NOTES

This section is devoted to brief research and expository articles, notes on methodology and other short items.

THE DETECTION OF DEFECTIVE MEMBERS OF LARGE POPULATIONS

By Robert Dorfman

Washington, D. C.

The inspection of the individual members of a large population is an expensive and tedious process. Often in testing the results of manufacture the work can be reduced greatly by examining only a sample of the population and rejecting the whole if the proportion of defectives in the sample is unduly large. In many inspections, however, the objective is to eliminate all the defective members of the population. This situation arises in manufacturing processes where the defect being tested for can result in disastrous failures. It also arises in certain inspections of human populations. Where the objective is to weed out individual defective units, a sample inspection will clearly not suffice. It will be shown in this paper that a different statistical approach can, under certain conditions, yield significant savings in effort and expense when a complete elimination of defective units is desired.

It should be noted at the outset that when large populations are being inspected the objective of eliminating all units with a particular defect can never be fully attained. Mechanical and chemical failures and, especially, man-failures make it inevitable that mistakes will occur when many units are being examined. Although the procedure described in this paper does not directly attack the problem of technical and psychological fallibility, it may contribute to its partial solution by reducing the tediousness of the work and by making more elaborate and more sensitive inspections economically feasible. In the following discussion no attention will be paid to the possibility of technical failure or operators' error.

The method will be described by showing its application to a large-scale project on which the United States Public Health Service and the Selective Service System are now engaged. The object of the program is to weed out all syphilitic men called up for induction. Under this program each prospective inductee is subjected to a "Wasserman-type" blood test. The test may be divided conveniently into two parts:

1. A sample of blood is drawn from the man,
2. The blood sample is subjected to a laboratory analysis which reveals the presence or absence of "syphilitic antigen." The presence of syphilitic antigen is a good indication of infection.

When this procedure is used, N chemical analyses are required in order to detect all infected members of a population of size N.

The germ of the proposed technique is revealed by the following possibility. Suppose that after the individual blood sera are drawn they are pooled in groups of, say, five and that the groups rather than the individual sera are subjected to chemical analysis. If none of the five sera contributing to the pool contains syphilitic antigen, the pool will not contain it either and will test negative. If, however, one or more of the sera contain syphilitic antigen, the pool will contain it also and the group test will reveal its presence.[1] The individuals making up the pool must then be retested to determine which of the members are infected. It is not necessary to draw a new blood sample for this purpose since sufficient blood for both the test and the retest can be taken at once. The chemical analyses require only small quantities of blood.

Two questions arise immediately:

1. Will the group technique require fewer chemical analyses than the individual technique and, if so, what is the extent of the saving; and
2. What is the most efficient size for the groups?

Both questions are answered by a study of the probability of obtaining an infected group. Let

$p =$ the prevalence rate per hundred, that is the probability that a random selection will yield an infected individual. Then

$1 - p =$ the probability of selecting at random an individual free from infection. And

$(1 - p)^n =$ the probability of obtaining by random selection a group of n individuals all of whom are free from infection. Then

$p' = 1 - (1 - p)^n =$ the probability of obtaining by random selection a group of n with at least one infected member.

Further

$N/n =$ the number of groups of size n in a population of size N, so

$p'N/n =$ the expected number of infected groups of n in a population of N with a prevalence rate of p.

The expected number of chemical analyses required by the grouping procedure would be

$$E(T) = N/n + n(N/n)p'$$

or the number of groups plus the number of individuals in groups which require retesting.[2] The ratio of the number of tests required by the group technique to the number required by the individual technique is a measure of its expected relative cost. It is given by:

$$C = T/N = 1/n + p'$$
$$= \frac{n+1}{n} - (1 - p)^n.$$

[1] Diagnostic tests for syphilis are extremely sensitive and will show positive results for even great dilutions of antigen.

[2] The variance of T is $\sigma_T^2 = nNp'(1 - p') = nN[(1 - p)^n - (1 - p)^{2n}]$. The coefficient of variation of T becomes small rapidly as N increases.

The extent of the savings attainable by use of the group method depends on the group size and the prevalence rate. Figure 1 shows the shape of the relative cost curve for five prevalence rates ranging from .01 to .15.[3] For a prevalence rate of .01 it is clear from the chart that only 20% as many tests would be required by group tests with groups of 11 than by individual testing. The attainable savings decrease as the prevalence rate increases, and for a prevalence rate of .15, 72% as many tests are required by the most efficient grouping as by individual testing. The optimum group size for a population with a known prevalence rate is the integral value of n which has the lowest corresponding value on the relative cost curve for that prevalence rate.

TABLE I

Optimum Group Sizes and Relative Testing Costs for Selected Prevalence Rates

Prevalence Rate (per cent)	Optimum Group Size	Relative Testing Cost	Percent Saving Attainable
1	11	20	80
2	8	27	73
3	6	33	67
4	6	38	62
5	5	43	57
6	5	47	53
7	5	50	50
8	4	53	47
9	4	56	44
10	4	59	41
12	4	65	35
13	3	67	33
15	3	72	28
20	3	82	18
25	3	91	9
30	3	99	1

Optimum group sizes and their costs relative to the cost of individual testing are given in Table I for selected prevalence rates.

This table, together with the description of the group testing technique as it might be applied to blood tests for syphilis, reveals the two conditions for the economical application of the technique:

1. That the prevalence rate be sufficiently small to make worth while economies possible; and

[3] The prevalence rate of syphilis among the first million selectees and volunteers was .0185 for whites and .2477 for other races. Geographically, the prevalence rate for whites ranged from .0505 in Arizona to .0051 in Wisconsin. See Parran, Thomas and Vonderlehr, R. A., *Plain Words about Venereal Disease*, Reynal and Hitchcock, New York.

2. That it be easier or more economical to obtain an observation on a group than on the individuals of the group separately.

Where these conditions exist, it will be more economical to locate defective members of a population by means of group testing than by means of individual testing.

The principle of group testing may be applied to situations where the interest centers in the degree to which an imperfection is present rather than merely in its presence or absence. For example, it could be applied to lots of chemicals where it is desired to reject all batches with more than a certain degree of impurity. If n samples of a chemical are pooled and subjected to a single analysis, the degree of impurity in the pool will be the average of the impurities in the separate samples. If the criterion were adopted that the members of a pool would be examined individually whenever the proportion of impurity in the pool is greater than $1/n$-th the maximum acceptable degree of impurity, clearly no excessively impure batches would get by. The extent of the saving accomplished by this means can be computed by letting p' equal the probability that the pool will be impure enough to warrant retesting its constituent batches and using the formulas given above. The probability, p', can be calculated easily from the probability distribution of impurities in the separate batches.

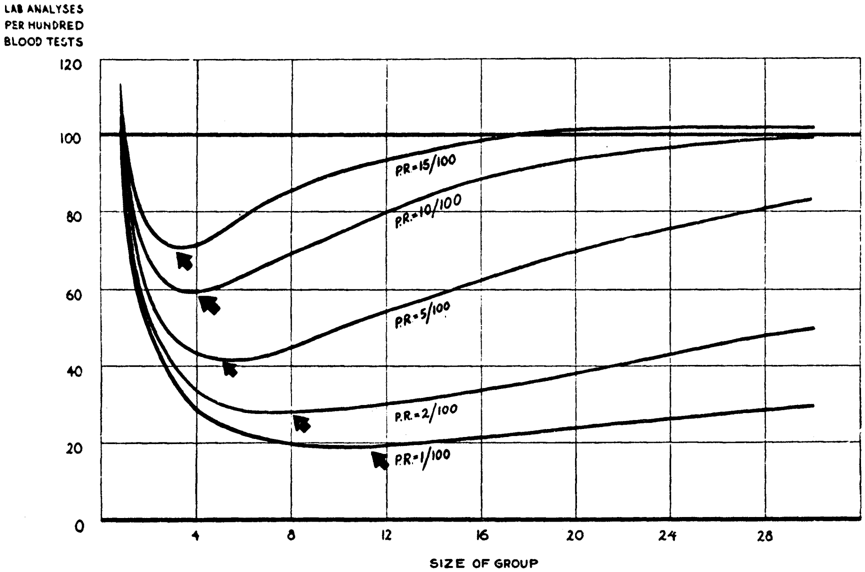

FIG. 1. Economies resulting from blood testing by groups
P.R. denotes prevalence rate

It is evident that this approach will produce worthwhile savings only if the limit of acceptability is liberally above the per cent of impurity encountered in the bulk of the batches. It is also evident that under this scheme many of the retests will indicate that all the batches in the pool are acceptable and that the retesting was not really needed. The criterion for retesting can be raised above $1/n$-th the limit of acceptability at the cost of a relatively small risk of accepting overly impure batches. The probability of failing to detect a defective batch when the retest criterion is raised in this manner will depend upon the form and parameters of the distribution of imperfection in single batches, as well as upon the number of batches in the pool. No simple general solution for this problem has been found.

FURTHER POINTS ON MATRIX CALCULATION AND SIMULTANEOUS EQUATIONS

By Harold Hotelling

Columbia University

Since the publication of "Some new methods in matrix calculation" in the *Annals of Mathematical Statistics* (March, 1943, pp. 1–34), the following relevant points have come to the attention of the author.

A. T. Lonseth has improved substantially the limit of error for the efficient method of inverting a matrix described on p. 14. He writes:

> "Your use of the 'norm' of a matrix in the *Annals* paper especially interests me, as I was recently led to use it in solving the errors problem for infinite linear systems which are equivalent to Fredholm-type integral equations.
>
> "It is possible to replace the term $p^{\frac{1}{2}}$ in your inequality (7.5) by one, so that
>
> $$N(C_m - A^{-1}) \leqq N(C_0)k^{2^m}/(1 - k).$$
>
> To see this, one observes that from the developments on the bottom of p. 13 it follows that $(I - D)^{-1} = I + D^*$, where $N(D^*) < k/(1 - k)$. Then
>
> $$C_0(I - D)^{-1} = C_0 + C_0D^*$$
>
> so that
>
> $$N[C_0(I - D)^{-1}] \leqq N(C_0) + N(C_0)\,N(D^*) = N(C_0)\{1 + N(D^*)\},$$
>
> from which the result stated is seen to follow. I happen to have noticed this because the same thing has cropped up often in my recent work, and for the infinite case a bound $p^{\frac{1}{2}}$ is no bound at all.
>
> "Your paper has suggested improvements in my own proofs, for which I am grateful."

Dr. Lonseth's first formula above might well be written at the bottom of p. 14 of my article as a substitute for (7.5). It both simplifies and reduces the limit of error.

A method of solving normal equations by iteration, in which trial values of the unknown regression coefficients were applied to the values of the predictors