ON THE DETECTION OF DEFECTIVE MEMBERS OF LARGE POPULATIONS[1]

By Andrew Sterrett

Denison University

1. Introduction. *The Annals of Mathematical Statistics* contained a note by Robert Dorfman [2] explaining an efficient method for eliminating all defective members of certain types of large populations. In particular, the application considered was the weeding out of all syphilitic men called up for induction into the armed forces. Instead of testing each blood sample individually, Dorfman proposed to pool k samples for a single analysis. The presence of syphilitic antigen in the pool led Dorfman to make k individual tests; the absence of the syphilitic antigen allowed him to clear k men with one test. One purpose of the note was to find the optimum k and the efficiency of the method for various prevalence rates of defectives. The purpose of this paper is to increase the efficiency of detection.

Rather than analyze each sample of a defective pool, it is proposed to make individual tests only until a defective is found. For small prevalence rates of defective members it is likely that a new pool formed from the untested samples will prove to be negative. If so, the work is finished for that pool; if not, one should test individuals again—but only until a defective is found. Continuing this procedure until a negative pool is found will increase Dorfman's efficiencies by about 6 per cent (from a savings over individual inspection of 80 per cent to a savings of 86 per cent for a prevalence rate of defectives equal to 0.01).

2. Notation. The probability that a pool containing k samples has exactly i defective members is given by $\mathrm{Pr}_k(i)$; the expected value of the number of analyses required to isolate the i defectives by the proposed method is $E_k(i)$.

Given a universe of N elements with p per cent defective, $E(N, k, p)$ is the total expected value of the number of analyses required to investigate the universe by pooling k samples at a time.

3. Procedure. Using the definition of expectation of a random variable,

$$E(N, k, p) = \frac{N}{k} \sum_{i=0}^{k} \{\mathrm{Pr}_k(i)\, E_k(i)\}. \tag{1}$$

Before $E(N, k, p)$ can be evaluated it must be shown that

$$E_k(i) = \frac{i}{i+1} k + i + 1 + \frac{i}{i+1} - 2i \cdot \frac{1}{k}. \tag{2}$$

When there are no defective elements in a pool, one laboratory analysis will suffice. That is, $E_k(0) = 1$ as Eq. (2) verifies.

Received January 8, 1957.

[1] The material in this article is derived from a dissertation submitted to the University of Pittsburgh, June, 1956.

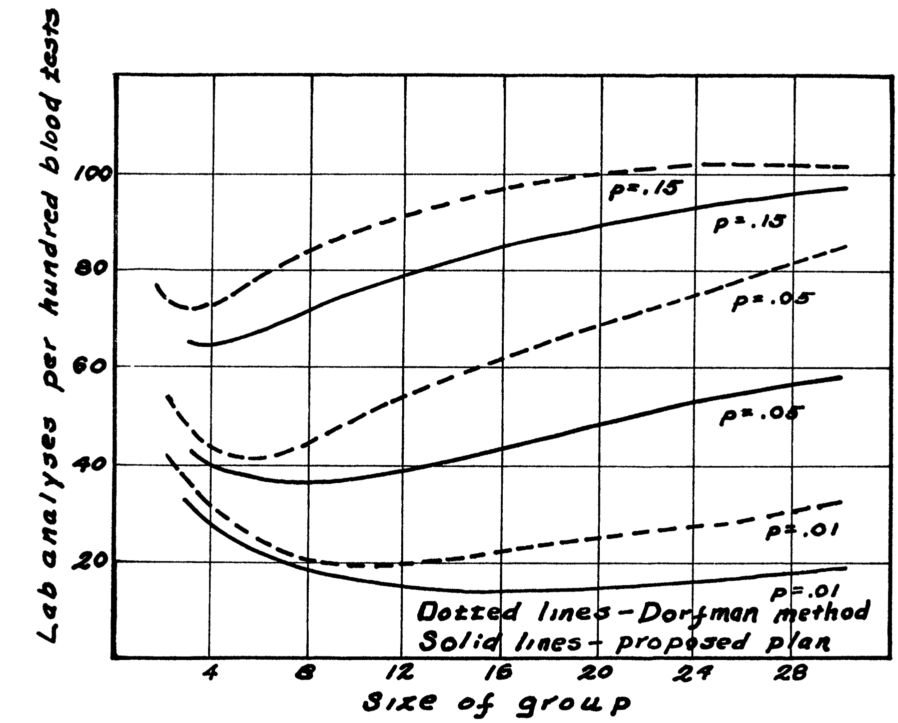

FIG. 1. Comparison of economies resulting from testing by two group methods

Now

$$
(3)\quad \begin{aligned} E_k(n) &= 1 + \frac{n}{k}\{1 + E_{k-1}(n-1)\} \\ &+ \sum_{j=1}^{k-n}\left[\left(\prod_{i=1}^{j}\frac{k-(i+n-1)}{k-(i-1)}\right)\frac{n}{k-j}\{(j+1) + E_{k-[j+1]}(n-1)\}\right]. \end{aligned}
$$

The first term on the right-hand side of Eq. (3) represents the initial group test. The factor n/k in the next term is the probability that the first sample tested is defective; the factor $\{1 + E_{k-1}(n - 1)\}$ is the sum of the number of tests required to find a defective on the first trial and the average number of tests needed to find $(n - 1)$ defectives in the remaining pool of $k - 1$ members. The probability that the first j samples are not defective is

$$
\prod_{i=1}^{j}[k-(i+n-1)]/[k-(i-1)],
$$

while the probability that the $(j + 1)$st element tested is defective is $n/(k - j)$. The number of tests required to find the first defective is $(j + 1,)$ and $E_{k-[j+1]}(n - 1)$ is the expected number of tests required to find the remaining $n - 1$ defectives among the $k - [j + 1]$ members.

Equation (3) reduces to the form given by Eq. (2) when values of $E_{k-[j+1]}(n - 1)$ obtained from Eq. (2) are properly substituted. The proof, then, of the formula for $E_k(i)$ follows by induction.

4. An approximation to $E(N, k, p)$. The probabilities connected with all but the first few terms of $E(N, k, p)$ are insignificant for small p. Therefore an approximation to $E(N, k, p)$ is defined as

$$E'(N, k, p) = \left(\frac{N}{k}\right) \sum_{i=0}^{m} \{\Pr_k(i) E_k(i)\},$$

where m is the smallest integer such that $\sum_{i=0}^{m} \Pr_k(i) > 0.99$.

TABLE I

Comparison of efficiencies by grouping under the Dorfman plan and the new method

Dorfman Plan				New Method		
p	Optimum k	Lab analyses per hundred	$(m + 1)$* or k	Optimum k	Lab analyses per hundred $E'(k, p)$	Difference
0.001	32	6	2	47	4	2
0.003	19	11	2	30	8	3
0.005	15	14	2	22	10	4
0.007	12	16	2	20	12	4
0.01	11	20	2	16	14	6
0.02	8	27	3	11	22	5
0.03	6	33	3	9	27	6
0.04	6	38	3	8	32	6
0.05	5	43	3	7	35	8
0.06	5	47	3	7	39	8
0.07	5	50	3	6	42	8
0.08	4	53	3	6	45	8
0.09	4	56	3	5	48	8
0.10	4	59	3	5	51	8
0.11			3	4	54	
0.12	4	65	3	4	57	8
0.13	4	67	3	4	59	8
0.14			3	4	61	
0.15	3	72	4	4	65	7
0.20	3	82	3	3	74	8
0.23			3	3	80	
0.25	3	91	3	3	84	7
0.26			3	3	86	
0.27			2	2	87	
0.30	3	99	2	2	90	9
0.32			2	2	93	
0.35	3	106	2	2	96	10
0.38			2	2	100	

* $(m + 1)$ is the number of subdivisions into which each member of the pool should be subdivided in order to be 99 per cent sure of knowing the history of the pool before exhausting any member.

The number of terms required to calculate $E'(N, k, p)$ is $m + 1$. This is also the minimum number of subdivisions into which an element must be divided by a laboratory technician to be at least 99 per cent confident that he will know the history of the group before exhausting any element. Values of $m + 1$ corresponding to many p's will be found in Table I.

5. An error expression. Define δ to be $E(N, k, p) - E'(N, k, p)$. In other words, $\delta = (N/k)\sum_{i=m+1}^{k} \{\Pr_k(i)E_k(i)\}$.

Since $E_k(k) = 2k - 1$, it follows that $\delta \leqq (2k - 1)\ (N/k)\sum_{i=m+1}^{k} \Pr_k(i)$.

Arbitrarily, m is chosen large enough to make $\sum_{i=0}^{m} \Pr_k(i)$ greater than 0.99. Therefore, $\sum_{i=m+1}^{k} \Pr_k(i)$ is less than 0.01. Consequently,

$$\delta < (2k - 1)(N/k)(0.01) = [2 - (1/k)]/100\cdot N.$$

That is, δ is less than $2 - (1/k)$ for each 100 items of the universe. This is a generous error since it was assumed that every pool containing more than m defectives contains k defective elements.

6. Conclusions. Using $E'(N, k, p)$, the optimum k and their corresponding economies are determined for many prevalence rates in the range $0.001 \leqq p \leqq 0.38$. Values of $E'(N, k, p)$ are calculated for $k = 4, 8, 12, \cdots$ and at the intermediate integral values necessary to insure that the minimum value is found. Results of this work and comparison with Dorfman's efficiencies are found in Table I.

REFERENCES

[1] STERRETT, ANDREW, "An efficient method for the detection of defective members of large populations," Ph.D. dissertation, University of Pittsburgh, 1956.

[2] DORFMAN, ROBERT, "The detection of defective members of large populations," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 436-440.

MAXIMUM LIKELIHOOD ESTIMATES IN A SIMPLE QUEUE

BY A. BRUCE CLARKE[1]

University of Michigan

0. Summary. The problem of obtaining maximum likelihood estimates for the parameters involved in a stationary single-channel, Markovian queuing process is considered. A method of taking observations is presented which simplifies this problem to that of determining a root of a certain quadratic equation. A useful and even simpler rational approximation is also studied.

1. Introduction. By a simple queue is meant a queue having a Poisson input and a negative exponential service time (type $M/M/1$ in the notation of Kendall

Received November 19, 1956.

[1] Research under contract with the Office of Naval Research carried out at Cornell University.