SEQUENTIAL MULTIVARIATE QUALITY CONTROL TESTS
USING TOLERANCE REGIONS

by

Ronald L. Boase and John E. Walsh

Technical Report No. 110
Department of Statistics ONR Contract

April 1, 197

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

DEPARTMENT OF STATISTICS
Southern Methodist University

TABLE OF CONTENTS

only this class the subtests are independent, thus the significance level of each subtest is not affected by the outcomes of preceding subtests.

Some of the desirable properties of the sequential significance test for quality control uses are: (1) the data considered may be multivariate, (2) the test permits legitimate reuse of previous data, (3) the permissible subtests are independent providing accurate evaluation of significance levels, (4) the random selection of previous data at each subtest level can be effectively used to emphasize the more recent data, and (5) the permutation-randomization approach yields subtests that are generally applicable.

In Chapter III the new proposed method of forming multivariate two-sample tolerance tests is introduced. This method yields tests that have a permutation basis and satisfy the requirements for subtests in the sequential significance tests proposed in Chapter II. A well known existing method is shown to be a special case of the proposed method. The data required by the proposed method need not be independent random samples but must have a symmetric joint null distribution. Also, any univariate two-sample tolerance test can be considered as a multivariate two-sample tolerance test. These include all run and rank tests.

The construction process of forming tolerance regions for the proposed method is outlined in Chapter IV. This process is a systematically staged procedure for establishing a set of tolerance regions for the two-sample problem. Certain symmetric information is available for use at each stage. This information basically includes knowledge of the combined observed vector values as long as they are not identified with the population from which they were taken. This information accumulates as the process continues providing an excellent source for determining the shapes of the desired tolerance regions. Thus, the desired tolerance regions can be constructed with the goal of making the selected test statistic as significant as possible.

Some suggested techniques for applying the proposed method of forming multivariate two-sample tolerance tests are presented in Chapter V. Also included, are certain practical considerations for using the new construction process effectively and for selecting appropriate univariate tests. A general outline of a suggested operational procedure is included.

Certain areas of possible application of the sequential significance test using subtests formed by the proposed method are considered in Chapter VI. Most of the discussion is devoted to applications in the medical field.

The last chapter, Chapter VII, contains statements of the basic theory verifying the results claimed for the proposed method of forming multivariate two-sample tolerance tests. The important results are stated in the form of a theorem and corollary. Because of its unusual length, the proof of the theorem has been relegated to the appendices. All other results claimed are verified in the discussion.

CHAPTER II

QUALITY CONTROL TESTS

Sequential randomization tests for univariate one-way analysis of variance have been developed by Walsh [15] and [16]. Presented here, is an analogous extension of Walsh's tests using multivariate data. Such tests possess desirable properties making them directly applicable to quality control uses.

A class of sequential significance tests which consist of a pre-specified number of subtests is developed. Each subtest in the sequence reuses all or part of the total data used in the preceding subtest in a manner which establishes independence among all subtests.

The data are taken in sets representing independent (finite) random samples[1] of multivariate observations which, under the null hypothesis, are from the same unknown, but partially-continuous[2] distribution. Each subtest is a two-sample test using as one population sample (previous data) a set of data vectors randomly chosen[3] from the totality of observation data vectors used in the preceding subtest and as its second

---

[1]It suffices to require the combined observations to have a symmetric joint null distribution.

[2]A random vector is defined to have a partially-continuous distribution if at least one component of the random vector has a continuous marginal distribution.

[3]The previous data for any two-sample subtest may then include the totality of observation data used in the previous subtest.

population sample (new data) one of the remaining unused data sets. The subtests are performed sequentially until either significance is obtained at a subtest level or a specified maximum number of subtests have been made. Significance for the overall test is obtained only when a subtest in the sequence proves significant. Thus, the overall test will not be significant if, and only if, all subtests in the sequence are not significant. Exact significance levels can be obtained by using appropriate randomization-permutation probability models and subtest statistics possessing a special property which insure independence between subtests.

Perhaps the most desirable feature of these tests is their ability to legally use (in a probabilistic sense) data of preceding subtests. The outcomes of the preceding subtests in many similar sequential tests produces a conditional effect on the significance level of succeeding subtests; however, in the tests studied here, the subtests are independent and no such conditional effects exist among them. Therefore, if $\alpha_1, \alpha_2, \ldots, \alpha_k$ denote the significance levels of the k subtests composing the overall test, the significance level, α, of the overall test can be computed directly:

$$\alpha = 1 - \prod_{i=1}^{k} (1 - \alpha_i).$$

Another feature, which is highly desirable in applied sequential quality control tests, is the ability to maintain a limited control of the emphasis placed on preceding data sets at each subtest stage. This is accomplished by properly selecting the sizes of each new data set used in the sequence and the size of the previous data set obtained by randomization at each subtest stage. Repeated randomization of the preceding data will emphasize the most recent data sets, while no randomization, that is, using all the preceding data, will tend to deemphasize the most recent data sets.

A third feature is that the randomization-permutation model yields subtests of general applicability which may be one or two-sided tests and can be oriented toward many forms of the alternative hypothesis.

The randomization contribution to the model insures, under the null hypothesis, that the observations selected as previous data at each subtest stage constitute a random sample from the population representing the totality of data used in the previous subtest. If no significance is obtained at the subtest stage, the previous and new data sets used are combined and represent a random sample from the population yielding the combined data. This combined data set becomes the data available for randomization (if any) defining the previous data set for the next subsequent subtest. The process is continued until either significance is obtained at some subtest level or a specified number of subtests have been made.

The permutation model is used to establish the conditional probability spaces on which the distribution of each subtest statistic is determined. If the observed vectors are ordered in some definite but arbitrary manner (e.g. the order in which they were obtained) then the sample space, induced by the permutation model, constitutes the set of all permutations of the observed vectors. Under the null hypothesis, the probability of any permutations is the same. The permutation sample space associated with the two-sample problem can be reduced by considering the set of all possible assignments of the ordered observed vectors into two sets; one of size equal to the new data set and the other whose size corresponds to the previous data set. Then, under the null hypothesis, all possible assignments made in this manner are equally-likely.

Now, any function, symmetric in the totality of observed vectors, is clearly a constant on all points of the associated permutation sample space. That is, under any hypothesis, this function is a constant with probability one. Therefore, the function, a statistic, is independent of any other statistic defined on the same permutation sample space. This fact motivates the method used for selecting appropriate subtest statistics having the property that the subtest significance level is not conditionally affected by the outcomes of previous subtests.

For each two-sample subtest consider the class of statistics which are symmetrical in the totality of observed vectors defined by the previous data set. Then, for any fixed set of observed vectors constituting a possible ordering of the new data set, the value of such a statistic remains unchanged over all permutations of the observed vectors in the previous data set. Also, for any fixed ordered set of observed vectors in the new data set, this statistic is defined on the permutation sample space obtained from the previous data set and on any permutation sample space constructed on a set of observed vectors contained in or containing the previous data set.

In order to verify that this class of statistics possess the property that in the sequential process each subtest statistic is independent of the results of the preceding subtests, two mutually exclusive cases are are considered.

First consider a sequence of subtests where at each subtest stage the previous data was taken to be the totality of data used in the preceding subtest. That is, none of the previous data sets defined at each subtest stage were obtained by randomization. The totality of data used for each subtest is then a proper subset of the totality of data used in the next subsequent subtest. Thus, any permutation of the observed vectors that could occur for any preceding subtest corresponds to a subtest of the permutations of the observed vectors in the previous data set. For any fixed order of observed vectors in the new data set the subtest statistic, by choice, is a constant over all permutations, thus all subsets of permutations, of the observed vectors in the previous data set. Therefore, the subtest statistic is a constant, with probability one, on each permutation sample space of the preceding subtests, and is independent of the permutation observed for each preceding subtest. Since the outcome of a subtest is determined by the actual permutation of vectors observed, then the subtest statistic is independent of the outcome of all preceding subtests.

Now, consider a subtest whose previous data set was obtained by randomization - randomly selecting a subset from the totality of data used in the preceding subtest. Since the new data sets obtained after randomization are independent of the data sets used prior to the randomization, it is only necessary to show that the previous data set is independent of all outcomes of the preceding subtests. To verify this, it suffices to consider only those preceding subtests occurring after the most recent previous subtest using randomization, for induction can be used to justify the remainder of the assertion. Verification then follows, since the totality of observation vectors in the preceding subtest is not affected by any permutations that could occur or any subsets of them. Thus, the subtest statistic is independent of the outcomes of these preceding subtests.

If randomization is used in the next following subtest to obtain its previous data set the above verification holds. However, if the next subsequent subtest does not use randomization, the situation is essentially the same as that cited in the first case. The proof then follows by induction.

The sequential significance test description given by Walsh [16] for the univariate case is analogous for the multivariate case. However, in this paper, the number of observed vectors in each new data set is permitted to be <u>one</u> or more provided the first data set used is of sufficient size to insure that the desired significance levels of all subtests and the overall test can be obtained (or approximately obtained). In like manner, for exact and approximate permutation tests, the determination of sharp lower bounds for the subtest significance levels, and the considerations on the sample sizes used in each subtest discussed by Walsh also hold for the multivariate use as well. This material has been thoroughly and clearly presented in the above reference.

To establish a sequential significance test, having all the properties outlined above, would first require finding appropriate two-sample multivariate subtests. These subtests not only should be selected to emphasize the alternative hypotheses, but be feasible in application. That is, under the permutation model, the null distribution of the subtest statistic should either be well approximated by some known distribution or easily determined. The two-sample tests considered in this paper are based on tolerance regions. It will be shown that a two-sample tolerance test has a permutation basis and the associated test statistic is symmetric in the observations in both samples separately. Thus all two-sample tolerance tests are permissible as subtests.

CHAPTER III

MULTIVARIATE TWO-SAMPLE TOLERANCE TESTS

A new method of forming two-sample tolerance tests is proposed. As an introduction to this method an existing method is presented first. The existing method is a special case of the new proposed method. The basic difference between them is the manner in which tolerance regions are formed. The new method is shown to overcome several of the major disadvantages common to the existing method. Both methods yield tests that are suitable as subtests in the quality control tests presented in the preceding chapter.

The same basic philosophy is used in both methods to form two-sample tolerance tests. A test is fundamentally established on a set of tolerance regions. The number of observations from one population falling within each tolerance region is counted. These frequency counts are then used to determine the outcome of the test.

The existing method ([19],[1], and [3]) requires that the data be two independent random samples. One of these samples is used to construct a set of disjoint nonparametric tolerance regions. The other sample is reserved for determining the region frequency counts. The construction process used to establish the tolerance regions ([17],[13],[12], and [9]) is a systematically staged procedure (referred to here as the "standard one-sample process").

In the first stage the sample space is partitioned into two disjoint tolerance regions, called blocks. This is accomplished by choosing some real-valued function whose null distribution is continuous and selecting some order statistic on the set of function values defined on the sample. These choices can be based on any independent information that is available prior to taking the observations. The value of this order statistic and the observed vector yielding it is all the additional information permitted ([6],[7],[8], and [9]). The function equated to this given value of the chosen order statistic defines a cut on the sample space producing two blocks. This partitioning also separates the sample into two conditionally independent subsamples [3]. Using this limited information, one of the two blocks formed is considered for division in the second stage. The above process is repeated only for the subsample of observations associated with the chosen block. However, the function considered in the second stage may differ from the function selected at the first stage level but its null distribution must also be continuous. The chosen block is then partitioned into two new blocks. Thus, by the end of the second stage three disjoint blocks have been formed. Again the only new permissible knowledge are the values of the order statistics and the observed vectors yielding them. The process is continued until the desired number of tolerance regions or blocks has been obtained. The content of the tolerance regions can be determined by the number of unused observations lying within them. If the process were continued until all observations were used to define cuts (each time producing two new blocks) the resulting blocks are called basic blocks or statistically equivalent blocks. Thus, the blocks obtained earlier in the process would consist of a fixed number of basic blocks. A more common measure of tolerance region content is the number of basic blocks contained within the tolerance region.

Finally, after the desired number and basic block contents of the tolerance regions have been established on the first sample, the corresponding set of frequency counts on the observations in the second independent sample are made yielding the outcome of the test. The null distribution of the test statistic is determined from the joint null distribution of the block frequency counts. This latter distribution has been established ([18] and [19]).

All tests based on this method have a permutation basis. This fact is shown by the corollary given in Chapter VII. The corresponding test statistics are symmetric in the totality of sample observations on which the blocks were defined; also, they are symmetric in the totality of sample observations used to determine the block frequency counts. This means that any two-sample test obtained by this method can be used as a subtest in the previous chapter. Furthermore, either the "previous" or "new" data set can be used to define the tolerance regions.

Unfortunately, this existing method has a major disadvantage; namely, the limited freedom in selecting the shapes of the tolerance regions so as to emphasize the alternative hypotheses of interest. At each stage in the above process only a small amount of knowledge was available on which to choose the shapes of the cutting functions. No knowledge or consideration of the second sample data was used. By combining both sets of observations and considering only certain permissible information (mainly symmetric) a vast amount of knowledge can be obtained on which to base the shapes of tolerance regions. This concept is explored by the next method.

The proposed new method of forming two-sample tolerance tests requires only that the two sets of data have a symmetrical joint null distribution. Both sets of observations are used to construct a set of disjoint tolerance regions on one of the two sets of observations. The process of constructing the tolerance region is similar to the standard one-sample process but allows knowledge of all observed vectors excluding their set association (i.e. knowledge of the observations within each of the two original data sets is forbidden). This construction procedure will be referred to as the "generalized block construction process for the two-sample problem" or simply as the B^* process.

First, one of the two data sets is designated as the set on which the tolerance regions (or blocks) will be formed [This set will be referred to as the "designated set"]. Prior to the formation of any blocks on the sample space a certain amount of information is available. This includes all information which is symmetric with respect to the totality of random vectors yielding the combined set of observations. Thus, knowledge of the observed vector values is permitted as long as all vectors are not associated (or identified) with either of the two data sets. This means that the totality of the unidentified observed vectors can be "looked-at" and treated numerically and/or graphically in any manner. Any function defined on any subset of the combined set of unidentified observations would be symmetric information permitted. All of this information plus any independent information available prior to taking the observations can then be used to select the first stage cutting function. This function must be real-valued and must either have a continuous null distribution or be selected in such a way as to guarantee that it will never pass through more than one observed vector for all the values in its range. The location of the cut is then determined by selecting some order statistic on the set of function values defined on the designated set. Only the value of this order statistic and the observed vector yielding it can be identified. Thus, one of the observed vectors, the one associated with the cut, is identified and all other observations remain unidentified. The additional information that now becomes available includes the two sets of unidentified observations falling within the two new blocks formed and all information which is symmetric with respect to both sets of random vectors yielding these two sets of unidentified observations. Then a block is chosen for the next stage division (it must contain at least one observation from the designated set). A decision must be made at this time to either reserve the remaining block for possible future division at some later stage in the $B*$ process or to never divide it at any stage. If the latter decision were made, all observations lying within the remaining block can be identified with the original data sets. All this information is then used to select the next cutting function. This process is continued until the desired tolerance regions have been formed. The joint null distribution of the block frequency counts at any stage is the same as the corresponding distribution determined by the standard one-sample process (see Chapter

VII). Thus, if the process were continued to the last possible stage the resulting blocks are statistically equivalent blocks in the sense that the joint null distribution of the basic block frequency counts is the same as that determined by the standard process[1].

In Chapter VII it is shown that any test produced by this new method also has a permutation basis and its associated statistic is symmetric in the set of observations on which the tolerance regions were defined and on the set of observations used to establish the block frequency counts. Then all two-sample tolerance tests defined by this new method are usable as subtests in Chapter II. Also, either the "previous" or "new" data sets can be used to define the block frequency counts.

One primary advantage the B^* process has over the standard one-sample process is the vast amount of information made available for forming tolerance regions. This advantage allows one to select desirable shapes of the tolerance regions so as to make the test statistic, previously chosen, as significant as possible. This is equivalent to emphasizing the alternative hypotheses for which the test accentuates.

A second important advantage of the B^* process is that it is not necessary to proceed through all stages once sufficient information to

---

[1]The term "statistically equivalent blocks," originally defined by Tukey [12], actually referred to a set of tolerance regions obtained at the completion of the process described by the first method. This concept basically defined statistically equivalent blocks as a set of $N + 1$ tolerance regions, formed by the first method, whose joint coverages represent the barycentric coordinates of a random point uniformly distributed in an N-dimensional simplex (where, N denotes the sample size). In this paper, statistically equivalent blocks will be defined as a set of tolerance regions on which the joint null distribution of the frequency counts determined by a second sample is the same as if it were determined by a set of statistically equivalent blocks defined by Tukey.

evaluate the outcome of the test is made available. The process B^* allows some observations to be completely identified at various stages. For example, if a certain block either has been designated as a block "never-to-be-divided" or is a basic block, then all observations lying within are completely identified with their original data sets. Thus, sufficient information about the block frequency counts may possibly be determined early in the process so as to conclude the outcome of the test.

Another advantage is the data need not be independent random samples. The corresponding requirement is that the joint null distribution of the combined data sets must be a symmetric function.

It is also shown in Chapter IV that the standard one-sample process is a special case of the B^* process. This implies that the standard process can be extended by requiring only that the combined data have a symmetric joint null distribution function.

The type of tests that may be formed by either method is considered next. Anderson [1] shows that any two-sample tolerance test adapted for the univariate case is also directly applicable for the multivariate case using the existing method. Tests based on multivariate data are obtained by relating each multivariate tolerance region to a univariate tolerance region containing the same number of basic blocks. Since the joint null distribution of the block frequency counts is unchanged whether the multivariate or univariate case is considered, then all univariate two-sample tolerance tests are usable. However, the joint null distribution of the block frequency counts for the new method is the same as that for the existing method; thus, univariate two-sample tolerance tests apply equally well to the new method. The types of nonparametric tests available for consideration include all run and rank tests (see pages 34 through 80 in reference [14]). Some of these tests are described later in Chapter VI.

A detailed description of the proposed B^* process is given in the next chapter. This is followed by two chapters devoted to suggested techniques and areas of practical application. The final chapter establishes verification of all results claimed.

CHAPTER IV

A GENERALIZED BLOCK CONSTRUCTION PROCESS, B^*, FOR THE TWO-SAMPLE PROBLEM

The process presented is a systematically staged method of constructing a set of distribution-free tolerance regions (blocks) for the two-sample problem. The data consist of two sets of at least partially continuous multivariate observations which are defined on the same sample space. Under the null hypothesis, the combined data must have a symmetric joint cumulative distribution function. The sets and the observations within each set are not required to be independent. At each stage in the process specific information on certain subsets of the combined set of observations is used to establish two new blocks (regions) within one of the blocks previously formed. This information collectively increases as the process continues, providing a better basis on which the shapes and locations of future blocks may be controlled. This freedom greatly increases the ability to construct blocks so as to emphasize the alternative hypotheses - thereby potentially increasing the power of any two-sample tolerance test.

The first two stages and a general stage are discussed in detail stressing the amount of permissible information available. Some suggested techniques for exploiting the information obtained are presented in Chapter V.

Let $O_n = \{X_1, X_2, \ldots, X_n\}$ and $O_m = \{Y_1, Y_2, \ldots, Y_m\}$ be two sets of p-component random vectors defined on a sample space X. These random vectors must be at least partially continuous and have a symmetric joint null cumulative distribution function. The process B^* will be demonstrated by forming a set of blocks (tolerance regions) on the observations on O_n.

For convenience, a few basic terms and symbols will be defined and used through the remaining text.

Definition 1

An observation is said to be identified if it can be associated with the set of random vectors which yielded it; that is, associated with either O_n or O_m. Thus, a set of observations is identified if each observation within the set is identified.

Definition 2

Information on a set of observations is said to be symmetric with respect to a block (or union of blocks) if the information is unaffected by interchanging the roles (relabeling the identities) of the random vectors yielding the observations falling within the block(s). For example, if $V_1, V_2, \ldots, V_k$ denote the set of random vectors yielding observations which fell within block B, then the information, I, defined on some set of observations which may or may not contain those in B, is symmetric with respect to B if for any reordering $(V_1', V_2', \ldots, V_k')$ of $(V_1, V_2, \ldots, V_k)$ I is unchanged.

Definition 3

The symbols $\bar{o}_n$, $\bar{o}_m$ and $\bar{o}_{n+m}$ will denote the sets of observations on

O_n, O_m, and O_{n+m}, respectively, where O_{n+m} is the combined set of random vectors $\{X_1, X_2, \ldots, X_n, Y_1, \ldots, Y_m\}$. Likewise, O_j^k will denote the set of random vectors in O_{n+m} yielding the set of observations $\bar{o}_j^k$ in O_{n+m} which fall within block B_j^k at the k^{th} stage in B^*, for $j = 1,2,\ldots,k+1$.

Definition 4

The symbols $Z_1, Z_2, \ldots, Z_{n+m}$ are a relabeling of the random vectors in O_{n+m}. Also, the symbols $Z_1^j, Z_2^j, \ldots, Z_{k(j)}^j$ will represent the random vectors in O_j^k and $X_1^j, X_2^j, \ldots, X_{i(j)}^j$ and $Y_1^j, Y_2^j, \ldots, Y_{t(j)}^j$ are those random vectors in $O_j^k \cap O_n$ and $O_j^k \cap O_m$, respectively, defined at the k^{th} stage. All lower case letters x, y, and z will denote corresponding observed vectors.

Stage 1

At the beginning of the first stage all information on $\bar{o}_{n+m}$ which is symmetric with respect to X (considering the space, X, as a non-random block) is permitted as well as all prior information (i.e. information available prior to taking the observations: previous observations, etc.). The totality of this information is denoted by I_1.

If at least one observation in $\bar{o}_{n+m}$ is identified, then any information on $\bar{o}_{n+m}$ is not symmetric with respect to X. For example, suppose $z^* \varepsilon \bar{o}_{n+m}$ is identified with O_n. By interchanging the roles of the random vectors in O_n with those in O_m the information that z^* is identified with O_n is no longer true. Also, any information on the set $\bar{o}_{n+m}-\{z^*\}$ is not symmetric with respect to X for the same reason. Trivially then, since $\bar{o}_n$ and $\bar{o}_m$ are completely identified, their knowledge is forbidden.

Suppose $\bar{o}_{n+m}$ is unidentified, then the knowledge of the observed vectors is $\bar{o}_{n+m}$ is symmetric with respect to X, since any interchanging of the roles of the random vectors in O_{n+m} does not in any way affect the values of the observed vectors in $\bar{o}_{n+m}$, and $\bar{o}_{n+m}$ still remains unidentified. Similarly, any subset of $\bar{o}_{n+m}$ is unidentified and is symmetric with respect to X. Furthermore, any function on any subset of observations in $\bar{o}_{n+m}$ would be symmetric with respect to X. Since the sample sizes m and n would be considered as prior information, it is permissible to select two subsets in $\bar{o}_{n+m}$ of sizes n and m to be likely candidates for the sets $\bar{o}_n$ and $\bar{o}_m$.

After I_1 has been established, the next step is to select some integer $i_1 \varepsilon \{1,2,\ldots,n\}$ and some real-valued measurable function $\phi_1(z,I_1)$ which either has a continuous null distribution or is chosen such that no ties exist in the set $\{\phi_1(z,I_1)|z \varepsilon \bar{o}_{n+m}\}$, if possible.

Through some independent source[1], having full knowledge of $\bar{o}_n$, the i_1^{st} largest value in the set $\{\phi_1(z,I_1)|z \varepsilon \bar{o}_n\}$ is determined. This value and the observed vector in $\bar{o}_n$, say x_1^*, yielding it are available information. Suppose $\phi_1(x_1^*,I_1) = c_1$, then the function $\phi_1(x,I_1) = c_1$ is used to divide the sample space X into two open regions:

$$B_1^1 = \{x \varepsilon X|\phi_1(x,I_1) < c_1\}$$

and

$$B_2^1 = \{x \varepsilon X|\phi_1(x,I_1) > c_1\} .$$

The new information, that is permitted after the blocks B_1^1 and B_2^1 have been formed, consists of c_1, x_1^* and the two subsets of observations $\bar{o}_1^1$ and $\bar{o}_2^1$ contained in $\bar{o}_{n+m}$. Additional information permitted at the end

---

[1]In this case, an independent source could be an assistant or a digital computer.

of stage 1 also includes all information on $\bar{o}_1^1$, $\bar{o}_2^1$, and on $\bar{o}_{n+m}$ (including x_1^*) which is symmetric with respect to both B_1^1 and B_2^1. That is, any information which remains unchanged when the roles of the random vectors in O_1^1 are interchanged and also is unchanged whenever the roles of the random vectors in O_2^1 are interchanged. This type of information will be defined as information which is "symmetric separately" with respect to a set of blocks. Now, if $i_1 = 1$, the block B_1^1 cannot contain any observation in $\bar{o}_n$, hence $\bar{o}_1^1$ is completely identified (although $\bar{o}_2^1$ is unidentified if $n > 1$). Similarly, if $i_1 = n$ then $\bar{o}_2^1$ is completely identified and $\bar{o}_1^1$ is unidentified for $n > 1$. Whenever $i_1 \neq 1, n$ $(n > 2)$, then both $\bar{o}_1^1$ and $\bar{o}_2^1$ are unidentified. Information of this type may be used advantageously in choosing the shape and possibly the location of the next new blocks (see Chapter V). By definition of B_1^1, B_2^1 and c_1, it is easy to deduce that there are exactly i_1-1 and $n-i_1$ observations in $\bar{o}_n$ lying within B_1^1 and B_2^1, respectively. Since $\bar{o}_1^1$ and $\bar{o}_2^1$ are permitted then the exact number of observations in $\bar{o}_m$ falling within B_1^1 and B_2^1 can be determined.

Stage 2

After considering all permissible information available at the end of stage 1, the next step is to select either block B_1^1 or block B_2^1 for division at the second stage level. The particular block chosen must contain at least one observation in $\bar{o}_n$. In addition, the other block not chosen for stage 2 division must be considered. One of two actions are required: (1) the block is reserved for potential division at some later stage (it must contain at least one observation in $\bar{o}_n$) or (2) the block will never be divided in the process. The first action does not necessarily imply that the block will eventually be divided but that it may be considered for division at some future stage. If the second action were taken, it is permissible to identify all observations falling within; although, some of these observations may be in $\bar{o}_n$. This action could provide considerable information especially if there were only a few observations in $\bar{o}_n$ and many observations in $\bar{o}_m$ lying within. Then a better choice of candidate sets could be made.

The next step is to determine the information I_2 that can be used to select the next cutting function, $\phi_2(z,I_2)$. The type of information permitted in I_2 depends upon the action taken above. If both blocks were considered for division, one at the second stage level and the other at some future stage, then I_2 is defined to consist of all previous information and all information on $\bar{o}_1^1$, $\bar{o}_2^1$ and $\bar{o}_{n+m}$ which is symmetric separately with respect to both B_1^1 and B_2^1. That is, the information must be symmetric with respect to B_1^1 and also symmetric with respect to B_2^1. Consider the unidentified sets $\bar{o}_1^1$ and $\bar{o}_2^1$. Since all observations in $\bar{o}_1^1$ and all observations in $\bar{o}_2^1$ are unidentified, then any information on $\bar{o}_1^1$ is symmetric with respect to B_1^1 and is trivially symmetric with respect to B_2^1, similarly any information on $\bar{o}_2^1$ is symmetric separately with respect to B_1^1 and B_2^1. Then any information on any subsets of $\bar{o}_1^1$ and $\bar{o}_2^1$ is symmetric separately with respect to B_1^1 and B_2^1. Trivially x_1^* is symmetric with respect to both blocks. Therefore, any information on any subset of observations in $\bar{o}_{n+m}$ is permitted.

If the second action above were selected, then I_2 is defined to contain all previous information and all information $\bar{o}_1^1$, $\bar{o}_2^1$ and $\bar{o}_{n+m}$

which is symmetric with respect to only the block chosen for division. For discussion, suppose block B_1^1 were chosen for division and it was decided that block B_2^1 would never be divided. Then the set $\bar{o}_2^1$ can be completely identified and any information on $\bar{o}_2^1$ or any subset of $\bar{o}_2^1$ is clearly symmetric with respect to B_1^1. Also, in line with the above discussion, any information on the unidentified set $\bar{o}_1^1$ or on any subset of $\bar{o}_1^1$ or on x_1^* is symmetric with respect to B_1^1. It follows then, any information on any subset in $\bar{o}_{n+m}$ given that $\bar{o}_1^1$ is unidentified is symmetric with respect to B_1^1.

In summary, I_2 contains all previous information and all information on any subset of observations in $\bar{o}_{n+m}$ which is symmetric when considering each block chosen for future division.

Again, suppose block B_1^1 were chosen for division at the second stage level. Then, using I_2, select an integer $i_2 \in \{1,2,\ldots,i_1-1\}$ and some real-valued measurable function $\phi_2(z,I_2)$ such that either it has a continuous null distribution or that there are no ties within the set $\{\phi_2(z,I_2) \mid z \in \bar{o}_1^1\}$. Through some independent source having full knowledge of $\bar{o}_1^1 \cap \bar{o}_n$, the i_2^{nd} largest value, say $\phi_2(x_2^*,I_2) = c_2$, in the set $\{\phi_2(x,I_2) \mid x \in \bar{o}_1^1 \cap \bar{o}_n\}$ is determined. The vector x_2^* and value c_2 constitute new permissible information. The cutting function $\phi_2(x,I_2) = c_2$ is used to obtain two open regions in B_1^1:

$$B_1^2 = \{x \in B_1^1 \mid \phi_2(x,I_2) < c_2\}$$

$$= \{x \in X \mid \phi_2(x,I_2) < c_2,\ \phi_1(x,I_1) < c_1\}$$

and

$$B_2^2 = \{x \in B_1^1 \mid \phi_2(x,I_2) > c_2\}$$

$$= \{x \in X \mid \phi_2(x,I_2) > c_2,\ \phi_1(x,I_1) < c_1\}$$

To standardize the notation let $B_3^2 = B_2^1$. Then the three blocks defined at the second stage are B_1^2, B_2^2, and B_3^2.

If B_2^1 were selected for division at stage 2, then i_2 would be selected from the integers $\{1,2,\ldots,n-i_1\}$ and an appropriate $\phi_2(z,I_2)$ function would be chosen. The vector x_2^* and the value c_2 would be such that $\phi_2(x_2^*,I_2) = c_2$ is the i_2^{nd} largest value in the set $\{\phi_2(x,I_2) | x \varepsilon \bar{o}_2^1 \cap \bar{o}_n\}$. The blocks formed would be

$$B_2^2 = \{x \varepsilon B_2^1 | \phi_2(x,I_2) < c_2\} ,$$

$$B_3^2 = \{x \varepsilon B_2^1 | \phi_2(x,I_2) > c_2\} ,$$

and

$$B_1^2 = B_1^1 .$$

The information that is available for entering the third stage (if desired) constitutes I_2, x_2^*, c_2, the observation sets $\bar{o}_1^2$, $\bar{o}_2^2$, and $\bar{o}_3^2$, and all information on $\bar{o}_{n+m}$ which is symmetric separately with respect to all the blocks whose corresponding observation sets are not identified. By the above definition, I_2 contains I_1, x_1^*, c_1, and all other information available at the beginning of the second stage.

<u>Stage r ($r \leq n$)</u>

At the beginning of the r^{th} stage the totality of permissible information consists of I_{r-1}, x_{r-1}^*, c_{r-1}, the blocks B_1^{r-1}, $B_2^{r-1}, \ldots, B_r^{r-1}$, and the corresponding observation sets $\bar{o}_1^{r-1}, \bar{o}_2^{r-1}, \ldots, \bar{o}_r^{r-1}$ some of which may be completely identified.

Within the set of blocks, that have not previously been designated as "blocks never to be divided", one block is selected for division at the r^{th} stage. Again, this block must contain at least one unidentified observation in $\bar{o}_n$. Each of the remaining blocks must be classified either as a block considered for future division or as a block never to be divided. None of the blocks $B_1^{r-1}, B_2^{r-1}, \ldots, B_r^{r-1}$ which has been categorized as a block never to be divided can at any stage be reclassified as a block considered for future division.

The information I_r is defined to consist of all previous information and all information on $\bar{o}_{n+m}$ or on any subset in $\bar{o}_{n+m}$ which is symmetric separately with respect to all blocks which could be chosen for division either at the r^{th} stage or at some future stage.

If block B_j^{r-1} (for some $j = 1,2,\ldots,r$) were chosen for division at the r^{th} stage and the number of observations in $\bar{o}_n$ lying within B_j^{r-1} is e_j, then, using I_r, select an integer i_r in $\{1,2,\ldots,e_j\}$ and a real-valued measurable function $\phi_r(z,I_r)$ such that either it has a continuous null distribution or there are no ties within the set $\{\phi_r(z,I_r) | z \varepsilon \bar{o}_j^{r-1}\}$. Through an independent source, the vector $x_r^* \varepsilon \bar{o}_n$ and the value $\phi_r(x_r^*, I_r) = c_r$ are provided where c_r is the i_r^{th} largest value in the set $\{\phi_r(x,I_r) | x \varepsilon \bar{o}_j^{r-1} \cap \bar{o}_n\}$. The cutting function $\phi_r(x,I_r) = c_r$ is used to define two new blocks in this block B_j^{r-1}:

$$B_j^r = \{x \varepsilon B_j^{r-1} | \phi_r(x,I_r) < c_r\}$$

and

$$B_{j+1}^r = \{x \varepsilon B_j^{r-1} | \phi_r(x,I_r) > c_r\}$$

The remaining blocks in $\{B_1^{r-1}, B_2^{r-1}, \ldots, B_r^{r-1}\}$ are relabeled as

$$B_i^r = B_i^{r-1} \quad \text{for } i = 1,2,\ldots,j-1$$

$$B_{i+1}^r = B_i^{r-1} \quad \text{for } i = j+1,\ldots,r .$$

The total permissible information available for entering the $(r+1)^{st}$

stage consists of I_r, x_r^*, c_r, the blocks $B_1^r, B_2^r, \ldots, B_{r+1}^r$, the corresponding observation sets $\bar{o}_1^r, \bar{o}_2^r, \ldots, \bar{o}_{r+1}^r$, and all information which is symmetric separately with respect to the set of all blocks in $\{B_1^r, B_2^r, \ldots, B_{r+1}^r\}$ having corresponding observation sets which have not been identified.

The process B^* may be continued through the n^{th} stage if all blocks designated "never to be divided" contain no observations in $\bar{o}_n$. Or the process may be stopped at any stage level if it has been decided that a sufficient number of blocks have been obtained to properly evaluate the two-sample tolerance test statistic considered. However, the test, previously selected, may dictate the number of blocks to be formed and possibly the number of observations in $\bar{o}_n$ which must lie within each block formed. Most two-sample tolerance tests are analogs of two-sample univariate rank tests and would possibly require the process to continue through the n^{th} stage, if it were not apparent at some earlier stage that all observations in $\bar{o}_m$ have been identified.

If the process B^* were permitted to continue through the n^{th} stage all blocks formed $B_1^n, B_2^n, \ldots, B_{n+1}^n$ are called basic blocks and are equivalent (in the probability sense defined in Chapter III) to statistically equivalent blocks formed by the standard one sample block construction process outlined in Chapter III.

Finally, it should be noted that the standard one-sample block construction process is a special case of B^*. Let both sets of random vectors O_n and O_m denote independent random samples which under the null hypothesis have the same distribution function, $F(x)$. Then the joint null distribution of O_{n+m} is symmetric. At each stage r, $r = 1,2,\ldots,n$ in B^* let the information I_r contain only the observed vectors $x_1^*, x_2^*, \ldots, x_{r-1}^*$ and $c_1, c_2, \ldots, c_{r-1}$ and all prior information available before obtaining the observations within the samples. Then this restricted version of B^* is identical with the standard one-sample process.

CHAPTER V

PRACTICAL CONSIDERATIONS

Several practical techniques are suggested for effectively applying the proposed method to forming multivariate two-sample tolerance tests. Also included are special considerations when using univariate two-sample tolerance tests, a suggested operational procedure, and a discussion on the potential problem of bias associated with the quality control tests presented in the second chapter. The terminology and notation defined in Chapters II and IV are used in this discussion. The dimensionality of all data vectors will be denoted by p.

A suggested preliminary procedure that should be considered before applying the B^* process (i.e. before looking at the data) begins by either selecting an appropriate univariate two-sample tolerance test or developing a multivariate test which apparently best applied to the given problem. The next step is to determine the number and basic block contents of the tolerance regions to be formed. These values are defined directly by the test selected for use. The third important step requires a specific description of a construction plan showing the general order or layout of the blocks to be formed at various stages. This construction plan is not intended to dictate the shapes of the desired tolerance regions but merely to state a means for identifying each particular desired region once they all have been formed by the B^* process. In other words, it is forbidden to first construct a set of blocks then decide on how the desired tolerance regions will be identified and/or perhaps formed by a combination of blocks (e.g. basic blocks). The final step, which appears to only be required by certain tests, is to associate each desired region, identified by the construction plan, with a unique frequency count statistic. That is, certain tests may require that the desired regions be preassigned a fixed order. These preliminary considerations are further discussed for specific tests given as examples in this chapter.

Once a tolerance test has been selected the proposed operational objective in using the B^* process is to form the desired blocks according to the construction plan so as to make the test statistic as significant as possible. Thus, it would appear that this objective can best be satisfied by visually considering the set of unidentified observed vectors. However, if the dimensionality, p, of the data vectors is large an actual "look-at" the data situation may prove to be impractical as well as confusing. To alleviate this problem the principal component technique is suggested. This technique will usually permit the analyst to consider only a two-dimensional plot of transformed data.

The statistical method of obtaining principal components ([2], Chapter II) can be used as a numerical technique for transforming the original coordinate system orthogonally onto another p-dimensional coordinate system. This new coordinate system is constructed by choosing the first coordinate to have maximum dispersion among the transformed data vectors. The second coordinate, orthogonal to the first is chosen to have the next largest dispersion among the transformed data, etc.

Technically, the first new coordinate is defined by an eigenvector associated with the largest eigenvalue of the scatter matrix determined from the set of observed vectors. Then the second coordinate is defined by an eigenvector, orthogonal to the first eigenvector, associated with the second largest eigenvalue of the scatter matrix, etc. Thus, the first two coordinates formed by the principal component technique describe the greatest amount of dispersion among the transformed data. This should be a convenient and valuable aid for selecting candidate sets and cutting functions.

Other numerical or statistical techniques can also be used for this purpose. For example, the statistical method for determining canonical correlations ([2], Chapter 12) also provide a new coordinate system. Actually, any continuous transformation on the original p-dimensional space could be considered.

The proposed practice of reducing the dimensionality of the multivariate situation by means of various transformation schemes offers a promising approach for considering the data. However, the data characteristics may also be studied by actually increasing the dimensionality of the problem. For example, considering the mean vectors and variance-covariance matrices of various subsets of the totality of unidentified observations could provide invaluable information for selecting the candidate sets and appropriate cutting functions. Also any other numerical methods can be used to analyze the data situation. This extended freedom can be used to further describe the generality of the permissible information on the multivariate observations defined by the B^* process.

In order to clarify the use of the proposed method of forming multivariate two-sample tolerance tests, four examples are provided. A different test is considered in each example. In these examples the observations on O_m will be used to establish the block frequency counts determined by the blocks formed on the observations on O_n. All schematic drawings used to display the data situations assume that either $p = 2$ or the data has been transformed so that a two-dimensional space suffices.

As the first example, suppose $m = 1$ and $n > 1$. A two-sample tolerance test can be developed by establishing one tolerance region (block) containing most of the observations on O_n. If the one observation on O_m (new observation) falls outside this region the null hypothesis is rejected, otherwise it is not rejected.

The basic block content of this region, say $n + 1 - v$, depends on the significance level chosen for the test. The exact significance level of this test is the null probability that the new observation falls outside the desired tolerance region. This probability is computed to be $v/(n + 1)$. If α denotes the chosen significance level of the test, then the value of v is determined to be the smallest integer such that

$$v \geq \alpha(n + 1).$$

After v has been determined, the objective of the proposed method is to construct according to some plan a tolerance region of content $n + 1 - v$ on the observations on O_n which apparently best emphasizes any difference between the set of observations on O_n and the new observation. This desired region can be constructed on one or in as many as n stages using the B^* process.

If it were considered to use only one stage to construct this region then the two blocks formed must be of content v and n + 1 - v, respectively. In this case, the block of content v would actually be the critical region of the test. The permissible information available to form this first stage cut would include all information on the combined observations which is symmetric with respect to O_{n+m} and all independent information available prior to taking the observations. However, this approach to forming the desired tolerance region would not exploit all the advantages of the $B*$ process.

A suggested approach, which apparently makes better use of the $B*$ process, requires that all n + 1 basic blocks be formed in a particular way. The objective of this approach is to form the shape of the desired tolerance region (also determines the shape of the critical region of this test) which best defines the difference between the new observation and the observations on O_n. This objective may be accomplished in the following manner. For the first stage, select one unidentified vector as the candidate for the new observation. This candidate may be selected as the vector which lies the "furthest" away from the remaining n vectors. Using these n remaining vectors for the observations on O_n, determine a real valued function which for some value in its range separates the new observation candidate from the remaining n vectors by enclosing the n vectors. Set $i_1 = 1$ (if the value of the function is directly proportional to its enclosed volume) and obtain the first stage cut using this function. This cut will define a basic block which should be located somewhat near the center of the empirical distribution of the observations on O_n. Maintaining the same objective used in the first stage the process may be continued up to the $(n-v)^{th}$ stage. Of course, at each succeeding stage a new observation candidate and functions may be chosen differently. However, if at any stage the new basic block formed contains an observation, this observation must be the new observation and the null hypothesis cannot be rejected concluding the test. If the new observation has not been identified by the end of the $(n-v)^{th}$ stage it must lie within the remaining tolerance region of content $v + 1$ and the process must continue.

Throughout these $n-v$ stages the objective is to select cutting functions so as to exclude the new observation candidate from the basic blocks formed. Since more information is provided by the B^* process at each successive stage, the cutting function defined at the $(n-v)^{th}$ stage should reasonably well represent the shape of the empirical distribution defined by the observations on O_n. A schematic drawing showing the general appearances of the $n-v$ cutting function is given in Figure 1a. Note at this point the $(n + 1 - v)^{th}$ cutting function which will define the desired tolerance region has not been formed. It could be formed in the next stage; however, more information pertaining to the "best" shape of this region can be obtained by forming $v - 1$ more basic blocks.

In the next stage the integer i_{n+1-v} is set equal to v and the cutting function is chosen in a similar way used in obtaining the cutting functions in the previous $n-v$ stages but with the new objective to include the new observation candidate within the basic block formed. The basic block formed at this stage includes all points in the sample space lying outside the region enclosed by the cut. The same objective is used to construct the next $v - 2$ basic blocks. If at any stage an observation

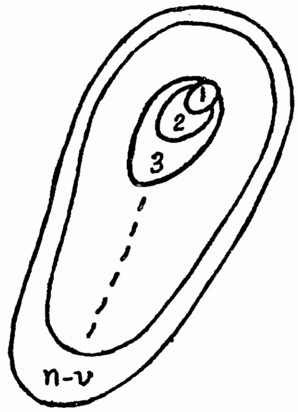

Figure 1a.

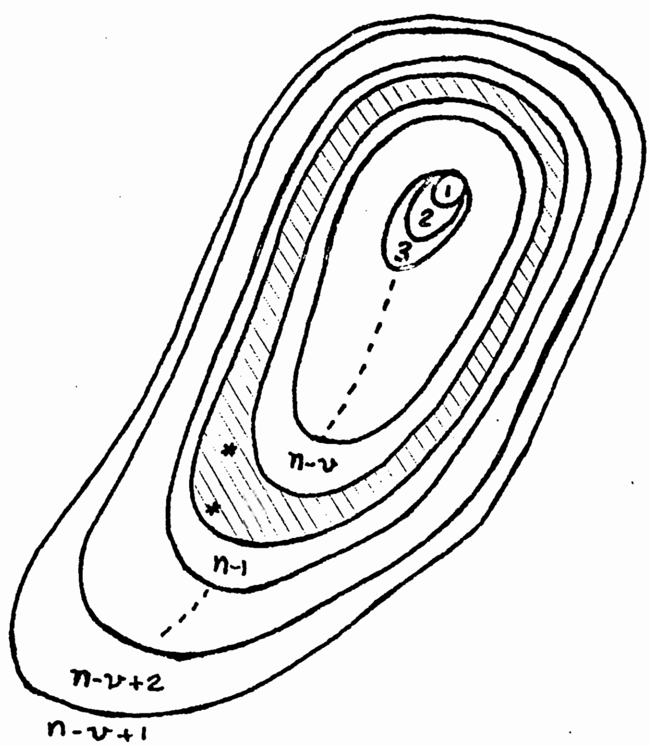

Figure 1b.

lies within the basic block formed, this observation must be the new observation and the null hypothesis is rejected concluding the test.

If after these n-1 stages the new observation has not been identified, the remaining region (basic block content is two) must contain exactly two vectors. One of these unidentified vectors is some observation on $\mathcal{O}_n$ and the other vector must be the new observation. A schematic of this situation is given in Figure 1b showing the remaining region. At the end of the $(n-1)^{st}$ stage two reasonably well shaped cutting functions defining the remaining region can then be used advantageously to define the shape of the final cut.

This latter approach may not seem too practical especially if n were very large. A similar method could be used in which blocks of content two or more are considered at each stage. This method would require fewer stages of the B^* process and would yield a favorable tolerance region. However, regardless of the approach used it must be decided (before applying the B^* process) how the desired tolerance regions will eventually be defined and how they occur in the test statistic.

Tests of this type have direct application in the quality control tests (Chapter II). For example, suppose the first data sets contain n observations and all other data sets consist of exactly one observation each. If it were decided that for each subtest the critical region would be of basic block content v, then the exact significance level of the j^{th} subtest would be $v/(n+j)$ if no randomization were used in all subtests. The significance level of the overall test would in this case be

$$\alpha = 1 - \prod_{j=1}^{k} \left(1 - \frac{v}{n + j}\right) = 1 - \binom{n}{v} \Big/ \binom{n + k}{v}$$

where k denotes the maximum number of subtests. From this expression, the value of either k, α, n, or v may be determined if the other values are specified. For example, if v = 1 for all subtests, then the value of k would be determined as the smallest integer satisfying

$$k \geq \frac{\alpha}{1 - \alpha} n.$$

Other similar quality control tests could be established using two-sample tests of this type.

As the next example, suppose that the Wilcoxon- Mann - Whitney test [10] has been selected for application. The statistic associated with this test is

$$U = \sum_{i=1}^{m} r_i - m(n + m + 1)/2$$

where r_i, i = 1,2,...,m are the "ranks" of the observations on O_m. Some properties of this test, considered for the univariate case, are presented in reference [14] on pages 61 through 68. The null distribution of this two-sample test for certain ranges of n and m are tabulated in reference [4]. For rather large n and m the null distribution of U can be approximated by the normal distribution.

In the univariate case, this test best emphasizes alternative hypotheses inferring slippage in the location between the two populations. That is, it will detect with rather high efficiency whether one population is statistically less than (or greater than) the other population. This test, however, is not efficient for testing a difference in dispersions if the two populations have nearly the same location.

To study the multivariate analogy of this test (or any other test) it is highly recommended that the statistic be expressed in terms of the appropriate block frequency counts. In the case of the Wilcoxon-Mann-Whitney test, the "ranks" of the observations on O_m can only be determined by basic blocks. Thus, the U statistic should be rewritten in terms of $m_1, m_2, \ldots, m_{n+1}$, the basic block frequency counts.

First note that the "ranks" of the observations on O_n are

$$m_1 + 1,\ m_1 + m_2 + 2, \ldots, \sum_{i=1}^{n} m_i + n.$$

Then the "ranks" of the observations on O_m would be the remaining integers in the set $\{1,2,\ldots,n+m\}$. Therefore, the sum of the ranks r_i of the observations on O_m is equivalent to the difference between the sum of all ranks in the combined set of observations on O_{n+m} and the sum of the ranks of the observations on O_n. This gives

$$\sum_{i=1}^{m} r_i = \sum_{j=1}^{n+m} j - \sum_{j=1}^{n} \{ \sum_{i=1}^{j} m_i + j \}$$

$$= m(m - n - 1)/2 + nm/2 + \sum_{j=1}^{n+1} j\, m_j$$

since $m - m_{n+1} = \sum_{j=1}^{n} m_i$.

Thus the U-statistic expressed in terms of the basic block frequency counts simply becomes

$$U = \sum_{j=1}^{n+1} j\, m_j - \frac{m(n+2)}{2} .$$

This implies that the value of U depends not only on the basic block frequency counts but also on the way in which the blocks are ordered.

To appropriately use the Wilcoxon-Mann-Whitney test as a multivariate two-sample tolerance test, a method of ordering the basic blocks to be determined by the B^* process must be prespecified. This additional consideration could be resolved by specifying the general manner in which the basic blocks are to be formed.

Another important consideration is the interpretation of the U-statistic when defined on some set of pre-ordered basic blocks. In the univariate case the basic blocks are determined by the order statistics and are ordered in the natural way. This ordering provides the basis on which the U-statistic was originally interpreted. That is, if the U-statistic obtained a value near either its lowest or highest possible values, then this would be interpreted correctly to mean that the simple two-sample null hypothesis was probably not true. However, in the multivariate case the interpretation of the U-statistic would depend largely on the ordering and relative locations of the basic blocks. If the basic blocks were ordered in any haphazard way, then any logical interpretation of various values of the U-statistic would be difficult to express.

If it were desired to interpret the U-statistic in the same concept used in the univariate case, two situations must be considered.

First, suppose that a two-sided U-test were selected. In the multivariate case the alternative hypotheses to be emphasized should reflect that the two populations differ in location in some direction in the p-dimensional space. The following is a suggested procedure for establishing and ordering the basic blocks for this two-sided U-test.

(1) Using some numerical procedure (e.g. least squares) determine

the best fit of the combined unidentified observed vectors to a straight line. This line can be assumed to represent the "most-likely" direction of any location difference between the two populations.

(2) The first stage cut is made by a hyperplane orthogonal to the line (established in (1)) passing through the (a) median (with respect to the hyperplane) of the observations on O_n, that is, i_1 is selected to be an integer nearest to $(\frac{n+1}{2})$.

(3) The two blocks defined at the first stage are then divided into basic blocks by forming a series of blocks radiating out from this hyperplane. This can be accomplished by choosing cutting functions nearest the center of the empirical distribution of the candidate vectors for the observations on O_n in each subsequent stage (see Figure 2a). This construction approach insures that the cutting functions will describe the shapes of the empirical probability surfaces of the observations on O_n in the "tails". The order established by the B^* process given by the subscripts in the set $\{B_1^n, B_2^n, \ldots, B_{n+1}^n\}$ would be a natural basic block ordering suitable for the desired interpretation of the U-statistic.

Now suppose that a one-sided U-test were desired. The alternative hypotheses associated with this test in the univariate case would reflect one population is stochastically larger (or smaller) than the other population. In the multivariate case the alternative hypotheses could be either that one population is stochastically larger (or smaller) than the other in some direction or that the two populations differ in location where the direction is not specified. To test the first form of the al-

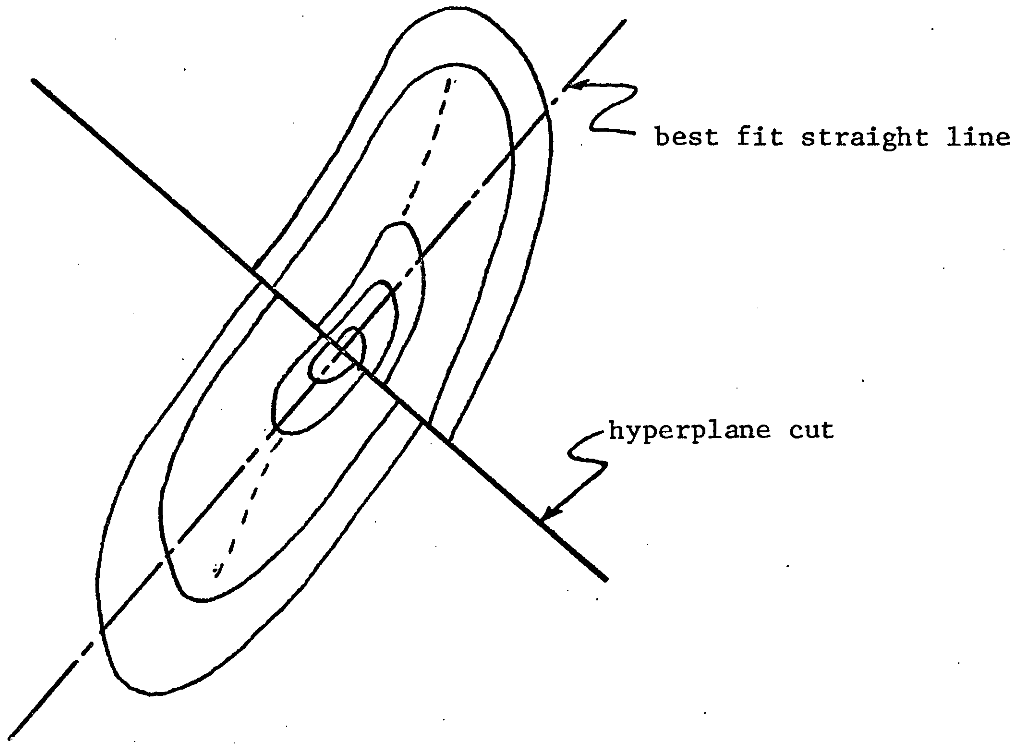

Figure 2a.

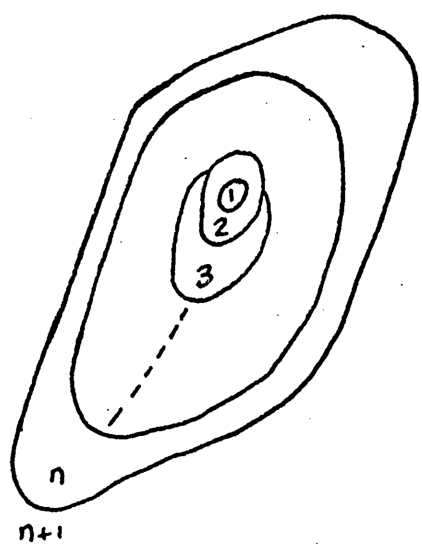

Figure 2b.

ternative hypothesis the same procedure outlined above for the two-sided test could be used. To test the second form of the alternative hypothesis another method is suggested (see Figure 2b).

(1) Determine the mean vector of the candidate vectors that are associated with the observations on O_n. If the candidate sets are difficult to determine the mean vector of the totality of observed vectors can be used.

(2) In the first stage take $i_1 = 1$ and make the first cut centered about the mean vector. The shape of the cut should perhaps be determined by the empirical distribution established on the candidate set for observations on O_n.

(3) Redefine (if necessary) the candidate set for the observations on O_n based on the information made available by the first stage. Then repeat (1) and (2) for $i_2 = 1$.

(4) Continue the procedure to the n^{th} stage. The resulting order of the basic blocks defined by the B^* process can then be used.

Note: This last procedure does not necessarily result in concentric cutting functions but the cutting functions will tend to radiate out from the center of the empirical distribution of the observations on O_n.

Again, the shapes of the cutting functions determined in the latter stages of the B^* process closely describe the shape of the empirical distribution of the observations on O_n. That is, the differences between the observations in the "tails" of this empirical distribution may be emphasized by this approach.

Another approach, which appears to be better for forming these basic blocks for the one-sided U-test and may be considered for other tests, seems to contradict one's natural intuition. This approach is as follows:

(1) Using the totality of unidentified observations determine the new coordinate system of principal components.

(2) For the first stage obtain an acceptable real-valued function on this new coordinate system which apparently best describes the general contour of the empirical distribution defined by the candidate set for the observations on $\mathcal{O}_n$ (or if not possible, on the totality of observation).

(3) Rotate this function through its center by making appropriate transformations and interchanging the roles of the i^{th} principal component with the $(p - i + 1)^{st}$ principal component for $i = 1,2,\ldots,[p/2]$ ($[x]$ denotes the largest integer less than or equal to x).

(4) Set $i_1 = 1$ and determine the first cut using this rotated function.

(5) Repeat steps (2) through (4) setting $i_j = 1$ for $j = 2,3,\ldots,n$ and possibly redefining new functions and candidate sets at each stage.

A schematic picture of this approach is given in two-dimensions in Figure 3.

Since the first principal component contains the greatest amount of dispersion among the transformed observations, this component axis may represent the most likely direction showing any differences in location between the two populations. The second principal component axis indicates the next most likely direction of location differences, etc. The

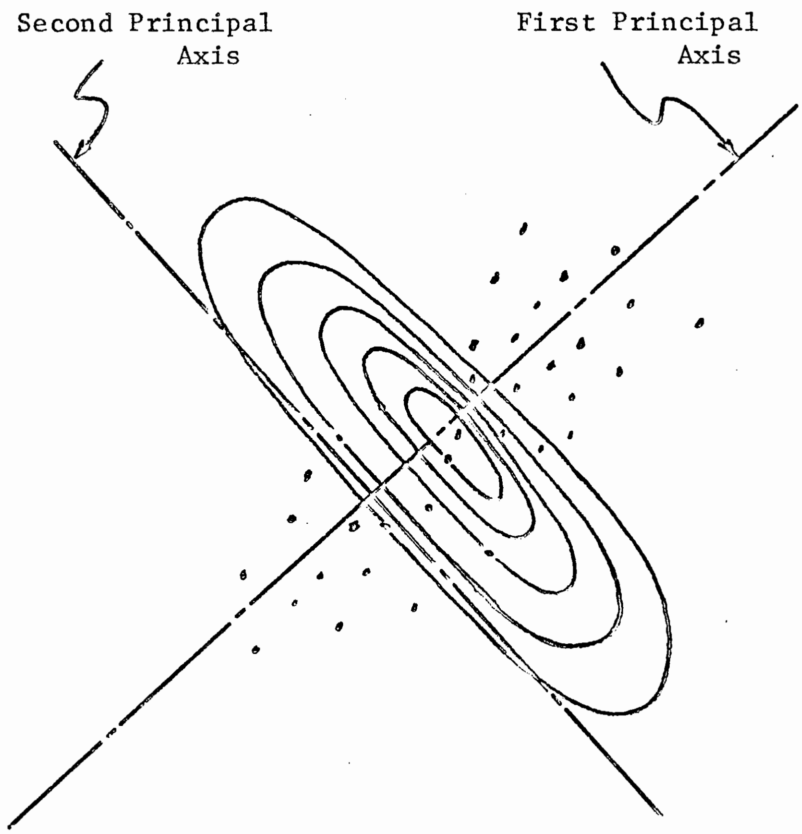

Figure 3.

objective of this approach is to shape the blocks in such a way so that the least number of blocks (in the "tail" for the one-sided U-statistic) contain the greatest number of observations on O_m. By transforming and rotating the functions according to the above method, the chances of accomplishing this objective appears to be rather good for emphasizing the alternative hypotheses.

Since this one-sided test can, in this case, be used to emphasize general two-sided alternatives that are associated with the U-test, then it would appear to have greater power than the two-sided test. The same construction techniques suggested for the multivariate one-sided U-test applies directly to the univariate case as well. Thus, all univariate two-sided alternatives can be treated by a univariate one-sided U-test.

This example was selected to emphasize the fact that not all appropriate univariate two-sample tests can readily be applied to the multivariate situation by disregarding the interpretation of the test statistic and any other consideration to be imposed on the content and use of the blocks. Some tests may be applied directly without any restriction on the block usage. This is shown by the next example.

Suppose that the Dixon C^2 test [5] is considered for application. In the univariate case this test is consistent and moderately efficient for virtually all alternatives of interest. The test statistic is expressed directly in terms of the basic block frequency counts:

$$C^2 = \sum_{i=1}^{n+1} \left[\frac{1}{n+1} - \frac{m_i}{m} \right]^2$$

A thorough description of the properties of the C^2 test is provided in reference [14] on pages 153 and 154. The null distribution of C^2 is tabulated for some n and m in reference [5]. If $\alpha < 1/2$ denotes the selected significance level of the test and if $nm/(n+m) \geq 6$ and $(n+m)/(4nm) \leq \alpha$ the null distribution of C^2 can be approximated by the chi-square distribution. The Dixon C^2 test is always one-sided.

Since the C^2-statistic does not depend on any preordering of the basic blocks, then the interpretation of the C^2 statistic for the multivariate case remains unchanged from the univariate case as long as a logical procedure for forming the blocks has been established. Thus, large values of C^2 will indicate in either case that the null hypothesis is probably not true. If the exact form of the alternative hypothesis cannot be specified, it would appear that the Dixon C^2 test would be a most appropriate choice.

Next, a few particular data situations are considered for applying the C^2 test (or most any other appropriate test whose statistic does not depend on an ordering of the blocks formed).

As stated earlier, the operational objective of the proposed method is to construct the desired tolerance regions (blocks) by trying to make the test statistic as significant as possible subject to the rules defined by the B^* process and any other additional considerations. The decision of which blocks to divide or not to divide, the choice of candidate sets, and the selection of a cutting function at each stage should naturally depend on the test statistic, the set sizes n and m, and on the significance level of the test. For example, suppose $n = m = 4$ and the test significance level was chosen to be 0.10, then the Dixon C^2 test would be to reject the null hypothesis if $C^2 \geq 0.8$, otherwise it is not rejected. This critical region is only obtained whenever one of the five basic blocks contain all four of the observations on O_m. If the plotted unidentified (transformed) observation vectors yielded the data situation given in Figure 4a, then the best intuitive procedure is to select four points lying in what appears to be a cluster as the candidates for the observations on O_n, the other points are then candidates for the observations on O_m. In the first stage take $i_1 = 4$ and a real-valued function which best describes (for convenience circles are used in Figure 4) a boundary about the candidate set for O_n. If after establishing the first stage cut the two blocks take the form given in Figure 4b, then all observations on O_m are clearly identified since they lie within one basic block - then the null hypothesis is rejected at the first stage. If the first stage blocks take the form shown in Figure 4c, then only x_1^* can be identified and a second stage is required. A new set of candidate points for the observations on O_n are selected, perhaps those nearest the identified observation x_1^*. A second function emcompassing these points is used to determine the second stage cut (for $i_2 = 3$). Then if the result given by Figure 4d is obtained the observations on O_m lie in a basic block rejecting the null hypothesis. If the result described by Figure 4e occurred then the null hypothesis cannot be rejected. Also, if at the first stage the resulting blocks took the general form given by Figure 4f the null hypothesis could not be rejected. A similar approach could be used whenever the data yielded two reasonably well defined clusters of sizes n and m.

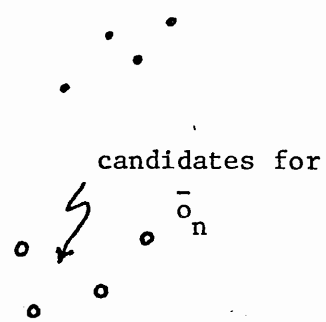

Figure 4a.

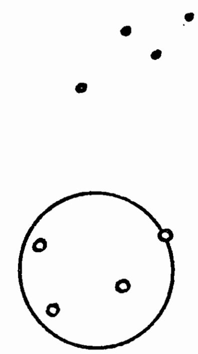

Figure 4b.

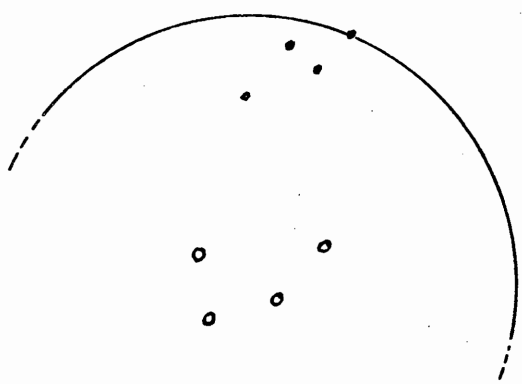

Figure 4c.

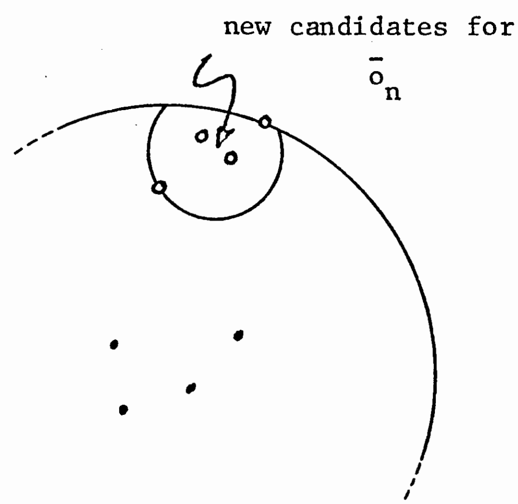

Figure 4d.

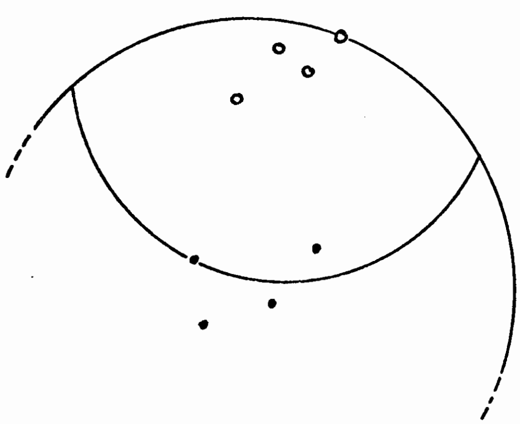

Figure 4e

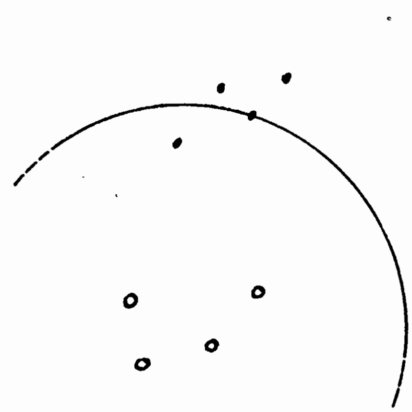

Figure 4f.

Now suppose the data situation yields one cluster of points. In this case, it may be rather difficult to select the candidate sets. The suggested technique for handling this situation is to first determine the mean vector for the totality of unidentified observed vectors. Take $i_1 = 1$ and define a first stage cutting function whose center is at the computed mean vector, as in Figure 5a. Considering all the permissible information available at the end of the first stage, in particular the relative positions of the vector x_1^* and the remaining unidentified observed vectors, it may now be possible to select the apparently "most-likely" candidate sets. Then continue the *B** process selecting cutting functions which tend to radiate out from the center (see Figure 5b). However, if candidate sets cannot be reasonably defined, set $i_2 = 1$ and select the second stage cutting function to have its center at x_1^*, etc.

Of course, there are many other data situations that could occur (e.g. three or more distinct data clusters). Similar procedures for handling these situations could be established as long as the basic operational objective remains unchanged.

The final example considers the use of Mathisen's quartile test [11]. This test is based on the frequency counts determined by tolerance regions representing 25 percent regions defined by the observations on O_n. The test requires that $(n+1)/4$ and $m/4$ are both integers. The test statistic is given by

$$B = 16 \sum_{i=1}^{4} [m_i - \frac{m}{4}]^2/(9m^2)$$

where m_i, $i = 1,2,3,4$ are the block frequency counts of the four 25 percent regions. Some properties of this test are presented on page 152 of

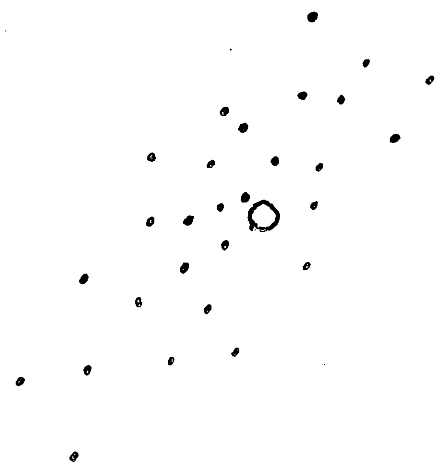

Figure 5a.

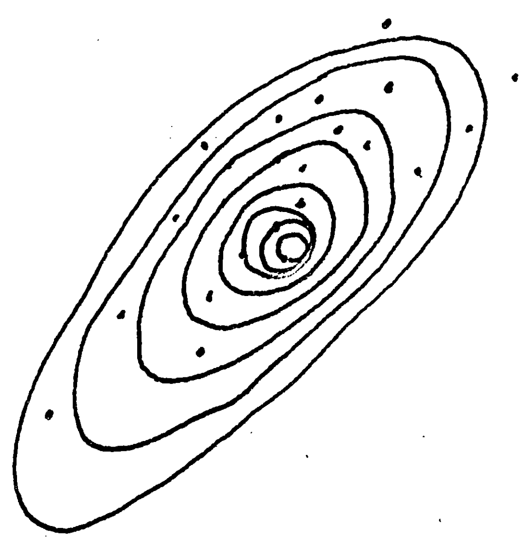

Figure 5b.

reference [14]. A table of its null distribution is given in reference [11]. For certain n and m the null distribution can be approximated by the beta distribution. In the univariate case this test emphasizes simultaneously differences in location, dispersion and skewness. The test is one-sided and is rather similar to the Dixon C^2 test for $n = 3$.

Since the B-statistic does not depend on any ordering of the blocks (25 percent regions) no special considerations of block orderings are necessary. Two approaches for forming the desired regions are given. The first approach uses only three stages of the $B*$ process in the most effective manner, while the other approach requires all n stages.

The first suggested procedure is to select $i_1 \varepsilon \{\frac{n+1}{4}, \frac{3(n+1)}{4}\}$ and choose some appropriate cutting function depending on the data situation. This first stage cut divides the sample space into a block of content $\frac{n+1}{4}$ and another block of content $\frac{3(n+1)}{4}$. Since the desired tolerance regions are 25 percent regions the block of content $\frac{n+1}{4}$ cannot be divided at any later stage. By the rules defined by the $B*$ process, all observations in this block can be identified. This information (if not sufficient to conclude the test) should prove extremely valuable in selecting the second stage candidate sets and cutting function. In the second, the remaining block is bisected by choosing $i_2 \varepsilon \{\frac{n+1}{4}, \frac{n+1}{2}\}$. Again, one block of content $\frac{n+1}{4}$ is formed (the other block is of content $\frac{n+1}{2}$). All observations in this block can be identified. All this information can now be used to select the third stage cutting function (for $i_3 = \frac{n+1}{2}$) to bisect the remaining block of content $\frac{n+1}{2}$.

The second suggested procedure is to predefine some scheme or construction plan for forming the desired tolerance regions from basic blocks. This scheme must be defined before "looking-at" the observed data. One simple approach is to use the $B*$ process by selecting cutting functions to form basic blocks whose centers are the apparent centers of the candidate sets for the observations on O_n. These cutting functions will tend to radiate out from the actual center of the observations on O_n, in a fashion shown by Figure 5b. The $\frac{n+1}{4}$, $\frac{n+1}{2}$, and $\frac{3(n+1)}{4}$ stage cutting functions will then define the desired 25 percent tolerance regions.

The null distributions referenced for the above statistics were derived under the unconditional probability model. By the corollary given in Chapter VII, these distributions also hold for the conditional permutation model.

A formal outline of the suggested operational procedure can now be stated.

(1) Consider all independent information that is available prior to taking the observations. This information may include previous observations, the modal characteristics of the underlying distributions, etc.

(2) Select or construct an appropriate two-sample tolerance test and specify the desired significance level.

(3) Carefully consider the test statistic expressed in terms of the appropriate block frequency counts. From this, determine the number and basic block contents of the tolerance regions to be formed and any additional considerations (e.g. construction plan, block order, etc.) that may be imposed on the block usage.

(4) Interpret the test statistic for the proposed method of using and ordering the blocks. Then define the critical region of the test.

(5) Collect the two observation sets.

(6) If the dimensionality, p, of the observed vectors is greater than two consider possible numerical techniques for transforming the data to a two-dimensional coordinate system. Then "look-at" the plotted (transformed) data points. This data is completely unidentified.

(7) If possible select two candidate sets for O_n and O_m.

(8) Using the rules specified by the B^* process and any other additional restrictions cited in (3) and (4), apply the B^* process constructing the desired blocks so as to make the test statistic as significant as possible.

(9) After (or during) the application of the B^* process evaluate the test statistic and determine the outcome of the test.

The final consideration is devoted to a potential problem that may arise when using the proposed method for determining subtests in the quality control tests presented in Chapter II. This problem occurs from the carryover of human bias from preceding subtests when "looking-at" the combined data. A subset or the set of the combined data of the subtest becomes the previous data set of the next subsequent subtest. Thus, knowledge of this previous data, especially if p = 1 or 2, may directly influence the choice of candidate samples. This knowledge is not permitted by the B^* process. The bias that may occur could be considerably large unless appropriate safeguards are taken. If this bias is not eliminated the B^* process is violated and the subtests are no longer independent. Some suggestions for alleviating this bias are presented.

One approach is to list (if possible) a set of fixed rules which generally apply for all data situations that could occur at any stage in the B^* process. These rules are used to make decisions for selecting the candidate sets and possibly the cutting functions at each stage. One elementary example is provided.

Suppose all new data sets, except the first data set, consists of only one observation. Then the rule to always select, as the candidate for the new observation, the one unidentified observation which lies the "furthest" away from the set of remaining unidentified observed vectors. The concept of distance in this rule may be defined by fitting, say, an ellipsoid function to the set of all data points. Then the observed vector yielding the largest value of this function will be the candidate for the new observation. This rule can be repeated at each stage of the B^* process using only the unidentified observed vectors.

The approach of alternating or employing different analysis for each subtest should reduce or eliminate all bias. This approach appears to be direct and simple to apply.

In the next chapter several areas of application are discussed. Most of these areas are limited to medical applications.

CHAPTER VI

SOME AREAS OF APPLICATION

The sequential significance tests having multivariate two-sample tolerance subtests presented in this paper appear to be generally applicable to most quality control as well as to other testing situations. A few specific medical applications are cited. Other areas apparently submissive to these tests are listed at the end of this chapter.

The first application to be considered is the quality control of a system used to determine the electrophoretic analysis of serum protein. This method of characterizing serum proteins has provided better understanding of associated clinical disorders and in some instances has aided in recognizing new diseases complicated by serum protein abnormalities.

Electrophoretically separated serum proteins are classified in five rather distinct groups: albumin, α_1-, α_2-, β-, and γ- globulins. The basic results of an electrophoretic serum analysis are given by the concentrations of these protein groups. These measurements are usually expressed in terms of the fractions of total protein concentration. The systems used to obtain these measurements is influenced by several factors: human, mechanical, chemical, and electrical.

Present quality control techniques used to test the system operation consider each of five univariate measurements independently of the other four. This method of testing would be highly questionable if there exist any dependent relationships among the five variables. This would not be a problem if multivariate tests were used.

The standard method of testing the quality of this system is a sequential quality control test which is very similar in structure to that presented in Chapter II. That is, the previous data is continually reused; however, there is no regard for independence between subtests. A reserve bank supplies the source of serum used to conduct the quality control tests. The serum in this bank is replenished periodically by sampling from the excess of serum tested over previous days. The serum samples are combined, homogeneously mixed, and frozen for preservation. For each subtest one or two samples are taken from the serum reserve bank and electrophoretically processed in the system. The results are analyzed, then tested against previous results to determine if the system is in or out of control.

Since the multivariate observations consist of five continuous variables, the proposed method of forming two-sample tolerance tests trivially holds. The test situation then appears to conform well to the quality control test presented in the first example of Chapter V.

The next application is to clinical trials. The objective of clinical trials is to compare the effect of some treatment (e.g. a drug) to some standard. This standard may be described by measurements on untreated patients or on patients subjected to a different treatment. The measurements used for comparison are in the form of symptoms, signs, and/or clinical findings. One approach to clinical trials is to enter one patient at a time into the experiment. A set of measurements on the treated patient is used to compare against the standard measurements. The sequential testing is continued until significance occurs or a maximum number of tests have been conducted.

All measurements of symptoms and signs are usually considered discrete; however, most clinical findings (e.g. temperature, weight, blood chemistry, serum protein analysis, etc.) are continuous variables. If at least one of these continuous variables are included in the measurement of treatment response, then the proposed method of forming two-sample tolerance tests can be used to determine subtests. Again, the quality control test given in the first example of Chapter V can be applied to this problem.

A rather common problem in medical research (similar to problems in other scientific and engineering disciplines) is cited next. Suppose three groups of patients are involved in an experiment where each group is subjected to a different treatment. From limited independent previous knowledge, there is reason to believe that two of these treatments, say A and B, do not differ in their measured responses, while the third treatment, C, response affect is either unknown or is believed to differ from the other two. The desired test procedure is to first test the null hypothesis that treatments A and B yield the same effects, then if this hypothesis cannot be rejected test if treatment C differs from the combined affects of treatments A and B. If the observations, used to measure the treatment effects, are at least partially continuous, then the sequential significance tests having the proposed multivariate two-sample tolerance tests as its subtests can be made to apply to this procedure.

An extension to this problem follows. Suppose there are k ($k \geq 3$) groups of patients. Each group is subjected to a different treatment. Then prior to taking the observations, the treatments are ordered according to their believed differences. That is, those treatments considered first in this ordering are assumed to produce near similar responses, etc. Again, if the observations are at least partially continuous, the tests proposed in this paper can be used.

Some other general areas of possible application are: water and waste treatment plant quality control, traffic studies, scientific and engineering research, industrial quality control, market and other sampling surveys.

CHAPTER VII

STATEMENT OF BASIC RESULTS

Two major results are presented in the form of a theorem and corollary. All other basic results, some of which are direct consequences of the theorem or corollary, are verified in an informal format.

The following theorem proves that the joint null distribution of block frequency counts obtained at any stage in the B^* process is the same as if it were obtained by the standard one-sample process. The proof is given in Appendix II.

<u>Theorem</u>

Let $O_n = \{X_1, X_2, \ldots, X_n\}$ and $O_m = \{Y_1, Y_2, \ldots, Y_m\}$ be two sets of random vectors, not necessarily independent, defined on a sample space X. These sets are such that there is a probability one of the construction process B^* being unique (no ties in the cutting function values) which occurs in particular whenever the random vectors are at least partially continuous. Under the null hypothesis, let the combined set of random vectors $O_{n+m} = \{X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m\}$ have a symmetric joint cumulative distribution function denoted by $F = F(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m)$. Given a set of $r + 1$ blocks $\{B_1^r, B_2^r, \ldots, B_{r+1}^r\}$ formed at the r^{th} stage of the generalized block construction process, B^*, on a set of observations on O_n, then the joint null distribution of $m_1, m_2, \ldots, m_{r+1}$, the respective block frequency counts on O_m, is given by

$$P(m_1,m_2,\ldots,m_{r+1}) = \left\{ \prod_{i=1}^{r+1} \binom{m_i + k_i - 1}{m_i} \right\} \Big/ \binom{m+n}{m}$$

for non-negative integers m_i, $i = 1,2,\ldots,r+1$ such that $m_1 + m_2 + \ldots + m_{r+1} = m$. Here, k_i denotes the "number of basic blocks" contained within the block B_i^r, with $k_i - 1$ being the number of observations on $\mathcal{O}_n$ in B_i^r.

<u>Corollary</u>

The above theorem holds under the permutation probability model whenever the sets of random vectors are such that there is probability one of the construction process $\mathcal{B}^*$ being unique.

<u>Proof of Corollary</u>

Let $\{x_1,x_2,\ldots,x_n\}$ and $\{y_1,y_2,\ldots,y_m\}$ denote fixed sets of observations on $\mathcal{O}_n$ and $\mathcal{O}_m$, respectively. Let S be the conditional permutation sample space given the combined fixed set of observations $\{x_1,x_2,\ldots,x_n, y_1,y_2,\ldots,y_m\}$. Then S is equivalent to the set of all $(n + m)$-tuples obtained by permuting this given combined set of observations. Let X_i, $i = 1,2,\ldots,n$ represent the random vector (in the permutation model) yielding the vector value located in the i^{th} coordinate position and Y_j, $j = 1,2,\ldots,m$, denote the random vector yielding the vector value located in the $(n + j)^{th}$ coordinate position in the $(n + m)$-tuples in S. Then the joint null cumulative distribution of the random vectors $\{X_1,X_2,\ldots,X_n, Y_1,Y_2,\ldots,Y_m\}$ is symmetric. Set $S = \mathcal{X}$ and the proof follows.

This corollary implies that any test statistic associated with a two-sample tolerance test formed by the proposed new method has the same null distribution whether it was considered on an unconditional or permutation probability basis. Thus, all null distributions that have been obtained, under the unconditional model, for any appropriate two-sample tolerance test statistic are directly usable under the permutation model.

An immediate consequence of the theorem is considered next. Suppose O_n and O_m are independent random samples. Then, under the null hypothesis, their joint distribution is symmetric.

In Chapter IV it was shown that the standard one-sample process was a special case of the $B*$ process. Then it follows, from the above remark, that the existing method is a special case of the proposed method for establishing multivariate two-sample tolerance tests. The existing method can then be extended to consider two data sets, not necessarily independent, whose joint null distribution function is symmetric. Then, from the above corollary, any two-sample tolerance test formed by the existing method has a permutation probability basis.

It remains to show that any statistic associated with a two-sample tolerance test formed by the proposed method is symmetric in the observations on which the tolerance regions were defined and is symmetric in the observations used to establish the block frequency counts. This is equivalent to showing that any statistic is symmetric with respect to both O_n and O_m. It was never required at any stage in the $B*$ process to associate an observation with the particular random vector yielding it. The process only permitted an observation to be identified with the <u>set</u> of observations from which it came. Then all identified observations (in this sense), thus the set of all observations, are symmetric with respect to both O_n and O_m. Hence, any statistic is symmetric with respect to both O_n and O_m.

APPENDIX I

Definition I

If $V = (V^{(1)}, V^{(2)}, \ldots, V^{(p)})^T$ is a p-component random vector and $v = (v^{(1)}, v^{(2)}, \ldots, v^{(p)})^T$ is a p x 1 vector having real components, then the event that $V^{(i)} \leq v^{(i)}$ simultaneously for all $i = 1,2,\ldots,p$ is represented by $V \leq v$.

Definition II

Let $V_1, V_2, \ldots, V_t$ be a set of p-component random vectors, then the joint cumulative distribution function of $V_1, V_2, \ldots, V_t$ is given by

$$P(V_1 \leq v_1,\ V_2 \leq v_2, \ldots, V_t \leq v_t)$$

where $v_1, v_2, \ldots, v_t$ are real component p x 1 vectors and the events $V_i \leq v_i$ $i = 1,2,\ldots,t$ are defined in Definition 1.

Lemma I-1

Let $W_1, W_2, \ldots, W_q$ be a set of p-component random vectors which are defined on the sample space $\mathcal{W}$ and have a joint cumulative distribution function

$$F(w_1, w_2, \ldots, w_q) = P(W_1 \leq w_1,\ W_2 \leq w_2, \ldots, W_q \leq w_q).$$

If $F(w_1, w_2, \ldots, w_q)$ is a symmetric function on all sets $\{w_1, w_2, \ldots, w_q\}$ of real component p x 1 vectors, then for any set of measurable functions (real or vector valued) $g_i(W_1, W_2, \ldots, W_q)$ $i = 1,2,\ldots,k$ and any reordering $(i_1, i_2, \ldots, i_q)$ of the integers $(1,2,\ldots,q)$

$$P[g_1(W_1,W_2,\ldots,W_q) \leq a_1,\ldots,g_k(W_1,W_2,\ldots,W_q) \leq a_k]$$

$$= P[g_1(W_{i_1},W_{i_2},\ldots,W_{i_q}) \leq a_1,\ldots,g_k(W_{i_1},W_{i_2},\ldots,W_{i_q}) \leq a_k]$$

where a_i is a real component vector of the same dimensionality as $g_i(W_1,W_2,\ldots,W_q)$ for $i = 1,2,\ldots,k$.

Proof

Let σ: $(1,2,\ldots,q) \rightarrow (i_1,i_2,\ldots,i_q)$,

$$A = \{(w_1,w_2,\ldots,w_q)|g_1(w_1,w_2,\ldots,w_q) \leq a_1,\ldots,$$
$$g_k(w_1,w_2,\ldots,w_q) \leq a_k;\ w_i \varepsilon R^p \quad i = 1,2,\ldots,q\},$$

$$A_\sigma = \{(w_{i_1},w_{i_2},\ldots,w_{i_q})|g_1(w_1,w_2,\ldots,w_q) \leq a_1,\ldots,$$
$$g_k(w_1,w_2,\ldots,w_q) \leq a_k;\ w_i \varepsilon R^p \quad i = 1,2,\ldots,q\},$$

and

$$B_\sigma = \{(w_1,w_2,\ldots,w_q)|g_1(w_{i_1},w_{i_2},\ldots,w_{i_q}) \leq a_1,\ldots,$$
$$g_k(w_{i_1},w_{i_2},\ldots,w_{i_q}) \leq a_k;\ w_i \varepsilon R^p \quad i = 1,2,\ldots,q\}$$

where R^p is the p-dimensional euclidian space.

It suffices to show that $P(A) = P(B_\sigma)$.

$$P(A) = \int_A dF(w_1,w_2,\ldots,w_q) = \int_{A_\sigma} dF(w_{i_1},w_{i_2},\ldots,w_{i_q})$$

since $F(w_1,w_2,\ldots,w_q)$ is symmetric.

Consider the transformation $U_{\sigma^{-1}(i_j)} = W_{i_j}$ for $j = 1,2,\ldots,q$, then the transformation $W_i = U_i$ $i = 1,2,\ldots,q$ in the last expression. This gives

$$\int_{A_\sigma} dF(w_{i_1},w_{i_2},\ldots,w_{i_q}) = \int_{B_\sigma} dF(w_1,w_2,\ldots,w_q) = P(B_\sigma)$$

proving the lemma.

Lemma I-2

Let $W_1, W_2, \ldots, W_q$ be a set of p-component random vectors defined on a sample space $\mathcal{W}$ and have a symmetric joint cumulative distribution function, $F(w_1, w_2, \ldots, w_q)$. Let $\{Z_1, Z_2, \ldots, Z_t\}$ be a subset of the random vectors $W_1, W_2, \ldots, W_q$ and $\mathcal{F}$ represent the totality of information on the observations on $W_1, W_2, \ldots, W_q$ which is symmetric with respect to $\{Z_1, Z_2, \ldots, Z_t\}$[1]. If $g(W, \mathcal{F})$ is a measurable function on $\mathcal{F}$ and $W = W_1$ for any $i = 1, 2, \ldots, q$, then the joint cumulative distribution function of $g(Z_1, \mathcal{F}), g(Z_2, \mathcal{F}), \ldots, g(Z_t, \mathcal{F})$ is symmetric.

Proof

By definition, $\mathcal{F}$ is invariant under any relabeling of the identities in $(Z_1, Z_2, \ldots, Z_t)$.

If the roles of the random vectors Z_i and Z_j are interchanged, the statistic $g(Z_i, \mathcal{F})$ becomes $g(Z_j, \mathcal{F})$ and vice versa. Then by an interchange of the roles of the random vectors in $\{Z_1, Z_2, \ldots, Z_t\}$, the set of statistics $\{g(Z_1, \mathcal{F}), g(Z_2, \mathcal{F}), \ldots, g(Z_t, \mathcal{F})\}$ is mapped onto itself.

Let $g_i(W_1, W_2, \ldots, W_q) = g(Z_i, \mathcal{F})$, $i = 1, 2, \ldots, t$ in Lemma I-1 for $k = t$ and the proof follows.

---

[1] Information on a set of observations is defined to be symmetric with respect to a set of random vectors if, and only if, the information is unchanged by interchanging the roles of the random vectors. (See Chapter IV and Appendix II).

APPENDIX II

Theorem

Let $O_n = \{X_1, X_2, \ldots, X_n\}$ and $O_m = \{Y_1, Y_2, \ldots, Y_m\}$ be two sets of random vectors, not necessarily independent, defined on a sample space X. These sets are such that there is a probability one of the construction process B^* being unique (no ties in the cutting function values) which occurs in particular whenever the random vectors are at least partially continuous. Under the null hypothesis, let the combined set of random vectors $O_{n+m} = \{X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m\}$ have a symmetric joint cumulative distribution function denoted by $F = F(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m)$. Given a set of r + 1 blocks $\{B_1^r, B_2^r, \ldots, B_{r+1}^r\}$ formed at the r^{th} stage of the generalized block construction process, B^*, on a set of observations on O_n, then the joint null distribution of $m_1, m_2, \ldots, m_{r+1}$, the respective block frequency counts on O_m, is given by

$$P(m_1, m_2, \ldots, m_{r+1}) = \{ \prod_{i=1}^{r+1} \binom{m_i + k_i - 1}{m_i} \} / \binom{m+n}{m}$$

for non-negative integers m_i, $1 = 1, 2, \ldots, r+1$ such that $m_1 + m_2 + \ldots + m_{r+1} = m$ and where k_i denotes the "number of basic blocks" contained within the block B_i^r, $i = 1, 2, \ldots, r+1$.

Proof

The formal method of proof uses induction on the number of stages in the generalized block construction process, B^*, presented in Chapter IV.

The joint null distribution of the block frequency counts on O_m is determined at each stage. This is arrived at by deriving the conditional null distribution for the block frequency count on O_m for one of the two new blocks formed at the stage being considered given the joint frequency counts observed at the previous stage. This proof makes repetitive applications of Lemmas I-1 and I-2 presented in Appendix I. For convenience, a few terms and symbols defined in Chapter IV are restated.

Definition 1

An observation is said to be identified if it can be associated with the set of random vectors which yielded it; that is, associated with O_n or O_m. Thus, a set of observations is identified if each observation within the set is identified.

Definition 2

Information on a set of observations is said to be symmetric with respect to a block (or union of blocks) if the information is unaffected by interchanging the roles (relabeling the identities) of the random vectors yielding the observations falling within the block(s). For example, if $V_1,V_2,\ldots,V_k$ denotes the set of random vectors yielding observations falling within the block B, then the information, I, defined on some set of observations which may or may not contain those in B, is symmetric on B if for any reordering $(V_1',V_2',\ldots,V_k')$ of $(V_1,V_2,\ldots,V_k)$ I is unchanged.

Definition 3

The symbols $\bar{o}_n$, $\bar{o}_m$, and $\bar{o}_{n+m}$ will denote the observations on the random vectors in O_n, O_m, and O_{n+m}, respectively. Likewise O_j^k will denote the set of random vectors yielding the set of observations $\bar{o}_j^k$ which fall within block B_j^k at the k stage of the process B^*.

<u>Definition 4</u>

The symbols $Z_1, Z_2, \ldots, Z_{n+m}$ will denote the random vectors in the combined set O_{n+m}. Also, the symbols $Z_1^j, Z_2^j, \ldots, Z_{k(j)}^j$ will represent the random vectors in O_j^k and $X_1^j, X_2^j, \ldots, X_{i(j)}^j$ and $Y_1^j, Y_2^j, \ldots, Y_{t(j)}^j$ will be those random vectors in $O_j^k \cap O_n$ and $O_j^k \cap O_m$, respectively. All lower case symbols x, y, and z will denote corresponding observed vectors.

At the first stage in B^* all symmetric information with respect to X is available. This includes the set of unidentified observations in $\bar{o}_{n+m}$. Then I_1 consists of the total information available on $\bar{o}_{n+m}$ which is symmetric on O_{n+m}.

Let $i_1 \varepsilon \{1, 2, \ldots, n\} = J_n$ be selected based on I_1. Then using I_1, determine a real-valued measurable function $\phi_1(Z, I_1)$ which either has a continuous distribution function whenever $Z \varepsilon O_{n+m}$ and the random vectors in O_{n+m} have the joint null distribution function, F, or was selected in such a way that there are no ties in the set of values $\{\phi_1(z, I_1) | z \varepsilon \bar{o}_{n+m}\}$.

Let $x_1^* \varepsilon \bar{o}_n$ be such that $\phi_1(x_1^*, I_1) = c_1$ is the i_1^{st} largest value in the set of real numbers $\{\phi_1(x, I_1) | x \varepsilon \bar{o}_n\}$. [Note: at this point only x_1^* and c_1 are available information not the entire set of values $\{\phi_1(x, I_1) | x \varepsilon \bar{o}_n\}$]. Now, using the cutting function $\phi_1(x, I_1) = c_1$ divide the sample space X into the first stage blocks denoted by the open regions $B_1^1 = \{x \varepsilon X | \phi_1(x, I_1) < c_1\}$ and $B_2^1 = \{x \varepsilon X | \phi_1(x, I_1) > c_1\}$. There are exactly $i_1 - 1$ and $n - i_1$ observations in $\bar{o}_n$ falling within B_1^1 and B_2^1, respectively.

To determine the null probability that t_1 observations in $\bar{o}_m$ will fall within block B_1^1 ($m-t_1$ in B_2^1) is equivalent to determining the null probability that exactly t_1 of the observed variables $\phi_1(y_i,I_1)$ $i = 1,2,\ldots,m$ have values less than c_1. To evaluate this probability it is necessary to establish the joint null distribution of the random variables $\phi_1(Z_1,I_1)$, $\phi_1(Z_2,I_1),\ldots,\phi_1(Z_{n+m},I_1)$.

Appeal to Lemma I-2 (replacing the W_i's with Z_i's, set $t = q = n+m$, and $\phi_1(Z,I_1) = g(Z,F)$). Then it follows that the joint cumulative distribution of $\phi_1(Z_1,I_1)$, $\phi_1(Z_2,I_1),\ldots,\phi_1(Z_{n+m},I_1)$ is symmetric. But this implies

$$P[\phi_1(Z_1,I_1) < \phi_1(Z_2,I_1) < \ldots < \phi_1(Z_{n+m},I_1)]$$

$$= P[\phi_1(Z_1',I_1) < \phi_1(Z_2',I_1) < \ldots < \phi_1(Z_{n+m}',I_1)]$$

for all reorderings $(Z_1',Z_2',\ldots,Z_{n+m}')$ of $(Z_1,Z_2,\ldots,Z_{n+m})$.

By the choice of $\phi_1(Z,I_1)$ all such ϕ_1-orderings are unique with probability one, then the null probability of each ϕ_1-ordering is $1/(m+n)!$.

Define

$$S_1 = \{[\phi_1(Z_1',I_1) < \phi_1(Z_2',I_1) < \ldots < \phi_1(Z_{n+m}',I_1)] \mid (Z_1',Z_2',\ldots,Z_{n+m}') \text{ is a reordering of } (Z_1,Z_2,\ldots,Z_{n+m})\}.$$

Then the null probability of each element in S_1 is $1/(m+n)!$.

The null probability of observing t_1 observations in $\bar{o}_m$ within B_1^1 is then equivalent to the null probability of obtaining an element in S_1 which assigns the $(t_1 + i_1)^{st}$ ϕ_1-order position to O_n and exactly t_1 of the first $t_1 + i_1 - 1$ ϕ_1-order positions to O_m. This reduces to counting the number of elements in S_1 which satisfy the desired event, since all elements in S_1 are equally likely. There are $\binom{m}{t_1}$ ways of choosing t_1 random vectors in O_m and $\binom{n}{i_1-1,1}$ ways of selecting $i_1 - 1$ and 1 (for the $(t_1 + i_1)^{st}$ position) random vectors in O_n to lie within the first $t_1 + i_1$ ϕ_1-order positions. Within each chosen set of t_1 random vectors in O_m there are $t_1!$ ways of assigning them to a fixed set of t_1 ϕ_1-order positions; likewise, there are $(i_1-1)!$ and $1!$ of assigning the random vectors chosen from O_n. The remaining ϕ_1-order positions $t_1 + i_1 + 1,\ldots,n+m$ must contain $m-t_1$ and $n-i_1$ assignments to O_m and O_n, respectively. The total number of ways which these positions may be assigned is $(m-t_1 + n-i_1)!$. Hence the total number of elements in S_1 satisfying the desired event is

$$\binom{m}{t_1}\binom{n}{i_1-1,1}\, t_1!(i_1-1)!\,1!(m-t_1+n-i_1)!$$

$$= \binom{t_1+i_1-1}{t_1}\binom{m-t_1+n-i_1}{m-t_1} m!n!\ .$$

Then the null probability of observing exactly t_1 observations in $\bar{o}_m$ within B_1^1 is

$$P(t_1) = \frac{\binom{t_1+i_1-1}{t_1}\binom{m+n-t_1-i_1}{m-t_1} m!n!}{(m+n)!}$$

$$= \binom{t_1+i_1-1}{t_1}\binom{m-t_1+n-i_1}{m-t_1}\Big/\binom{m+n}{m}$$

for $t_1 = 0,1,\ldots,m$.

The joint null distribution function for t_1,t_2, the frequency counts in blocks B_1^1 and B_2^1, respectively, is obtained by transforming $m - t_1 = t_2$

$$P(t_1,t_2) = \binom{t_1+i_1-1}{t_1}\binom{t_2+n-i_1}{t_2}\Big/\binom{m+n}{m}$$

for non-negative t_i $i = 1,2$ such that $t_1 + t_2 = m$.

In order to clarify the method of proof the joint null frequency count distribution will be derived for the second stage.

The complete information available at the beginning of the second stage consists of I_1, c_1, x_1^*, and the two sets of unidentified observations $\bar{o}_1^1$ and $\bar{o}_2^1$.

At the start of stage 2, one of the two first stage blocks B_1^1 and B_2^1, is selected for division. This selection is based on the information available at the end of the first stage. Suppose B_1^1 were selected (the proof is analogous for B_2^1) and t_1 observations in $\bar{o}_m$ fell in B_1^1

If B_2^1 were to be decomposed at some later stage in the process $\mathcal{B}^*$, then I_2 contains all information symmetric separately with respect to B_2^1 and B_1^1. However, if B_2^1 were never to be decomposed at some later stage, then I_2 contains at least all information symmetric with respect to B_1^1. This allows I_2 to contain any information which is symmetric with respect to B_1^1 but not on B_2^1. For example, complete identification of $\bar{o}_2^1$. This is trivially true when B_2^1 is a basic block - all observations within B_2^1 have to be in $\bar{o}_m$. Clearly I_2 contains all information that was available at the end of stage 1.

Next, select an integer $i_2 \in J_{i_1-1}$ based on I_2 and choose a real-valued measurable function $\phi_2(Z,I_2)$ which either has a continuous null distribution for $Z \in O_1^1$ or is such that there are no ties in the set of values $\{\phi_2(z,I_2)/z \in \bar{o}_1^1\}$.

Let $x_2^* \in \bar{o}_n \cap \bar{o}_1^1$ such that $\phi_2(x_2^*,I_2) = c_2$ is the i_2^{nd} largest value in the set $\{\phi_2(x,I_2) | x \in \bar{o}_n \cap \bar{o}_1^1\}$. Again, only x_2^* and c_2 are determined. Using the cutting function $\phi_2(x,I_2) = c_1$, divide the block B_1^1 into two disjoint subsets.

$$B_1^2 = \{x \in B_1^1 | \phi_2(x, I_2) < c_2\}$$

and

$$B_2^2 = \{x \in B_1^1 | \phi_2(x, I_2) > c_2\}$$

and for standardizing notation, let $B_3^2 = B_2^1$. Then there are exactly $i_2 - 1$, $i_1 - i_2 - 1$, and $n - i_2$ observations in $\bar{o}_n$ falling within blocks B_1^2, B_2^2 and B_3^2, respectively.

Now, consider the conditional event, A_1, that t_1 observations in $\bar{o}_m$ fall within B_1^1 given c_1 and x_1^*. This event is the intersection of the following two events.

$$A_1^1 = \{\phi_1(Z_1^1, I_1) < c_1,\ \phi_1(Z_2^1, I_1) < c_1, \ldots, \phi_1(Z_{t_1+i_1-1}^1, I_1) < c_1\}$$

and

$$A_2^1 = \{\phi_1(Z_1^2, I_1) > c_1\ \ \phi_1(Z_2^2, I_1) > c_1, \ldots, \phi_1(Z_{n+m-t_1-i_1}^2, I_1) > c_1\}$$

where $\{Z_1^1, Z_2^1, \ldots, Z_{t_1+i_1-1}^1\}$ and $\{Z_1^2, Z_2^2, \ldots, Z_{n+m-t_1-i_1}^2\}$ are the random vectors in O_1^1 and O_2^1, respectively. Then $A_1 = A_1^1 \cap A_2^1$.

For any relabeling of the identities within the set O_1^1:

(1) by definition I_1 remains unchanged,

(2) event A_1^1 is unaffected since the set of all ϕ_1 functions in A_1^1 is mapped onto itself, and

(3) each ϕ_1 function in A_2^1 is unchanged.

Hence the event A_1, and the null probability of A_1, is unaffected over all such relabelings in O_1^1

The objective at this point is to establish the conditional null probability of any ϕ_2-ordering on O_1^1 given event A_1 has occurred. Since the null probability of A_1 is not changed under any relabeling in O_1^1, then it is not changed over all ϕ_2-orderings on O_1^1. Therefore, it suffices to determine the joint null probability

$$p_1 = P[Z_1^1 \le z_1,\ Z_2^1 \le z_2, \ldots, Z_{t_1+i_1-1}^1 \le z_{t_1+i_1-1},\ A_1]$$

for any set of real-component vectors $z_1, z_2, \ldots, z_{t_1+i_1-1}$. Now, p_1 is actually a joint probability of functions involving all random vectors in O_{n+m} since I_1, hence $\phi_1(Z, I_1)$ is defined on O_{n+m}.

If p_1 can be shown to be symmetric in $z_1, z_2, \ldots, z_{t_1+i_1-1}$ then the joint conditional null probability

$$P[Z_1^1 \le z_1,\ Z_2^1 \le z_2, \ldots, Z_{t_1+i_1-1}^1 \le z_{t_1+i_1-1} \mid A_1]$$

would be a symmetric function in $z_1, z_2, \ldots, z_{t_1+i_1-1}$. Then by applying Lemma I-2 it can be shown that all ϕ_2-orderings on O_1^1 given A_1 are equally-likely. Then the determination of the frequency count distribution within one of the two new blocks formed at the second stage given the event A_1 can be made.

First, to show that p_1 is indeed symmetric in $z_1, z_2, \ldots, z_{t_1+i_1-1}$ an application of Lemma I-1 will be made. Using the notation established in Lemma I-1 (also replacing the W_i's with Z_1^1's and setting $q = n+m$) define the following $K = n + m + t_1 + i_1 - 2$ functions:

$$g_i(Z_1, Z_2, \ldots, Z_{n+m}) = \begin{cases} Z_i^1 & i = 1, 2, \ldots, t_1 + i_1 - 1 \\ \phi_1(Z_{i-t_1-i_1+1}^1, I_1) & i = t_1+i_1, \ldots, 2t_1+2i_1-2 \\ \phi_1(Z_{i-2t_1-2i_1+2}^2, I_1) & i = 2t_1+2i_1-1, \ldots, n+m+t_1+i_1-2 \end{cases}$$

Appealing directly to Lemma I-1, it follows p_1 is symmetric in $z_1, z_2, \ldots, z_{t_1+i_1-1}$. Thus the joint conditional null distribution

$$P[Z_1^1 \le z_1,\ Z_2^1 \le z_2, \ldots, Z_{t_1+i_1-1} \le z_{t_1+i_1-1} \mid A_1]$$

is symmetric in $z_1, z_2, \ldots, z_{t_1+i_1-1}$.

Now apply Lemma I-2 (replacing the W_i's with the conditional Z_i^1's given A_1, set $t = q = t_1 + i_1 - 1$, and $\phi_2(Z, I_2) = g(Z, F)$). This proves that the joint conditional null distribution of $\phi_2(Z_1^1, I_2)$, $\phi_2(Z_2^1, I_2), \ldots, \phi_2(Z_{t_1+i_1-1}^1, I_2)$ given A_1 is a symmetric function. Therefore, all possible ϕ_2-orderings on O_1^1 given A_1 are unique (by choice of ϕ_2) and equally-likely.

Let S_2 denote the set of all ϕ_2-orderings on O_1^1 given A_1. Then each element in S_2 has a conditional null probability of $1/(t_1 + i_1 - 1)!$ of occurring. Then the conditional null probability that exactly s_1 observations on O_m fall within B_1^2 given that t_1 observations on O_m fell within B_1^1 is exactly the same as the null probability of observing an element in S_2 which assigns the $(s_1 + i_2)^{nd}$ ϕ_2-order position to O_n and exactly s_1 of the first $s_1 + i_2 - 1$ ϕ_2-order positions to O_m. By the same type argument used at the first stage, this probability becomes

$$P(s_1 \mid t_1) = \binom{s_1+i_2-1}{s_1}\binom{t_1-s_1+i_1-1-i_2}{t_1-s_1}\Big/\binom{t_1+i_1-1}{t_1}$$

$$\text{for } s_1 = 0, 1, \ldots, t_1 .$$

Multiplying by $P(t_1)$, derived in stage 1, and substituting $t_1 - s_1 = s_2$ and $m - t_1 = s_3$, then the joint null distribution of s_1, s_2, s_3, the respective block frequency counts on O_m, is

$$P(s_1, s_2, s_3) = \binom{s_1+i_2-1}{s_1}\binom{s_2+i_1-i_2-1}{s_2}\binom{s_3+n-i_1}{s_3}\Big/\binom{m+n}{m}$$

for non-negative integers s_i, $i = 1,2,3$ such that $s_1 + s_2 + s_3 = m$.

Similarly, if block B_2^1 had been chosen for division, the I_2 would contain at least all information symmetric with respect to B_2^1. Then for $i_2 \in I_{n-i}$ and $\phi_2(Z,I_2)$ the new blocks formed would be a division of B_2^1 into B_2^2 and B_3^2 where the first stage block B_1^1 would be designated as B_1^2. There are i_1-1, i_2-1, and $n-i_2$ observations in $\bar{o}_n$ falling within the respective blocks. Then if s_1,s_2,s_3 denote the respective block frequency count on O_m, their joint null distribution would become

$$P(s_1,s_2,s_3) = \binom{s_1+i_1-1}{s_1}\binom{s_2+i_2-1}{s_2}\binom{s_3+n-i_2}{s_3}\Big/\binom{m+n}{m}$$

for all non-negative integers s_i, $i = 1,2,3$ such that $s_1 + s_2 + s_3 = m$.

The information now available for starting the third stage in B^* consists of I_2, c_2, x_2^*, and the unidentified observation sets $\bar{o}_1^2$, $\bar{o}_2^2$, and $\bar{o}_3^2$ if all blocks are to be further divided at later stages. Otherwise, if the decision were made at stage 2 to never divide any one (or two) of the blocks B_1^2, B_2^2, B_3^2 at any later stage, then the observations set(s) associated with the block(s) selected can be identified completely. This information is symmetric on the remaining blocks and would then be made available at the start of stage 3. It should be noted also, that the information I_2 contains all information available at the start of stage 2, namely: I_1, c_1, x_1^* and the unidentified observation sets $\bar{o}_1^1$ and $\bar{o}_2^1$.

In the r^{th} stage ($r \leq n$) in the process B^* let the blocks

$$B_1, B_2, \ldots, B_{r+1}^r$$

be formed and e_i, $i = 1,2,\ldots,r+1$ denote the number of observations in $\bar{o}_n$ lying within the respective blocks. Assert that the joint null distribution of $m_1, m_2, \ldots, m_{r+1}$, the respective block frequency counts on O_m is

$$P(m_1, m_2, \ldots, m_{r+1}) = \{\prod_{i=1}^{r+1} \binom{m_i + e_i}{m_i}\}/\binom{m+n}{m}$$

for all non-negative integers m_i, $i = 1,2,\ldots,r+1$ such that $m_1 + m_2 + \ldots + m_{r+1} = m$. This assertion is verified by mathematical induction.

The fact that this assertion holds for $r = 1,2$ has been shown above. Now, assume it holds true for the $(r-1)^{st}$ stage and all previous stages. Let $B_1^{r-1}, B_2^{r-1}, \ldots, B_r^{r-1}$ denote the blocks formed at the $(r-1)^{st}$ stage and h_i is the number of observations in $\bar{o}_n$ contained in block B_i^{r-1}, $i=1,2,\ldots,r$. Then if $s_1, s_2, \ldots, s_r$ denote the respective block frequency counts on O_m, by assumption, the joint null distribution of $s_1, s_2, \ldots, s_r$ is

$$P(s_1, s_2, \ldots, s_r) = \{\prod_{i=1}^{r} \binom{s_i + h_i}{s_i}\}/\binom{m+n}{m}$$

for all non-negative s_i, $i = 1,2,\ldots,r$ such that $s_1 + s_2 + \ldots + s_r = m$.

The information available at the start of the r^{th} stage consists of I_{r-1}, c_{r-1}, x^*_{r-1}, and the two new unidentified observation sets obtained on the two new blocks formed at the $(r-1)^{st}$ stage. Now, I_{r-1} contains all information that was available at all previous stages. In particular, $I_{r-1} \supset I_{r-2} \supset ,,, \supset I_2 \supset I_1$. Thus, if any blocks formed at some previous stage were chosen to never be divided in B^* then the corresponding identified observation sets is information contained in I_{r-1}. Furthermore, the information on the identified observation sets was symmetric with respect to all blocks which were divided at later stages, and hence, symmetric with respect to all blocks available for division at the $(r-1)^{st}$ and r^{th} stages.

Now suppose block B_j^{r-1} (for some $j = 1,2,\ldots,r$) is available and selected for division at the r^{th} stage.

Then determine I_r-containing all information which is symmetric with respect to B_j^{r-1} and symmetric separately with respect to all blocks in the set $\{B_1^{r-1}, B_2^{r-1}, \ldots, B_r^{r-1}\}$ that are intended to be decomposed at some later stage.

Using I_r, select $i_r \varepsilon J_{h_j}$ and a real-valued measurable function $\phi_r(Z, I_r)$ either having a continuous null distribution for $Z \varepsilon O_j^{r-1}$ or is such that there are no ties within the set of values $\{\phi_r(z, I_r) | z \varepsilon \bar{o}_j^{r-1}\}$

Let $x_r^* \varepsilon \bar{o}_n \cap \bar{o}_j^{r-1}$ be such that $\phi_r(x_r^*, I_r) = c_r$ is the i_r^{th} largest value in the set $\{\phi_r(x, I_r) | x \varepsilon \bar{o}_j^{r-1}\}$. The cutting function $\phi_r(x, I_r) = c_r$ divides the block B_j^{r-1} into

$$B_j^r = \{x \varepsilon B_j^{r-1} | \phi_r(x, I_r) < c_r\}$$

and

$$B_{j+1}^r = \{x \varepsilon B_j^{r-1} | \phi_r(x, I_r) > c_r\}$$

For consistency in notation, the remaining blocks defined at the $(r-1)^{st}$ stage are relabeled:

$$B_i^r = B_i^{r-1} \quad i = 1,2,\ldots,j-1$$

and

$$B_i^r = B_{i-1}^{r-1} \quad i = j+2,\ldots,r+1$$

The event A_{r-1} that the respective block frequency counts at the $(r-1)^{st}$ stage was $s_1, s_2, \ldots, s_r$ is determined by considering the following facts.

In the construction process B^*, each block B_i^{r-1} $i = 1,2,\ldots,r$ was originally established either at the $(r-1)^{st}$ stage or some earlier stage by dividing some block previously established, and each of those blocks were formed by dividing some block established yet earlier, etc. For each block B_i^{r-1} consider only those stages in the process B^* for which one of the two blocks newly formed at the stage contains the block B_i^{r-1}. Then each block B_i^{r-1} can be associated with a unique subset of integers in $\{1,2,\ldots,r-1\}$ such that each integer within the subset represents a stage level in which two new blocks were defined (from the set of blocks established at previous levels), one of which contains block B_i^{r-1}. If j_1 were in the subset associated with block B_i^{r-1} then either

$$B_i^{r-1} \subset \{x \in X \mid \phi_{j_1}(x, I_{j_1}) < c_{j_1}\}$$

or

$$B_i^{r-1} \subset \{x \in X \mid \phi_{j_1}(x, I_{j_1}) > c_{j_1}\} .$$

Thus the subset of integers associated with B_i^{r-1} can be partitioned into two unique subsets $\{a_1(i), a_2(i), \ldots, a_{u(i)}(i)\}$ and $\{b_1(i), b_2(i), \ldots, b_{v(i)}(i)\}$ such that the block B_i^{r-1} is defined by

$$B_i^{r-1} = \{x \in X \mid \phi_{a_1(i)}(x, I_i) < c_{a_1(i)},\ \phi_{a_2(i)}(x, I_i) < c_{a_2(i)}, \ldots,$$

$$\phi_{a_{u(i)}(i)}(x, I_i) < c_{a_{u(i)}(i)},\ \phi_{b_1(i)}(x, I_i) > c_{b_1(i)}, \ldots,$$

$$\phi_{b_{v(i)}(i)}(x, I_i) > c_{b_{v(i)}(i)}\} \quad \text{for } i = 1,2,\ldots,r.$$

As before define $O_i^{r+1} = \{Z_1^i, Z_2^i, \ldots, Z_{s_i+h_i}^i\}$ as the set of random vectors yielding observations in block B_i^{r-1}, $i = 1,2,\ldots,r$. Then consider the events defined at the $(r-1)^{st}$ stage

$$D_{i,k}^- = \{\phi_{a_1(i)}(Z_k^i, I_i) < c_{a_1(i)}, \ldots, \phi_{a_{u(i)}(i)}(Z_k^i, I_i) < c_{a_{u(i)}(i)}\}$$

and

$$D_{i,k}^+ = \{\phi_{b_1(i)}(Z_k^i, I_i) > c_{b_1(i)}, \ldots, \phi_{b_{v(i)}(i)}(Z_k^i, I_i) > c_{b_{v(i)}(i)}\}$$

for $k = 1,2,\ldots,s_i + h_i$ and $i = 1,2,\ldots,r$.

Then the event of obtaining s_i observations in $\bar{o}_m$ in B_i^{r-1} is equivalent to the event that $D_{i,k}^-$ and $D_{i,k}^+$ hold simultaneously for $k = 1,2,\ldots,$ $s_i + h_i$. This event can be expressed as

$$A_i^{r-1} = \bigcap_{k=1}^{s_i+h_i} [D_{i,k}^- \cap D_{i,k}^+].$$

The event of obtaining $s_1,s_2,\ldots,s_r$ observation in $\bar{o}_m$ in the blocks $B_1^{r-1},B_2^{r-1},\ldots,B_r^{r-1}$, respectively, is

$$A_{r-1} = \bigcap_{i=1}^{r} A_i^{r-1} \; .$$

Now, consider the random vectors $(Z_1^j,Z_2^j,\ldots,Z_{s_j+h_j}^j) = \mathcal{O}_j^{r-1}$ which yield observations in B_j^{r-1}. The information sets $\mathcal{I}_1,\mathcal{I}_2,\ldots,\mathcal{I}_{r-1}$ are defined to be symmetric on B_j^{r-1}, otherwise B_j^{r-1} would not be available for cutting at the r^{th} stage. Then the ϕ-functions within the events A_i^{r-1} for $i = 1,2,\ldots,r$ and $i \neq j$ are unchanged over any relabeling within $\mathcal{O}_j^{r-1}$, thus the events are unchanged. The event A_j^{r-1} is mapped onto itself by relabeling within $\mathcal{O}_j^{r-1}$. Hence, A_{r-1} and the probability of A_{r-1} is unaffected by any such relabeling within $\mathcal{O}_j^{r-1}$. Then the joint null probability

$$P[Z_1^j \leq z_1,\ Z_2^j \leq z_2,\ldots,Z_{s_j+h_j}^j \leq z_{s_j+h_j},\ A_{r-1}]$$

is symmetric in the vectors $z_1,z_2,\ldots,z_{s_j+h_j}$ by Lemma I-1. Therefore, the joint conditional null cumulative distribution of $Z_1^j,Z_2^j,\ldots,Z_{s_j+h_j}^j$ given A_{r-1} is symmetric in the vectors $z_1,z_2,\ldots,z_{s_j+h_j}$. Then by applying Lemma I-2 the joint null distribution of $\phi_r(Z_1^j,\mathcal{I}_r)$, $\phi_r(Z_2^j,\mathcal{I}_r),\ldots,$ $\phi_r(Z_{s_j+h_j}^j,\mathcal{I}_r)$ given A_{r-1} is found to be symmetric. It follows that all possible ϕ_r-orderings on $\mathcal{O}_j^{r-1}$ are equally-likely. Using the same type argument employed before, the conditional null probability that m_j observations in $\bar{o}_m$ fall within block B_j^r given the block frequency counts $s_1,s_2,\ldots,s_r$ determined at the $(r-1)^{st}$ stage becomes

$$P(m_j|s_1,s_2,\ldots,s_r) = \binom{m_j + i_{r-1}}{m_j}\binom{s_j-m_j+h_j-i_r}{s_j-m_j}\bigg/\binom{s_j + h_j}{s_j}$$

for $m_j = 0,1,\ldots,s_j$.

Observe that (in terms of the notation defined for the r^{th} stage) the following equalities hold:

$$e_i = h_i \text{ and } m_i = s_i \quad \text{for } i = 1,2,\ldots,j-1,$$

$$e_i = h_{i-1} \text{ and } m_i = s_{i-1} \quad \text{for } i = j+1,\ldots,r+1,$$

$$m_j + m_{j+1} = s_j, \quad i_r-1 = e_j, \quad \text{and } h_j - i_r = e_{j+1}$$

Multiplying the above conditional null probability by the joint probability $P(s_1,s_2,\ldots,s_r)$ and using the above equalities, the joint null distribution of $m_1,m_2,\ldots,m_{r+1}$ becomes

$$P(m_1,m_2,\ldots,m_{r+1}) = \left\{\prod_{i=1}^{r+1}\binom{m_i + e_i}{m_i}\right\}\Big/\binom{m + n}{m}$$

for all non-negative m_i, $i = 1,2,\ldots,r+1$ such that $m_1 + m_2 + \ldots + m_{r+1} = m$.

This completes the proof of the assertion.

Now by the definition given in Chapter IV, the basic blocks are the n + 1 statistically equivalent blocks obtained if the process B^* could be continued through the n^{th} stage. In this case, each B_i^r would be further divided until all the observations in $\bar{o}_n$ lying in B_i^r were consumed. Thus, if block B_i^r contained e_i observations in $\bar{o}_n$ then there would eventually be $e_i + 1$ basic blocks formed within B_i^r. Substituting $k_i = e_i + 1$ for $i = 1,2,\ldots,r+1$ in the above probability expression gives the desired results.

LIST OF REFERENCES

[1] Anderson, T. W., "A Method of Constructing Nonparametric Multivariate Tests (Preliminary Report)," Annals of Mathematical Statistics, 26, 773, 1955.

[2] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, New York: John Wiley & Sons, Inc., 1958.

[3] Anderson, T. W., "Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks," Multivariate Analysis (Proceedings of an International Symposium Held in Dayton, Ohio, June 14-19, 1965), New York: Academic Press, 5-27, 1966.

[4] Auble, D., "Extended Tables for the Mann-Whitney Statistic," Bulletin of the Institute of Educational Research at Indiana University, 1, No. 2, 1-39, 1953.

[5] Dixon, W. J., "A Criterion for Testing the Hypothesis that Two Samples are from the Same Population," Annals of Mathematical Statistics, 11, 199-204, 1940.

[6] Fraser, D.A.S., "Sequentially Determined Statistically Equivalent Blocks," Annals of Mathematical Statistics, 22, 372-381, 1951.

[7] Fraser, D.A.S., "Nonparametric Tolerance Regions," Annals of Mathematical Statistics, 24, 44-55, 1953.

[8] Fraser, D.A.S. and Gutman, I., "Tolerance Regions," Annals of Mathematical Statistics, 27, 162-79, 1956.

[9] Kemperman, J. H. B., "Generalized Tolerance Limits," Annals of Mathematical Statistics, 27, 180-6, 1956.

[10] Mann, H. B. and Whitney, D. R., "On a Test of Whether One of Two Variables is Stochastically Larger Than the Other," Annals of Mathematical Statistics, 18, 50-60, 1947.

[11] Mathisen, H. C., "A Method of Testing the Hypothesis That Two Samples are from the Same Population," Annals of Mathematical Statistics, 14, 188-94, 1943.

[12] Tukey, J. W., "Nonparametric Estimation, II. Statistically Equivalent Blocks and Tolerance Regions - the Continuous Case," Annals of Mathematical Statistics, 18, 529-39, 1947.

[13] Wald, A., "An Extension of Wilks' Method for Setting Tolerance Limits," Annals of Mathematical Statistics, 14, 45-55, 1943.

[14] Walsh, J. E., Handbook of Nonparametric Statistics, II, Princeton, N.J.: D. Van Nostrand Co., Inc., 1965.

[15] Walsh, J. E., "Generally Applicable Limited Length Sequential Permutation Tests for One-Way ANOVA," THEMIS Report No. 80, 1970.

[16] Walsh, J. E., "Always Applicable Sequential Randomization Tests for One-Way ANOVA That Emphasize the More Recent Data," THEMIS Report No. 83, 1970.

[17] Wilks, S. S., "On the Determination of Sample Sizes for Setting Tolerance Limits," Annals of Mathematical Statistics, 12, 91-96, 1941.

[18] Wilks, S. S., "Statistical Prediction with Special Reference to the Problem of Tolerance Limits," Annals of Mathematical Statistics, 13, 400-9, 1942.

[19] Wilks, S. S., Mathematical Statistics, New York: John Wiley & Sons, Inc., 1962.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* SOUTHERN METHODIST UNIVERSITY	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP UNCLASSIFIED

3. REPORT TITLE

Sequential Multivariate Quality Control Tests Using Tolerance Regions

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Ronald L. Boase

John E. Walsh

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1, 1971	86	19

8a. CONTRACT OR GRANT NO. N00014-68-A-0515 b. PROJECT NO. NR 042-260 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 110 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY OFFICE OF NAVAL RESEARCH

13. ABSTRACT

Two new developments, multivariate sequential significance tests and a method of forming multivariate two-sample tolerance tests, are proposed. The sequential significance test considered is a fixed-length succession of two-sample subtests where each subtest reuses some or all of the data for preceding subtests. By proper choice of the subtest statistics and use of a permutation basis the subtests are made independent.

The class of multivariate two-sample tolerance tests developed by the proposed method are directly applicable as subtests in the sequential significance tests. The proposed method is based on a new technique of constructing tolerance regions for the two-sample problem. Subject to certain mild limitations the analyst may actually look at the combined observed vector values in order to construct the desired tolerance regions. This advantage can be used effectively in choosing the shapes of the tolerance regions so as to emphasize the alternative hypotheses associated with the selected test statistic. The only requirements are that the joint null distribution of the combined data be a symmetric function and that the observations be such that the proposed construction process is unique with probability one.