The Annals of Statistics
1989, Vol. 17, No. 1, 436–439

A COUNTEREXAMPLE TO A CORRELATION INEQUALITY IN FINITE SAMPLING[1]

BY KENNETH S. ALEXANDER

University of Southern California

Each individual in a population of size n is assigned a positive number as its weight. k of the n are sampled without replacement, with the individuals remaining at the time of each selection chosen with probabilities proportional to their weights. It is shown that for two fixed individuals, the events that each is in the sample can be positively correlated.

Consider the following situation: Lottery tickets numbered 1 through n are sold. Various entrants, including ourselves, buy one or more tickets. k winning numbers are then selected at random, but with the provision that no person is allowed to win more than one prize—numbers are drawn until k distinct individuals are winners. The drawing is held, and before learning of our fortune, we find out that a particular individual, whom we will call the Rival, was not one of the k winners. Our reaction is to say, "Fine with us; having the Rival out of the picture can only improve *our* chances of winning."

However, we would be wrong.

In fact, it is possible to allocate the tickets among the contestants in such a way that our chances of winning given that the Rival lost are actually smaller than our unconditional chance of winning.

More generally, consider the problem of weighted sampling from a finite population. Each of the n individuals is assigned a positive number as its weight. k of the n are then sampled without replacement; whatever individuals remain before each of the k selections are chosen with a probability proportional to their weight. This is called *successive sampling*. Fix two individuals and consider the events that each is in the sample. The surprising fact is that these two events can be positively correlated. Such sampling schemes are discussed by Hájek [(1981), Chapter 9] and Rao (1963), among many others. Sampford (1969) gives reasons why it is desirable to choose sampling schemes in which the correlation is negative, so our result perhaps reduces the desirability for theoretical purposes of successive sampling in comparison to other weighted sampling techniques.

Here is another formulation of the problem. Balls are placed at random in n bins. The placements are done independently, with each ball having probability μ_j of being put in bin j for $j \leq n$. We say a bin is occupied if it contains at least one ball. We can then list the bins in the order in which they become occupied and define $r(A)$ to be the rank of bin A on this list, for each bin A. Thus, for example, $r(A) = 3$ if A is the third bin to become occupied. A plausible (to us)

Received December 1987.
[1]Research supported by NSF Grant DMS-87-02906.
AMS 1980 *subject classifications*. 62D05, 60C05.
Key words and phrases. Weighted sampling, sampling without replacement, finite sampling, successive sampling.

conjecture might be that if we know that a given bin A has a small rank, this reduces the probability that any fixed second bin B has small rank, i.e.,

$$(1) \quad P[r(A) \le k | r(B) \le k] \le P[r(A) \le k] \quad \text{for all distinct bins } A, B.$$

But intuition is not necessarily a good guide, for in this note we will show that such a conjecture is false.

A third formulation of this problem, due to Gordon (1983) and readily seen to be equivalent, is the following. It will be useful for analyzing certain limiting cases. Let $(S, \mathcal{S}, \mu)$ be a probability space, let λ denote Lebesgue measure on $\mathbb{R}^+$, and let N be a Poisson process on $\mathbb{R}^+ \times S$ with intensity measure $\lambda \times \mu$. For brevity, set

$$N(t, A) := N([0, t] \times A), \qquad A \in \mathcal{S}.$$

If $N(t, \{x\}) = 1$ we say a point appears at x at time t. Define

$$T_A := \min\{t > 0 : N(t, A) > 0\},$$

with $T_A = \infty$ if there is no such t. We think of T_A as the time at which A becomes occupied. Let $\mathcal{E}$ be a partition of S into disjoint measurable sets and set

$$r_{\mathcal{E}}(F) := \operatorname{card}\{E \in \mathcal{E} : T_E \le T_F\} \quad \text{for } F \in \mathcal{E}.$$

Thus as in the bin formulation $r_{\mathcal{E}}(F) = k$ if F is the kth set in $\mathcal{E}$ to become occupied. The conjecture (1) now becomes

$$(2) \quad P[r_{\mathcal{E}}(A) \le k, r_{\mathcal{E}}(B) \le k] \le P[r_{\mathcal{E}}(A) \le k] P[r_{\mathcal{E}}(B) \le k]$$
for every finite $\mathcal{E}$, every distinct A, B in $\mathcal{E}$ and every $k \ge 1$.

To violate (2), we must first consider a limiting case in which the partition is infinite. Take $(S, \mathcal{S}, \mu)$ to be an atomless probability space in which single points are measureable (the unit interval will do fine) and partition S into three sets A, B and C. For brevity we write a, b and c for $\mu(A)$, $\mu(B)$ and $\mu(C)$. Later we will further divide C into a large number of small sets. For now, we approximate this by completely pulverizing C; that is, we define the partition

$$\mathcal{F} := \{A, B\} \cup \{\{x\} : x \in C\}$$

and set

$$\tau_j := \min\{t > 0 : N(t, C) \ge j\}.$$

Then

$$\begin{aligned} P[r_{\mathcal{F}}(A) \le k, r_{\mathcal{F}}(B) \le k] &= P[\tau_{k-1} > \max(T_A, T_B)] \\ &= 1 - P[\tau_{k-1} < T_A] - P[\tau_{k-1} < T_B] \\ &\quad + P[\tau_{k-1} < \min(T_A, T_B)]. \end{aligned}$$

Now $[\tau_{k-1} < T_A]$ is (a.s.) the event that the first $k-1$ points to appear in $A \cup C$ all fall in C, which has probability $(c/(a+c))^{k-1}$. Similarly, $P[\tau_{k-1} < T_B] = (c/(b+c))^{k-1}$. Also $[\tau_{k-1} < \min(T_A, T_B)]$ is the event that the

first $k-1$ points anywhere all fall in C, which has probability c^{k-1}. Thus

$$P[r_{\mathscr{F}}(A) \le k, r_{\mathscr{F}}(B) \le k] = 1 - \left(\frac{c}{c+a}\right)^{k-1} - \left(\frac{c}{c+b}\right)^{k-1} + c^{k-1}.$$

Second,

$$P[r_{\mathscr{F}}(A) \le k] = P[T_A < \tau_{k-1}] + P[\tau_{k-1} < T_A < \min(T_B, \tau_k)].$$

This last event is the event that the first $k-1$ points anywhere all fall in C and the kth falls in A. This has probability ac^{k-1} and $P[T_A < \tau_{k-1}]$ was determined above, so

$$P[r_{\mathscr{F}}(A) \le k] = 1 - \left(\frac{c}{c+a}\right)^{k-1} + ac^{k-1}.$$

Similarly,

$$P[r_{\mathscr{F}}(B) \le k] = 1 - \left(\frac{c}{c+b}\right)^{k-1} + bc^{k-1}.$$

Combining, we obtain

$$\begin{aligned} &P[r_{\mathscr{F}}(A) \le k, r_{\mathscr{F}}(B) \le k] - P[r_{\mathscr{F}}(A) \le k]P[r_{\mathscr{F}}(B) \le k] \\ &\quad = (a+c)^{-(k-1)}(b+c)^{-(k-1)} \\ &\quad\quad \times \Big\{c^k(a+c)^{k-1}(b+c)^{k-1} \\ &\quad\quad\quad - c^{2k-2}\big(1 - a(a+c)^{k-1}\big)\big(1 - b(b+c)^{k-1}\big)\Big\}. \end{aligned} \tag{3}$$

Now take $k \ge 3$, so $k < 2k-2$. Then fix a_0 and b_0 in $(0,1)$ with $a_0 + b_0 = 1$. Let $a \to a_0$, $b \to b_0$ and $c \to 0$ in such a way that $a + b + c = 1$. When c becomes sufficiently small, the right side of (3) is clearly positive.

We now approximate this situation with a finite partition. Choose a, b and c so that (3) is a positive number ε and fix an integer $m > k^2/\varepsilon$. Divide C into m sets $C_1, \ldots, C_m$ each of probability c/m and let $\mathscr{E}$ be the partition $\{A, B, C_1, \ldots, C_m\}$. Then

$$\begin{aligned} P[r_{\mathscr{F}}(A) \le k] &\ge P[r_{\mathscr{E}}(A) \le k] - P\Big[\max_{j \le m} N(\tau_{k-1}, C_j) > 1\Big] \\ &\ge P[r_{\mathscr{E}}(A) \le k] - m^{-1}\binom{k-1}{2} \\ &> P[r_{\mathscr{E}}(A) \le k] - \varepsilon/2 \end{aligned}$$

and similarly for B, so that

$$\begin{aligned} P[r_{\mathscr{E}}(A) \le k, r_{\mathscr{E}}(B) \le k] &\ge P[r_{\mathscr{F}}(A) \le k, r_{\mathscr{F}}(B) \le k] \\ &= P[r_{\mathscr{F}}(A) \le k]P[r_{\mathscr{F}}(B) \le k] + \varepsilon \\ &> P[r_{\mathscr{E}}(A) \le k]P[r_{\mathscr{E}}(B) \le k]. \end{aligned}$$

As a numerical example, we can take $a = b = 0.45$, $c = 0.1$ and $k = 3$. Then

$$P[r_{\mathscr{F}}(A) \leq k, r_{\mathscr{F}}(B) < k] = 0.94388,$$

$$P[r_{\mathscr{F}}(A) \leq k]\, P[r_{\mathscr{F}}(B) \leq k] = 0.94370.$$

If $m = 100{,}000$, then (2) is violated. In our original lottery example, this means that if there are to be three winners, if the Rival and ourselves each hold 450,000 tickets, and if 100,000 other people hold one ticket each, we should be disappointed to learn that the Rival was not a winner, since this fact reduces our own chances.

REMARK 1. To understand this result, which at first sight may seem surprising, we return to the balls-in-urns formulation. Consider the following way of allocating balls to bins A, B and C independently with probabilities a, b and c each, respectively. First, a list is made, with independent random entries "B" and "C," with probabilities $b/(b + c)$ and $c/(b + c)$, respectively, for each entry. Then balls are allocated independently between bin A and a temporary holding bin "$B \cup C$" with probabilities a and $b + c$, respectively. (As before, $a + b + c = 1$.) Each time the ball enters "$B \cup C$" the next entry B or C is read from the list and the ball is moved to the corresponding bin. C consists of infinitely many subbins $C_1, C_2, \ldots$, each of which can hold only one ball; these become occupied sequentially.

Suppose now that we have placed a bet that A will be one of the first k bins (among $A, B, C_1, C_2, \ldots$) to become occupied. We would then hope that the list starts off with lots of B's and few C's, since the second, third, fourth, etc., balls to fall in B are harmless to our cause, unlike those in C. Imagine learning that B was not among the first k bins occupied. This would tell us that the list started off with at least $k - 1$ C's, which is bad news! This makes it plausible that the events, "A is among the first k occupied" and "B is among the first k occupied" are positively correlated.

Acknowledgments. The author would like to thank Lou Gordon and Richard Arratia, who suggested the problem and Ronald Pyke, who contributed a large part of Remark 1.

REFERENCES

GORDON, L. (1983). Successive sampling in large finite populations. *Ann. Statist.* **11** 702–706.

HÁJEK, J. (1981). *Sampling from a Finite Population.* Dekker, New York.

RAO, J. N. K. (1963). On three procedures of unequal probability sampling without replacement. *J. Amer. Statist. Assoc.* **58** 202–215.

SAMPFORD, M. (1969). A comparison of some possible methods of sampling from smallish populations, with units of unequal size. In *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, Jr., eds.). Wiley, New York.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-1113