COMPLETENESS FOR SEQUENTIAL SAMPLING PLANS

Ken-ichi Koike and Masafumi Akahira
Institute of Mathematics
University of Tsukuba
Tsukuba, Ibaraki 305, Japan

ABSTRACT

In the sequential multinomial sampling case, a sufficient condition for a non-randomized sequential procedure to be complete is given, and also a necessary and sufficient condition for a randomized sequential procedure to be complete is obtained.

1. INTRODUCTION

In the sequential binomial sampling case necessary and sufficient conditions on the sequential procedure for completeness are obtained (e.g. see Girshick *et al.* (1946) and Lehmann and Stein (1950)). In the sequential multinomial case, a sufficient condition on the non-randomized sequential procedure for completeness is given by Kremers (1990), but, in the sequential binomial case, the condition seems to be a little stronger than that of Girshick *et al.* (1946). In this paper a sufficient condition for completeness is given so that it coincides with that of Girshick *et al.* in the sequential binomial case. The condition is weaker than that of Kremers (1990). Further a necessary and sufficient condition for a randomized sequential procedure to be complete is obtained.

2. DEFINITIONS AND NOTATIONS

Suppose that $U_1, U_2, \ldots$ is a sequence of independent and identically distributed k-dimensional multinomial trials, that is , for each $i = 1, 2, \ldots$, $U_i = (U_{i1}, \ldots, U_{ik})$ is a random vector with $U_{ij} \in \{0, 1\}$ $(j = 1, \ldots, k)$, $\sum_{j=1}^{k} U_{ij} = 1$, and, for $p = (p_1, \ldots, p_k)$ with $0 < p_j < 1$ $(j = 1, \ldots, k)$ and $\sum_{j=1}^{k} p_j = 1$,

$$P_p(U_{ij} = 1) = p_j \quad (j = 1, \ldots, k).$$

In the sequential sampling a decision whether or not to sample U_{n+1} is based upon $U_1, \ldots, U_n$ for each positive interger n. The sample size may be a random variable specified by a sampling plan under consideration. So whenever a capital letter is used, it denotes a random variable (or vector). Since the random vector $X_N = \sum_{i=1}^{N} U_i$ is a sufficient statistic for p (e.g. see Ferguson (1967)), we consider only estimators based upon X_N, and assume that any decision whether or not to sample is also based upon X_N. Then let the stopping rule φ be a sequence

$$\varphi(z) = (\varphi_0, \varphi_1(x^{(1)}), \varphi_2(x^{(2)}), \ldots),$$

where $z = (x^{(1)}, x^{(2)}, \ldots)$ with φ_j defined on the sample space of X_j and $0 \leq \varphi_j \leq 1$ for all $j = 0, 1, \ldots$. For each $j = 1, 2, \ldots$, the function $\varphi_j(x^{(j)})$ represents the conditional probability that a statistician stop sampling , given that he has taken $X_j = x^{(j)}$, and φ_0 is a constant representing the probability of taking no observations at all. To avoid the case where the sampling continues forever , we assume that the stopping rule is closed , that is, $P_p(N < \infty) = 1$ for all p.

For a given stopping rule φ, the probability mass function of X_N is given by

$$P_p(X_N = x) = c(x) \prod_{j=1}^{k} {p_j}^{x_j},$$

where $0 \leq c(x) \leq N! / \prod_{j=1}^{k} x_j!$ with $x = (x_1, \ldots, x_k)$.

The outcome of such a X_N can be represented as a random walk in the k-fold direct product $\mathbf{N}_0^k$ of a set $\mathbf{N}_0$ of the all non-negative integers. The walk starts at origin and moves a unit to the direction according to the first trial's result. From the resulting point it again moves a unit in the same manner, and continues in this way until the stopping rule tells it to stop. The sequential procedure is said to be *complete* if X_N is complete as a statistic. For the purpose of the present paper it is enough to restrict to the space $\mathbf{N}_0^k$.

3. NON – RANDOMIZED SEQUENTIAL PROCEDURE

In this section we consider the case when a stopping rule is non-randomized, that is, in the above φ, each φ_n takes on only the values 0 or 1. For a point $x = (x_1, \ldots, x_k) \in \mathrm{N}_0^k$ the sum $\sum_{j=1}^k x_j$ of its coordinates is called *the index of* x. The point x is said to be *accessible* if $P_p(X_m = x, N \geq m) > 0$, otherwise it is said to be *inaccessible*. The point x is said to be *a continuation point* if it is accessible and $\varphi_m(x) = 0$, and a set of the all points is called a *continuation region*. The point x is called *a boundary point* if $\varphi_m(x) = 1$, and a set of the all points is called a *boundary region*. A stopping rule is said to be *bounded* if there exists some constant c with $0 < c < \infty$ such that the index of any accessible point is less than c, and simple if the convex hull of the continuation region on each index contains no points except for continuation points. In order to get a sufficient condition for completeness of the sequential procedure, we have the following lemma.

Lemma 1. *If the stopping rule φ has, as a boundary region, either*

$$\mathcal{B} = \{x = (x_1, \ldots, x_k) \ : \ \sum_{j=1}^k x_j A_j \geq c, \ \sum_{j=1}^k x_j = n\},$$

or

$$\mathcal{C} = \{x = (x_1, \ldots, x_k) \ : \ \sum_{j=1}^k x_j = m\},$$

where $c, A_1, \ldots, A_k$ are constants, and m and n are non-negative integers with $m > n$, then the sequential procedure is complete.

Proof. Without loss of genenality, we may assume that $A_k = \max\{A_1, \ldots, A_k\}$. At first we translate $\mathcal{B}$ over $\mathcal{C}$ parallel by $m - n$ in the positive direction of the axis of x_k, and we denote it by $\mathcal{B}'$. Then the point of $\mathcal{B}'$ is inaccessible. Indeed, otherwise, we have points $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ with $0 \leq y_j \leq x_j \quad (j = 1, \ldots, k)$ such that on the index n

$$\sum_{j=1}^k y_j A_j < c, \quad \sum_{j=1}^k y_j = n,$$

and on the index m

$$\sum_{j=1}^k x_j A_j \geq c + (m-n)A_k, \quad \sum_{j=1}^k x_j = m.$$

Since, by the assumption, $A_j \leq A_k \quad (j = 1, \ldots, k)$, we obtain $(x_j - y_j)A_j \leq (x_j - y_j)A_k \quad (j = 1, \ldots, k)$. Taking their sum , we have

$$\sum_{j=1}^{k}(x_j - y_j)A_j \leq (m-n)A_k. \tag{1}$$

On the other hand, if we translate the point y in the positive direction of the axis of x_k so that its coordinate is $(y_1, \ldots, y_{k-1}, y_k + m - n)$, then it is contained in $\mathcal{C} - \mathcal{B}'$. Hence we have

$$\sum_{j=1}^{k-1} y_j A_j + (y_k + m - n)A_k < c + (m-n)A_k. \tag{2}$$

Since, by the assumption,

$$c + (m-n)A_k \leq \sum_{j=1}^{k} x_j A_j, \tag{3}$$

we obtain from (2) and (3)

$$\sum_{j=1}^{k}(x_j - y_j)A_j > (m-n)A_k.$$

This contradicts the inequality (1). Hence the point of $\mathcal{B}'$ is inaccessible. If we redefine $\mathcal{X} = \mathcal{B} \cup (\mathcal{C} - \mathcal{B}')$ as a boundary region, a stopping rule with $\mathcal{X}$ is essentially the same as the previous one. Suppose that for a point $x = (x_1, \ldots, x_{k-1}, x_k) \in \mathcal{B}$ there exists a point x' satisfying $x' = (x_1, \ldots, x_{k-1}, x_k{}') \in \mathcal{C} - \mathcal{B}'$. Since, from $x \in \mathcal{B}$, $\sum_{j=1}^{k} x_j = n$, it follows that

$$\sum_{j=1}^{k-1} x_j A_j + x_k A_k = \sum_{j=1}^{k-1} x_j A_j + (n - \sum_{j=1}^{k-1} x_j)A_k \geq c,$$

hence

$$\sum_{j=1}^{k-1} x_j A_j - A_k \sum_{j=1}^{k-1} x_j \geq c - nA_k. \tag{4}$$

On the other hand, since, by $x' \in \mathcal{C} - \mathcal{B}'$, $\sum_{j=1}^{k-1} x_j + x_k{}' = m$, we similarly have by a straightforward computation

$$\sum_{j=1}^{k-1} x_j A_j - A_k \sum_{j=1}^{k-1} x_j < c - nA_k.$$

Then this fact contradicts the inequality (4). Since for a point $x = (x_1, \ldots, x_{k-1}, x_k) \in \mathcal{B}$ there does not exist a point x' satisfying $x' = (x_1, \ldots, x_{k-1}, x_k') \in \mathcal{C} - \mathcal{B}'$, it follows that $T = (x_1, \ldots, x_{k-1})$ and $X = (x_1, \ldots, x_{k-1}, x_k)$ are one-to-one.

To show the completeness of the sequential procedure, let f be any unbiased estimator of 0 , i.e., $E_p[f(X_N)] = 0$ for all p. Then it it enough to show that $f(x) = 0$ for all $x \in \mathcal{X}$. First we have, from $\sum_{j=1}^{k} p_j = 1$,

$$\begin{aligned} E_p[f(X_N)] &= \sum_x f(x)c(x)(1 - \sum_{j=1}^{k-1} p_j)^{x_k} \prod_{j=1}^{k-1} {p_j}^{x_j} \\ &= \sum_x f(x)c(x)\{1 + \sum_{l=1}^{x_k} \binom{x_k}{l} (-\sum_{j=1}^{k-1} p_j)^l\} \prod_{j=1}^{k-1} {p_j}^{x_j} \\ &= 0 \end{aligned} \tag{5}$$

for all p. Since $(p_1, \ldots, p_{k-1})$ moves over an open set, it follows from (5) that the coefficients of polynomials on $(p_1, \ldots, p_{k-1})$ is equal to 0. Now we use an induction method. Since T and X are one to one, there exists at most one $x = (x_1, \ldots, x_k) \in \mathcal{X}$ such that $\sum_{j=1}^{k-1} x_j = 0$. Denote it by x', and the constant term of (5) is only $f(x')c(x')$. Since $c(x') > 0$, we have $f(x') = 0$. Next assume that $f(x) = 0$ for any $x = (x_1, \ldots, x_k) \in \mathcal{X}$ satisfying $\sum_{j=1}^{k-1} x_j \leq l - 1$. Taking any $x'' = ({x_1}'', \ldots, {x_k}'') \in \mathcal{X}$ with $\sum_{j=1}^{k-1} {x_j}'' = l$, we see that the coefficient on $\prod_{j=1}^{k-1} {p_j}^{{x_j}''}$ is only $f(x'')c(x'')$ since T and X are one-to-one. Indeed, we obtain from (5)

$$\begin{aligned} E_p[f(X_N)] &= \sum_{x_1 + \cdots + x_{k-1} \geq l} f(x)c(x)(1 - \sum_{j=1}^{k-1} p_j)^{x_k} \prod_{j=1}^{k-1} {p_j}^{x_j} \\ &= \sum_{x_1 + \cdots + x_{k-1} \geq l} f(x)c(x)\{1 + p_1 + p_2 + \cdots\} \prod_{j=1}^{k-1} {p_j}^{x_j} \end{aligned} \tag{6}$$

Considering the term of $\prod_{j=1}^{k-1} {p_j}^{{x_j}''}$ in (6), we can get it from a product of the inner part of $\{\cdots\}$ and $\prod_{j=1}^{k-1} {p_j}^{x_j}$. Then it follows from the assumption that the degree of each term

of $\{\cdots\}\prod_{j=1}^{k-1} {p_j}^{x_j}$ in (6) is equal or greater than that of $\prod_{j=1}^{k-1} {p_j}^{x_j''}$. Hence the coefficient for $\prod_{j=1}^{k-1} {p_j}^{x_j''}$ is only $f(x'')c(x'')$. Since $c(x'') > 0$, it follows that $f(x'') = 0$. Thus we complete the proof.

It is noted that the induction part of the above proof is similar to that of Kremers(1990). The following lemma is due to Lehmann and Stein (1950).

Lemma 2. *Let $X_1, X_2, \ldots$ be a sequence of random variables such that for each positive integer m the set $X_1, \ldots, X_m$ admits a real valued sufficient statistic $T_m = t_m(X_1, \ldots, X_m)$. Let $\Sigma_1, \Sigma_2, \ldots, \Sigma_r$ each be a complete, closed, sequential procedure based on these sufficient statistics. Let $\Sigma_1 \cup \Sigma_2 \cup \cdots \cup \Sigma_r$ denote the sequential procedure according to which we continue taking observations until at least one of the stopping rules $\Sigma_1, \Sigma_2, \ldots, \Sigma_r$ telles us to stop. Then the procedure $\Sigma_1 \cup \Sigma_2 \cup \cdots \cup \Sigma_r$ is complete.*

Combining Lemma 1 and Lemma 2 , we get the following theorem.

Theorem 1. *If a bounded stopping rule is simple , then the sequential procedure is complete.*

Proof. It follows from the supporting hyperplane theorem that the convex hull of the continuation region for each index can be expressed as an intersection of a finite number of supporting half-spaces . Further, by the boundedness of the stopping rule, we can take a sufficiently large m which X_N never fail to stop . From Lemma 1 and Lemma 2 we complete the proof .

Remark. It is known that, in the sequential binomial case, the sequential procedure is complete if and only if simple (see Girshick *et al.* (1946)). As a sufficient condition for completeness in the sequential multinomial case, Kremers (1990) also stated that the stopping rule is bounded and the convex hull of the continuation region does not contain a boundary point.

The following example shows that Theorem 1 holds but not the Kremers condition. Assume that a boundary region consists of $(0,2),(1,1),(2,2),(3,1)$ and $(3,0)$ in the sequential binomial case. Since the continuation points are $(0,1),(1,0),(2,0)$ and $(2,1)$, it

is easily seen that the stopping rule is bounded and the convex hull of the continuation region contains the boundary point $(1,1)$. Hence the Kremers condition is not satisfied, but Theorem 1 holds since the stopping rule is easily seen to be simple.

4. RANDOMIZED SEQUENTIAL PROCEDURE

Let $\{q_n\}$ be a sequence such that $q_n \geq 0$ $(n = 0,1,2,\ldots)$ and $\sum_{n=0}^{\infty} q_n = 1$. For a stopping rule φ in the section 2 we assume that $\varphi_n = q_n$ for each non-negative integer n . Then the stopping rule is randomized . We have for any estimator f

$$E_p\left[f(X_N)\right] = \sum_{n} \sum_{x_1+\cdots+x_k=n} q_n n! f(x_1,\ldots,x_k) \prod_{j=1}^{k} {p_j}^{x_j}/x_j!, \tag{7}$$

where $\sum_{x_1+\cdots+x_k=n}$ means that we take the sum with respect to the all possible combination with $\sum_{j=1}^{k} x_j = n$. Then we have the following.

Theorem 2. *A necessary and sufficient condition for the sequential procedure to be complete is that there exists a unique non-negative integer n satisfying $q_n = 1$.*

Proof. Sufficiency. Since, in this case, X_N is distributed according to the k-dimensional multinomial distribution $\mathrm{M}(n, p_1, \ldots, p_k\ )$ with a mass function $n! \prod_{j=1}^{k} \frac{{p_j}^{x_j}}{x_j!}$ for all non-negative integers satisfying $\sum_{j=1}^{k} x_j = n$, it is easily seen that the sequential procedure is complete.

Necessity. Suppose that $q_m, q_n > 0$ for $m < n$ with non-negative integers m and n. It suffices to prove that unbiased estimator of 0 based on only the points over the indices m and n can be constructed. The number of the possible combination $(x_1,\ldots,x_k)$ such that $\sum_{j=1}^{k} x_j = n$ is equal to $\binom{k+n-1}{n}$. On the other hand, since $\sum_{j=1}^{k} p_j = 1$ yields

$$\prod_{j=1}^{k} {p_j}^{x_j} = (1 - \sum_{j=1}^{k-1} p_j)^{x_k} \prod_{j=1}^{k-1} {p_j}^{x_j},$$

we may regard the right-hand side of (7) as a polynomial in $(p_1,\ldots,p_{k-1})$. If the polynomial is identically equal to 0, then each coefficient must be equal to 0. Since the number

of terms of the polynomial with degree r is $\binom{k+r-2}{r}$ in $\prod_{j=1}^{k-1} {p_j}^{x_j}$, it follows that the total number of its terms is $\sum_{r=0}^{n} \binom{k+r-2}{r}$. It is also seen that the coefficients of the polynomial of the right-hand side of (7) are linear functions of $f(x_1, \ldots, x_k)$'s. Since each coefficient of the polynomial must be 0, we may regard it as the simultaneous linear equations of $f(x_1, \ldots x_k)$'s. Then there exist non-trivial solution of these equations. In fact, since

$$\begin{aligned}\binom{k+n-1}{n} &= \binom{k+n-2}{n-1} + \binom{k+n-2}{n} \\ &= \binom{k+n-3}{n-2} + \binom{k+n-3}{n-1} + \binom{k+n-2}{n} \\ &= \cdots = \sum_{r=0}^{n} \binom{k+r-2}{r},\end{aligned}$$

it is easily seen that

$$\binom{k+n-1}{n} + \binom{k+m-1}{m} > \sum_{r=0}^{n} \binom{k+r-2}{r}.$$

Moreover the left and right sides of the above inequality are equal to the number of unknown variables $f(x_1, \ldots, x_k)$ and that of linear restrictions, respectively. Consequently we can get a non-trivial unbiased estimator of 0 based on the points over the indices m and n. Thus we complete the proof.

The above result may be applied to the following estimation problem. For each $n = 0, 1, 2, \ldots$, let δ_n be an estimator of a function $g(p)$ of p. Then our purpose is to minimize $\sum_{n=0}^{\infty} E_p\left[q_n\{\delta_n - g(p)\}^2\right]$ under the condition $\sum_{n=0}^{\infty} E_p\left[q_n\delta_n\right] = g(p)$ for all p. In order to do so, it is enough to obtain δ_n minimizing $E_p\left[\{\delta_n - g(p)\}^2\right]$ for each n under the condition $\sum_{n=0}^{\infty} q_n E_p\left[\delta_n\right] = g(p)$, since, for each n, the randomized stopping rule φ_n is equal to q_n which is independent of X_n. Hence it is seen that the above problem is similar to that in the case of a fixed size of a sample. In this case it is also easy to check the necessary and sufficient condition for the sequential procedure to be complete in Theorem 2.

REFERENCES

Ferguson, T. S. (1967). *Mathematical Statistics.* Academic Press, New York.

Girshick, M. A., F. Mosteller & L. J. Savage (1946). Unbiased estimates for certain binomial sampling problems with applications. *Annals of Mathematical Statistics*, 17. 13-23.

Kremers, W. K. (1990). Completeness and unbiased estimation in sequential multinomial sampling. *Sequential Analysis,* 9. 43-58.

Lehmann, E. L. & C. Stein (1950). Completeness in the sequential case. *Annals of Mathematical Statistics*, 21. 376-385.