AMERICAN UNIVERSITY OF BEIRUT

SOME PROBLEMS IN INFERENCE FOR MULTINOMIAL POPULATIONS

By

Sad T. Bakir

Approved:

Assoc. Prof.
Advisor

Assist. Prof.
Member of Committee

Assist. Prof.
Member of Committee

Member of Committee

Date of Thesis Presentation: February 3, 1968.

SOME PROBLEMS IN INFERENCE FOR MULTINOMIAL POPULATIONS

By

Sad T. Bakir

Submitted in Partial Fulfillment for the Requirements

of the Degree Master of Science

in the Mathematics Department of the

American University of Beirut

Beirut, Lebanon

1968

SOME PROBLEMS IN INFERENCE FOR MULTINOMIAL POPULATIONS

By

Sad T. Bakir

PREFACE

The multinomial distribution appears to be an appropriate model for a variety of practical problems where observations are made on categorical data. Its wide applicability raises a number of new inference problems, some of which are direct extensions of standard procedures in bivariate populations, while others require special treatment.

This thesis treats two problems of inference connected with several multinomial populations. One problem is to decide whether or not several random samples belong to the same multinomial population. If they do not, the question arises as to which subset, if any, of those samples may be considered as belonging to the same population. This problem is then essentially the problem of homogeneity of multinomial populations.

The other problem arises in situations where one wishes to select a subset of the populations which, at a preassigned confidence level, contains a "best" population. The selection of such population has recently received considerable attention from statisticians.

The solutions offered in this thesis to the two problems lend themselves easily to numerical applications.

CONTENTS

CHAPTER I

REVIEW OF RELATED LITERATURE

The first of the two problems considered in this thesis is the problem of homogeneity of several multinomial populations. In particular, we consider m multinomial populations, each consisting of k classes, with parameters π_{ij} $(i = 1,\ldots,m;\ j = 1,\ldots,k)$ where π_{ij} is the probability that an individual from the ith population falls in the jth class. The hypothesis that the m populations are homogeneous is then the hyposthesis H: $\pi_{1j} = \ldots = \pi_{mj}$ for $j = 1,\ldots,k$. Given independent random samples of sizes $n_{i.}(i = 1,\ldots,m)$ from the populations, the usual test for H is based on the statistic $X^2 = \sum_{i,j} \frac{(n_{ij} - n_{i.}P_j)^2}{n_{i.}P_j}$ where n_{ij} is the frequency in the (i,j) th cell, and $P_j = \frac{n_{.j}}{N}$ with $n_{.j} = \sum_{i=1}^{m} n_{ij}$, $N = \sum_{i,j} n_{ij}$.

Under H, X^2 is asymptotically distributed as a chi-square variable with $(m-1)(k-1)$ degrees of freedom (see, for example, Cramer 1946).

Goodman (1964) proposed a test for the hypothesis H based on the statistic $Y^2 = \sum_{i,j} \frac{(n_{ij} - n_{i.}\,Q_j)^2}{n_{ij}}$, where the Q_j are obtained by minimizing Y^2 subject to $\sum_{j=1}^{k} Q_j = 1$. The values of Q_j are found to be $Q_j = \frac{\bar{P}_j}{\sum_{r=1}^{k} \bar{P}_r}$, where $\bar{P}_j = \frac{N}{\sum_{i=1}^{m} \frac{n_{i.}}{p_{ij}}}$ $(j = 1,\ldots,k)$, the weighted harmonic means of the proportions $p_{ij} = \frac{n_{ij}}{n_{i.}}$ occuring in the jth class over all the m populations. Under H, Y^2 is asymptotically distributed as a chi-square variable with $(m-1)(k-1)$

degrees of freedom, so that it is asymptotically equivalent to X^2. A useful computational formula of Y^2 is $Y^2 = N(\frac{1}{\sum_{j=1}^{k} \frac{1}{F_j}} - 1)$. It is important to note that Goodman's Y^2 - statistic is undefined in the presence of zero frequencies in some cells.

The admantages of Y^2 over X^2 lie in that the Y^2 - statistic admits certain properties similar to those of the F-test used in the analysis of variance context. These properties will be cited in the following pages.

In the event that the hypothesis of homogeneity is rejected, one is usually interested in more detailed decisions as to which populations are alike and which are not. In other words, it is then desired to divide the populations into distinct homogeneous subgroups.

Goodman's (1964) technique of multiple contrasts among the multinomial populations is one way to deal with the above problem, (contrasts in this problem were first discussed by Gold 1963). A contrast among the m multinomial populations is defined to be a linear function of the π_{ij}, $\theta = \sum_{i,j} c_{ij} \pi_{ij}$ subject to the condition that $\sum_{i=1}^{m} c_{ij} = 0$ for $j = 1, \ldots, k$. Then it is easy to see that the hypothesis of homogeneity H is equivalent to the hypothesis that all possible contrasts equal zero.

The maximum likelihood unbiased estimate of θ is $\hat{\theta} = \sum_{i,j} c_{ij} p_{ij}$ subject to the condition that $\sum_{i=1}^{m} c_{ij} = 0$ for $j = 1, \ldots, k$; and where the $p_{ij} = \frac{n_{ij}}{n_{i.}}$ are the maximum likelihood estimates of π_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, k)$. Simple calculations lead to: $\operatorname{var}(\hat{\theta}) = \sum_{i=1}^{m} \frac{1}{n_{i.}} \left[\sum_{j=1}^{k} c_{ij}^2 \pi_{ij} - \left(\sum_{j=1}^{k} c_{ij} \pi_{ij} \right)^2 \right]$ and its consistent

estimate $S^2(\hat{\theta}) = \sum_{i=1}^{m} \frac{1}{n_{i\cdot}} \left[\sum_{j=1}^{k} c_{ij}^2 p_{ij} - \left(\sum_{j=1}^{k} c_{ij} p_{ij} \right)^2 \right]$.

Goodman, in the same paper (1964), also proved two important theorems about the multiple contrasts and about the Y^2 - statistic. These theorems will be used to develop a procedure for dividing the set of the m populations into homogeneous subsets. The first theorem gives simultaneous confidence intervals, based on the chi-square distribution, for all possible contrasts θ. If the confidence interval for some contrast θ does not contain zero, the estimated contrast $\hat{\theta}$ is said to be significantly different from zero.

The second theorem relates the Y^2 - test to the simultaneous confidence intervals. It asserts that the Y^2 - statistic rejects the hypothesis H if and only if at least one estimated contrast is significantly different from zero.

Goodman points out that in view of the second theorem the simultaneous confidence intervals presented in the first theorem for all contrasts can be used to supplement the test for the hypothesis of homogeneity based on Y^2 in the following way: If the test based on Y^2 leads to acceptance of H, then all confidence intervals would contain zero. But if the test rejects H, we could then calculate the simultaneous confidence intervals to determine which particular contrasts are significantly different from zero and thus determine the particular ways in which the multinomial populations are not homogeneous. Before examining particular contrasts one should calculate Y^2 to determine whether there are significant contrast at all. Finally, Goodman points out why a result similar to the second theorem does not hold for the usual X^2 - statistics.

Theorems analogous to those of Goodman were earlier obtained by Scheffé (1953, 1959) in the analysis of variance context. In this case the F-test plays the role of the Y^2 test, and the F-distribution takes the place of the chi-square distribution. If the F-test rejects the hypothesis of homogeneity (equality of means) of normal populations, the Scheffé multiple contrast technique is used to detect the reasons for heterogeneity.

Both Scheffé's and Goodman's contrast procedures may be criticized on the grounds that one cannot examine all possible contrasts since there are infinitely many; hence one would be unable to decide whether a particular subset of the m populations (whether in the normal or multinomial case) is homogeneous or not.

In the analysis of variance context the difficultly of applying Scheffé's procedure is overcome by a technique proposed by Gabriel(1964). Essentially, he suggests that to determine the homogeneity, or lack of it,in a subset R of the normal populations, we need only examine a single statistic. The statistic is chosen such that some estimated contrast in R is significantly different from zero iff this statistic is significantly large. Gabriel's statistic is the between samples S.S. in the set R, S_R^2. R is judged significant iff $S_R^2 > s^2(m-1)\ F_\alpha;\ (m-1)(n-m)$ where s^2 is the sample estimate (error S.S) of the common unknown variance of the populations. Gabriel also describes a systematic way for classifying the populations into homogeneous subsets.

A technique analogous to Gabriel's is developed in Chapter II to deal with the similar problem in the case of multinomial populations.

A second direction in which a statistician may wish to go

beyond the traditional tests of homogeneity (e.g. Fisher's analysis of variance, Bartlett's test for homogeneity of variances, homogeneity in multinomial populations, etc.) is the problem of ordering the populations in some order of preference. In an agricultural experiment, for example, the prospect is to choose a best variety among several.

One of the first attempts to deal with such situations is that by Mosteller (1948) who considers the hypothesis of homogeneity against a specific alternative, known as the slippage alternative. Formally, given populations, $f(x - a_1)$, ..., $f(x - a_m)$ which are identical except for translation; it is required to test the hypothesis H: $a_1 = \ldots = a_m$ against the alternative K: for some i $a_i > \max(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_m)$. While the hypothesis H implies homogeneity, K implies that the ith population, which may be called the best population, has slipped farther to the right than any of the others. Mosteller gave a nonparametric test for this problem.

Paulson (1949,1952) devised, in the first paper, a rule for classifying m normal populations into a "superior", and an "inferior" group according to the values of their means. If all populations fall into one group, he calls it "neutral" and decides that the populations are homogeneous. In the second paper a procedure is given for determining the best among normal populations, and among binomial populations using the inverse sine transformation.

Bechhofer, in a series of papers starting in (1954), formulates the problem of ranking populations (with special attention paid to normal populations) as follows: The experimenter singles out a

parameter $u_i (i = 1, \ldots, m)$ according to which the populations should be ordered. He then specifies an indifference zone in the parameter space, $\frac{u_i}{u_r} < t$ (for some pair (i, r)). The main point in the procedure is to determine the common sample size n which guarantees that the probability of correct ranking is at least a preassigned value P. Then samples of size n are drawn from each population and the estimates $\hat{u}_i$ for u_i are calculated. The population associated with the largest $\hat{u}_i$ is asserted to be the best, the one with the second largest $\hat{u}_i$ is asserted to be the second best, and so on. Values of n are tabulated to carry out the procedure in different situations. Sequential procedures are given by Bechhofer and Sobel (1954), Bechhofer (1958), and Bechhofer and Blumenthal (1962).

A similar approach involving the selection of a subset of the given populations differs from Bechhofer's in that it assumes that the number of observations from each population is given. The approach is due to Gupta and Sobel(1958). The sample estimates $\hat{u}_i$ for the u_i $(i = 1, \ldots, m)$ are computed. Then the procedure is to retain every population for which $\hat{u}_i \geq \max_i (\hat{u}_i) - c$, where c is a constant chosen so that the probability of retaining the population with the largest u_i is at least a preassigned level P. The final decision is then, the selection of a subset of the populations which is asserted to contain the best population.

Gupta and Sobel (1960) have tabulated the different values of the constant c appropriate for carrying out the procedure in different situations.

Guttman (1963) supplements the Gupta-Sobel procedure with

a sequential procedure where at each stage we retain fewer populations until we are left with a single population asserted to be the best.

A Baysian approach to the best population problem is adopted by Guttman and Tiao (1964). Exponential and normal populations are considered with special attention.

Studden (1967) discusses the selection problem in terms of decision functions and characterizes optimal selection subset rules. Denoting by L_i the loss when the ith population is retained, and considering the additional loss when an incorrect selection is made (i.e. the selected subset does not contain the best population) he finds a formula for the risk R; and the problem is to find the procedure which minimizes R. Among invariant decision rules Studden characterizes the best invariant decision rule.

CHAPTER II

DIVIDING A SET OF MULTINOMIAL POPULATIONS INTO HOMOGENEOUS SUBSETS

Suppose that we have m multinomial populations of k classes each, with true parameters $(\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik})$, $i = 1, \ldots, m$. Given samples $(n_{i1}, n_{i2}, \ldots, n_{ik})$, $i = 1, \ldots, m$, of sizes $\sum_{j=1}^{k} n_{ij} = n_{i.}$ from each population, it is required to divide the populations into homogeneous subsets. By homogeneity in a subset R of the m populations is meant that the hypothesis H: $\pi_{rj} = \pi_{sj}$ holds true for all $j = 1, \ldots, k$ and all $r, s \in R$. A procedure for testing the homogeneity of any subset of the m populations is given here, and by this procedure it will be possible to distinguish the homogeneous subsets.

The procedure is parallel to that of Gabriel (1964) in the sense that Gabriel's procedure is based on Scheffé's contrasts among normal populations, while the present procedure is based on Goodman's contrasts among multinomial populations.

The Procedure:

As it was defined in chapter I, a contrast among the m multinomial populations is a linear function of the parameters π_{ij}, $\theta = \sum_{i,j} c_{ij} \pi_{ij}$ subject to the restriction that $\sum_{i=1}^{m} c_{ij} = 0$ for all j. The maximum likelihood unbiased estimate

of θ is $\hat{\theta} = \sum_{i,j} c_{ij} p_{ij}$, where $p_{ij} = \frac{n_{ij}}{n_{i.}}$ are the maximum likelihood unbiased estimates of π_{ij}, i = 1, ..., m; j = 1 ... k. Before examining subsets of the m populations, one should test the homogeneity of the whole set of populations to see whether differences exist at all. To accomplish this we propose to use Goodman's statistic $Y^2 = \sum_{i,j} \frac{(n_{ij} - n_{i.} Q_j)^2}{n_{ij}} = N(\frac{1}{\sum_{j=1}^{K} \bar{P}_j} - 1)$ where $N = \sum_{i,j} n_{ij}$ and $\bar{P}_j$, Q_j are as defined in chapter I. Under the hypothesis H: $\pi_{1j} = \pi_{2j} = \cdots = \pi_{mj}$ for all j, Y^2 has chi-square distribution with (m-1)(k-1) degrees of freedom.

We need the following two theorems due to Goodman (1964, pp. 718 and 721):

Theorem 1: As $\sum_{i=1}^{m} n_{i.} = N$ tends to infinity, the probability will approach $(1 - \alpha)$ that simultaneously for all functions θ

$$\hat{\theta} - S(\hat{\theta})\, L \leqslant \theta \leqslant \hat{\theta} + S(\hat{\theta})L \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where L is the positive square root of the upper α point of the chi-square distribution with (m-1)(k-1) degrees of freedom, and $S(\hat{\theta})$ is the estimated standard deviation of $\hat{\theta}$, whose explicit form was given in chapter I.

From the inequalities (1), it is seen that the confidence interval for a contrast θ does not contain zero if and only if $|\hat{\theta}| > S(\hat{\theta})L$ which is equivalent to $\hat{\theta}^2 > S^2(\hat{\theta})L^2$, or $\frac{\hat{\theta}^2}{S^2(\hat{\theta})} > L^2$. If $\frac{\hat{\theta}^2}{S^2(\hat{\theta})} > L^2$ it is said that the estimated contrast θ is significantly different from zero or, briefly, significant.

Theorem 2: The test of homogeneity based on the Y^2 statistic rejects the hypothesis of homogeneity of the m populations if and

only if at least one estimated contrast is significant.

Therefore, to determine the reasons for rejecting the hypothesis, one may interpret theorem 2 as suggesting a search for a significant estimated contrast; this, however, cannot be done by examining all the possible contrasts, since there are infinitely many.

Consider a subset R of the m populations. A contrast among the populations in R is a linear function $\theta = \sum_{i=1}^{m} \sum_{j=1}^{k} c_{ij} \pi_{ij}$ subject to the conditions: $\sum_{i \in R} c_{ij} = 0$ for all j and $c_{ij} = 0$ for all i not in R. So that a contrast in R is also a contrast in the set of all the m populations, say M.

The subset R is homogeneous if and only if all contrasts in R equal zero. For, assuming homogeneity, we have $\theta = \sum_{i=1}^{m} \sum_{j=1}^{k} c_{ij} \pi_{ij}$ $= \sum_{j} (\pi_{.j} \sum_{i \in R} c_{ij}) = 0$, where $\pi_{.j}$ is the common value of the probabilities in the jth class. Conversely, assume that all contrasts equal zero and suppose $\pi_{ej} \neq \pi_{sj}$ for some e, s in R and some j. Then the contrast $\pi_{ej} - \pi_{sj} \neq 0$ which contradicts the assumption that all contrasts equal zero.

Moreover, in view of theorem 1, if an estimated contrast is significantly different from zero, then the $(1 - \alpha)$ confidence interval for the true contrast does not contain zero; hence the probability is at least $(1 - \alpha)$ that the true contrast is not zero. So it is reasonable to judge the subset R heterogeneous if and only if at least one estimated contrast in R is significantly different from zero.

We recall that an estimated contrast $\hat{\theta}$ is significant if and only if $\frac{\hat{\theta}^2}{S^2(\hat{\theta})} > L^2$, so that to detect the existence of a singificant

estimated contrast we need only examine the maximum of the ratios $\frac{\hat{\theta}^2}{S^2(\hat{\theta})}$. If this maximum is smaller than L^2, then all other ratios are smaller than L^2 and, therefore, no estimated contrast is significant, which implies the homogeneity of R. On the other hand if the maximum is larger than L^2, then the particular contrast associated with this maximum ratio is significant, beside possibly other contrasts, which implies heterogeneity of R.

The problem, then, reduces to maximizing $\frac{\hat{\theta}^2}{S^2(\hat{\theta})}$

$$= \frac{(\sum_{i=1}^{m}\sum_{j=1}^{k} c_{ij}\, p_{ij})^2}{\sum_{i=1}^{m}\frac{1}{n_{i\cdot}}\sum_{j=1}^{k} c_{ij}^2 p_{ij} - \sum_{i=1}^{m}\frac{1}{n_{i\cdot}}(\sum_{j=1}^{k} c_{ij}p_{ij})^2}$$ with respect to c_{ij}

subject to the constraints: $\sum_{i\in R} c_{ij} = 0$ for all j, and $c_{ij} = 0$ for $i \notin R$. Let $w_1, w_2, \ldots, w_k$ and b_{ij}, $i \notin R$, $j = 1 \ldots k$, be constant multipliers. Let

$$F(c_{ij}) = \frac{\hat{\theta}^2}{S^2(\hat{\theta})} + w_1 \sum_{i\in R} c_{i1} + \cdots + w_s \sum_{i\in R} c_{is} + \cdots + w_k \sum_{i\in R} c_{ik} + \sum_{i\notin R}\sum_{j=1}^{k} b_{ij}\, c_{ij}.$$

Differentiating $F(c_{ij})$ with respect to c_{rs} for all $r \in R$ and $s = 1, \ldots, k$, and equating the results to zero, we get:

$$\frac{\partial F}{\partial c_{rs}} = \frac{1}{S^4(\hat{\theta})}\left[2\hat{\theta}\, S^2(\hat{\theta})\, p_{rs} - 2\hat{\theta}^2\left(\frac{c_{rs}p_{rs}}{n_{r\cdot}} - \frac{p_{rs}}{n_{r\cdot}}\sum_{j=1}^{k} c_{rj}\, p_{rj}\right)\right] + w_s = 0 \quad \ldots(2)$$

multiplying (2) by $\frac{n_{r\cdot}}{p_{rs}}$ and summing over r in R we get:

$$\frac{2}{S^4(\hat{\theta})}\left[\hat{\theta}\, S^2(\hat{\theta}) \sum_{r\in R} n_{r\cdot} - \hat{\theta}^2\left(-\sum_{r\in R}\sum_{j=1}^{k} c_{rj}\, p_{rj}\right)\right] + w_s \sum_{r} \frac{n_{r\cdot}}{p_{rs}} = 0$$

Let

$$\frac{N_R}{\sum_{r\in R}\frac{n_{r\cdot}}{p_{rs}}} = \bar{p}_s, \text{ where } N_R = \sum_{r\in R} n_{r\cdot} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

We get:

$$\frac{2}{S^4(\hat{\theta})}\left[N_R\,\hat{\theta}\,S^2(\hat{\theta}) + \hat{\theta}^3\right] + w_s\,\frac{N_R}{\bar{P}_s} = 0.$$

Solving for w_s, we find $w_s = -\frac{2\bar{P}_s}{S^4(\hat{\theta})}\left[\hat{\theta}\,S^2(\hat{\theta}) + \frac{\hat{\theta}^3}{N_R}\right]$, and substituting this value in (2) we get:

$$\frac{1}{S^4(\hat{\theta})}\Big[2\hat{\theta}\,S^2(\hat{\theta})\,p_{rs} - 2\,\hat{\theta}^2\left(\frac{c_{rs}\,p_{rs}}{n_{r.}} - \frac{p_{rs}}{n_{r.}}\sum_{j=1}^{k} c_{rj}\,p_{rj}\right)$$

$$- 2\,\bar{P}_s\left(\hat{\theta}\,S^2(\hat{\theta}) + \frac{\hat{\theta}^3}{N_R}\right)\Big] = 0$$

which can be written as:

$$S^2(\hat{\theta})p_{rs} - \hat{\theta}\left(\frac{c_{rs}p_{rs}}{n_{r.}} - \frac{p_{rs}}{n_{r.}}\sum_{j=1}^{k} c_{rj}\,p_{rj}\right) - \bar{P}_s\left(S^2(\hat{\theta}) + \frac{\hat{\theta}^2}{N_R}\right) = 0 \quad \ldots\ldots \quad (4)$$

Summing (4) over S = 1, 2, ..., k we get:

$$S^2(\hat{\theta}) - \sum_{s=1}^{k} \bar{P}_s\left(S^2(\hat{\theta}) + \frac{\hat{\theta}^2}{N_R}\right) = 0$$

from which we see that

$$\frac{1}{\sum_{s=1}^{k} \bar{P}_s} = \frac{S^2(\hat{\theta}) + \frac{\hat{\theta}}{N_R}}{S^2(\hat{\theta})} = 1 + \frac{\hat{\theta}^2}{N_R\,S^2(\hat{\theta})} \quad \ldots\ldots\ldots \quad (5)$$

From (4) above we solve for c_{rs} to get:

$$\hat{\theta}\,c_{rs}\,p_{rs} = n_{r.}\left[S^2(\hat{\theta})p_{rs} + \frac{\hat{\theta}p_{rs}}{n_{r.}}\sum_{j=1}^{k} c_{rj}p_{rj} - \bar{P}_s\left(S^2(\hat{\theta}) + \frac{\hat{\theta}^2}{N_R}\right)\right.$$

$$= n_{r.}\left[S^2(\hat{\theta})(p_{rs}-P_s) + \frac{\hat{\theta}\,p_{rs}}{n_{r.}}\sum_{j=1}^{k} c_{rj}p_{rj} - \frac{\bar{P}_s}{N_R}\hat{\theta}^2\right] \quad \ldots\ldots \quad (6)$$

Now $$\frac{\hat{\theta}^2}{S^2(\hat{\theta})} = \frac{\left(\sum_{r\in R}\sum_{s=1}^{k} c_{rs}p_{rs}\right)\left(\sum_{i=1}^{m}\sum_{j=1}^{k} c_{ij}p_{ij}\right)}{S^2(\hat{\theta})} = \frac{\sum_{r\in R}\sum_{s=1}^{k} \hat{\theta}c_{rs}p_{rs}}{S^2(\hat{\theta})}$$

and it attains its maximum when $\hat{\theta}c_{rs}p_{rs}$ is given by (6).

Substituting for $\hat{\theta}c_{rs}p_{rs}$ from (6) in the expression for $\frac{\hat{\theta}^2}{S^2(\hat{\theta})}$ we

have

$$\max_{c_{ij}} \frac{\hat{\theta}^2}{S^2(\hat{\theta})} = \frac{1}{S^2(\hat{\theta})} \sum_{r \in R} \sum_{s=1}^{k} \Big[S^2(\hat{\theta}) n_{r.}(p_{rs} - \bar{P}_s) + \hat{\theta} p_{rs} \sum_{j=1}^{k} c_{rj} p_{rj} - \frac{n_{r.}}{N_R} \bar{P}_s \hat{\theta}^2 \Big]$$

we note that:

$$\sum_{r \in R} \sum_{s=1}^{k} n_{r.}(p_{rs} - \bar{P}_s) = \sum\sum n_{r.}\left(\frac{n_{rs}}{n_{r.}} - \bar{P}_s\right) = N_R\left(1 - \sum_{s=1}^{k} \bar{P}_s\right),$$

$$\sum_{r \in R} \sum_{s=1}^{k} \hat{\theta} p_{rs}\left(\sum_{j=1}^{k} c_{rj} p_{rj}\right) = \hat{\theta}^2,$$

and

$$\sum_{r \in R} \sum_{s=1}^{k} \frac{n_{r.}}{N_R} \bar{P}_s \hat{\theta}^2 = \hat{\theta}^2 \sum_{s=1}^{k} \bar{P}_s.$$

So that

$$\max \frac{\hat{\theta}^2}{S^2(\hat{\theta})} = \frac{1}{S^2(\hat{\theta})} \Big[N_R S^2(\hat{\theta})\Big(1 - \sum_{s=1}^{k} \bar{P}_s\Big) + \hat{\theta}^2 - \hat{\theta}^2 \sum_{s=1}^{k} \bar{P}_s \Big]$$

$$= \frac{1}{S^2(\hat{\theta})} \Big(1 - \sum_{s=1}^{k} \bar{P}_s\Big)\big(N_R S^2(\hat{\theta}) + \hat{\theta}^2\big)$$

$$= N_R\Big(1 - \sum_{s=1}^{k} \bar{P}_s\Big)\Big(1 + \frac{\hat{\theta}^2}{N_R S^2(\hat{\theta})}\Big).$$

Using (5) we get:

$$= N_R\Big(1 - \sum_{s=1}^{k} \bar{P}_s\Big)\Big(\frac{1}{\sum_{s=1}^{k} \bar{P}_s}\Big)$$

That is

$$\max \frac{\hat{\theta}^2}{S^2(\hat{\theta})} = N_R\Big(\frac{1}{\sum_{s=1}^{k} \bar{P}_s} - 1\Big) \equiv Y_R^2 .$$

Hence the procedure is: "The set R is judged heterogeneous if and only if $Y_R^2 > L^{2}{}_{\alpha}$"; where $L^{2}{}_{\alpha}$ is the upper α point of the chi-square distribution with $(m-1)(k-1)$ degrees of freedom. When R is the whole set of the m populations, it is clear that Y_R^2 is identical with the Y^2 statistic used to test the homogeneity of all the m populations. So that the procedure contains the Y^2-test as a component. Moreover, if R_1 is a

subset of R, then $Y^2_{R_1} \leqslant Y^2_R$ since a contrast in R_1 is also a contrast in R. This result leads to the following property of transitivity of the procedure: Any set containing a heterogeneous subset, is itself heterogeneous. Otherwise stated: If a set is homogeneous then all its subsets are homogeneous.

However, a heterogeneous set need not contain a proper subset which is heterogeneous.

Overall size of the test and type I errors:

Suppose that the Y^2-test for all the m populations is carried out at level α. A type I error is committed whenever a subset R is judged heterogeneous while it is not. But judging R heterogeneous implies that the set M of all the m populations is heterogeneous. Hence under the hypothesis of homogeneity:

$$p_r(\text{R judged heterogeneous}) \leqslant p_r(\text{M judged heterogenous}) \equiv \alpha$$

This means that the overall size of the test (supremum of type I errors) is exactly α.

To find the exact probabilities of type I errors, we assume that R contains q populations. By the above procedure, R is heterogeneous if and only if $Y^2_R > L^2_{\alpha;(m-1)(k-1)}$. Hence the probability of type I error is equal to $p_r(Y^2_R > L^2_{\alpha;(m-1)(k-1)} \mid H) \equiv \alpha_q$, say. But under H, Y^2_R is asymptotically distributed as a chi-square variable with (q-1)(k-1) degrees of freedom (this is because Y^2_R is identical with Goodman's Y^2 defined for q populations only).

Hence

$$\alpha_q \equiv p_r\,(Y^2_R > L^2_{\alpha_q;(q-1)(k-1)} \mid H).$$

Therefore, the point $L^2_{\alpha;(m-1)(k-1)}$ is the upper α_q point

of the chi-square distribution with $(q-1)(k-1)$ degrees of freedom. Hence, we have the relation $L^2_{\alpha_q;(q-1)(k-1)} = L^2_{\alpha;(m-1)(k-1)}$. The right hand side is known, and $(q-1)(k-1)$ is known, therefore we can use tables of the chi-square distribution with $(q-1)(k-1)$ degrees of freedom to determine the point satisfying the above relation, and hence obtain α_q. This may be done for any $q = 2, 3, \ldots, m$.

Application:

To divide the set M of all the m populations into homogeneous subsets, we first calculate the overall Y^2_M to test the homogeneity of the whole set M. Upon the rejection of the hypothesis of homogeneity, theorem 2 asserts that there is at least one significant estimated contrast, which in terms of our procedure means that there exists at least one heterogeneous subset of the set M(the only one may be M itself). With M containing m populations, there are $2^m - m - 1$ subsets of M, each containing at least two populations; and we have to examine all these subsets. However, the transitivity of the procedure saves us the trouble of calculating the relevant Y^2 for every individual of the $2^m - m - 1$ subsets. For example, if one of them is found homogeneous then we need not calculate Y^2 for any of its subsets because they must all be homogeneous; while if one set is heterogeneous, the Y^2 for any set containing it need not be calculated since transitivity implies that it is heterogeneous too.

Following Gabriel (1964, pp. 46), we call a set minimal heterogeneous if all its proper subsets are homogeneous; and a set is said to be maximal homogeneous if any set containing it properly is heterogeneous (notethat Gabriel calls a heterogeneous set significant,

and a homogeneous set nonsignificant). Given either one of the class of minimal heterogeneous sets or the class of maximal homogeneous sets, the homogeneity or heterogeneity of any set of populations can be established immediately. There are two methods that can be adopted to accomplish the division of M into homogeneous subsets:

(i) Augmenting sets: calculate Y^2 for each of the $\binom{m}{2}$ pairs of populations, extend all homogeneous pairs to triplets, and all homogeneous triplets to quartets, and so on until no sets remain to be tested. This method leads to identifying the maximal homogeneous sets.

(ii) Reducing sets: calculate Y^2 for each set of size (m-1), reduce all heterogeneous sets to sets of size (m-2) and so on. This method leads to identifying minimal heterogeneous sets.

The following relations are given to simplify computations:

For a set M_1 consisting of m_1 populations, $Y^2_{M_1} = N_{M_1}\left(\frac{1}{\sum_{j=1}^{k} \bar{P}_j} - 1\right)$;

where $N_{M_1} = \sum_{i=1}^{m_1} n_{i\cdot}$ with $n_{i\cdot}$ being the sample size from popilation i, and

$$\bar{P}_j = \frac{N_{M_1}}{\sum_{i=1}^{m_1} \frac{n_{i\cdot}}{p_{ij}}} = \frac{N_{M_1}}{\sum_{i=1}^{m_1} \frac{n_{i\cdot}^2}{n_{ij}}} .$$

So that, one first calculates $\bar{P}_1$, $\bar{P}_2$, ... $\bar{P}_k$ and it is easy then to calculate $Y^2_{M_1}$.

Denote $\bar{P}_j$ for the set M_1 by $\bar{P}_{m_1,j}$ to indicate that it is based on m_1 populations.

1. A relation suitable for the method of augmenting sets:

Suppose R_1 is a set of q populations, and R is exactly R_1 plus an additional population.

Then $$\bar{P}_{q,j} = \frac{N_{R_1}}{\sum_{i=1}^{q} \frac{n_{i\cdot}^2}{n_{ij}}}$$

and

$$\bar{P}_{q+1,j} = \frac{N_R}{\sum_{i=1}^{q+1} \frac{n_{i\cdot}^2}{n_{ij}}} = \frac{N_R}{\sum_{i=1}^{q} \frac{n_{i\cdot}^2}{n_{ij}} + \frac{n_{q+1\cdot}^2}{n_{q+1,j}}}$$

$$= \frac{N_R}{\frac{N_{R_1}}{\bar{P}_{q,j}} + \frac{n_{q+1\cdot}^2}{n_{q+1,j}}}$$

This allows us to calculate the $\bar{P}$'s for the larger set from the $\bar{P}$'s of the smaller set, which have been obtained previously. And the new Y^2_R would be

$$Y_R^2 = N_R\left(\frac{1}{\sum_{j=1}^{k} \bar{P}_{q+1,j}} - 1\right).$$

2. A relation suitable for the method of reducing sets:

If the $\bar{P}$'s for the larger set R have been calculated, then the $\bar{P}$'s for the smaller set R_1 should be calculated from the following relation:

$$\bar{P}_{q,j} = \frac{N_{R_1}}{\sum_{i=1}^{q} \frac{n_{i\cdot}^2}{n_{ij}}} = \frac{N_{R_1}}{\sum_{i=1}^{q+1} \frac{n_{i\cdot}^2}{n_{ij}} - \frac{n^2_{q+1\cdot}}{n_{q+1,j}}} = \frac{N_{R_1}}{\frac{N_R}{\bar{P}_{q+1,j}} - \frac{n^2_{q+1\cdot}}{n_{q+1,j}}}.$$

And the smaller

$$Y^2_{R_1} = N_{R_1}\left(\frac{1}{\sum_{j=1}^{k} \bar{P}_{q,j}} - 1\right).$$

A Numerical Example

To illustrate the method obtained above, we apply it to a set

of data taken out of a bulletin entitled "Agriculture and Arts and Sciences Distribution of Grades" and issued by the Registrar's Office of the American University of Beirut for the second semester of the academic year 1966-67.

Our set M of populations consists of the six courses: Chemistry 101, 206, 208, 210, 216 and 218, and the observations on them are given by frequencies of students in the four final grade groups A, B, C, and D corresponding to grades in the 90's, 80's, 70's and 60's, respectively. Here, $m = 6$ and $k = 4$. The first four columns of the table below give the frequency distribution of students in the six courses and the four grade groups. The additional five columns of the table contain results of calculations that will be used to obtain the various $\bar{P}$'s. The results are included in the table, because they will be used again and again in the calculations. The subscript i appearing in the table refers to the six rows. The subscript j refers to the first four columns. The n_{ij} ($i = 1, \ldots, 6$; $j = 1, \ldots, 4$) are the frequencies of students in the respective cells.

The $\bar{P}_j$ ($j = 1, \ldots, 4$), based on the six samples, are found to be:

$$\bar{P}_1 = 0.0862, \quad \bar{P}_2 = .2638, \quad \bar{P}_3 = .3299, \quad \bar{P}_4 = 0.1534.$$

Then the over all

$$Y^2 = N\left(\frac{1}{\sum_{j=1}^{4} \bar{P}_j} - 1\right) = (438)(.2000) = 87.6000.$$

The upper (.05) point of the chi-square distribution with $5 \times 3 = 15$ degrees of freedom is 24.9958. Hence, the set of the six populations is heterogeneous.

FREQUENCY DISTRIBUTION OF STUDENTS IN SIX

CHEMISTRY COURSES AND FOUR FINAL GRADE GROUPS

Population i	Class j A	B	C	D	Total $n_{i.}$	$n_{i.}^2$	$n_{i.}^2/n_{i1}$	$n_{i.}^2/n_{i2}$	$n_{i.}^2/n_{i3}$	$n_{i.}^2/n_{i4}$
I	11	22	20	2	55	3025	275.0000	1375.0000	151.2500	1512.5000
II	11	44	35	36	126	15,876	1443.2727	360.8181	453.6000	441.0000
III	6	8	11	9	34	1156	192.6667	144.5000	105.0909	128.4444
IV	8	13	25	31	77	5929	741.1250	456.0769	237.1600	191.2581
V	7	41	49	31	128	16,384	2340.5710	399.6097	334.3673	528.5161
VI	3	2	7	6	18	324	108.0000	162.0000	46.2857	54.0000
Total	..	..	..	..	438	..	..	..	..	..

We reduce the six populations to sets of five elements each. The new $\bar{P}_i$'s are calculated easily, from those previously obtained, by means of the appropriate relation; it is found that all these sets are heterogeneous except the set R = (II, III, IV, V, VI). For the set R it is found that:

$$N_R = 383;\ \bar{P}_1 = 0{,}0797,\ \bar{P}_2 = .2515,\ \bar{P}_3 = .3256,\ \bar{P}_4 = 0.2852,$$

and $Y_R^2 = 23.5928$ which is less than the significant value 24.9958.

Further, we examine pairs of populations. Because of the homogeneity of the set R, the only pairs that need to be examined are: (I, II), (I, III), (I, IV), (I, V), and (I, VI). The results of calculations are the following:

(I, II): $N_{(I,II)} = 181;\ \bar{P}_1 = 0.1054,\ \bar{P}_2 = .3632,\ \bar{P}_3 = .2992,\ \bar{P}_4 = 0.0927;$

$Y^2_{(I,II)} = 29.3401$. Hence, (I, II) is heterogeneous.

(I, III): $N_{(I,III)} = 89;\ \bar{P}_1 = 0{,}1903,\ \bar{P}_2 = .3156,\ \bar{P}_3 = .3472,\ \bar{P}_4 = 0.0542;$

$Y^2_{(I,III)} = 9.0958$. Hence, (I,III) is homogeneous.

(I, IV): $N_{(I,IV)} = 132;\ \bar{P}_1 = 0.1299,\ \bar{P}_2 = .2224,\ \bar{P}_3 = .3398,\ \bar{P}_4 = 0.0775;$

$Y^2_{(I,IV)} = 39.5208$. Hence, (I,IV) is homogeneous.

(I, V): $N_{(I,V)} = 183;\ \bar{P}_1 = 0.0700,\ \bar{P}_2 = .3407,\ \bar{P}_3 = .3768,\ \bar{P}_4 = 0.0896;$

$Y^2_{(I,V)} = 25.6383$. Hence, (I,V) is heterogeneous.

(I, VI): $N_{(I,VI)} = 73;\ \bar{P}_1 = 0.1906,\ \bar{P}_2 = .2437,\ \bar{P}_3 = .3695,\ \bar{P}_4 = 0.0466;$

$Y^2_{(I,VI)} = 12.8407$. Hence, (I,VI) is homogeneous.

Therefore, the pairs (I,II), (I,IV), (I,V) are heterogeneous. And the pairs (I,III), (I,VI) are homogeneous; we augment these two

pairs to triplets. The only triplet that need to be examined is (I, III, VI), since any other triplet would contain a heterogeneous pair which implies that it is heterogeneous.

We obtain the $\bar{P}_i$'s of the set (I, III, VI) from the $\bar{P}_i$'s of the set (I, III), or the set (I, VI) by applying the appropriate relation. The results are the following:

(I, III, VI): $N_{(I,III,VI)} = 107$; $\bar{P}_1 = 0.1859$, $\bar{P}_2 = .2410$, $\bar{P}_3 = .3536$, $\bar{P}_4 = 0.0631$; $Y^2_{(I,III,VI)} = 19.8378$. Hence, (I,III,VI) is homogeneous.

There is no need for augmenting (I, III, VI) to sets of four elements, since any such set would contain a heterogeneous pair which implies that it is heterogeneous. That finishes computations, and we conclude that: the sets (I, II), (I, IV), (I,V) constitute the class of minimal heterogeneous sets, and the sets (I, III, VI), (II, III, IV, V, VI) constitute the class of maximal homogeneous sets. The homogeneity or lack of it of any other set can be inferred from either of the two classes.

The above results indicate that population I is the main reason for heterogeneity of the six populations. Recalling that population I corresponds to the course chemistry 101 which is a freshman course, our results agree with the general idea that the freshman class is different, in some sense, from the higher classes.

CHAPTER III

SELECTING A SUBSET CONTAINING THE BEST OF SEVERAL MULTINOMIAL POPULATIONS

This chapter and the next deal with the problem of selecting a subset of several multinomial populations which is asserted to contain the "best" population with probability greater or equal to a preassigned value P. The definition of a best population is somewhat arbitrary. In the present chapter an exact procedure is given for a rather restricted definition of "best". With a more general definition, an approximate procedure is given in the next chapter. The two procedures are of the same nature as the one given by Gupta and Sobel (1960) for selecting a subset containing the best of several binomial populations, where they define the best binomial population as the one with the highest probability of success.

Suppose we have m multinomial populations of k classes each, with true parameters $(\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik})$, $i = 1, 2, \ldots, m$. Let a_1 and a_2 be real numbers such that $a_1 > a_2$. Define the linear functions $h_i = a_1(\pi_{i1} + \cdots + \pi_{ir}) + a_2(\pi_{i,r+1} + \cdots + \pi_{ik})$, where $i = 1, \ldots, m$ and r is a positive integer such that $1 \leq r < k$.[1] The ranked h_i are denoted by $h_{(1)} \leq h_{(2)} \leq \cdots \leq h_{(m)}$. It is assumed that the correct pairing of the $h_{(i)}$ with the m populations is

---

[1] There is no loss of generality in assuming that $a_1 > a_2$, sine one can renumber the classes of each population so that the first r classes are always associated with the larger $a_i (i=1,2)$ which we call a_1.

not known. The best population is defined to be the one associated with $h_{(m)}$.

An instance of the application of this definition of best multinomial population is where one is interested in comparing the m populations with respect to a subset, say the first r, of the k classes. Then one chooses a_2 small enough in absolute value, may be zero, so that the contribution due to the remaining $(k - r)$ classes to the function h is negligible.

For example, if the first class is the only class of interest and the population with the highest probability in this class is considered to be the best, then the choice $a_1 = 1$, $a_2 = 0$ and $r = 1$ gives $h_i = \pi_{i1} (i = 1, \ldots, m)$, which is what we need to compare.

Adopting the "linear function" definition of best multinomial population, a procedure is given below for selecting a subset which contains the best population with probability at least P, regardless of the true parameter values.

Suppose that independent random samples $(n_{i1}, \ldots, n_{ik})$ of sizes $n_i = \sum_{j=1}^{k} n_{ij}$ $(i = 1, \ldots, m)$ are drawn from the multinomial populations. Let the unbiased estimates of the h_i be $v_i = a_1(p_{i1} + \cdots + p_{ir}) + a_2(p_{i,r+1} + \cdots + p_{ik})$, where $p_{ij} = \frac{n_{ij}}{n_i}$ $(i = 1, \ldots, m;\ j = 1, \ldots, k)$. Let $v_{max} = \max(v_1, \ldots, v_m)$. Then the procedure D is: Retain the ith population if and only if $v_i \geqslant v_{max} - c$, where c is a constant depending on m, $n_i (i = 1, \ldots, m)$ and P. We say that a correct selection (CS) is made if and only if the retained subset contains the best population. And it is required that $\mathrm{pr}(CS \mid \pi_{ij}) \geqslant P$ for all possible configurations of the true

parameters $\pi_{ij}(i = 1, \ldots, m;\ j = 1, \ldots, k)$. The constant c is chosen to be the smallest non-negative number such that the infimum of pr(CS), taken over all n_i and $\pi_{ij}(i = 1, \ldots, m;\ j = 1, \ldots, k)$, is greater than or equal to P. In order to find pr(CS), we adopt the convention that when there is more than one population associated with $h_{(m)}$ (i.e., more than one best population) we consider one particular "tagged" population as being the best.

To determine the probability of a correct selection, we write

$$v_i = a_1(p_{i1} + \cdots + p_{ir}) + a_2(p_{i,r+1} + \cdots + p_{ik})$$

as

$$\frac{1}{n_i}\left[a_1(n_{i1} + \cdots + n_{ir}) + a_2(n_{i,r+1} + \cdots + n_{ik})\right]$$

$$= \frac{1}{n_i}\left[(a_1 - a_2)(n_{i1} + \cdots + n_{ir}) + a_2 n_i\right], \quad i = 1, \ldots, m.$$

Define $u_i = \dfrac{n_i(v_i - a_2)}{a_1 - a_2} = (n_{i1} + \cdots + n_{ir})$. Then it is known that u_i has the binomial distribution $B(n_i;\ Q_{i,r})$, where

$$Q_{i,r} = \pi_{i1} + \cdots + \pi_{ir} = \frac{h_i - a_2}{a_1 - a_2}, \quad i = 1, \ldots, m.$$

For simplicity we denote $Q_{i,r}$ by Q_i.

Let $Q_{(i)}$, $n_{(i)}$, $v_{(i)}$ and $u(i)$ be those particular quantities Q_i, n_i, v_i and u_i, respectively, which are associated with the population corresponding to $h_{(i)}$, $i = 1, \ldots, m$. Then, using the procedure D, a correct selection is made if and only if

$v_{(m)} \geq v_{max} - c$ which is equivalent to

$v_{(i)} \leq v_{(m)} + c$ for all $i < m$ (for $i = m$ the inequality is satisfied with probability 1)

or

$$\frac{n_{(i)}(v_{(i)} - a_2)}{a_1 - a_2} \leq \frac{n_{(i)}(v_{(m)} - a_2)}{a_1 - a_2} + \frac{n_{(i)}c}{a_1 - a_2}$$

or

$$u_{(i)} \leqslant n_{(i)}\left(\frac{u_{(m)}}{n_{(m)}} + \frac{c}{a_1 - a_2} \right) \text{ for all } i < m \quad \ldots\ldots\ldots \quad (1)$$

Since the $u_{(i)}$'s are independent binomial $B(n_{(i)};\ Q_{(i)})$ variables, it can be seen that the probability that (1) holds true, i.e. $pr(CS \mid \Pi_{ij})$ is equal to:

$$\sum_{u=0}^{n_{(m)}} \binom{n_{(m)}}{u} Q_{(m)}^{u} (1 - Q_{(m)})^{n_{(m)}-u} \prod_{i=1}^{m-1} \left\{ \sum_{x=0}^{\left[n_{(i)}\left(\frac{u}{n_{(m)}} + \frac{c}{a_1 - a_2}\right)\right]} \binom{n_{(i)}}{x} Q_{(i)}^{x} (1-Q_{(i)})^{n_{(i)}-x} \right\}$$

$\ldots\ldots\ldots\ldots\ldots$ (2)

Here, $[z]$ denotes the largest integer less than or equal to z.

The problem of interest, now, is to minimize (2).

Each one of the (m-1) factors in the braces appearing in (2) is a non-increasing function of $Q_{(i)}$ as can be seen by expressing each factor as an incomplete beta function. Recalling that $Q_{(i)} = \frac{h_{(i)} - a_2}{a_1 - a_2}$ together with the assumption that $a_1 - a_2 > 0$, we see that the ranking $h_{(1)} \leqslant h_{(2)} \leqslant \cdots \leqslant h_{(m)}$ is equivalent to the ranking $Q_{(1)} \leqslant Q_{(2)} \leqslant \cdots \leqslant Q_{(m)}$. Therefore, for a fixed $Q_{(m)}$, each factor in the braces is minimized by taking $Q_{(i)} = Q_{(m)}$, $i = 1, \ldots, m$. Then we consider the infimum of (2) over the range of $Q_{(m)} = Q$, say, which is $0 \leqslant Q \leqslant 1$. To achieve the absolute minimum of pr(CS), we must further minimize (2) with respect to $n_{(m)}$ which is an element of the set $\{n_1, n_2, \ldots, n_m\}$. Because all the Q_i are taken equal to Q, any manner of pairing the other $(m - 1)$ $n_{(i)}$'s with the remaining n_i's (after selecting a value for $n_{(m)}$) will give exactly the same minimum value for the product of the $(m - 1)$ factors. The condition that the infimum of pr(GS) is greater than or equal to P, is then:

$$\min_{n_{(m)} \in \{n_1, \ldots, n_m\}} \left\{ \inf_{0 \leqslant Q \leqslant 1} \sum_{u=0}^{n(m)} \binom{n(m)}{u} Q^u (1-Q)^{n(m)-u} \prod_{s \neq (m)} \left[\sum_{x=0}^{n_s \left(\frac{u}{n_{(m)}} + \frac{c}{a_1 - a_2} \right)} \binom{n_s}{x} Q^x (1-Q)^{n_s - x} \right] \right\} \geqslant P \quad \ldots\ldots\ldots\ldots \quad (3)$$

For equal sample sizes $n = n_i (i = 1, \ldots, m)$, (3) becomes:

$$\inf_{0 \leqslant Q \leqslant 1} \left\{ \sum_{u=0}^{n} \binom{n}{u} Q^u (1-Q)^{n-u} \left[\sum_{x=0}^{u + \frac{nc}{a_1 - a_2}} \binom{n}{x} Q^x (1-Q)^{n-x} \right]^{m-1} \right\} \geqslant P \quad \ldots\ldots \quad (4)$$

The constant $\frac{nc}{a_1 - a_2} = d$ is taken to be the smallest non-negative integer such that (4) is satisfied. The values of d satisfying (4) have been tabulated by Gupta and Sobel (1960, pp. 242 - 45) to carry out their procedure for selecting a subset containing the best binomial population, where they define the best binomial population as the one with the highest probability of success. The tables give the values of d for $P = 0.75, .90, .95, 0.99$;

$n = 1\ (1)20,\ 25(5)50,\ 50(10)100,\ 100(25)200,$ and $200(50)500$;

$m = 1(1)20,\ 25(5)50$.

Then, after obtaining $d = \frac{nc}{a_1 - a_2}$ from the tables, we solve for c which enables us to carry out the procedure D, that is to retain only those populations for which $v_i \geqslant v_{max} - c$. Better still, the procedure D in the case of equal sample sizes can be put in the form: Retain the ith population if and only if $u_i \geqslant u_{max} - \frac{nc}{a_1 - a_2}$, where $u_i = \frac{n(v_i - a_2)}{a_1 - a_2} = (n_{i1} + \ldots + n_{ir})$ and $u_{max} = \max(u_1, \ldots, u_m)$. This last form of the procedure D is more convenient for computations than the first, because it is easier to compute the u_i rather than the $v_i (i = 1, \ldots, m)$.

In the case of unequal sample sizes, there is no general rule as to which particular value of $n_{(m)}$ minimizes the left hand side of (3), above. Gupta and Sobel (1960, pp. 230-231) empirically found that for some interval of the constant-value, $[0, d_0]$, the left hand side of (3) is minimized by taking $n_{(m)}$ to be the largest of the $n_i (i = 1, \ldots, m)$. But this is not true when d is larger than d_0. In practical applications, they suggested to take the arithmetic mean $\bar{n}$ of the sample sizes as the common sample size and to use the appropriate table with $n = \bar{n}$ to obtain the value of the constant; this value may be improved by further computations depending on the specific situation. Since the present case of multinomial populations is similar to that of binomial populations, we follow this approach of taking $n = \bar{n}$ to deal with situation in which the sample sizes are not equal.

Finally we comment on the expected size of the retained subset, $E(S)$. In the case of binomial populations and where the best population is defined to be the one with the highest probability of success p_i, Gupta and Sobel (1960, pp. 231-34) derived an expression for $E(S)$ and tabulated the values of $\frac{E(S)}{m}$ for the particular case where it is assumed that $p_{(m-1)} = \ldots = p_{(1)} = p$ and $p_{(m)} = p + \delta$, where the $p_{(i)}$ are the ranked p_i and $0 \leqslant \delta \leqslant 1$, $0 \leqslant p \leqslant 1 - \delta$. In addition, it is assumed that the sample sizes are equal. In order to control the size of the retained subset one should consult the tables for the values of $\frac{E(S)}{m}$ to determine the nessecary value of n which guarantees that the size of the retained subset is at most a preassigned positive integer. Then this values of n is used to find the constant required to carry

out the procedure. In the present case of multinomial populations we first note that $h_i = a_1(\pi_{i1} + \cdots + \pi_{ir}) + a_2(\pi_{i,r+1} + \cdots + \pi_{ik}) \leqslant a_1$. Then we assume that $h_{(m-1)} = \cdots = h_{(1)} = h$ and $h_{(m)} = h + e$, where $h \leqslant a_1 - e$ and e is a specified positive number. Hence, $Q_{(m)} = Q + \frac{e}{a_1 - a_2}$, where $Q_{(i)} = \frac{h_{(i)} - a_2}{a_1 - a_2} = \pi_{i1} + \cdots + \pi_{ir}$ as was defined before. Noting that the Q_i correspond to the p_i in the binomial case, we see that $\frac{e}{a_1 - a_2}$ corresponds to δ ; hence the tables can be used with $\delta = \frac{e}{a_1 - a_2}$ to give the common sample size necessary to guarantee that the retained subset of the multinomial populations is of at most a preassigned size.

CHAPTER IV

A LARGE SAMPLE PROCEDURE FOR SELECTING A SUBSET CONTAINING THE BEST MULTINOMIAL POPULATION

Suppose we have m multinomial populations of k classes each, with true parameters $(\pi_{i1}, \ldots, \pi_{ik})$, $i = 1, \ldots, m$. Let $a_1, \ldots, a_k$ be non-negative real numbers such that $\sum_{j=1}^{k} a_j = 1$ and $a_r \neq a_s$ for at least one pair $r,s = 1, \ldots, k$. Define the linear functions $h_i = a_1 \pi_{i1} + \ldots + a_k \pi_{ik}$ $(i = 1, \ldots, m)$, and denote the ordered h_i by $h_{(1)} \leqslant \ldots \leqslant h_{(m)}$. The correct pairing of the $h_{(i)}$ with the m populations is not known. We define the best multinomial population to be any particular population that is associated with $h_{(m)}$. This definition is more general than the one adopted in chapter III. However, because of the difficulty in working with multinomial distributions, a large sample procedure based on the normal approximation to the multinomial distribution is given here. The final decision in the procedure is the selection of a subset of the populations that is asserted to contain the best population with a preassigned probability P, regardless of the true values of the parameters $\pi_{ij}(i = 1, \ldots, m,; j = 1, \ldots, k)$.

The h_i may be interpreted as weighted sums of the class probabilities in each population, and the interest is assumed to be in the population with the largest such weighted sum. As

an application of this interpretation of the h_i's, consider m population strataand the following categories of expenditure: food, clothing, medical care, and schooling, i.e.,k = 4. Suppose that the true expenditure proportions on each of the four categories are: π_{i1}, π_{i2}, π_{i3}, π_{i4}, i = 1, ..., m. Suppose, further, that with respect to a certain base year, the price indices are now: 110 for food, 105 for clothing, 100 for medical care, and 90 for schooling. The question, now, is: which stratum is most affected (i.e., suffers the most) by this change in prices? An answer may be provided by considering the quantities

$$h_i = \frac{1}{405}(110\,\pi_{i1} + 105\,\pi_{i2} + 100\,\pi_{i3} + 90\,\pi_{i4}) \quad (i = 1, \ldots, m),$$

where we have divided by the sum of the indices in order to render the sum of the coefficients of the parameters π_{ij} equal to one. Then, we may regard the stratum with the highest h_i as the one that suffers the most by such change in prices.

Suppose that independent random samples $(n_{i1}, \ldots, n_{ik})$, where n_{ij} denotes the frequency in the jth class of the ith population $(i = 1, \ldots, m;\ j = 1, \ldots, k)$, have been drawn from the populations. Then, the maximum likelihood unbiased estimates of the $h_i = \sum_{j=1}^{k} a_j \pi_{ij}$ are: $v_i = \sum_{j=1}^{k} a_j p_{ij}$, where $p_{ij} = \frac{n_{ij}}{n_i}$ and $n_i = \sum_{j=1}^{k} n_{ij} (i = 1, \ldots, m;\ j = 1, \ldots, k)$.

Let $v_{max} = \max(v_1, \ldots, v_m)$. Then the procedure R is given by: retain the ith population if and only if $v_i \geq v_{max} - c$, where c is a non-negative constant to be determined such that the probability of a correct selection (CS) is at least a preassigned value P regardless of the true configuration of the parameters $\pi_{ij} (i=1,\ldots,m;\ j=1,\ldots,k)$.

To determine the probability of a correct selection, we need to know the distribution of the v_i(i = 1, ..., m).

Mood (1950, pp. 215) proved that for large samples, the vector $(p_{i1}, \ldots, p_{i,k-1})'$ is approximately distributed according to the multivariate normal distribution $N(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$, where $\underset{\sim}{\mu} = (\pi_{i1}, \ldots, \pi_{i,k-1})'$ and $\underset{\sim}{\Sigma} = \| \frac{1}{n_i} \pi_{ij}(\delta_{js} - \pi_{is}) \|$ (j,s = 1,..., k-1) and $\delta_{js} = 1, 0$ when $j = s$, $j \neq s$, respectively. Furthermore, by the definition of the singular normal distribution (Anderson, 1958, pp. 26), it follows that the vector $(p_{i1}, \ldots, p_{in})'$ is approximately distributed according to the singular normal distribution $N(\underset{\sim}{\mu}^*, \underset{\sim}{\Sigma}^*)$, where $\underset{\sim}{\mu}^* = (\pi_{i1}, \ldots, \pi_{ik})'$ and $\underset{\sim}{\Sigma}^* = \| \frac{1}{n_i} \pi_{ij}(\delta_{js} - \pi_{is}) \|$ (j,s = 1, ..., k). With simple algebraic computations, we obtain the result that $v_i = \sum_{j=1}^{k} a_j p_{ij}$ is approximately distributed according to the univariate normal distribution $N(h_i, \frac{\sigma_i^2}{n_i})$, where $h_i = \sum_{j=1}^{k} a_j \pi_{ij}$ and $\sigma_i^2 = \sum_{j=1}^{k} a_j^2 \pi_{ij} - (\sum_{j=1}^{k} a_j \pi_{ij})^2$; this is true for i = 1, ..., m.

Let $\sigma_{(i)}^2$, $n_{(i)}$, and $v_{(i)}$ be those particular quantities σ_i^2, n_i, and v_i, respectively, that correspond to the population which is associated with $h_{(i)}$ for i = 1, ..., m. Then, using the procedure R, a correct selection is made if and only if $v_{(m)} \geq v_{max} - c$.

Or, equivalently, a correct slection is made if and only if

$$v_{(i)} \leq v_{(m)} + c \quad \text{for all } i < m \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Upon standardizing the variables $v_{(i)}$ (i = 1, ..., m), (1) becomes:

$$\frac{v_{(i)} - h_{(i)}}{\sigma_{(i)} / n_{(i)}^{\frac{1}{2}}} \leq \frac{v_{(m)}}{\sigma_{(i)} / n_{(i)}^{\frac{1}{2}}} + \frac{c - h_{(i)}}{\sigma_{(i)} / n_{(i)}^{\frac{1}{2}}} \quad \text{for all } i < m,$$

or,

$$\frac{v_{(i)} - h_{(i)}}{\sigma_{(i)}/n_{(i)}^{\frac{1}{2}}} \leq \frac{v_{(m)} - h_{(m)}}{\sigma_{(m)}/n_{(m)}^{\frac{1}{2}}} \cdot \left(\frac{n_{(i)}}{n_{(m)}}\right)^{\frac{1}{2}} \cdot \frac{\sigma_{(m)}}{\sigma_{(i)}} + \frac{c + (h_{(m)} - h_{(i)})}{\sigma_{(i)}/n_{(i)}^{\frac{1}{2}}}$$

for all $i < m$ ..... (2)

Denote the standard normal density and cummulative distribution functions by $f(z)$ and $F(z)$, respectively. Then, given π_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, k)$, the probability that (2) holds true (i.e., $pr(CS \mid \pi_{ij})$) is equal to:

$$\int_{-\infty}^{\infty} f(z) \prod_{i=1}^{m-1} \left\{ F\left[z\left(\frac{n_{(i)}}{n_{(m)}}\right)^{\frac{1}{2}} \cdot \frac{\sigma_{(m)}}{\sigma_{(i)}} + \frac{c + (h_{(m)} - h_{(i)})}{\sigma_{(i)}/n_{(i)}^{\frac{1}{2}}} \right] \right\} dz \quad \ldots\ldots\ldots \quad (3)$$

As in the problem of chapter III, we want to choose the constant c such that the minimum value, taken over all possible values of the π_{ij}'s, of (3) is equal to the preassigned value P. Thus, we are interested in minimizing (3). First of all, it is seen that each factor in the braces appearing in (3) is an increasing function of $h_{(m)} - h_{(i)}$, $i = 1, \ldots, m - 1$. Recalling that $h_{(m)}$ is the largest of the $h_{(i)}$'s, it follows that the whole product of these factors is minimized by taking $h_{(m)} - h_{(i)} = 0$ for all $i < m$; that is taking all the $h_{(i)}$'s to be equal to each other.

Moreover, for fixed (but arbitrary) $i, s = 1, \ldots, m$, taking $h_i = \sum_{j=1}^{k} a_j \pi_{ij}$ to be equal to $h_s = \sum_{j=1}^{k} a_j \pi_{sj}$ for all possible values of the coefficients a_j, implies that $\pi_{ij} = \pi_{sj}$ for all j. This is because, a necessary and sufficient condition that $\sum_{j=1}^{k} a_j \pi_{ij} = \sum_{j=1}^{k} a_j \pi_{sj}$ holds true for all values a_j, is that $\pi_{ij} = \pi_{sj}$ for all j. But, $\pi_{ij} = \pi_{sj}$ for all j implies that $\sigma_i = \sigma_s$; hence taking $h_i = h_s$ for all pairs (i,s), implies that $\sigma_i = \sigma_s$

for all i, s = 1, ..., m. With these simplifications, (3) becomes:

$$\int_{-\infty}^{\infty} f(z) \prod_{i=1}^{m-1} F(z(\frac{n(i)}{n(m)})^{\frac{1}{2}} + \frac{n(i)^{\frac{1}{2}} c}{\nabla(i)})dz \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

The above expression can be further reduced by substituting an upper bound for the $\nabla_{(i)}$ (i = 1, ..., m-1). Dropping the subscript i, we are interested in finding an upper bound for $\nabla = \sum_{j=1}^{k} a_j^2 \pi_j - (\sum_{j=1}^{k} a_j \pi_j)^2$, where ∇ is considered as a function of the π_j's, while the a_j's are fixed, Because of the form of ∇, it was not possible to maximize ∇ subject to the condition that $\sum_{j=1}^{k} \pi_j = 1$. It was also not possible to find a very sharp upper bound that can hold irrespective of the values of the a_j's. The following two upper bounds for ∇ were found to be the best that I could achieve:

$$B_1 = \sum_{j=1}^{k} a_j^2.$$

This upper bound neglects the term $(\sum_{j=1}^{k} a_j \pi_j)^2$. It should be admitted that B_1 is a rather rough upper bound since the neglected term is an appreciable quantity compared to $\sum_{j=1}^{k} a_j^2 \pi_j$, so that is really much smaller than B_1. Using B_1 may lead to unnecessarily large values of the constant c. However, B_1 has the advantage that it holds irrespective of the values of the a_j's.

The second upper bound is $B_2 = \frac{1}{4}(\sum_{j=1}^{k} a_j^2 - \frac{(k-2)^2}{\sum_{j=1}^{k} a_j^{-2}})$. B_2 holds as an upper bound for ∇, provided that $a_j \neq 0$ for all j. It was obtained by maximizing $\sum_{j=1}^{k} a_j^2 \pi_j - \sum_{j=1}^{k} a_j^2 \pi_j^2$ subject to the condition that $\sum_{j=1}^{k} \pi_j = 1$. Thus, B_2 is sharper than B_1 provided that no a_j is equal to zero. It is seen that B_2 overestimates ∇ by as much

as $\sum_{j \neq s} a_j a_s \pi_j \pi_s$ which is a small quantity.

Replacing $\sigma_{(i)}$ by B, where B is either B_1 or B_2, the expression in (4), above, becomes:

$$\int_{-\infty}^{\infty} f(z) \prod_{i=1}^{m-1} F\left(z\left(\frac{n(i)}{n(m)}\right)^{\frac{1}{2}} + \frac{n(i)^{\frac{1}{2}} c}{B}\right) dz \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

Assuming that the sample sizes are all equal, $n_i = n (i = 1, \ldots, m)$, (5) becomes:

$$\int_{-\infty}^{\infty} f(z) \left[F\left(z + \frac{n^{\frac{1}{2}} c}{B}\right)\right]^{m-1} dz.$$

Hence, the constant c is obtained from the equation:

$$\int_{-\infty}^{\infty} f(z) \left[F\left(z + \frac{n^{\frac{1}{2}} c}{B}\right)\right]^{m-1} dz = P \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

Here, P is the preassigned level of probability of a correct selection. The values of the constant $c' = \frac{n^{\frac{1}{2}} c}{B}$ satisfying (6) are given by Bechhofer (1954) for m = 2(1)10, and by Gupta (1956) for m = 2(1)50. Then, solving for c, $c = \frac{B}{(n)^{\frac{1}{2}}} c'$, we can carry out the procedure R: retain all populations for which $v_i \geqslant v_{max} - c$.

In the case of unequal sample sizes, it is seen that (5), above, is minimized by taking $n_{(i)} = \min(n_1, \ldots, n_m) = n_{min}$, and $n_{(m)} = \max(n_1, \ldots, n_m) = n_{max}$. So that, instead of equation (6), we obtain the equation:

$$\int_{-\infty}^{\infty} f(z) \left[F\left(z\left(\frac{n_{min}}{n_{max}}\right)^{\frac{1}{2}} + \frac{n_{min}^{\frac{1}{2}} c}{B}\right)\right]^{m-1} dz = P.$$

However, since tables that give the values of the constant that satisfy the above equation are not available, we suggest to take the arithmetic mean $\bar{n}$ of the sample sizes as the common sample size. Therefore, (6) would be exactly the same except that n is replaced by $\bar{n}$.

The same tables, now give $c'' = \frac{(\bar{n})^{\frac{1}{2}} c}{B}$ from which we solve for c, and carry out the procedure as before.

A Numerical Example

To illustrate the procedure numerically, consider again the example given in chapter II.

Let $a_j (j = 1, \ldots, 4)$ be a numerical description of the jth grade class, for instance $a_1 = 95$, $a_2 = 85$, $a_3 = 75$, $a_4 = 65$. Thus, for the ith chemistry course, $h_i = \sum_{j=1}^{4} a_j \pi_{ij}$ may be interpreted as an average numerical grade for the course. The "best" population would then be the chemistry course with the largest average numerical grade. However, we take

$$h_i = \frac{1}{\sum_{j=1}^{4} a_j} \left(\sum_{j=1}^{4} a_j \pi_{ij} \right) \quad (i = 1, \ldots, 6)$$

so that the sum of the new coefficients is equal to 1; and it is seen that this does not affect the comparison among the populations with respect to the h_i $(i = 1, \ldots, 6)$.

Calculations: The values of the $v_i = \frac{1}{\sum_{j=1}^{4} a_j} \sum_{j=1}^{4} a_j p_{ij}$ $(i = 1, \ldots, 6)$ are:

$$v_1 = 0.2582, \qquad v_2 = .2418, \qquad v_3 = .2445$$

$$v_4 = .2336, \qquad v_5 = .2402, \qquad v_6 = 0.2378.$$

Therefore, $v_{max} = 0.2582$. The upper bound for σ is: $B_2 = 0.0491$.

Referring to the tables (Bechhofer 1954, table I), the constant c' assumes the values : 3.1519, 2.7100, and 1.9674 corresponding to the probability levels P: 0.95, .90, and 0.75, respectively. The procedure R is given by the rule: retain only those populations for which $v_i \geqslant v_{max} - c$, where $c = \frac{B_2}{n^{\frac{1}{2}}} c'$.

Since the sample sizes are not equal, the value of n is taken to be $\bar{n} = \frac{438}{6} = 73$, so that $n^{\frac{1}{2}} = 8.5440$.

Results:

Probability level	c	$v_{max}-c$	Populations retained
0.75	0.0113	0.2469	I, only
.90	.0125	.2457	I, only
0.95	0.0182	0.2400	I, II, III, V

Therefore, we can claim with 90% confidence that population I is the best population, and with 95% confidence that the best population is one of the four populations: I, II, III, V.

It is seen that these results are in fair agreement with the results obtained in chapter II. There, it was shown that the six populations (i.e., chemistry courses) could be divided into two homogeneous subsets one of which consists of population I(i.e., chemistry 101), and the other consists of the remaining populations. Similary, the present results indicate that population I is different from the others, in the sense that it has the highest numerical average grade. This may indicate that (in some situations, as the present one) the selection procedure gives the same information that was given by the homogeneity test and, in addition, it orders the populations which is a desired information in many situations.

REFERENCES

Anderson, T. W., An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons, Inc., 1958.

Bechhofer, R. E., "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances", Ann. Math. Stat., 25 (1954), 16 - 39.

_____, "A Sequential Multiple-Decision Procedure for Selecting the Best One of Several Normal Populations with a Common Unknown Variance, and Its Use with Various Experimental Designs", Biometrics, 14(1958), 408 - 429.

Bechhofer, R. E., and Blumenthal, S. "A Sequential Multiple-Decision Procedure for Selecting the Best one of Several Normal Populations with a Common Unknown Variance, II: Monte Carlo Sampling Results and New Computing Formulae", Biometrics, 18(1962), 52 - 67.

Bechhofer, R. E., and Sobel, M. "A Single-Sample Multiple Decision Procedure for Ranking Variances of Normal Populations", Ann. Math. Stat., 25 (1954), 273 - 289.

Cramer, H., Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946.

Gabriel, K. R. "A Procedure for Testing the Homogeneity of All Sets of Means in Analysis of Variance", Biometrics, 20(1964), 459-477.

Gold, R. Z., "Tests Auxiliary to χ^2 Tests in a Markov Chain", Ann. Math. Stat., 34(1963), 56 - 74.

Goodman, L.A., "Simultaneous Confidence Intervals for Contrasts among Multinomial Populations", Ann. Math. Stat., 35(1964), 717-725.

Gupta, S. S., On a Decision Rule for a Problem in Ranking Means, Institute of Statistics, Mimeograph Series Report No. 150 (Chapel Hill, N.C. University of North Carolina, May 1956).

Gupta, S. S., and Sobel, M., "On Selecting a Subset Which Contains All Populations Better than a Standard", Ann. Math. Stat., 29 (1958), 235 - 244.

Gupta, S. S., and Sobel, M. Selecting a Subset Containing the Best of Several Binomial Populations, Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, ed. by Ingram Olkin and others (Stanford, California: Stanford University Press, 1960), pp. 224.

Guttman, I., "A Sequential Procedure for the Best Population", Sankhyā, Ser. A., 25 (1963), 25 - 28.

Guttman, I., and Tiao, G. C., "A Baysian Approach to Some Best Population Problems", Ann. Math. Stat., 35(1964), 825 - 835.

Mood, A. M., Introduction to the Theory of Statistics. New York: John Wiley and Sons, Inc., 1950.

Mosteller, F., "A k-Sample Slippage Test for an Extreme Population", Ann. Math. Stat., 19(1948), 58 - 65.

Paulson, E., "A Multiple Desision Procedure for Certain Problems in Analysis of Variance", Ann. Math. Stat., 20 (1949), 95-98.

_____ , "On the Comparison of Several Experimental Categories with a Control", Ann. Math. Stat., 23(1952), 239 - 246.

Scheffé, H., "A Method for Judging All Contrasts in the Analysis of Variance", Biometrika, 40(1953), 87-104.

_____ , The Analysis of Variance. New York: John Wiley and Sons, Inc., 1959.

Studden, W. J., "On Selecting a Subset of k Populations Containing the Best", Ann. Math. Stat., 38(1967), 1072 - 1078.