Biom. J. 29 (1987) 7, 857–863

On a Mixed-Type Randomized Rule for Selecting Superior Binomial Models

WEN-TAO HUANG and A. K. MD. EHSANES SALEH

Academia Sinica Taipei, Taiwan, R.O.C.
Carleton University Ottawa, Canada.

Summary

We propose a new class of selection rules for selecting superior models from finite Binomial models. This new class of rules extends the classes of classical rules and shows its superiority to the classical selection rules by some Monte Carlo results. This new class of rules is easier and more flexible to apply than these known classical rules.

Key words: Binomial models; Randomized selection rules; Mixed-type selection rules; Preliminary tests.

1. Introduction

In many occasions, an experimenter is often confronted with choosing some processes or models which are considered superior than others in some sense when there is significant evidence to support that these models are not homogeneous (equivalent). For example, in k different processes of producing some product, we are interested in selecting which process has a least probability of producing defective items. Or, $k\,(\geq 2)$ different clinical treatments for a certain symptons (diseases), we are interested in selecting a subset (or some best) of them in the sense that has the highest probability of curing the symptons (diseases). For each experiment, an observation (response) can be classified as success or non-success. This kinds of Binomial processes or models frequently occur in clinics, Biostatistics, Engineering and Social Sciences etc. The problem of selecting a Binomial model associated with the largest probability of success has been formulated in two different types, which are, respectively, called the indifference zone formulation (SOBEL and HUYETT, 1957) and the subset type approach (GUPTA and SOBEL, 1960).

Let $\pi_1, \pi_2, \ldots, \pi_k$ denote k Binomial models such that π_i is associated with parameter p_i (probability of success) and common n (number of experiments). In other words, π_i denotes the i-th clinical treatment with n experiments. Let $p_{[1]} \leqq p_{[2]} \leqq \leqq \ldots \leqq p_{[k]}$ denote the ordered values of p_i. Let X_i denote a sample from π_i (i.e. total number of successful responses in n experiments of i-th treatment).

For the indifference zone formulation, the experimenter specifies Δ $(0<\Delta<1)$ and requires to do experiment at least n_0 times $\left(n_0\right.$ is needed to be determined and which depends on Δ, k and some prefixed value p^* $\left.\left(\frac{1}{k}<P^*<1\right)\right)$ so that under the assumption that $p_{[k]}-p_{[k-1]}\geqq\Delta$, the probability that the model having the largest observation is associated with $p_{[k]}$ is at least P^*, a prefixed value. In most cases, Δ is difficult to specify and in many situations, for economic reason or sampling restrictions etc., it is difficult to take a sample at least n_0. Especially, when Δ is small and P^* is large, n_0 would be large. On the other hand, when the assumption $p_{[k]}-p_{[k-1]}\geqq\Delta$ is not made, GUPTA and SOBEL (1960) considered to select all model π_i whenever $X_i\geqq\max_{1\leqq j\leqq k} X_j-c$, where the constant c is so chosen so that the probability that at least one best is included in the selected subset is at least P^* (usually call it P^*-condition), a prefixed value. As can be expected, when P^* is large and each p_i is close to each other, the size of the selected subset may be large and it may contain all models which is undesirable.

In order to adjust these two factors, the minimum sample size n_0 and the size of selected subset, we are naturally first of all to ask whether all models are sufficiently close to each other (all p_i are close to each other). If it is, we just select any one of them since they are almost equally good. If not, we just select the smallest cluster that are sufficiently close to each other containing one best. Based on this idea of preliminary test, we propose a socalled mixed-type randomized rule which includes these two classical rules as extreme cases. We can control and adjust the factors among minimum sample size n_0, size of selected subset and the experimenter's choice of some value δ to a compromise which is desirable or acceptable to meet the experimenter's demand. In this sense, the proposed new rule is more flexible for application in Biostatistics and others.

2. A Mixed-Type Randomized Rule R

Let $\pi_1, \pi_2, \ldots, \pi_k$ denote k Binomial models such that π_i is associated with unknown p_i and common known n. Let X_i denote a sample from π_i. For a specified value δ $(0<\delta<1$, here δ is up to the experimenter's decision) we call π_i is good if $p_{[k]}-p_i<\delta$. If all models are good (there is only one cluster), we are satisfied to select any one of them. If not, we desire to select the one associated with $p_{[k]}$ (the best model). It is of course possible that there are more than one model that are associated with $p_{[k]}$, hence, we may selected a subset as small as possible to include one best. A correct selection (CS) occurs if at least one best (any one associated with $p_{[k]}$) is selected.

Let a and b $(0\leqq a, b\leqq n)$ denote two non-negative integers and let $X_{\max}=\max_{1\leqq j\leqq n} X_j$,

$X_{\min} = \min_{1 \leqq j \leqq n} X_j$. We propose a mixed-type randomized rule $R(a, b)$ as follows.

(2.1) $R(a, b)$: (i) If $X_{\max} - X_{\min} \leqq a$,

select the one associated with $X_{\max}$ and break a tie by a random mechanism.

(ii) If $X_{\max} - X_{\min} > a$, select π_j, if and only if, $X_j \geqq X_{\max} - b$ $(b < a \leqq n)$.

We note that when $a = n$, $R(n, b)$ becomes the Sobel-Huyett rule (1957), call it R_{SH}. When $a = 0$, $R(0, b)$ becomes the Gupter-Sobel rule (1960), as $\delta \to 0^+$, call it R_{GS}.

For a given values of δ and P^*, using some recursive method (in k), the exact values of constants a and b in the proposed rule can be obtained such that the P^*-condition ($P(\mathrm{CS}) \geqq P^*$ for any parameters) is satisfied. However, here we propose some other method for the approximations of a and b. We have the following formula to obtain the constants a and b in the proposed rule $R(a, b)$. When n is large enough, and if a and b satisfy the following

$$(2.2) \qquad 1 + (\Phi(\sqrt{2}c) - \Phi(-\sqrt{2}c))^{k+1} - \Phi^{k-1}\left(2\left(b + n\delta + \frac{1}{2}\right)\Big/(n(1-\delta^2))^{\frac{1}{2}}\right) \leqq \alpha$$

then, $P(\mathrm{CS}) \geqq 1 - \alpha$, where $c = 2\left(a + \frac{1}{2}\right)/(n(1-\delta^2))^{\frac{1}{2}}$.

3. Tables, Monte Carlo Studies and Examples

In table 1, we tabulate some values for a and b associated with $R(a, b)$ corresponding, respectively, to $P^* = 0.90$, $P^* = 0.95$, $k = 2(1)6$ and $\delta = 0.01$, 0.05, 0.1. For instance, when $k = 5$, $\delta = 0.01$, the first row in the entry of Table 1 is given by $0(5-18)2(5)$, which means when $n = 5$, $10(2)18$, $a = 0$ and when $n = 5$, b is given by $b = 2$. Also, when $k = 6$, $\delta = 0.1$, the second row of entry is given by $1(16-30)$ $3(16-30)$, and which means when $n = 16(2)30$, a and b are given by $a = 1$, $b = 3$ respectively.

According to $R(a, b)$, when the constant a is taken to be n, the sample size, $R(n, b)$ becomes the Sobel-Huyett rule. In Monte Carlo studies, we therefore consider only comparisons between R_{GS} and R, i.e. comparisons between subset-type rule and the mixed-type rule.

We consider sampling from four ($k = 4$) Binomial models with respective success probabilities, $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.4$ and $p_4 = 0.5$. By computer simulation of sampling from these four Binomial models, we simultaneously apply, respectively, R_{GS} and R to the observed data for our selection. We both observe the event of correct selection (selection of π_4 is a correct selection) and the size of the selected subset. Repeating 1000 times of sampling, we take the frequency of the event of correct selection and the average size of selected subset for our probability of correct selection and the expected size of selected subset and denote them,

Table 1

$P^* = 0.90$ $n = 5,10(2)30$

k	δ	0.01	0.05		0.1	
2	0(n)	1(5) 2(10–14) 3(16–26) 4(28–30)	0(n)	1(5) 2(10–30)	0(n)	1(5–30)
3	0(n)	2 (5–10) 3(12–18) 4(20–28) 5(30)	0(n)	1(5) 2(10–14) 3(16–30)	0(n)	1(5) 2(10–30)
4	0(n)	2(5) 3(10–14) 4(16–24) 5(26–30)	0(n)	2 (5–10) 3(12–24) 4(26–30)	0(n)	1(5) 2(10–30)
5	0 (5–18) 1(20–30)	2(5) 3(10–12) 4(14–20) 5(22–30)	0 (5–20) 1(22–30)	2(5) 3(10–30) 4(22–30)	0 (5–20) 1(22–30)	1(5) 2(10–20) 3(22–30)
6	0 (5–14) 1(16–30)	2(5) 3(10–12) 4(14–20) 5(22–28) 6(30)	0 (5–14) 1(16–30)	2(5) 3(10–16) 4(18–30)	0 (5–14) 1(16–30)	2 (5–14) 3(16–30)

Table 1 (continued)

$P^* = 0.95$ $n = 5,10(2)30$

k	δ	0.01	0.05		0.1	
2	0(n)	2(5–10) 3(12–18) 4(20–28) 5(30)	0(n)	1(5) 2(10–14) 3(16–30)	0(n)	1(5) 2(10–30)
3	0(n)	2(5) 3(10–12) 4(14–20) 5(22–30)	0(n)	2(5) 3(10–18) 4(20–30)	0(n)	1(5) 2(10–20) 3(22–30)
4	0(n)	2(5) 3(10) 4(12–18) 5(20–26) 6(28–30)	0(n)	2(5) 3(10–16) 4(18–28) 5(30)	0(n)	2 (5–12) 3(14–30)
5	0(n)	2(5) 3(10) 4(12–16) 5(18–24) 6(26–30)	0(n)	2(5) 3(10–14) 4(16–24) 5(26–30)	0(5–28) 1(30)	2 (5–10) 3(12–30)
6	0 (5–18) 1(20–30)	2(5) 3(10) 4(12–16) 5(18–22) 6(24–30)	0 (5–18) 1(20–30)	2(5) 3(10–12) 4(14–22) 5(24–30)	0 (5–18) 1(20–30)	2(5) 3(10–22) 4(24–30)

respectively, by PCS_{GS} and ES_{GS} when R_{GS} is applied and by PCS and ES when R is applied. We define a measure of efficiency of R_{GS} (R) by $\text{Ef}_{GS} \equiv PCS_{GS}/ES_{GS}$ ($\text{Ef} \equiv PCS/ES$). We consider efficiency of R to R_{GS} by the quantity $\text{Eff} = Ef/Ef_{GS}$. For some special values of n, the sample size, a, b (constants for R) and c (constant for R_{GS}), the associated PCS_{GS}, ES_{GS}, Ef_{GS} (PCS, ES, Ef) and Eff are computed and given in Table 2. For fixed $n = 20$ (first block of Table 2), we have considered 6 different rules of R_{GS} and R by taking 6 different constants for each rule. Taking ratio of max Ef to max Ef_{GS}, we define this quantity to be $\text{Eff}_n = \max Ef/\text{Max } Ef_{GS} = 0.743/0.536 = 1.386$ which is tabulated in the first row of last column, where the maximum is taken over 6 values of Ef and Ef_{GS}, respectively.

In Table 2, we have considered only small and large values of a for R, however, the most advantageous range of c for R_{GS} has been covered in Table 2. Accordingly,

Table 2

Monte Carlo Results

n	a	b	c	PCS_{GS}	PCS	ES_{GS}	ES	Ef_{GS}	Ef	Eff	Eff_n
20	0	1	2	0.924	0.862	1.740	1.427	0.531	0.604	1.138	1.386
	0	2	3	0.953	0.912	2.053	1.715	0.464	0.532	1.146	
	1	1	4	0.985	0.862	2.446	1.423	0.403	0.606	1.504	
	1	2	5	0.985	0.914	2.732	1.725	0.361	0.530	1.470	
	8	1	2	0.916	0.779	1.729	1.049	0.530	0.743	1.402	
	8	2	2	0.906	0.777	1.691	1.094	0.536	0.710	1.326	
30	0	1	2	0.919	0.873	1.525	1.311	0.603	0.666	1.105	1.205
	0	2	3	0.953	0.933	1.672	1.480	0.570	0.630	1.106	
	0	3	4	0.975	0.956	1.940	1.701	0.503	0.562	1.118	
	1	1	5	0.990	0.878	2.192	1.267	0.452	0.693	1.534	
	1	2	6	0.992	0.904	2.456	1.495	0.404	0.605	1.497	
	1	3	7	0.996	0.954	2.726	1.732	0.365	0.551	1.508	
	11	2	2	0.924	0.823	1.487	1.078	0.621	0.763	1.229	
	11	3	2	0.921	0.839	1.454	1.099	0.633	0.763	1.205	
40	0	1	2	0.935	0.893	1.326	1.181	0.705	0.756	1.072	1.201
	0	2	3	0.946	0.922	1.511	1.370	0.626	0.673	1.075	
	0	3	4	0.963	0.948	1.670	1.501	0.577	0.632	1.095	
	1	1	5	0.982	0.881	1.860	1.196	0.528	0.737	1.395	
	1	2	6	0.992	0.929	2.062	1.356	0.481	0.685	1.424	
	1	3	7	0.993	0.945	2.255	1.542	0.440	0.613	1.392	
	14	2	2	0.930	0.871	1.320	1.029	0.705	0.846	1.201	
	14	3	2	0.929	0.866	1.343	1.070	0.692	0.809	1.170	
50	0	2	2	0.940	0.940	1.264	1.264	0.744	0.744	1.000	1.133
	0	3	3	0.947	0.947	1.382	1.382	0.685	0.685	1.000	
	0	4	4	0.967	0.967	1.544	1.544	0.626	0.626	1.000	
	1	2	5	0.978	0.919	1.659	1.271	0.590	0.723	1.227	
	1	3	6	0.990	0.962	1.808	1.409	0.548	0.683	1.247	
	1	4	7	0.996	0.963	1.937	1.519	0.514	0.634	1.233	
	2	2	8	1.000	0.941	2.145	1.279	0.466	0.736	1.578	
	2	3	9	1.000	0.958	2.246	1.369	0.445	0.700	1.572	
	2	4	10	1.000	0.982	2.483	1.502	0.403	0.654	1.623	
	17	2	2	0.940	0.873	1.258	1.031	0.747	0.847	1.33	
	17	3	2	0.936	0.891	1.260	1.071	0.743	0.832	1.133	

the superiority of R to R_{GS} in our sense of efficiency is strongly supported by this Monte Carlo results given in Table 2.

Example. In a clinical experiment, data are obtained from five different treatments ($k=5$) such that each response is either success or non-success. In each treatment 20 experiments ($n=20$) are observed and the total number of successes for each treatment are respectively given by 11, 10, 17, 15, and 12. In order to select those treatments associated with the largest success probability, let the probability of correct selection be fixed by $P^*=0.90$. Then, following Table 1 of GUPTA and SOBEL (1960), the constant of the subset selection rule R_{GS} is given by $c=6$. Now, $X_{max}=17$, $X_{min}=10$. Hence, according to R_{GS}, all these treatments whose total success is bigger than or equal to $X_{max}-c=17-6=11$ are selected. Thus, treatment 1, 3, 4 and 5 are all selected. On the other hand, if we consider $\delta=0.01$ (i.e. we consider these treatments i are superior if $\max_j p_j-p_i\leqq\delta$), then, by Table 1, we have $a=1$ and $b=4$. Since $X_{max}-X_{min}=17-10=6>a=1$, we select those treatments if their total success numbers are not less than $X_{max}-b=17-4=13$, i.e. we select treatments 3, 4. Since $\delta=0.01$ is small and the selected subset size using our mixed-type rule R is just half of that using R_{GS}, we prefer R to R_{GS} for this case. To use R_{SH}, there is some difficulty since we do not know exact value of Δ, the lower bound of the difference between the largest and the second largest p_i's. However, if we estimate this value by $\hat{\Delta}=\frac{17}{20}-\frac{15}{20}=0.10$. Then, we can not claim that treatment 4 ($X_4=X_{max}=17$) is associated with $\max p_i$ since according to SOBEL and HUYETT (1957), it needs n to be bigger than 20 if $P^*=0.90$.

If the observed data are given respectively by 10, 10, 11, 11 and 11, then, according to R_{GS}, treatment 3, 4 and 5 are all selected, however, any one of treatment 3, 4 or 5 is selected if the mixed-type rule R is applied with $\delta=0.01$ and $P^*=0.90$. The size of selected subset is 3 for R_{GS} and is one for R. If R_{SH} is applied using $\hat{\Delta}=\frac{11}{20}-\frac{10}{20}=0.05$, we can not claim that either treatment 3, 4 or 5 is associated with $\max p_i$ with probability of correct selection $P^*=0.90$, since it needs $n>20$.

References

GUPTA, S. S. and SOBEL, M., 1960: Selecting a subset containing the best of several Binomial populations. In Contributions To Probability And Statistics (Ed. I. OLKIN) Stanford University Press, California, 224–248.

SOBEL, M. and HUYETT, M. J., 1957: Selecting the best one of several Binomial populations. Bell System Tech. J. 537–576.

Manuscript received: July 4, 1985

Authors' addresses:

W. T. HUANG
Institute of Statistics
Academia Sinica
Taipei, Taiwan, R.O.C.

A. K. Ehsanes SALEH
Department of Mathematics and Statistics
Carleton University
Ottawa, CANADA