AN EMPIRICAL DISTRIBUTION FUNCTION FOR SAMPLING WITH INCOMPLETE INFORMATION

By Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman

Sandia Corporation, University of Missouri, Northwestern University

Summary. For $i = 1, 2, \cdots, n$, let N_i independent trials be made of an event with probability p_i, and suppose that the probabilities p_i are known to satisfy the inequalities $p_1 \geqq p_2 \geqq \cdots \geqq p_n$. Let a_i denote the number of successes in the i-th trial, and p_i^* the ratio $a_i/N_i (i = 1, 2, \cdots, n)$. Then the maximum likelihood estimates $\bar{p}_1, \cdots, \bar{p}_n$ of the numbers $p_1, \cdots, p_n$ may be found in the following way. If $p_1^* \geqq p_2^* \geqq \cdots \geqq p_n^* \geqq 0$, then $\bar{p}_i = p_i^*, i = 1, 2, \cdots, n$. If $p_k^* \leqq p_{k+1}^*$ for some $k (k = 1, 2, \cdots, n - 1)$, then $\bar{p}_k = \bar{p}_{k+1}$; the ratios $p_k^* = a_k/N_k$ and $p_{k+1}^* = a_{k+1}/N_{k+1}$ are then replaced in the sequence $p_1^*, p_2^*, \cdots, p_n^*$ by the single ratio $(a_k + a_{k+1}) / (N_k + N_{k+1})$, obtaining an ordered set of only $n - 1$ ratios. This procedure is repeated until an ordered set of ratios is obtained which are monotone non-increasing. Then for each i, $\bar{p}_i$ is equal to that one of the final set of ratios to which the original ratio a_i/N_i contributed. It is seen that this method of calculating the $\bar{p}_1, \cdots, \bar{p}_n$ depends on a grouping of observations which might very well appeal to an investigator on purely intuitive grounds. It seems of interest to note that it yields the maximum likelihood estimates of the desired probabilities.

Particular examples of this situation are found in bio-assay [3] and in the proximity fuze problem discussed by M. Friedman ([1], Chapter 11).

The last section is devoted to a consistency property of the maximum likelihood estimators.

1. Introduction. In ordinary sampling one observes directly values of a random variable. There are, however, certain investigations, of which examples are to be found in a number of different fields, in which the result of each observation is not a sample value of the random variable being tested, but only a number, together with the information that the sample value is less than, or is greater than, that number. Bio-assay furnishes an example ([3]; for further references see [3], p. 416; [1], p. 352). Certain other examples occurring in the biological sciences have been suggested to the authors. Still another situation of this kind is mentioned by M. Friedman ([1], Chapter 11). Given a population of proximity fuzes, one is interested in the distribution of the random variable t, maximum distance from target at which a proximity fuze will operate. The result of a test of an individual proximity fuze is the distance of its nearest approach to the target and the information that it did or did not operate (we assume that the proximity fuze will not operate before reaching its point of closest approach to

Received March 8, 1954.

the target. We do not know whether or not this is in fact true of any actual proximity fuze; cf. [1], Chapter 11). The result of such an observation is therefore not a sample value of the random variable, **t**, but rather a distance, t_0, and the information that the sample value of **t** corresponding to the particular proximity fuze is less than t_0 (if it did not operate) or at least t_0 (if it did operate).

Let $F(t) = \Pr\{\mathbf{t} < t\}$; $F(t)$ is the distribution function of the random variable **t**. Let $p(t) = 1 - F(t) = \Pr\{\mathbf{t} \geqq t\}$; $p(t)$ represents the probability that the fuze will operate if its minimum distance from the target is t. Suppose R fuzes are tested, and observed to pass within distances $t_1, t_2, \cdots, t_n$ of the target ($n \leqq R$; several may have the same minimum distance from target); for convenience suppose the $\{t_i\}_1^n$ are arranged in increasing order. The R tests may be regarded as a set of R independent trials of events having probabilities $p_i = p(t_i)(i = 1, 2, \cdots, n)$ of success (those observed at the same minimum distance from target having the same a priori probability of operating), if the term "success" is used to signify that the proximity fuze operated. The problem is to estimate the probabilities $\{p_i\}_1^n$ from the results of the R trials.

In a typical bio-assay situation, a large number of trials is made at each parameter value $t_i(i = 1, 2, \cdots, n)$. In such a situation the ratios, number of successes divided by number of trials, each determined for a particular parameter value, will with high probability be in monotone non-increasing order (assuming $t_1 \leqq t_2 \leqq \cdots \leqq t_n$). The "best" estimates of the probabilities are then these ratios, and if $\bar{p}(t)$ is a non-increasing function assuming these values at the points $\{t_i\}_1^n$ then $\bar{F}(t) = 1 - \bar{p}(t)$ is an obvious empirical distribution function. In other cases, such as that discussed above, one might expect a small number of trials corresponding to each parameter value, so that the average numbers of successes could not be expected to be in monotone order. It is for such situations that the maximum likelihood estimators of the probabilities $\{p(t_i)\}_1^n$ are determined in this paper. If $\{\bar{p}_i\}_1^n$ denotes the set of maximum likelihood estimates, and if $\bar{p}(t)$ is a monotone non-increasing function such that $\bar{p}(t_i) = p_i(i = 1, 2, \cdots, n)$ then $\bar{F}(t) = 1 - \bar{p}(t)$ will be termed an *empirical distribution function*.

In bio-assay situations it is often assumed that the random variable in question (perhaps after an elementary transformation) is normally distributed. Methods of probit analysis ([1], [2], [3]) have been developed for use with such an assumption. While it is true that an empirical distribution function may be useful in determining parameters of a normal distribution under such an assumption, the primary purpose of this paper is to present estimators of the probabilities $\{p(t_i)\}_1^n$ without reference to any assumption as to the distribution of the random variable being tested. These estimators are derived in section 2. The calculations required for their computation are extremely simple and rapid. In section 3, the consistency of the estimators is considered. A theorem is proved which states that the empirical distribution function, $\bar{F}(t) = 1 - \bar{p}(t)$, converges in probability to the distribution function $F(t)$ as the number of tests or trials becomes infinite in an appropriate way.

2. Maximum likelihood estimators of the probabilities. Let $a_i + b_i$ independent trials be made corresponding to the same parameter value, or observation point, t_i, of which a_i are successes ($i = 1, 2, \cdots, n$). If $p_i = p(t_i)$ denotes the probability of success in a trial corresponding to the parameter value $t_i (i = 1, 2, \cdots, n)$, then the a priori probability of the event that, for each integer i, $1 \leqq i \leqq n$, a specified a_i of the trials will result in success is

$$\prod = \prod_{i=1}^{n} p_i^{a_i}(1 - p_i)^{b_i}. \tag{2.1}$$

Since $p(t)$ is non-increasing, the $\{p_i\}_1^n$ are known to satisfy the relations

$$1 \geqq p_1 \geqq p_2 \geqq \cdots \geqq p_n \geqq 0. \tag{2.2}$$

The maximum likelihood estimates of $\{p_i\}_1^n$ are those numbers, $\{\bar{p}_i\}_1^n$, which maximize the probability $\prod$ subject to the relations (2.2). (These estimates also maximize the probability,

$$\prod_{i=1}^{n} \binom{a_i + b_i}{a_i} p_i^{a_i}(1 - p^i)^{b_i},$$

that for each i, $1 \leqq i \leqq n$, there will be a_i successes among the $a_i + b_i$ trials at the observation point t_i).

In the context of the above discussion the numbers a_i and $b_i (i = 1, 2, \cdots, n)$ are non-negative integers. In section 3 they will be so regarded. However, the discussion of this section requires only that they be non-negative real numbers, such that $a_i + b_i > 0$ $(i = 1, 2, \cdots, n)$.

Let $\mathfrak{P}_n$ denote the class of sets of real numbers $\{p_i\}_1^n$ satisfying the inequalities (2.2). The problem is to determine a set $\{\bar{p}_i\}_1^n$ in $\mathfrak{P}_n$ affording a maximum value to $\prod$:

$$\prod_{i=1}^{n} \bar{p}_i^{a_i}(1 - \bar{p}^i)^{b_i} = \max_{\{p_i\} \in \mathfrak{P}_n} \prod_{i=1}^{n} p_i^{a_i}(1 - p^i)^{b_i}. \tag{2.3}$$

LEMMA 2.1. *There is a maximizing set* $\{\bar{p}_i\}_1^n$.

This follows immediately from the observation that the product is a continuous function of its arguments $p_1, p_2, \cdots, p_n$ and hence assumes its maximum on the closed, bounded set described by inequalities (2.2).

Set

$$p_i^* = a_i/(a_i + b_i) \qquad (i = 1, 2, \cdots, n). \tag{2.4}$$

THEOREM 2.1. *If* $\{\bar{p}_i\}_1^n$ *is a maximizing set, and if* $\bar{p}_k > \bar{p}_{k+1}$ *for some* k, $1 \leqq k \leqq n$, *then* $p_k^* \geqq \bar{p}_k > \bar{p}_{k+1} \geqq p_{k+1}^*$. *Also,* $p_1^* \leqq \bar{p}_1$, *and* $p_n^* \geqq \bar{p}_n$.

PROOF. We prove first that $p_k^* \geqq \bar{p}_k$. The basis of the proof is the observation that the function $p^a(1 - p)^b$ increases for $0 \leqq p < a/(a + b)$ and decreases for $a/(a + b) < p \leqq 1$. Suppose $\bar{p}_k > p_k^*$. Choose $p_k' = \max(p_k^*, \bar{p}_{k+1})$. Then $\bar{p}_1 \geqq \bar{p}_2 \geqq \cdots \geqq \bar{p}_{k-1} > p_k' \geqq \bar{p}_{k+1} \geqq \cdots \geqq \bar{p}_n$, while

$$p_k'^{a_k}(1 - p_k')^{b_k} > \bar{p}_k^{a_k}(1 - \bar{p}_k)^{b_k}.$$

This means that $\prod$ is increased by replacing $\bar{p}_k$ by p_k' (note that $\max \prod > 0$), contrary to (2.3). Therefore $p_k^* \geqq \bar{p}_k$. Similarly $\bar{p}_{k+1} \geqq p_{k+1}^*$. Hence $p_k^* \geqq \bar{p}_k > \bar{p}_{k+1} \geqq p_{k+1}^*$. The proof of the last statement of the theorem is similar.

For integers r, s, with $1 \leqq r \leqq s \leqq n$, define

$$\begin{cases} \alpha(r, s) = \sum_{\nu=r}^{s} a_\nu, \qquad \beta(r, s) = \sum_{\nu=r}^{s} b_\nu, \\ A(r, s) = \alpha(r, s)/[\alpha(r, s) + \beta(r, s)]. \end{cases} \tag{2.5}$$

THEOREM 2.2. *For* $1 \leqq i \leqq n$, *we have*

$$\bar{p}_i = \min_{1 \leqq r \leqq i} \max_{i \leqq s \leqq n} A(r, s) = \max_{i \leqq s \leqq n} \min_{1 \leqq r \leqq i} A(r, s)$$
$$= \min_{1 \leqq r \leqq i} \max_{r \leqq s \leqq n} A(r, s) = \max_{i \leqq s \leqq n} \min_{1 \leqq r \leqq s} A(r, s).$$

The original proof, based on Theorem 2.1, is omitted. The reader is referred to the following paper for a simpler proof.

COROLLARY 2.1. *The maximizing set* $\{\bar{p}_i\}_1^n$ *is unique.* Each $\bar{p}_i (i = 1, 2, \cdots, n)$ is determined uniquely by any of the formulas in Theorem 2.2.

Theorem 2.2 gives explicit formulas for the determination of the $\{\bar{p}_i\}$, but these are not recommended for calculation. Theorem 2.1 provides a means of calculating the maximizing set, $\{\bar{p}_i\}_1^n$, as outlined in the summary, which is very fast even for moderately large n.

The following interesting inequality was mentioned by a referee:

$$\sum_k (p_k^* - p_k)^2 (a_k + b_k) \geqq \sum_k (\bar{p}_k - p_k)^2 (a_k + b_k).$$

Here p_k^* and $\bar{p}_k$ are as defined above, while $p_1, p_2, \cdots, p_n$ is any set of numbers such that $1 \geqq p_1 \geqq p_2 \geqq \cdots \geqq p_n \geqq 0$. Indeed, one has

$$\sum_k (p_k^* - p_k)^2 (a_k + b_k) \geqq \sum_k (\bar{p}_k - p_k)^2 (a_k + b_k) + \sum_k (p_k^* - \bar{p})^2 (a_k + b_k),$$

as was shown by two of the authors, independently, in more general contexts, subsequent to the submission of the manuscript. These inequalities show that the numbers $\bar{p}_k$ are, on the average (in an obvious sense), closer to the numbers p_k respectively than are the numbers p_k^*.

3. The consistency of the estimators. Let $F(t)$ be the distribution function of the random variable $\mathbf{t}$ (see Section 1). The probability that $\mathbf{t}$ will assume a value t or greater is given by $p(t) = 1 - F(t)$. The method discussed in Section 2 provides the maximum likelihood estimates, $\bar{p}_i$, of $p(t)$ at specified parameter values, or observation points, $t_i (i = 1, 2, \cdots, n)$. Let $\bar{p}(t)$ denote any non-increasing function, $0 \leqq \bar{p}(t) \leqq 1$, assuming the values $\bar{p}_i$ at the points $t_i (i = 1, 2, \cdots, n)$, and $\bar{F}(t) = 1 - \bar{p}(t)$ an empirical distribution function associated with trials at the observation points $t_i (i = 1, 2, \cdots, n)$. If the points $t_1, \cdots, t_n$ were to remain fixed and the number of trials at each to increase in-

definitely, it would follow from the strong law of large numbers that for $k = 1, 2, \cdots, n$, p_k^* and $\bar{p}_k$ converge with probability 1 to p_k. In the following theorem, however, neither n nor the points $t_1, t_2, \cdots, t_n$, nor hence the probabilities p_k need remain fixed. For a fixed t_0, the number of trials made at t_0 need not become infinite, nor need any at all be made at t_0. We shall have $\bar{p}(t_0)$ near $p(t_0)$ with high probability if only enough trials are made at points *near* t_0, even if only one trial is made at each point.

The following theorem of Kolmogorov (strong law of large numbers) will be useful in establishing such a result.

LEMMA 3.1. (Kolmogorov) *Let $\mathbf{y}_j$ be a sequence of independent random variables having expected values $E(\mathbf{y}_j)$ and variances $V_j (j = 1, 2, \cdots)$. Let ϵ be an arbitrary positive number, and M a positive integer. Then*

$$\Pr\left\{\sup_{k \geqq M} \left| \frac{1}{k} \sum_{j=1}^{k} [\mathbf{y}_j - E(\mathbf{y}_j)] \right| \leqq \epsilon \right\} > 1 - \frac{16}{\epsilon^2} \left[\sum_{j=M}^{\infty} \frac{V_j}{j^2} + \frac{1}{4M^2} \sum_{j=1}^{M} V_j \right] \tag{3.1}$$

([4], p. 203).

THEOREM 3.1. *Let t_0 be a continuity point of the distribution function $F(t)$. Let ϵ, η be arbitrary positive numbers. Let t', t'' be chosen so that $t' < t_0 < t''$ and so that $|F(t) - F(t_0)| < \epsilon/2$ for $t' \leqq t \leqq t''$. Then*

$$\Pr\{|\bar{F}(t_0) - F(t_0)| < \epsilon\} > 1 - \eta \tag{3.2}$$

provided that at least N trials are made between t' and t_0 and at least N trials are made between t_0 and t'', where N is chosen so that

$$\sum_{j=N}^{\infty} \frac{1}{j^2} + \frac{1}{4N} < \epsilon^2 \eta / 32. \tag{3.3}$$

PROOF. We shall prove first that $\Pr\{\bar{F}(t_0) > F(t_0) - \epsilon\} > 1 - \eta/2$ or

$$\Pr\{\bar{p}(t_0) < p(t_0) + \epsilon\} > 1 - \eta/2 \tag{3.4}$$

provided that at least N trials are made between t' and t_0. It can be shown similarly that $\Pr\{\bar{F}(t_0) < F(t_0) + \epsilon\} > 1 - \eta/2$, or

$$\Pr\{\bar{p}(t_0) < p(t_0) - \epsilon\} > 1 - \eta/2, \tag{3.5}$$

provided that at least N trials are made between t_0 and t''. Inequality (3.2) follows from (3.4) and (3.5).

In order to establish (3.4), let $t^* = t_0$ if t_0 is an observation point. If not, let t^* denote the first observation point to the left of t_0. Since $\bar{F}(t_0) \geqq \bar{F}(t^*)$, or $\bar{p}(t_0) \leqq \bar{p}(t^*)$, it suffices to prove

$$\Pr\{\bar{p}(t^*) < p(t_0) + \epsilon\} > 1 - \eta/2. \tag{3.6}$$

Let the observation points be $\{t_i\} (i = 1, 2, \cdots, n)$ with $t_1 \leqq t_2 \leqq \cdots \leqq t_n$. Let t_m be the first observation point to the right of t'. Let M be the number of trials at observation points $t_m, t_{m+1}, \cdots, t_u = t^*$. By hypothesis, $M \geqq N$; that is, $\sum_{i=m}^{u} (a_i + b_i) \geqq N$. Order the trials at observation points $t_m, t_{m+1}, \cdots,$

t_n in the order of increasing t_i, ordering in an arbitrary way those occurring at the same observation point. Let $T_1, T_2, \cdots, T_M, T_{M+1}, \cdots, T_R$, where R is the total number of trials, denote the trials so ordered. Let $\{\mathbf{y}_j\}$ denote the number of successes in the trial $T_j (j = 1, 2, \cdots, R)$ so that $\mathbf{y}_j = 1$ with probability $p(t_i)$ and $\mathbf{y}_j = 0$ with probability $1 - p(t_i)$, where t_i is the observation point at which the trial T_j occurs. For $j > R$, let $\{\mathbf{y}_j\}$ be independent random variables, each assuming the value $p(t_0)$ with probability 1. Set $\mathbf{s}_k = \sum_{j=1}^{k} \mathbf{y}_j$. By Theorem 2.2,

$$\bar{p}(t^*) = \bar{p}(t_u) = \min_{1 \leqq r \leqq u} \max_{u \leqq s \leqq n} A(r, s).$$

Hence

$$\bar{p}(t^*) \leqq \max_{u \leqq s \leqq n} A(m, s)$$

(t_m is the first observation point to the right of t'). The symbol $A(m, s)$ represents the average number of successes in trials starting at t_m and terminating at t_s. Hence as s varies ($s \geqq u$) these ratios form a subsequence of the sequence $\mathbf{s}_k/k (k \geqq M)$. This implies that

$$\bar{p}(t^*) \leqq \sup_{k \geqq M} \mathbf{s}_k/k. \tag{3.7}$$

By Lemma 3.1,

$$\Pr\left\{\sup_{k \geqq M} \left[\frac{\mathbf{s}_k}{k} - \frac{1}{k}\sum_{j=1}^{k} E(\mathbf{y}_j)\right] \leqq \frac{\epsilon}{2}\right\} \geqq \Pr\left\{\sup_{k \geqq M} \left|\frac{\mathbf{s}_k}{k} - \frac{1}{k}\sum_{j=1}^{k} E(\mathbf{y}_j)\right| \leqq \frac{\epsilon}{2}\right\}$$

$$> 1 - \frac{64}{\epsilon^2}\left[\sum_{j=M}^{\infty} V_j/j^2 + \frac{1}{4M^2}\sum_{j=1}^{M} V_j\right].$$

But $V_j = \mathrm{Var}(\mathbf{y}_j) = p(t_i)[1 - p(t_i)] \leqq \frac{1}{4}$, t_i being the observation point at which the trial T_j occurs. Hence by hypothesis (3.3),

$$\Pr\left\{\sup_{k \geqq M}\left[\frac{\mathbf{s}_k}{k} - \frac{1}{k}\sum_{j=1}^{k} E(\mathbf{y}_j)\right] \leqq \frac{\epsilon}{2}\right\} > 1 - \frac{16}{\epsilon^2}\left[\sum_{j=M}^{\infty}\frac{1}{j^2} + \frac{1}{4M}\right] > 1 - \eta/2,$$

since $M \geqq N$. Further, if $1 \leqq j \leqq R$, then $E(\mathbf{y}_j) = p(t_i) < p(t_0) + \epsilon/2$; if $j > R$, then $E(\mathbf{y}_j) = p(t_0)$. Hence

$$\Pr\{\sup_{k \geqq M} \mathbf{s}_k/k < p(t_0) + \epsilon\} > 1 - \eta/2.$$

By (3.7) it then follows that

$$\Pr\{\bar{p}(t^*) < p(t_0) + \epsilon\} > 1 - \eta/2. \tag{3.6}$$

The proof of Theorem 3.1 is completed as indicated immediately following its statement.

REFERENCES

[1] E. EISENHART, M. W. HASTAY, W. A. WALLIS (Editors), *Techniques of Statistical Analysis*, McGraw-Hill, New York, 1947.

[2] D. J. FINNEY, *Probit Analysis*, Cambridge University Press, 2nd edition, 1951.

[3] C. H. GOULDEN, *Methods of Statistical Analysis*, John Wiley and Sons, 2nd edition, New York, 1952.

[4] J. V. USPENSKY, *Introduction to Mathematical Probability*, McGraw-Hill, New York, 1937.