The Annals of Statistics
1979, Vol. 7, No. 5, 1140–1143

ON A SCREENING PROBLEM[1]

BY JOSEPH A. YAHAV
The Hebrew University, Jerusalem

Fisher considered the problem of constructing sequentially a "better" finite population from a given infinite one. The purpose of this paper is to prove the optimality of Fisher's procedure.

1. Introduction. Fisher, in his paper on sequential experimentation [1], considered a problem of sequential screening. In this problem one is interested in selecting a finite number of female mice which have a specific genetic trait. Fisher stated the problem and prescribed a sequential probability ratio test as a solution. It is the purpose of this paper to prove the optimality of Fisher's procedure and to explain and generalize his ideas.

We consider a dichotomized infinite population. The proportion of A-elements is Π and the proportion of $\bar{A}$-elements is $1 - \Pi$. Given an element, we are not able to tell if it is an A-element or an $\bar{A}$-element. However, we can test the element, with a given test procedure, so that we get a positive or a negative result, where

$$P(+|A) = \alpha, \qquad P(+|\bar{A}) = \beta. \tag{1.1}$$

We assume Π, α and β to be known and $\beta < \alpha$. Furthermore, we assume that an element can be tested repeatedly with independent and identically distributed results, conditional on being A or $\bar{A}$, satisfying (1.1).

Our target is to construct a finite population, of size N, consisting of elements from the original population so that the proportion of A-elements in the new population exceeds Π^*, where $\Pi^* > \Pi$.

As is easily seen, this condition can never be satisfied with a finite number of tests unless $\alpha = 1$ and $\beta = 0$. Hence we reduce our standard somewhat, and ask that, for any element in the newly constructed population, the conditional probability of the element being an A-element (given that the element was selected) is greater or equal to Π^*.

The selection takes place in a sequential manner. There is a fixed cost C ($C > 0$) for each test; there are no additional costs in the process. The objective is to minimize the expected cost, subject to satisfying the standard.

Fisher prescribed the following procedure: take an element from the original population, keep testing it so long as

$$\Pi \leqslant P(A|\text{the results on testing}) < \Pi^*, \tag{1.2}$$

Received July 1977; revised October 1977.
[1]Research was supported in part by the Office of Naval Research contract N00014-75-C-0560 P00003 at Columbia University, New York.
AMS 1970 *subject classifications.* 62L05, 62L15.
Key words and phrases. Sequential probability ratio test, screening.

and stop the first time the inequality in (1.2) is violated. If $P(A|\text{the results on testing}) \geqslant \Pi^*$, then select this element. If $P(A|\text{the results on testing}) < \Pi$, then reject this element. Continue the testing until N elements are selected.

2. The case N = 1. For $N = 1$ we have to select one element subject to

$$P(A|\text{the element was selected}) \geqslant \Pi^*. \tag{2.1}$$

We restrict ourselves at this stage to procedures that do not permit recall of an element that was tested and rejected. Thus, we test elements $\varepsilon_1, \varepsilon_2, \cdots$ until an element satisfying (2.1) is selected. A sequential selection procedure is defined by a sequence of bivariate random variables $(T_1, I_1), (T_2, I_2), \cdots, (T_K, I_K)$. Where T_i is a stopping time for tests on element ε_i and

$$\begin{aligned} I_i &= 1 \quad \text{if element } i \text{ is selected} \\ &= 0 \quad \text{if element } i \text{ is rejected.} \end{aligned} \tag{2.2}$$

Hence, if ε_K is selected, K is a stopping time we have

$$I_i = 0, \qquad i = 1, 2, \cdots, K-1; \quad I_K = 1. \tag{2.3}$$

The total cost is then given by

$$\text{Cost} = C \cdot \Sigma_{i=1}^{K} T_i, \tag{2.4}$$

where $C > 0$ is the cost per test.

Our objective is to find stopping times $T_1, T_2, \cdots,$ and K so that the expected cost is minimized subject to

$$P(\varepsilon_K \text{ is an } A\text{-element}|I_K = 1) \geqslant \Pi^*. \tag{2.5}$$

Since the problem can be formulated as a negative dynamic programming problem, it is enough to consider stationary procedures. For stationary procedures we have (T_i, I_i) are i.i.d., so that

$$E\left[\Sigma_{i=1}^{K} T_i\right] = E\left[T_1\right].E\left[K\right]. \tag{2.6}$$

Since $E[T_1] = \infty$ or $E[K] = \infty$ implies $E[\Sigma_{i=1}^{K} T_i] = \infty$, it is enough to exhibit a stationary procedure for which $E[T_i] < \infty$ and $E[K] < \infty$ in order to conclude that an optimal stationary procedure (if it exists) satisfies (2.6).

Since the I_i are Bernoulli variables, K is geometrically distributed and we have

$$E[K] = \frac{1}{P(I_1 = 1)}. \tag{2.7}$$

Let $X_j(\varepsilon_i)$ denote the result on the jth test of the ith element:

$$\begin{aligned} X_j(\varepsilon_i) &= 1 \quad \text{if the result on test } j \text{ with element } i \text{ is positive} \\ &= 0 \quad \text{if the result on test } j \text{ with element } i \text{ is negative.} \end{aligned} \tag{2.8}$$

Let $S_r(\varepsilon_i) = \Sigma_{i=1}^{r} X_j(\varepsilon_i)$. Suppose $T_i = n$, and $S_{T_i} = m$. Then (2.5) is equivalent to

$$\frac{\alpha^m(1-\alpha)^{n-m}.\Pi}{\alpha^m(1-\alpha)^{n-m}.\Pi + \beta^m(1-\beta)^{n-m}.(1-\Pi)} \geqslant \Pi^* \tag{2.9}$$

or

$$LR_n(\varepsilon_i) = \left(\frac{\alpha}{\beta}\right)^m \left(\frac{1-\alpha}{1-\beta}\right)^{n-m} \geqslant \frac{\Pi^*(1-\Pi)}{\Pi(1-\Pi^*)}. \tag{2.10}$$

$LR_n(\varepsilon_i)$ is the likelihood ratio for ϵ_i after n tests.

We can conclude now that whenever we have n trials on element ε_i with m positive results, so that (2.10) is satisfied, then there is no need for additional testing and $i = K$.

We can therefore identify the stopping and selection region as follows: stop and select ε_i whenever $LR_n(\varepsilon_i) \geqslant \Pi^*(1-\Pi)/\Pi(1-\Pi^*)$. That is,

$$\{T_i = n, I_i = 1\} = \left\{LR_n(\varepsilon_i) \geqslant \frac{\Pi^*(1-\Pi)}{\Pi(1-\Pi^*)}\right\}.$$

It still remains to identify the stopping and rejection region. Consider the problem of testing the hypothesis H_0: "ε_1 is an A-element" versus the alternative H_1: "ε_1 is an $\bar{A}$-element."

LEMMA 2.1. *Using a Bayesian framework with* $P(H_0) = \Pi$, *we have that* P_{H_1}*(accepting* $H_0) = (\gamma/(1-\Pi))(1-\Pi^*)$ *and* P_{H_0}*(rejecting* $H_0) = 1 - (\Pi^*/\Pi)\gamma$ *are equivalent to* $P(H_0|$*accepting* $H_0) = \Pi^*$ *and* $P($*accepting* $H_0) = \gamma$.

PROOF. P(accepting H_0) $= \Pi P_{H_0}$(accepting H_0) $+ (1-\Pi)P_{H_1}$(accepting H_0) and

$$P(H_0|\text{accepting } H_0) = \frac{P(H_0)\cdot P_{H_0}(\text{accepting } H_0)}{P(\text{accepting } H_0)}.$$

Let (T_1, I_1) denote the variables for any stationary procedure in our original problems and let $\gamma = P(I_1 = 1)$. Consider a W.S.P.R.T. with P(type I error) $= 1 - (\Pi^*/\Pi)\gamma$ and P(type II error) $= \gamma(1 - \Pi^*/1 - \Pi)$. Applying Lemma 2.1 to this test, we have $P(H_0/\text{accepting } H_0) \geqslant \Pi^*$ and P(rejecting H_0) $= \gamma$. Let T^* be the W.S.P.R.T. stopping time. Then we have $E[T^*] \leqslant E[T]$ and so $E[T^*]/\gamma \leqslant E[T]/\gamma$. We can conclude that for any stationary procedure that satisfies (2.9) there is a W.S.P.R.T. which satisfies (2.9) and which does at least as well in terms of expected cost.

The W.S.P.R.T. is defined by two constants, say B_0, B_1, $B_0 \leqslant 1 < B_1$; sampling is continued so long as

$$B_0 \leqslant LR_n < B_1. \tag{2.11}$$

We have already identified the upper bound as

$$B_1 = \frac{\Pi^*(1-\Pi)}{\Pi(1-\Pi^*)}. \tag{2.12}$$

To identify B_0 we need two lemmas.

LEMMA 2.2. *For the* W.S.P.R.T., $E_{H_0}[T]$, $E_{H_1}[T]$, *and* $E[T]$ *are nonincreasing functions of* B_0.

PROOF. Immediate.

LEMMA 2.3. *For the* W.S.P.R.T., *P(accepting* $H_0|LR_n$*) is a nondecreasing function of* LR_n.

PROOF. Note first that we have defined LR_n as the likelihood under H_0 divided by the likelihood under H_1, and that the acceptance region of H_0 was $\{LR_n \geqslant B_1\}$. Given LR_n, the continuation region can be viewed as a new W.S.R.R.T. with boundaries B_0/LR_n, B_1/LR_n for observations $n+1, n+2, \cdots$. Since both new boundaries are decreasing functions of LR_n, the probability of exiting through the upper boundary is an increasing function of LR_n. For $LR_n < B_0$ or $LR_n \geqslant B_1$ the result is immediate.

THEOREM 2.1. *Among all procedures satisfying* (2.5) *the* W.S.P.R.T. *with* $B_0 = 1$ *and* $B_1 = (\Pi^*(1-\Pi)/\Pi(1-\Pi^*))$ *minimizes* $E[T]/P(\text{selecting } \varepsilon_1)$.

PROOF. Follows from Lemmas 2.1, 2.2 and 2.3.

3. N $\geqslant$ 2. For $N \geqslant 2$ we have $(T_{11}, I_{11}), (T_{12}, I_{12}), \cdots, (T_{1K_1}, I_{1K_1})$; $(T_{21}, I_{21}), (T_{22}, I_{22}), \cdots, (T_{2K_2}, I_{2K_2})$; $(T_{N_1}, I_{N_1}), (T_{N_2}, I_{N_2}), \cdots (T_{NK_N}, I_{NK_N})$. Our expected cost is $E[C \cdot \Sigma_{m=1}^{N}\Sigma_{j=1}^{K_m} T_{mj}]$.

Our problem is to minimize the expected cost subject to (2.5) for each of the N elements. Since N is a constant, we have an N-times repeated problem of choosing one element at a time, minimizing $E[\Sigma_{j=1}^{K_m} T_{mj}]$ subject to (2.5). If we use a weaker criterion, namely that (2.5) is to be satisfied on the average, then due to overshoot over the boundaries we can economize and adjust the upper boundary B_1 according to the accumulated average odds.

4. Remarks. It is clear from the analysis that the results are general enough to include the case for which the test response is a random variable with a distribution which depends on A and on $\bar{A}$. If tests are not conditionally independent the theory may fail.

REFERENCES

[1] FISHER, R. A. (1952). Sequential experimentation. *Biometrics* **8** 183–187.

DEPARTMENT OF STATISTICS
THE HEBREW UNIVERSITY
JERUSALEM, ISRAEL