On Selecting a Subset Containing the Best of Several Discrete Distributions

by

Klaus Nagel

Purdue University*

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series No. 66

March 1966

*This research was supported in part by Contract AF 33(657)11737 with the Aerospace Research Laboratories. Reproduction in whole or in part permitted for any purposes of the United States Government.

On Selecting a Subset Containing the Best of Several Discrete Distributions

by

Klaus Nagel

Purdue University*

1. Introduction and Formulation of the Problem

In the problem of selecting a subset, which contains the best of several binomial distributed populations (c.f. [1]) the question arises to find the "worst" configuration i.e., that vector $p = (p_1,\ldots,p_k)$ for which the probability of correct selection using procedure R proposed and studied in [1], P(CS/R), attains its minimum. Procedure R is defined as follows: We take n independent observations on each of the k populations $\Pi_1,\ldots,\Pi_k$. Let x_i denote the number of successes in the i th sample. Then the decision rule R is:

"Select Π_i iff $x_i \geq \max\limits_{j=1,\ldots,k} x_j - d$"

where d is a given non-negative integer. In [1] d is chosen to satisfy the requirement that the probability of a correct selection is at least equal to a specified value P^*. By a correct selection we mean the selection of any subset which contains the best population i.e., the population with maximal p.

*This research was supported in part by Contract AF 33(657)11737 with the Aerospace Research Laboratories. Reproduction in whole or in part permitted for any purposes of the United States Government.

[In the case where several populations have equal maximal p-values one of these will be "tagged" as the best one.] We assume that the number of observations from each population is equal to n. Then $P\{CS;k,n,p,R\}$, the probability of a correct selection in using R is given by

$$(1.1)\qquad P\{CS;k,n,p,R\} = \sum_{m=0}^{n}\binom{n}{m}p_{[k]}^{m}(1-p_{[k]})^{n-m}\prod_{\alpha=1}^{k-1}\left\{\sum_{j=0}^{m+d}\binom{n}{j}p_{[\alpha]}^{j}(1-p_{[\alpha]})^{n-j}\right\}$$

where

$$p_{[k]} \geq p_{[k-1]} \geq \cdots \geq p_{[1]}$$

are the ordered values of the unknown parameters p_i's. As shown in [1] this expression attains its minimum for the configuration

$$(1.2)\qquad p_{[1]} = p_{[2]} = \cdots = p_{[k]} = p.$$

It was also shown there that for $k=2$ this common value p is equal to $1/2$ and that for a fixed k $p \to 1/2$ as $n \to \infty$. However, in general, it is not known what the above common value of p is. In the case where all p_i's are equal to p we have

$$(1.3)\qquad P(CS/R) = \sum_{i=0}^{n}\binom{n}{i}p^{i}(1-p)^{n-i}\left[\sum_{j=0}^{i+d}\binom{n}{j}p^{j}(1-p)^{n-j}\right]^{k-1}.$$

The "worst" configuration is of this form (1.2) for any distribution with the property TP_2 i.e., total positivity of order 2 which is equivalent to the property of monotone likelihood ratio (see [2]). Therefore it is reasonable to try to minimize the expression

$$(1.4)\qquad S = \sum_{i=0}^{n} a_i \Big(\sum_{j=0}^{i+d} a_j\Big)^{k-1}$$

only under the condition

$$(1.5)\qquad \sum_{i=0}^{n} a_i = 1,\ a_i \geq 0 \qquad i=0,\dots,n.$$

(1.4) includes (1.3) as a special case, and by minimizing (1.4) we will get a lower bound for the probability of correct selection for any finite discrete distribution with the property TP_2.

2. Minimization for the Class of Discrete Distributions; d=0

If we denote

$$(2.1)\qquad A_i = \sum_{j=0}^{i} a_j \qquad i=0,\dots,n$$

$$A_i = 0 \qquad i < 0$$

$$A_i = A_n \qquad i > n$$

(1.4) can be rewritten

$$(2.2)\qquad S = \sum_{i=0}^{n} (A_i - A_{i-1}) A_{i+d}^{k-1}$$

where the A_i must be nondecreasing and the condition $A_n = 1$ must hold. At first we will consider the case $d=0$:

$$(2.3)\qquad S = \sum_{i=0}^{n} (A_i - A_{i-1}) A_i^{k-1}.$$

By differentiation we obtain the conditions

$$\text{(2.4)} \qquad \frac{\partial S}{\partial A_i} = kA_i^{k-1} - (k-1)A_{i-1}A_i^{k-2} - A_{i+1}^{k-1} = 0 \qquad i=0,1,2,\ldots,n-1$$

which gives us the recursion formula

$$\text{(2.5)} \qquad A_{i+1} = \sqrt[k-1]{kA_i^{k-1} - (k-1)A_{i-1}A_i^{k-2}} \qquad i=0,1,\ldots,n-1.$$

Because $A_{-1}=0$ is given, the sequence $\{A_i\}$ would be defined, if A_0 is known. The problem is to determine A_0 such that $A_n = 1$ is satisfied. We define c_i, $i=0,\ldots,n$, by

$$\text{(2.6)} \qquad A_{i-1} = c_i A_i$$

and it will become evident that the c_i's are independent of the initial value A_0: Substituted (2.6) in (2.5) gives us

$$\text{(2.7)} \qquad A_{i+1} = \sqrt[k-1]{k-(k-1)c_i}\, A_i = \frac{1}{c_{i+1}} A_i$$

which yields for the sequence $\{c_i\}$ the recursion formula

$$\text{(2.8)} \qquad c_{i+1} = [k-(k-1)c_i]^{\frac{-1}{k-1}}$$

where c_0 follows from $A_{-1} = 0$.

Hence

$$\text{(2.9)} \qquad A_n = \frac{A_0}{c_1 \cdot c_2 \cdot \quad \cdot c_n} = 1$$

and

$$\text{(2.10)} \qquad A_0 = \prod_{i=1}^{n} c_i .$$

To evaluate the minimum value of S in (2.3) we insert (2.4) and obtain

$$(2.11)\qquad S_{MIN}(k,n) = \sum_{i=0}^{n-1} \left\{A_i - \frac{kA_i^{k-1} - A_{i+1}^{k-1}}{(k-1)A_i^{k-2}}\right\} A_i^{k-1} + 1 - A_{n-1}$$

$$= \frac{1}{k-1} \sum_{i=0}^{n-1}\left(A_{i+1}^{k-1} - A_i^{k-1}\right) A_i + 1 - A_{n-1}$$

$$= \frac{1}{k-1} \left\{ \sum_{i=0}^{n-1} A_{i+1}^{k-1} A_i - \sum_{i=0}^{n-1} A_i^{k-1} A_i \right\} + 1 - A_{n-1}$$

$$= \frac{1}{k-1} \left\{ \sum_{i=0}^{n} A_i^{k-1} A_{i-1} - \sum_{i=0}^{n-1} A_i^{k-1} A_i \right\} + 1 - A_{n-1}$$

$$= \frac{1}{k-1} (1 - S_{MIN}) + 1 - A_{n-1}.$$

Hence

$$(2.12)\qquad S_{MIN}(k,n) = \frac{1}{k} + \frac{k-1}{k}(1 - A_{n-1}) = \frac{1}{k} + \frac{k-1}{k} a_n .$$

Since $A_{n-1} = c_n A_n = c_n$ holds

$$(2.13)\qquad S_{MIN}(k,n) = \frac{1}{k} + \frac{k-1}{k} (1 - c_n)$$

(2.8) and (2.11) yield the recursion formula

$$(2.14)\qquad S_{MIN}(k,n+1) = 1 - \frac{k-1}{k \sqrt[k-1]{k S_{MIN}(k,n)}}$$

with $S_0 = 1$. If we denote $b_k = \dfrac{k-1}{k \sqrt[k-1]{k}}$ then (2.14) becomes

$$(2.15)\qquad S_{MIN}(k,n+1) = 1 - \frac{b_k}{\sqrt[k-1]{S_{MIN}(k,n)}} .$$

It must be shown that for the so determined A_i's S really attains its minimum and that $A_{i-1} \leq A_i$, $i=0,\ldots,n$, holds. If $A_{i-1} \leq A_i$, then from (2.5) follows

$$A_{\nu+1}^{k-1} = kA_{\nu}^{k-1} - (k-1)A_{\nu-1}A_{\nu}^{k-2}$$

$$\geq kA_{\nu}^{k-1} - (k-1)A_{\nu}A_{\nu}^{k-2} = A_{\nu}^{k-1}.$$

Since $A_{-1} = 0$ and $A_0 \geq 0$ holds the A_i's form a nondecreasing sequence, and therefore from (2.12) follows, that $S_{MIN}(n,k)$ is decreasing in n. If we assume that the minimum is attained on the boundary i.e., for some i's $A_i = A_{i+1}$ must hold, then the corresponding item in (2.3) vanishes and the problem of finding the minimum value of S on the boundary, is equal to the original problem for all smaller values of n. Since $\{S_{MIN}(k,m)\}$ is a nonincreasing sequence in m, the minimum occurs for $m=n$. Hence $S_{MIN}(k,n)$ is indeed the minimum. Table 1 shows the values of $S_{MIN}(k,n)$ for some k and n.

3. The General d Case

For general d differentiation of (2.2)

$$S = \sum_{i=0}^{n} (A_i - A_{i-1})A_{i+d}^{k-1}$$

leads to the system of conditions

$$\frac{\partial S}{\partial A_j} = -A_{j+d+1}^{k-1} + A_{j+d}^{k-1} + (k-1)A_j^{k-2}(A_{j-d} - A_{j-d-1}) = 0 \tag{3.1}$$

where again $A_i = 0$ for $i<0$ and $A_i = 1$ for $i \geq n$ is assumed.

Explicitly written (3.1) becomes

$$(3.2a)\quad \begin{cases} A_d - A_{d+1} & = 0 \\ \vdots & \vdots \\ A_{2d-1} - A_{2d} & = 0 \end{cases}$$

$$(3.2b)\quad \begin{cases} A_{2d}^{k-1} - A_{2d+1}^{k-1} + (k-1)(A_0 - A_{-1})A_d^{k-2} & = 0 \\ \vdots & \vdots \\ A_{n-1}^{k-1} - A_n^{k-1} + (k-1)(A_{n-2d-1} - A_{n-2d-2})A_{n-d-1}^{k-2} & = 0 \end{cases}$$

$$(3.2c)\quad \begin{cases} A_{n-2d-1} - A_{n-2d} & = 0 \\ \vdots & \vdots \\ A_{n-d-2} - A_{n-d-2} & = 0 \end{cases}$$

Hence must hold

$$(3.3)\qquad A_d = A_{d+1} = \dots\dots\dots = A_{2d}$$

and

$$(3.4)\qquad A_{n-2d-1} = A_{n-2d} = \dots\dots = A_{n-d-1}.$$

From the conditions (3.2b) we obtain

$$(3.5)\qquad A_i = A_{i-1} \text{ iff } A_{i+2d+1} = A_{i+2d}.$$

(3.2), (3.3) and (3.4) yield the conditions

$$(3.6)\qquad A_i = A_{i+1} = \dots\dots = A_{i+d} \quad \text{for } i \equiv d \mod(2d+1)$$
$$\text{or } i \equiv n \mod(2d+1)$$

(3.7) j be the smallest non-negative integer such that either $i \equiv d$ or $i \equiv n \mod(2d+1)$. $(2d+1)$ will be denoted by c.

If we choose $A_i = 0$ for $i < j$ we obtain by (3.4)

$$
\begin{aligned}
&A_{-1} = A_0 = \cdots\cdots = A_{j-1} = 0 = B_{-1} \\
&A_j = A_{j+1} = \cdots\cdots = A_{j+2d} = B_0 \\
(3.8)\qquad &A_{j+2d+1} = \cdots\cdots = A_{j+4d+1} = B_1 \\
&\vdots \qquad\qquad\qquad \vdots \\
&A_{j+\nu(2d+1)} = \cdots\cdots = A_{j+(\nu+1)(2d+1)-1} = B_\nu \\
&\vdots \qquad\qquad\qquad \vdots \\
&A_{j+n^*(2d+1)} = \cdots\cdots = A_n = 1 = B_{n^*}
\end{aligned}
$$

where the B_i and n^* are defined by these conditions. From (3.1) we obtain

$$
\begin{aligned}
(3.9)\qquad B_\nu^{k-1} &= A_{j+\nu\cdot c}^{k-1} = A_{j+\nu c-1}^{k-1} + (k-1)[A_{j+\nu c-2d-1} - A_{j+\nu c-2d-2}]A_{j+\nu c-d-1}^{k-2} \\
&= B_{\nu-1}^{k-1} + (k-1)(B_{\nu-1} - B_{\nu-2})B_{\nu-1}^{k-2} .
\end{aligned}
$$

This recursion formula is identical to (2.5). Hence the results of the case $d=0$ can be applied if only n is restored by

$$(3.10)\qquad n^* = \left[\frac{n-j}{2d+1} \right]$$

where j is defined as in (3.6) and $[x]$ denotes the biggest integer less than or equal to x. If we insert these A_i's in (2.2)

$$S_{MIN} = \sum_{i=0}^{n} (A_i - A_{i-1})A_{i+d}^{k-1}$$

the difference in the parentheses will vanish for any $i \neq j+\nu\cdot c$ such that (2.2) can be rewritten

$$S_{MIN} = \sum_{i=0}^{n^*} (B_i - B_{i-1})B_i^{k-1} .$$

Hence the results of the case $d=0$ can also be applied to calculate S_{MIN}.

4. Some New Results for the Binomial Case

In this chapter some numerically results for the binomial case shall be given, which are not directly related to the previous chapters. The expression (1.3)

$$P(CS/R) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left[\sum_{j=0}^{i+d} \binom{n}{j} p^j (1-p)^{n-j} \right]^{k-1}$$

is a polynomial of degree $n \cdot k$ in p, say

$$P(CS/R) = Q_{k,n,d}(p) = \sum_{i=0}^{n \cdot k} c_i(k,n,d) p^i .$$

Since $Q_{k,n,d}(p)$ is a probability and since $Q_{k,n,d}(0) = Q_{k,n,d}(1) = 1$, the minimum is attained for some p_0, $0 < p_0 < 1$, for which

$$\left.\frac{dQ}{dp}\right|_{p_0} = 0 \qquad \text{holds.}$$

In order to do this differentiation, the coefficients $c_i(k,n,d)$ have been evaluated numerically for $k = 2,(1),7$, $n = 2,(1),7$ and $d = 0,(1),n-1$. It turned out that

$$c_0(k,n,d) = 1$$

$$c_i(k,n,d) = 0 \qquad 0 < i \leq d$$

$$c_{d+1}(k,n,d) = -(k-1) \cdot \binom{n}{d+1}$$

holds in all these cases. Hence, for $d>0$, the first through the d'th derivatives vanish at $p=0$. It also turned out that these derivatives are zero for $p=1$. Therefore the first derivative is of the form

$$\frac{dQ}{dp} = [p(1-p)]^{d-1} \cdot T(p)$$

where $T(p)$ is a polynomial of a smaller degree. Since the coefficients of Q, and especially that of $\frac{dQ}{dp}$, appeared to be very large, it is useful to divide $\frac{dQ}{dp}$ by $[p(1-p)]^{d-1}$ and find the zeros of the smaller polynomial $T(p)$. The computations showed that $T(p)$ does not necessarily have a single zero in $[0,1]$, which means that $Q(p)$ may have several local minima in the unit interval, one of which is the minimum of Q in $[0,1]$. Table 2 shows the minima values of Q for $d=0$ and $d=1$. The missing values were not calculated, because the degrees or coefficients of the polynomials $T(p)$ in these cases were too large for the computer program, that was used. It should be pointed out, that the d-values determined to guarantee a given probability of correct selection, p^*, as it was done in [1] by approximative methods, agree completely with the results of this paper.

5. Acknowledgement

I would like to express my appreciation to Professor S.S. Gupta for having suggested these problems and for all of his helpful advice.

Table 1 - S(k,n), b_k. (c.f. 2.14, 2.15)

P \ K	2	3	4	5	6	7	8
1	0.75000	0.61510	0.52753	0.46501	0.41754	0.38027	0.34988
2	0.66667	0.50923	0.41527	0.35214	0.30655	0.27190	0.24465
3	0.62500	0.46065	0.36671	0.30550	0.26225	0.23004	0.20504
4	0.60000	0.43283	0.33992	0.28040	0.23891	0.20828	0.18472
5	0.58333	0.41499	0.32301	0.26480	0.22457	0.19506	0.17248
6	0.57143	0.40251	0.31139	0.25421	0.21491	0.18622	0.16434
7	0.56250	0.39332	0.30294	0.24656	0.20793	0.17990	0.15854
8	0.55556	0.38627	0.29651	0.24073	0.20277	0.17517	0.15421
9	0.55000	0.38070	0.29147	0.23627	0.19871	0.17150	0.15086
10	0.54545	0.37618	0.28740	0.23264	0.19547	0.16857	0.14819
11	0.54167	0.37245	0.28406	0.22967	0.19281	0.16618	0.14602
12	0.53846	0.36931	0.28125	0.22720	0.19060	0.16419	0.14421
13	0.53571	0.36664	0.27886	0.22510	0.18874	0.16251	0.14269
14	0.53333	0.36434	0.27684	0.22330	0.18714	0.16107	0.14139
15	0.53125	0.36233	0.27507	0.22174	0.18575	0.15983	0.14027
16	0.52941	0.36056	0.27351	0.22037	0.18454	0.15875	0.13929
17	0.52778	0.35900	0.27214	0.21916	0.18348	0.15779	0.13842
18	0.52632	0.35761	0.27091	0.21809	0.18253	0.15695	0.13765
19	0.52500	0.35636	0.26982	0.21714	0.18168	0.15619	0.13697
20	0.52381	0.35523	0.26883	0.21627	0.18092	0.15551	0.13636
21	0.52273	0.35421	0.26794	0.21549	0.18023	0.15489	0.13581
22	0.52174	0.35328	0.26713	0.21478	0.17961	0.15434	0.13530
23	0.52083	0.35242	0.26639	0.21414	0.17904	0.15383	0.13485
24	0.52000	0.35164	0.26570	0.21354	0.17851	0.15336	0.13443
25	0.51923	0.35092	0.26507	0.21300	0.17803	0.15293	0.13404
B(K)	0.25000	0.33490	0.47247	0.53499	0.58236	0.61973	0.65012

Table 1 (Cont'd.)

N \ K	9	10	12	15	20	25	30
1	0.32459	0.30316	0.26869	0.23082	0.18358	0.16049	0.14031
2	0.22259	0.20435	0.17583	0.14589	0.11411	0.09400	0.08008
3	0.18505	0.16870	0.14351	0.11745	0.09037	0.07357	0.06211
4	0.16502	0.15031	0.12753	0.10366	0.07914	0.06407	0.05386
5	0.15463	0.14016	0.11811	0.09564	0.07268	0.05866	0.04919
6	0.14709	0.13314	0.11194	0.09042	0.06852	0.05519	0.04622
7	0.14174	0.12817	0.10760	0.08676	0.06583	0.05279	0.04417
8	0.13775	0.12443	0.10439	0.08407	0.06350	0.05104	0.04267
9	0.13467	0.12163	0.10191	0.08200	0.06188	0.04970	0.04153
10	0.13223	0.11937	0.09995	0.08037	0.06060	0.04864	0.04063
11	0.13023	0.11753	0.09836	0.07904	0.05953	0.04779	0.03991
12	0.12855	0.11601	0.09705	0.07795	0.05871	0.04709	0.03932
13	0.12719	0.11473	0.09594	0.07703	0.05800	0.04651	0.03882
14	0.12600	0.11364	0.09500	0.07625	0.05739	0.04601	0.03840
15	0.12498	0.11269	0.09419	0.07558	0.05687	0.04556	0.03804
16	0.12408	0.11187	0.09345	0.07499	0.05641	0.04521	0.03773
17	0.12329	0.11115	0.09285	0.07443	0.05601	0.04489	0.03745
18	0.12260	0.11051	0.09231	0.07403	0.05566	0.04460	0.03721
19	0.12197	0.10994	0.09182	0.07362	0.05535	0.04434	0.03699
20	0.12141	0.10942	0.09138	0.07326	0.05507	0.04412	0.03680
21	0.12091	0.10895	0.09096	0.07293	0.05481	0.04391	0.03662
22	0.12045	0.10854	0.09062	0.07264	0.05459	0.04372	0.03647
23	0.12004	0.10816	0.09029	0.07237	0.05438	0.04355	0.03632
24	0.11965	0.10781	0.08999	0.07212	0.05419	0.04340	0.03619
25	0.11930	0.10749	0.08972	0.07189	0.05401	0.04326	0.03607
B(K)	0.67541	0.69804	0.73131	0.76916	0.81142	0.83951	0.85969

Table 2

Min P(CS/R) (c.f. 1.3 and ch.4)

d = 0

N \ K	2	3	4	5	6	7
2	0.68750	0.54488	0.46129	0.40487	0.36244	0.32894
3	0.65625	0.50664	0.41970	0.36165	0.31963	0.28755
4	0.63672	0.48338	0.39596	0.33858		
5	0.62305	0.46730				
6	0.61279	0.45537				

d = 1

N \ K	2	3	4	5	6	7
2	0.93750	0.88960	0.85084	0.81898	0.79052	0.76617
3	0.89063	0.81734	0.76334	0.72120	0.68700	0.65346
4	0.85547	0.76552	0.70179	0.65317		
5	0.82813	0.72652	0.65662			
6	0.80615	0.69596				
7	0.78803	0.67124				

REFERENCES

[1] S.S. Gupta and M. Sobel.

Selecting a Subset Containing the Best of Several Binomial Populations. Contributions to Probability and Statistics. Ch. XX. Stanford University Press.

[2] S.S. Gupta

On Some Selection and Ranking Procedures for Multivariate Normal Populations Using Distance Functions. Purdue University, Dept. of Statistics. Mimeograph Series No. 43. June 1965. To appear in the Proceedings of the International Symposium on Multivariate Analysis, Academic Press, 1966.