Probability in the Engineering and Informational Sciences, **30**, 2016, 547–552.
doi:10.1017/S0269964816000267

ASYMPTOTIC ANALYSIS OF OPTIMAL NESTED GROUP-TESTING PROCEDURES

NABIL ZAMAN and NICHOLAS PIPPENGER
Harvey Mudd College, 301 Platt Blvd., Claremont, CA 91711, USA
E-mail: nzaman@g.hmc.edu; njp@math.hmc.edu.

We analyze a construction for optimal nested group-testing procedures, and show that, when individuals are independently positive with probability p, the expected number of tests per positive individual, $F^*(p)$, has, as $p \to 0$, the asymptotic behavior

$$F^*(p) = \log_2 \frac{1}{p} + \log_2 \log 2 + 2 + f\left(\log_2 \frac{1}{p} + \log_2 \log 2\right) + O(p),$$

where

$$f(z) = 4 \times 2^{-2^{1-\{z\}}} - \{z\} - 1,$$

and $\{z\} = z - \lfloor z \rfloor$ is the fractional part of z. The function $f(z)$ is a periodic function (with period 1) that exhibits small oscillations (with magnitude <0.005) about an even smaller average value (<0.0005).

1. INTRODUCTION

In 1943, Dorfman [2] introduced the study of *group testing*. In its simplest instantiation, there is a population of n individuals, each of which is independently *positive* with probability p, or *negative* with the complementary probability $1 - p$. In a single operation, it is possible to test any subset of the population, and the outcome of the test indicates whether or not one or more individuals in the tested set are positive. It is desired to find a testing strategy (a sequence of tests in which the later subsets tested may be depend upon the outcomes of previous tests) that determines, with a minimum expected number of tests, the status (positive or negative) of each individual. Let $G_n(p)$ denote this minimum expected number of tests. By concatenating strategies, we have subadditivity of $G_n(p)$: $G_{n+m}(p) \leq G_n(p) + G_m(p)$. This subadditivity implies (see Pólya and Szegö [7], Part I, Chapter 3, Section 1, Problem 98, pages 23, 198) that $G_n(p)/n$ tends to a limit as n tends to infinity. We shall denote this limit $G(p)$; it may be thought of as the asymptotic minimum expected number of tests per individual. For many purposes it is more convenient to work with $F(p) = G(p)/p$, which, by the following argument, may be thought of as the asymptotic minimum expected number of tests per positive individual. Suppose an infinite sequence of individuals is tested by runs of an optimal strategy achieving $G_n(p)$ for successive intervals of length n. As the number of runs of this strategy increases, the strong law of

© Cambridge University Press 2016 0269-9648/16 $25.00

https://doi.org/10.1017/S0269964816000267 Published online by Cambridge University Press

large numbers implies that #tests/#runs $\to G_n(p)$ and #runs/#positives $\to 1/np$. Thus, #tests/#positives $\to G_n(p)/np$. Letting $n \to \infty$, we obtain $G_n(p)/np \to G(p)/p = F(p)$.

Dorfman [2] suggested a strategy whereby individuals are first tested in batches. If the outcome of a batch test is negative, all the individuals in that batch are determined to be negative. If the outcome of a batch test is positive, the individuals in the batch are then tested one-at-a-time to determine the status of each. With an appropriate choice of the batch size, this strategy yields the bound $F(p) \le 2/p^{1/2} + O(1)$. Sterrett [10] gave a refinement of this strategy that yields the bound $F(p) \le (2/p)^{1/2} + O(1)$. Ungar [11] showed that $p_0 = (3 - \sqrt{5})/2 = 0.3819\ldots$ is the cut-off point for group testing, in the sense that for $p > p_0$, testing one-at-a-time is optimal, but for $p < p_0$, better strategies exist.

There is a lower bound for $F(p)$ that follows from the work of Huffman [5]. The sequence of outcomes of the tests in a strategy can be regarded as a sequence of binary symbols that forms a prefix-free code for the subset of positive individuals. Huffman showed that the expected length of a prefix-free code for a discrete random variable X is at least the entropy (in the sense of Shannon [8]) of that variable:

$$H(X) = -\sum_x \Pr[X = x] \log_2 \Pr[X = x].$$

In our case, the random variable X can be regarded as a sequence $X = (X_1, \ldots, X_n)$ of independent binary random variables, where X_i equals 1 (indicating that the ith individual is positive) with probability p, and equals 0 (indicating that the ith individual is negative) with probability $1 - p$. The entropy of each X_i is thus $H(X_i) = h(p)$, where $h(p) = -p \log_2 p - (1 - p) \log_2 (1 - p)$. Since the X_i are independent, the entropy of X is the sum of their entropies: $H(X) = n\, h(p)$. Thus, we have the lower bound $G_n(p) \ge n\, h(p)$, which implies

$$\begin{aligned} F(p) &\ge h(p)/p \\ &= \log_2 \frac{1}{p} + \log_2 e + O(p), \end{aligned} \tag{1.1}$$

where $e = 2.718\ldots$ is the base of natural logarithms and $\log_2 e = 1.442\ldots$. This argument is due in essence to Sobel and Groll [9], page 1212.

Sobel and Groll [9], page 1183, defined a class of group-testing strategies that have come to be called *nested* strategies. These strategies can be defined in several ways, but perhaps the simplest is this: at any point in the execution of a nested strategy, the population is partitioned into: (1) a set of individuals whose status has been determined, (2) a set of individuals about which it is known that at least one is positive, and (3) a set of individuals whose status is distributed exactly as it was at the outset of the execution. We define $G_n^*(p)$ by analogy with $G_n(p)$, with the minimum over strategies being restricted to nested strategies. By the arguments presented above, we then have that $G_n^*(p)/n$ tends to a limit $G^*(p)$, which may be thought of as the asymptotic minimum expected number of tests per individual, and that $F^*(p) = G^*(p)/p$ may be thought of as the asymptotic minimum expected number of tests per positive individual. It is not known whether there are nested strategies among the optimal group-testing strategies. But nested strategies have enough additional structure to allow $G_n^*(p)$ to be determined by a dynamic programming algorithm. Sobel and Groll [9] gave such an algorithm using $O(n^2)$ arithmetic steps; Hwang [6] gave an improved algorithm using only $O(n)$ steps. Chen, Hsu, and Sobel [1] have given a heuristic procedure for finding nested strategies with an expected number of tests per positive individual close to the lower bound (1.1) that applies to $F(p)$.

https://doi.org/10.1017/S0269964816000267 Published online by Cambridge University Press

Wolf [W] observed that there is a connection between nested strategies and certain prefix-free binary codes for the number K of negative individuals between two successive positive individuals. This number K has a geometric distribution:

$$\Pr[K = k] = p(1-p)^k \tag{1.2}$$

for $k \geq 0$. The codes in question are those we shall call "monotone". Suppose $\kappa(K)$ is a prefix-free binary code for K. We shall say that $\kappa(K)$ is *monotone* if each successive bit of $\kappa(K)$ is 1 or 0 according as $K \leq k$ or $K > k$, where k is a threshold that may depend on the preceding bits of $\kappa(K)$. Let $E^*(p)$ denote the minimum expected length of a monotone prefix-free binary code for K. Wolf observed that a sequence of nested strategies can be used to construct a monotone prefix-free binary code for K, with bits of the code corresponding to tests, so that $E^*(p) \leq F^*(p)$. Conversely, a monotone prefix-free binary code can be used to construct a sequence of nested strategies, with tests corresponding to bits of the code, so that $F^*(p) \leq E^*(p)$. Thus, $F^*(p) = E^*(p)$.

Gallager and Van Voorhis [3] (building on earlier work by Golomb [4]) found optimal prefix-free binary codes for geometrically distributed random variables, for any $0 < p < 1$. Some of these codes, as we shall see below, are monotone. Thus the value of $F^*(p)$ is in principle determined by the work of Wolf [12] and Gallager and Van Voorhis [3].

Our goal in this paper is to analyze the construction of Gallager and Van Voorhis [3] to give an asymptotic formula for $F^*(p)$. We shall show that

$$F^*(p) = \log_2 \frac{1}{p} + \log_2 \log 2 + 2 + f\left(\log_2 \frac{1}{p} + \log_2 \log 2\right) + O(p), \tag{1.3}$$

where the function $f(z)$ is a bounded periodic function of z with period 1. Specifically,

$$f(z) = 4 \times 2^{-2^{1-\{z\}}} - \{z\} - 1,$$

with $\{z\} = z - \lfloor z \rfloor$ denoting the fractional part of z. The function $f(z)$ exhibits small oscillations about its average value $\omega = \int_0^1 f(z)\,dz = 4(\log_2 e)\big(E_1(\log 2) - E_1(2\log 2)\big) - 3/2 = 0.0004547\ldots$, where $E_1(y) = \int_y^\infty (e^{-x}/x)\,dx$. It assumes its largest value of $f(z_1) = 0.004195\ldots$ at $z_1 = 1 + \log_2 \log 2 - \log_2 x_1 = 0.7680\ldots$, where $x_1 = 0.8140\ldots$ is the smaller solution of the equation $xe^{-x} = (\log_2 e)/4$, and its smallest value of $f(z_0) = -0.003438\ldots$ at $z_0 = 1 + \log_2 \log 2 - \log_2 x_0 = 0.1934\ldots$, where $x_0 = 1.2123\ldots$ is the larger solution of that equation. Comparing this result with (1.1), we see that the entropy bound (1.1) is not achievable by nested strategies: the average of $F^*(p)$ lies $\log_2 \log 2 + 2 + \omega - \log_2 e = 0.02899\ldots$ above the entropy. This difference is a bound on the improvement possible by considering non-nested strategies.

2. ANALYSIS

According to Gallager and Van Voorhis [3], optimal prefix-free binary codes for a geometric random variable K, distributed according to (1.2), can be constructed as follows. Set

$$m = \left\lceil \frac{\log(2-p)}{-\log(1-p)} \right\rceil. \tag{2.1}$$

(Note that $m = 1$ for $p > p_0 = (3 - \sqrt{5})/2 = 0.3819\ldots$.) Divide K by m to obtain a quotient $S \geq 0$ and a remainder $0 \leq R \leq m - 1$:

$$K = Sm + R.$$

https://doi.org/10.1017/S0269964816000267 Published online by Cambridge University Press

The distribution of S is geometric with parameter $q = 1 - (1-p)^m$; that is

$$\Pr[S = s] = (1-q)q^s$$

for $s \geq 0$. The distribution of R is "truncated geometric":

$$\Pr[R = r] = \frac{(1-p)^r}{1-(1-p)^m} \tag{2.2}$$

for $0 \leq r \leq m-1$. We shall take optimal prefix-free codes $\sigma(S)$ for S and $\varrho(R)$ for R, and concatenate them (as strings) to obtain an optimal code $\kappa(K) = \sigma(S)\,\varrho(R)$.

Since (2.1) implies $q \geq (3-\sqrt{5})/2$, an optimum prefix-free code for the quotient S is $\sigma(S) = 1^S\,0$. This code is monotone, and its expected length is

$$\mathrm{Ex}[\#\sigma(S)] = \mathrm{Ex}[S] + 1 = \frac{1}{q} = \frac{1}{1-(1-p)^m}. \tag{2.3}$$

The optimum prefix-free code for the remainder R is a Huffman code $\varrho(R)$ (see Huffman [5]). We observe that the ratio between the smallest and the largest of the probabilities given by (2.2) is $(1-p)^{m-1}$. Since (2.1) implies that $(1-p)^{m-1} \geq 1/(2-p) > 1/2$, it follows that all of the codewords in this Huffman code must be of at most two consecutive lengths. (If there were a codeword ξ of length i and two codewords $\eta\,0$ and $\eta\,1$ each of length $j \geq i+2$, then the probability of ξ would be strictly smaller than the sum of the probabilities of $\eta\,0$ and $\eta\,1$, and the code with η of length $j-1$ and $\xi\,0$ and $\xi\,1$ each of length $i+1$ would have strictly smaller expected length.) Take

$$l = \lfloor \log_2 m \rfloor$$

and

$$h = m - 2^l,$$

so that $0 \leq h \leq 2^l - 1$. Then there will be $2^l - h$ codewords of length l and $2h$ codewords of length $l+1$. Since the $2h$ longer codewords will have the $2h$ smallest probabilities, the expected codeword length for the Huffman code is

$$\mathrm{Ex}[\#\varrho(R)] = l + \sum_{m-2h \leq k \leq m-1} \frac{(1-p)^k}{1-(1-p)^m} = l + \frac{(1-p)^{m-2h} - (1-p)^m}{1-(1-p)^m}. \tag{2.4}$$

Since longer codewords are assigned to larger values of R, the codewords can be assigned to values of K so that the code is monotone.

The concatenation of the monotone codes $\sigma(S)$ and $\varrho(R)$ for S and R yields a monotone code $\kappa(K) = \sigma(S)\,\varrho(R)$ for K. Combining (2.4) for the expected length of the encoding of R with (2.3) for the expected length of the encoding of S, we obtain

$$\begin{aligned} \mathrm{Ex}[\#\kappa(K)] &= \mathrm{Ex}[\#\sigma(S)] + \mathrm{Ex}[\#\varrho(R)] \\ &= l + \frac{(1-p)^{m-2h} - (1-p)^m}{1-(1-p)^m} + \frac{1}{1-(1-p)^m} \end{aligned} \tag{2.5}$$

for the expected length of the encoding of K.

https://doi.org/10.1017/S0269964816000267 Published online by Cambridge University Press

Since l and h are defined in terms of m, we shall start by eliminating them in favor of

$$\vartheta = \{\log_2 m\},$$

the fractional part of $\log_2 m$. Then $l = \log_2 m - \vartheta$, and from $m = 2^l + h$ we obtain $1 - 2h/m = 2^{1-\vartheta} - 1$. Substituting these expressions into (2.5) yields

$$\mathrm{Ex}[\#\kappa(K)] = \log_2 m - \vartheta + \frac{(1-p)^{m(2^{1-\vartheta}-1)} - (1-p)^m}{1-(1-p)^m} + \frac{1}{1-(1-p)^m}. \tag{2.6}$$

Since ϑ is defined in terms of m, our next step will be to eliminate p in favor of m by using the relation

$$m = \frac{\log 2}{p} + O(1), \tag{2.7}$$

which follows from (2.1) and implies

$$p = \frac{\log 2}{m} + O\left(\frac{1}{m^2}\right).$$

Thus

$$(1-p)^m = \frac{1}{2} + O\left(\frac{1}{m}\right).$$

Substituting this expression into (2.6) yields

$$\mathrm{Ex}[\#\kappa(K)] = \log_2 m + 2 + f(\log_2 m) + O\left(\frac{1}{m}\right), \tag{2.8}$$

where

$$f(z) = 4 \times 2^{-2^{1-\{z\}}} - \{z\} - 1.$$

The function f is periodic with period 1. It is also continuous (because $\lim_{z\to 1} f(z) = f(0) = 0$), and has a continuous derivative (because $\lim_{z\to 1} f'(z) = f'(0) = 2(\log 2)^2 - 1$). These properties, combined with the relation

$$\log_2 m = \log_2 \frac{1}{p} + \log_2 \log 2 + O(p),$$

which follows from (2.7), allow us to deduce

$$f(\log_2 m) = f\left(\log_2 \frac{1}{p} + \log_2 \log 2\right) + O(p).$$

This in turn allows us to rewrite (2.8) in terms of p as

$$\mathrm{Ex}[\#\kappa(K)] = \log_2 \frac{1}{p} + \log_2 \log 2 + 2 + f\left(\log_2 \frac{1}{p} + \log_2 \log 2\right) + O(p).$$

It remains to examine the properties of the function $f(z)$. The continuity of the periodic function $f(z)$ ensures that it is bounded, and the continuity of its derivative ensures that its maxima and minima occur at values of z for which the derivative vanishes. The vanishing of the derivative is given by the equation

$$1 = 4 \times 2^{-2^{1-\vartheta}} \times 2^{1-\vartheta} (\log 2)^2.$$

The substitution $x = 2^{1-\vartheta} \log 2$ reduces this equation to $x\, e^{-x} = (\log_2 e)/4$, from which the numerical results mentioned in Section 1 follow. The same substitution also reduces the integral $\omega = \int_0^1 f(z)\, dz$ for the average value of $f(z)$ to the integral $4(\log_2 e) \int_{\log 2}^{2\log 2} (e^{-x}/x)\, dx$, leading again to the numerical results mentioned in Section 1.

https://doi.org/10.1017/S0269964816000267 Published online by Cambridge University Press

References

1. Chen, P., Hsu, L., & Sobel, M. (1987). Entropy-based optimal group testing procedures. *Probability in the Engineering and Informational Sciences* 1: 497–509.
2. Dorfman, R. (1943). The detection of defective members of a large population. *Annals of Mathematical Statistics* 14: 436–440.
3. Gallager, R.G. & Van Voorhis, D.C. (1975). Optimal source codes for geometrically distributed integer alphabets. *IEEE Transactions on Information Theory* 21: 228–229.
4. Golomb, S.W. (1966). Run-length encodings. *IEEE Transactions on Information Theory* 12: 399–401.
5. Huffman, D. (1952). A method for the construction of minimum redundancy codes. *Proceding of the IRE* 40: 1098–1103.
6. Hwang, F.K. (1976). An optimum nested procedure in binomial group testing. *Biometrics* 32: 939–943.
7. Pólya, G. & Szegö, G. (1972). *Problems and theorems in analysis.* Berlin: Springer-Verlag.
8. Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 127: 379–423, 623–656.
9. Sobel, M. & Groll, P.A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Technical Journal* 38: 1179–1252.
10. Sterrett, A. (1957). On the detection of defective members of a large population. *Annals of Mathematical Statistics* 28: 1033–1036.
11. Ungar, P. (1960). The cutoff point for group testing. *Communications in Pure and Applied Mathematics* 13: 49–54.
12. Wolf, J. (1985). Born again group testing: multiaccess communications. *IEEE Transactions on Information Theory* 31: 185–191.

https://doi.org/10.1017/S0269964816000267 Published online by Cambridge University Press