investigation's results. Also an inefficient allocation of the appropriate sample size of risk units to the primary sampling units in a multi-stage plan can cause

both wasted resources and diminished validity.

Frequently sufficient information is not available to be certain that a sample size is the best one. However, sampling theory does provide some measures by which the problem can be intelligently approached.

In applying the theory of sampling to this type of clustering problem it was found that there are several parameters that must be estimated independently of each other. That is, every stratum value of X_{gh}^* must be estimated individually. Since X_{gh}^* is given by equation (9), this implies that all strata values R_{gh} must be obtained. Too, it was seen in the previous section that \bar{Y}_g must frequently be estimated in every region.

For each R_{gh} that requires estimation there is a corresponding sample size, n_{gh} , which is necessary to insure that some preassigned measure of estimator precision is attained once the sample has been selected. In what follows, an estimate of the sample size necessary to achieve this precision is derived. It will be seen that this derivation requires knowledge concerning:

- (1) the sampling variation of r_{gh} ,
- (2) the measure of precision that is required, and
- (3) the confidence level, $(1 - \alpha)$.

Consider only a single stratum population at risk containing $N_{gh} = N$ members, where the mean for a particular characteristic is

$$\bar{X} = \frac{1}{N} \sum_j^N X_j \quad ,$$

and the variance is

$$\begin{aligned}\sigma_X^2 &= \frac{1}{N} \sum_j^N (X_j - \bar{X})^2 \\ &= \frac{N \sum_j^N X_j^2 - \left[\sum_j^N X_j \right]^2}{N^2} .\end{aligned}$$

In a stratum population where members are now characterized by a particular attribute; $X_j = 1$ if the member possesses the attribute, and $X_j = 0$ if the member does not possess the attribute. Therefore in a population where some N_1 members possess this attribute, it is true that

$$\sum_j^N X_j = N_1$$

and

$$\sum_j^N X_j^2 = N_1 .$$

Inserting these values in \bar{X} and σ_X^2 , the mean

$$\bar{X} = N_1/N = R$$

is the true proportion of the stratum population at risk possessing the attribute, and the variance becomes

$$\begin{aligned}\sigma_X^2 &= \frac{N(N_1) - N_1^2}{N^2} \\ &= \frac{N_1}{N} \left[1 - \frac{N_1}{N} \right] \\ &= R(1 - R) .\end{aligned}$$

Since the variance for the mean of a simple random selection of n observations is (Hansen, Hurwitz, Madow, 1962)

$$\sigma_{\bar{X}}^2 = \frac{N-n}{Nn} s^2 \quad (12)$$

where

$$s^2 = \frac{1}{N-1} \sum_j^N (X_j - \bar{X})^2$$

$$= \frac{N}{N-1} \sigma_X^2, \quad (13)$$

one is able to arrive at the variance for attribute sampling. If among n observations of a random sample there are n_1 members with the attribute, the variance of the sample proportion $r = n_1/n$ is the same as equation (12), where now $\bar{x} = r$, and equation (13) becomes

$$S^2 = \frac{N}{N-1} R(1-R).$$

Hence

$$\sigma_r^2 = \frac{N-n}{N-1} \frac{R(1-R)}{n}.$$

Now the sample size, which insures the fixed precision with preassigned confidence level can be determined. Assuming normality, it is true that errors smaller than $(u_{1-\alpha/2} \sigma_r)$ occur with probability $(1-\alpha)$, where $u_{1-\alpha/2}$ is the standard normal deviate. That is,

$$\Pr(|r - R| \leq u_{1-\alpha/2} \sigma_r) \geq 1 - \alpha.$$

Consider now a maximum error F , such that

$$F \leq u_{1-\alpha/2} \sigma_r$$

or

$$\left[\frac{F}{u_{1-\alpha/2}} \right]^2 \leq \frac{N-n}{N-1} \frac{R(1-R)}{n}.$$

Solving for n one obtains, for a particular stratum,

$$n_{gh} \geq \frac{\left[\frac{u_{1-\alpha/2}}{F} \right]^2 R_{gh} (1 - R_{gh})}{1 + \frac{1}{N_{gh}} \left[\left(\frac{u_{1-\alpha/2}}{F} \right)^2 R_{gh} (1 - R_{gh}) \right]} \quad (14)$$

which is an estimate of the sample size required in stratum U_{gh} .

In practice an approximate value for R_{gh} to use in equation (14)

will suffice. If no such value is known, by letting R_{gh} assume the value 0.5, equation (14) will give the maximum sample size necessary.

Further, since \bar{Y}_g often must be estimated, one also has available $n_g = \sum_h^{L_g} n_{gh}$, as an appropriately selected sample to use in a stratified design, in order to obtain the corresponding sample estimator, \bar{y}_g given by equation (11).

It was seen in a previous section (Sample Estimators) that in one instance the sample size n_{gh} had to be obtained by initially sampling m_{gh} primary sampling units, and then selecting n_{ghi} risk units from each included primary unit so that $\sum_i^{m_{gh}} n_{ghi} = n_{gh}$. Therefore the problem of determining the correct number m_{gh} is apparent. The following example is taken from an investigation concerning an infectious disease in Puerto Rico (Whorton 1964), and it demonstrates one method of solution.

Consider only a single geographic stratum, U_h , containing $N_h = 973$ children of a certain age group. An estimate of the true prevalence rate, R_h , for some particular characteristic (presence of the infectious disease) was desired. Assuming that $R_h = 0.5$, and requiring a maximum error of $F = .05$, the necessary sample size was found to be

$$n_h = \frac{\left[\frac{1.96}{.05} \right]^2 (.25)}{1 + \frac{1}{973} [.25 \left(\frac{1.96}{.05} \right)^2 - 1]}$$

$$\cong 210 .$$

Now assume that a list of these $N_h = 973$ children is not available from which to draw the sample. The M_h primary sampling units were defined as $M_h = 44$ classrooms, each containing N_{hi} children

of that age group. It was found that there was an average of $\bar{N}_h = N_h/M_h \cong 22$ children per classroom, and it was decided to let the sample size per included classroom be

$$n_{hi} = \frac{1}{2} (\bar{N}_h) .$$

It follows that since

$$\begin{aligned} n_h &= \sum_i^{m_h} n_{hi} \\ &= \sum_i^{m_h} \frac{1}{2} (\bar{N}_h) \\ &= 11.0 (m_h) , \end{aligned}$$

and $n_h = 210$, then

$$\begin{aligned} m_h &= \frac{210}{11} \\ &= 19 \end{aligned}$$

estimates the required number of classrooms to include in the sample.

The actual sample size which resulted was $n_h = \sum_i^{19} n_{hi} = 304$ children, and the estimate of R_h was computed from equation (10).

CHAPTER VI

TEST STATISTIC RECOMMENDED FOR USE WHEN $k < K$ REGIONS ARE SAMPLED AT RANDOM

It was discussed in an earlier part of this paper (CHAPTER II) that frequently it is desirable not to include every defined region of interest in a study, but only a sample of size $k < K$ regions selected at random. In fact limitations concerning costs, time allowed for the study, etc., often make such a sample mandatory. It was also stated that once the sample is drawn, the current Ederer-Myers-Mantel $\chi^2_{(1)}$ test should no longer be used for testing the null hypothesis when inferences are made to the larger total population. What is needed, as a consequence, is the availability of a test statistic that is appropriate in this situation. To this end, a test statistic is determined which may be used to test the hypothesis of no stratum clustering when k regions are selected at random.

Consider the following model

$$X_g = E(X_g) + C + e_g \quad g = 1, 2, \dots, K \quad (15)$$

where

$$X_g \equiv X_{gh_m}^*$$

$$E(X_g) \equiv E(X_{gh_m}^*)$$

C = the population clustering effect,

and

e_g = a random error where

$$E(e_g) = 0$$

$$E(e_g e_{g'}) = 0 \quad g \neq g' \quad ,$$

and

$$\begin{aligned} \text{Var}(e_g) &= \text{Var}(X_{gh_m}^*) \\ &= \sigma_g^2 \end{aligned}$$

Now let $(X_g - E(X_g)) = D_g$. Rewriting the above model one obtains

$$D_g = C + e_g \quad . \quad (16)$$

It is to be noticed here that D_g represents the magnitude of clustering in the g^{th} region because D_g is actually an observed maximum minus the maximum that is expected if the null hypothesis ($C \leq 0$) is true. This is somewhat similar to an experiment where each of K independent units serves as its own control in order to investigate the significance of some treatment effect in the total population. In the clustering problem the "control" is never actually administered. If the values of D_g are consistently large this implies that the null hypothesis is not true and clustering actually exists. However, when D_g varies a great deal from region to region one would hesitate to conclude that clustering exists unless the values for D_g are extremely large. Therefore, both the magnitude and variance of D_g in the defined population must be considered in making a decision about clustering.

Returning to the model in equation (16), it is seen that the expected value of D_g is

$$\begin{aligned} E(D_g) &= E(C) + E(e_g) \\ &= C \quad . \end{aligned}$$

Similarly

$$E \left[\sum_g^K D_g \right] = KC$$

and

$$E \left[\frac{1}{K} \sum_g^K D_g \right] = C$$

is the population mean.

Also since

$$\begin{aligned} E(D_g - E(D_g))^2 &= E(D_g - C)^2 \\ &= E(e_g^2) \\ &= \sigma_g^2, \end{aligned}$$

it follows that

$$\begin{aligned} E\left(\frac{1}{K} \sum_g^K (D_g - C)^2\right) &= \frac{1}{K} \sum_g^K \sigma_g^2 \\ &= \sigma^2 \end{aligned}$$

is the population variance. With these considerations and assuming normality of the sum,

$$\sum_g^K D_g \sim N(KC, K\sigma^2)$$

or equivalently

$$\frac{\sum_g^K D_g - KC}{\sqrt{K\sigma^2}} \sim N(0,1)$$

Moreover

$$z = \frac{\sum_g^K D_g}{\sqrt{K\sigma^2}} \quad (17)$$

is an appropriate test statistic for testing the significance of the hypothesis, $H_0: C \leq 0$. The hypothesis is rejected if the observed z is such that

$$z \geq z_{1-\alpha}$$

where $z_{1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ percentage point of the standard normal

distribution.

It should be noted that the observed test procedure considers only a one-tail rejection region, since if the numerator of z is negative, this would imply a lack of clustering. While a rejection in the lower tail would indicate an abnormal event, it would not relate to this particular hypothesis. It should be further noted that the square of z is

$$\begin{aligned} z^2 = \chi^2_{(1)} &= \frac{\left[\begin{array}{c} K \\ \sum_g D_g \end{array} \right]^2}{K \sigma^2} \\ &= \frac{\left[\begin{array}{c} K \\ \sum_g X_{gh_m}^* - \sum_g (E(X_{gh_m}^*)) \end{array} \right]^2}{K \sum_g \text{Var}(X_{gh_m}^*)} \end{aligned}$$

and is essentially the Ederer-Myers-Mantel statistic given previously, differing mainly in the recommended rejection region of the normal test. Hence, since a $\chi^2_{(1)}$ test is in effect equivalent to a two-tailed normal test, the z test and its corresponding rejection region seems to be a more appropriate test.

What is sought now is a test statistic which is an "approximation" to the z test when regions are selected at random. It is clear from equation (17) that when $k < K$ regions are to be selected by a random process, the procedure for estimating the parameters of the model must involve the probabilities of the sampling process as well as those inherent in the model.

Consider now a sample (d_1, d_2, \dots, d_k) drawn at random from (D_1, D_2, \dots, D_K) . If f_g is the density function for e_g , then the density function g for $[(d_1-C), (d_2-C), \dots, (d_k-C)]$ is given as

follows. Let

$$A = (J = j_1, j_2, \dots, j_k \mid 1 \leq j_1 < j_2 < \dots < j_k \leq K)$$

and f_J be the marginal density given by

$$f_J(x_1, x_2, \dots, x_k) = f_{j_1}(x_1) f_{j_2}(x_2) \dots f_{j_k}(x_k).$$

$$\text{Then } g(x_1, x_2, \dots, x_k) = \frac{1}{\binom{K}{k}} \sum_{J \in A} f_J(x_1, x_2, \dots, x_k).$$

It follows that

$$\begin{aligned} E\left(\sum_i^k (d_i - c)^2\right) &= \frac{1}{\binom{K}{k}} \sum_i^k \sum_{J \in A} \sigma_{j_i}^2 \\ &= \frac{\binom{K-1}{k-1}}{\binom{K}{k}} \sum_{g=1}^K \sigma_g^2 \\ &= \frac{k}{K} \sum_g^K \sigma_g^2 \\ &= k \sigma^2. \end{aligned}$$

Moreover

$$\begin{aligned} E(\bar{d} - c)^2 &= E\left(\frac{1}{k} \sum_i^k (d_i - c)\right)^2 \\ &= \frac{1}{k^2} E\left(\sum_i^k (d_i - c)^2 + 2 \sum_{i < i'} (d_i - c)(d_{i'} - c)\right) \\ &= \frac{1}{k^2} E\left(\sum_i^k (d_i - c)^2\right) \\ &= \frac{\sigma^2}{k}. \end{aligned}$$

Now let

$$s^2 = \frac{1}{k-1} \sum_i^k (d_i - \bar{d})^2.$$

Then

$$E(s^2) = E\left(\frac{1}{k-1} \sum_i^k (d_i - \bar{d})^2\right)$$

$$\begin{aligned}
&= \frac{1}{k-1} E\left(\sum_i^k (d_i - C)^2 - k(\bar{d} - C)^2 \right) \\
&= \frac{1}{k-1} E\left(\sum_i^k (d_i - C)^2 \right) - \frac{1}{k-1} E(k(\bar{d} - C)^2) \\
&= \frac{k}{k-1} \sigma^2 - \frac{\sigma^2}{k-1} \\
&= \sigma^2
\end{aligned}$$

is the unbiased sample estimator of the population variance σ^2 .

Now the Central Limit Theorem (Cramer, 1946) would indicate that it is reasonable to assume that \bar{d} is distributed

$$N\left(C, \frac{\sigma^2}{k}\right).$$

Moreover it would be expected, as it is here assumed, that the variate

$$\sqrt{\frac{\frac{\bar{d} - C}{k}}{\frac{\sum_i (d_i - \bar{d})^2}{k(k-1)}}}$$

is distributed as the Student's "t" with $(k-1)$ degrees of freedom.

Hence, the statistic recommended for testing the hypothesis $H_0: C \leq 0$

is given by

$$t_0 = \frac{\bar{d}_0}{\sqrt{\frac{s_0^2}{k}}} \quad (18)$$

where

$$\bar{d}_0 = \frac{1}{k} \sum_i^k d_{oi}$$

and the values d_{oi} are those values which correspond to the particular

$J_0 = (j_1, j_2, \dots, j_k)$ chosen such that

$$d_{oi} = (x_{j_i h_m}^* - E(X_{j_i h_m}^*)) .$$

Thus

$$\bar{d}_o = \frac{1}{k} \sum_{g \in J_o} (x_{gh_m}^* - E(X_{gh_m}^*))$$

and

$$s_o^2 = \frac{1}{k-1} \sum_i (d_{oi} - \bar{d}_o)^2 .$$

The hypothesis above is rejected whenever

$$t_o \geq t(1-\alpha), (k-1)$$

where $t(1-\alpha), (k-1)$ is the 100 $(1-\alpha)$ th percentage point of the Student's "t" distribution with $(k-1)$ degrees of freedom.

In order to determine experimentally how well this t_o test statistic performed regarding the probability of rejecting a true hypothesis of no clustering, a population composed of $K = 29$ regions and $L_g = 5$ strata per region was constructed. Each stratum per region contained a different risk population size, Y_{gh} , and for each region the number of units, X_g , possessing a certain characteristic was randomly determined. With these values as input to the computer each stratum value for X_{gh} was generated in such a way that the null hypothesis was true, except for random variations. Each resulting value X_{gh} was transformed to X_{gh}^* according to the methodology given in CHAPTER IV for analysis. The pertinent population values that resulted are

$$K = 29$$

$$s^2 = \frac{1}{K-1} \sum_g (D_g - \bar{D})^2 = 38.8840$$

$$\sigma^2 = 37.779$$

and $\bar{D} = 1.810$

From this population 100 independent random samples of size $k = 5$ regions per sample were selected from the $K = 29$ regions. The test statistic (equation 18) was calculated for each of the 100 samples. A portion of these results are listed in Table 10.

The EMM $\chi^2_{(1)}$ test statistic was applied to the transformed sample data as if a total population had been enumerated, thus disregarding any regional sampling variation. These tests resulted in rejecting the true hypothesis a total of 16 times in the 100 samples. This estimates the Type I error as $\hat{\alpha} = .16$, and is significantly greater than the theoretical $\alpha = .05$. The t_0 test statistic using the recommended one-tailed rejection region rejected this hypothesis only 6 times for $\hat{\alpha} = .06$. If the two-tailed t_0 test had been employed, the null hypothesis would have been rejected 11 times instead of the 6 rejections for the one-tail test. For example, notice that in sample number 17 of Table 10, \bar{d} is a large negative number, which implies some abnormal homogeneity among the strata per sampled region. The $\chi^2_{(1)}$ is highly significant, whereas the one-tailed t_0 test is, of course, not significant for the pertinent hypothesis.

Subsequently 50 independent samples of size $k = 10$ were selected from this same population, and similar tests were performed. In these 50 experiments the t_0 test statistic rejected the null hypothesis 2 times, while the $\chi^2_{(1)}$ statistic rejected it 6 times.

Finally 50 samples of size $k = 15$ were performed, and the t_0 rejection level was $\hat{\alpha} = .04$ while the $\chi^2_{(1)}$ rejection level was $\hat{\alpha} = .03$.

In summary it is concluded that the t_0 test statistic with the recommended one-tail rejection region may be used for testing the

hypothesis $H_0: C \leq 0$ when regions are selected at random from some larger population of inference.

TABLE 10
TEST STATISTIC VALUES AND RESULTS FOR 20 OF THE 100
INDEPENDENT SAMPLES OF SIZE $k = 5$ REGIONS

Sample Number	\bar{d}	$\chi^2_{(1)}$ EMM	Rejections.	t_0	Rejections
1	.108	.0018		.0264	
2	4.712	6.2440	*	1.6748	
3	-3.393	3.9781	*	-3.509	
4	3.821	1.7741		1.0069	
5	-1.783	.4223		-.6595	
6	8.440	10.3432	*	2.7016	*
7	-.181	.0051		-.0573	
8	2.484	.8729		.6847	
9	-2.542	1.0055		-1.4223	
10	1.964	.5925		.7576	
11	.110	.0017		.0296	
12	2.087	.5021		.6497	
13	-.266	.0083		-.1074	
14	1.822	.4843		.4309	
15	-3.049	1.1985		-1.5039	
16	.210	.0065		.1126	
17	-6.579	6.1591	*	-2.4505	
18	1.741	.3886		.3865	
19	5.391	3.7302		1.9259	
20	3.378	3.3089		1.7032	

* Denotes a rejected hypothesis

Rejection regions:

$$\chi^2 \geq \chi^2_{(1), .95} = 3.840$$

$$t_0 \geq t_{4, .95} = 2.132$$

CHAPTER VII

SUMMARY

The purpose of this dissertation has been to develop and demonstrate the applicability of some extensions and generalizations to the current Ederer-Myers-Mantel procedure and test statistic for the particular type of clustering problem discussed in CHAPTER II. To this end:

1. a transformation to allow for different strata risk population sizes,
2. sample estimators for estimating R_{gh} , \bar{Y}_g , and therefore X_{gh}^* ,
3. methodology for determining each stratum sample size required to achieve pre-assigned estimator precision, and
4. a test statistic recommended for use when only a sample of the total number of regions is included in a study

have been developed. These developments were necessary to overcome frequently encountered problems which tends to invalidate the EMM procedure. Each new concept has been discussed and experimentally evaluated where possible through the use of an electronic computer.

Through the use of the transformation on X_{gh} , each stratum per region no longer must contain equal risk population sizes. This particular transformation made it possible to estimate precisely each required transformed value without the necessity of pre-determining several population values in advance. These methods of parameter

estimation along with techniques to determine corresponding sample sizes are included.

In order to overcome the necessity of including every one of the K defined regions of interest in a particular study, a test statistic has been derived to allow for a random selection of these regions and still test the hypothesis of no clustering in the total population. The test statistic is investigated experimentally and the results support its appropriateness.

The allowable number of strata per region has been extended. Methods for calculating the expectation and variance of X_{gh_m} when $L_g = 2, 3, \dots, 10$ and $X_g \geq 100$ cases are given in the appendix.

The writer believes that each of the extensions made is based on the needs of researchers in a variety of important problems in biology, health, and medical areas where total sampling is not feasible and equal population at risk in strata are not available. To compensate for these problems these extensions are recommended as the need arises. Certainly each development is not purely a mathematical or statistical concept, but just as certainly it is a methodology for solving a particular type of problem. Hopefully these extensions and modifications will help to overcome the frequent necessity for compromising an investigator's objective in order to fit an existing analysis procedure. These extensions have been shown to be reasonably flexible and empirically successful.

LIST OF REFERENCES

- Bonner, R. E. On Some Clustering Techniques. IBM J. of Res. and Dev., Jan., 1964.
- Cochran, W. G. Sampling Techniques. John Wiley and Sons, Inc., New York, N. Y., 1963.
- Cramer, Harold Mathematical Methods of Statistics. Princeton University Press, Princeton, N. J., 1946.
- Ederer, F., Myers, M. H., Mantel, N. A Statistical Problem in Time and Space: Do Leukemia Cases Come in Clusters? Biometrics, Vol. 20, No. 3, Sept., 1964.
- Feller, W. An Introduction to Probability Theory and Its Applications. Wiley and Sons, Inc., 1957.
- Gower, J. C. A Comparison of Some Methods of Cluster Analysis. Biometrics, Vol. 23, No. 4, Dec., 1967.
- Hansen, M. H., Hurwitz, W. N., Madow, W. G. Sample Survey Methods and Theory, Vols. I and II, Wiley and Sons, Inc., New York, N. Y., 1962.
- Stark, C. R., Mantel, N. Lack of Seasonal-or Temporal-Spatial Clustering of Down's Syndrome Births in Michigan, Amer. J. of Epid., Vol. 86, No. 1, Aug., 1967.
- Sukhatme, P. W. Sampling Theory of Surveys with Applications, Iowa State Univ. Press, Ames, Iowa, 1960.
- Whorton, E. B. An Investigation of a Survey for Estimating the Prevalence Rates of an Infectious Disease, Unpublished Masters Thesis, Tulane University, New Orleans, Louisiana.

APPENDIX

EXPECTATIONS AND VARIANCES OF $X_{gh_m}^*$ WHEN
 $L_g = 2, 3, \dots, 10$ AND $X_g^* \geq 100$

L_g	$E(X_{gh_m}^*)$	$Var(X_{gh_m}^*)$
2	$X_g^*/2 + .39889 \sqrt{X_g^*}$	$.09084 X_g^*$
3	$X_g^*/3 + .48660 \sqrt{X_g^*}$	$7.892 + .07538 (X_g^* - 100)$
4	$X_g^*/4 + .51470 \sqrt{X_g^*}$	$6.595 + .06043 (X_g^* - 100)$
5	$X_g^*/5 + .5201 \sqrt{X_g^*}$	$5.604 + .04951 (X_g^* - 100)$
6	$X_g^*/6 + .5173 \sqrt{X_g^*}$	$.03665 X_g^*$
7	$X_g^*/7 + .51107 \sqrt{X_g^*}$	$.03558 X_g^*$
8	$X_g^*/8 + .5033 \sqrt{X_g^*}$	$.02808 X_g^*$
9	$X_g^*/9 + .4950 \sqrt{X_g^*}$	$.02735 X_g^*$
10	$X_g^*/10 + .4866 \sqrt{X_g^*}$	$.02427 X_g^*$

For values of $E(X_{gh_m}^*)$ and $Var(X_{gh_m}^*)$ when $L_g = 3, 4, 5$ and $X_g^* = 2, 3, \dots, 100$ see Ederer, Myers, and Mantel 1964 and Stark and Mantel 1967.