SNOWBALL SAMPLING[1]

By Leo A. Goodman

University of Chicago

1. Introduction and Summary. An s stage k name snowball sampling procedure is defined as follows: A random sample of individuals is drawn from a given finite population. (The kind of random sample will be discussed later in this section.) Each individual in the sample is asked to name k different individuals in the population, where k is a specified integer; for example, each individual may be asked to name his "k best friends," or the "k individuals with whom he most frequently associates," or the "k individuals whose opinions he most frequently seeks," etc. (For the sake of simplicity, we assume throughout that an individual cannot include himself in his list of k individuals.) The individuals who were not in the random sample but were named by individuals in it form the first stage. Each of the individuals in the first stage is then asked to name k different individuals. (We assume that the question asked of the individuals in the random sample and of those in each stage is the same and that k is the same.) The individuals who were not in the random sample nor in the first stage but were named by individuals who were in the first stage form the second stage. Each of the individuals in the second stage is then asked to name k different individuals. The individuals who were not in the random sample nor in the first or second stages but were named by individuals who were in the second stage form the third stage. Each of the individuals in the third stage is then asked to name k different individuals. This procedure is continued until each of the individuals in the sth stage has been asked to name k different individuals.

The data obtained using an s stage k name snowball sampling procedure can be utilized to make statistical inferences about various aspects of the relationships present in the population. The relationships present, in the hypothetical situation where each individual in the population is asked to name k different individuals, can be described by a matrix with rows and columns corresponding to the members of the population, rows for the individuals naming and columns for the individuals named, where the entry θ_{ij} in the ith row and jth column is 1 if the ith individual in the population includes the jth individual among the

Received June 4, 1959; revised September 28, 1960.

[1] Part of this research was carried out at the Statistical Research Center, University of Chicago, under sponsorship of the Statistics Branch, Office of Naval Research and part while the author was at the Statistical Laboratory of the University of Cambridge under a National Science Foundation Senior Postdoctoral Fellowship and a John Simon Guggenheim Memorial Foundation Fellowship. Reproduction in whole or in part is permitted for any purpose of the United States Government. For their helpful comments, the author is indebted to J. S. Coleman, who introduced him to the general topic [2], and to A. Barton, W. H. Kruskal, and J. H. Lorie.

k individuals he would name, and it is 0 otherwise. While the matrix of the θ's cannot be known in general unless every individual in the population is interviewed (i.e., asked to name k different individuals), it will be possible to make statistical inferences about various aspects of this matrix from the data obtained using an s stage k name snowball sampling procedure. For example, when $s = k = 1$, the number, M_{11}, of mutual relationships present in the population (i.e., the number of values i with $\theta_{ij} = \theta_{ji} = 1$ for some value of $j > i$) can be estimated.

The methods of statistical inference applied to the data obtained from an s stage k name snowball sample will of course depend on the kind of random sample drawn as the initial step. In most of the present paper, we shall suppose that a random sample (i.e., the "zero stage" in snowball sample) is drawn so that the probability, p, that a given individual in the population will be in the sample is independent of whether a different given individual has appeared. This kind of sampling has been called binomial sampling; the specified value of p (assumed known) has been called the sampling fraction [4]. This sampling scheme might also be described by saying that a given individual is included in the sample just when a coin, which has a probability p of "heads," comes up "heads," where the tosses of the coin from individual to individual are independent. (To each individual there corresponds an independent Bernoulli trial determining whether he will or will not be included in the sample.) This sampling scheme differs in some respects from the more usual models where the sample size is fixed in advance or where the ratio of the sample size to the population size (i.e., the sample size-population size ratio) is fixed. For binomial sampling, this ratio is a random variable whose expected value is p. (The variance of this ratio approaches zero as the population becomes infinite.) In some situations (where, for example, the variance of this ratio is near zero), mathematical results obtained for binomial sampling are sometimes quite similar to results obtained using some of the more usual sampling models (see [4], [7]; compare the variance formulas in [3] and [5]); in such cases it will often not make much difference, from a practical point of view, which sampling model is utilized. (In Section 6 of the present paper some results for snowball sampling based on an initial sample of the more usual kind are obtained and compared with results presented in the earlier sections of this paper obtained for snowball sampling based on an initial binomial sample.)

For snowball sampling based on an initial binomial sample, and with $s = k = 1$, so that each individual asked names just one other individual and there is just one stage beyond the initial sample, Section 2 of this paper discusses unbiased estimation of M_{11}, the number of pairs of individuals in the population who would name each other. One of the unbiased estimators considered (among a certain specified class of estimators) has uniformly smallest variance when the population characteristics are unknown; this one is based on a sufficient statistic for a simplified summary of the data and is the only unbiased estimator of M_{11} based on that sufficient statistic (when the population characteristics are unknown). This estimator (when $s = k = 1$) has a smaller variance than a

comparable minimum variance unbiased estimator computed from a larger random sample when $s = 0$ and $k = 1$ (i.e., where only the individuals in the random sample are interviewed) even where the expected number of individuals in the larger random sample ($s = 0$, $k = 1$) is equal to the maximum expected number of individuals studied when $s = k = 1$ (i.e., the sum of the expected number of individuals in the initial sample and the maximum expected number of individuals in the first stage). In fact, the variance of the estimator when $s = 0$ and $k = 1$ is at least twice as large as the variance of the comparable estimator when $s = k = 1$ even where the expected number of individuals studied when $s = 0$ and $k = 1$ is as large as the maximum expected number of individuals studied when $s = k = 1$. Thus, for estimating M_{11}, the sampling scheme with $s = k = 1$ is preferable to the sampling scheme with $s = 0$ and $k = 1$. Furthermore, we observe that when $s = k = 1$ the unbiased estimator based on the simplified summary of the data having minimum variance when the population characteristics are unknown can be improved upon in cases where certain population characteristics are known, or where additional data not included in the simplified summary are available. Several improved estimators are derived and discussed.

Some of the results for the special case of $s = k = 1$ are generalized in Sections 3 and 4 to deal with cases where s and k are any specified positive integers. In Section 5, results are presented about s stage k name snowball sampling procedures, where each individual asked to name k different individuals chooses k individuals at random from the population. (Except in Section 5, the numbers θ_{ij}, which form the matrix referred to earlier, are assumed to be fixed (i.e., to be population parameters); in Section 5, they are random variables. A variable response error is not considered except in so far as Section 5 deals with an extreme case of this.)

For social science literature that discusses problems related to snowball sampling, see [2], [8], and the articles they cite. This literature indicates, among other things, the importance of studying "social structure and . . . the relations among individuals" [2].

2. The Case $s = k = 1$. The term "sample" will be used throughout (except in Section 6) to refer to the "binomially sampled" sample; i.e., to the "zero stage" in the s stage k name snowball sample. The number of individuals in the population who enter mutual relationships is $2M_{11}$, We now consider the problem of estimating M_{11} when $s = k = 1$. Let y be the number of individuals in the sample who enter mutual relationships (with individuals in the population, and thus with individuals who are either in the sample or in the first stage). The random variable y has a binomial distribution with expected value $E\{y\} = 2M_{11}p$. (To see this, think of the population as divided into those individuals who enter mutual relationships, plus the others.) Thus an unbiased estimator of M_{11} is $y/(2p)$.

Let y_2 be the number of individuals in the sample who enter mutual relation-

ships with other individuals in the sample, and let y_1 be the number of individuals in the sample who enter mutual relationships with individuals who do not appear in the sample but who are, of course, in the first stage. Then $y = y_1 + y_2$. The random variable $y_2/2$ has a binomial distribution with expected value $E\{y_2/2\} = M_{11}p^2$. (To see this, think of the population as divided into those pairs of individuals who name each other, plus the others.) Thus an unbiased estimator of M_{11} is $y_2/(2p^2)$. The random variable y_1 has a binomial distribution with expected value $E\{y_1\} = M_{11}2pq$, where $q = 1 - p$. Thus an unbiased estimator of M_{11} is $y_1/(2pq)$. Of course,

$$E\{y\} = E\{y_1\} + E\{y_2\} = M_{11}[2pq + 2p^2] = 2M_{11}p.$$

Let x_{11} be the number of mutual relationships observed from these data; i.e., $x_{11} = \frac{1}{2}y_2 + y_1 = w_{112} + w_{111}$, where $\frac{1}{2}y_2 = w_{112}$ is the number of mutual relationships observed with both individuals in the sample and $y_1 = w_{111}$ is the number of mutual relationships observed with only one of the individuals in the sample. (We shall consider later in this section the number, w_{110}, of mutual relationships observed, where neither individual entering the relationship is in the sample, i.e., where both individuals are in the first stage; but at this point this number is to be ignored.) We have introduced the more cumbersome notation (i.e., the x_{11} and the w's) since a notation of this kind will be used in the generalizations presented later. The random variable x_{11} has a binomial distribution with expected value $E\{x_{11}\} = M_{11}(1 - q^2)$. Thus, an unbiased estimator of M_{11} is $x_{11}/(1 - q^2) = x_{11}/[p(2 - p)] = \hat{M}_{11}$.

Four unbiased estimators have been presented. In passing, we note that, when the observed values of y_1 and y_2 are both zero, all four estimators lead to an estimate of zero for M_{11}. In particular, if no individuals appear in the binomial sample, all four estimators of M_{11} yield zero. If the population size, N, is reasonably large, the probability of no individuals, q^N, is very small.

All four estimators are linear functions of y_1 and y_2. We now consider the class of all linear functions of y_1 and y_2. Writing $Y_1 = y_1/(2pq)$ and $Y_2 = y_2/(2p^2)$, all linear functions of y_1 and y_2 that are unbiased estimators of M_{11} must be of form $AY_1 + (1 - A)Y_2 = \hat{M}(A)$. The variance of Y_1 is $M_{11}(1 - 2pq)/(2pq)$, the variance of Y_2 is $M_{11}(1 - p^2)/p^2$, and the covariance between Y_1 and Y_2 is $-M_{11}$. These results follow from the fact that the sampling scheme divides the M_{11} pairs into a trinomial with probabilities p^2 (both individuals in the sample), $2pq$ (just one in), and q^2 (neither in); $y_2/2$ is the number in the first cell of the trinomial sample, y_1 is the number in the second cell, and the second moments of these random variables are then immediate from those of a trinomial. The variance of $\hat{M}(A)$ is thus

$$\begin{aligned}\sigma^2_{\hat{M}(A)} &= M_{11}\{A^2[(1 - 2pq)/(2pq)] + (1 - A)^2[(1 - p^2)/p^2] - 2A(1 - A)\} \\ &= M_{11}[A^2(p + 2q) - 4Aq + (1 - p^2)2q]/(2p^2q).\end{aligned}$$

It follows that $A = 2q/(p + 2q)$ minimizes the variance of $\hat{M}(A)$. Thus,

among the class of unbiased estimators, $\hat{M}(A)$, that are linear combinations of y_1 and y_2, the estimator with the smallest variance is

$$(Y_1 2q + Y_2 p)/(2q + p) = (y_1 + \tfrac{1}{2}y_2)/[p(2 - p)]$$
$$= x_{11}/[p(2 - p)] = \hat{M}_{11}.$$

The variance of $\hat{M}_{11}$ is $\sigma^2_{\hat{M}_{11}} = M_{11}q^2/[p(2 - p)] = M_{11}q^2/(1 - q^2)$. When $A = q$, the unbiased estimator is $Y_1 q + Y_2 p = (y_1 + y_2)/(2p) = y/(2p)$, and its variance is $M_{11}q/(2p)$.

The preceding comments dealt with all linear functions of $y_1 = w_{111}$ and $\frac{1}{2}y_2 = w_{112}$ that are unbiased estimators of M_{11}; we showed that $\hat{M}_{11}$ had the smallest variance among these. If we consider the class of all possible functions of w_{111} and w_{112} (not only linear functions) that are unbiased estimators of M_{11}, we shall prove below a more general result, from which it follows that the estimator $\hat{M}_{11}$ has the smallest variance among this class.

Let z_{11} be the number of individuals in the sample who do not enter mutual relationships in the population. Because the snowball sampling design has a first stage, z_{11} is observed. We shall refer to the set $(w_{111}, w_{112}, z_{11})$ as the simplified set of data for mutual relationships when $s = k = 1$. (We noted earlier that x_{11} and y were linear functions of w_{111} and w_{112}.) We shall now limit our consideration to $(w_{111}, w_{112}, z_{11})$, although, as we shall observe later in this section, it may sometimes be worthwhile to make use of additional available data. We now prove the following result:

THEOREM 1: *If the population characteristics (including its size) are completely unknown, then the estimator $\hat{M}_{11}$ has minimum variance among all unbiased estimators of M_{11} based on the simplified set of data when $s = k = 1$.*

PROOF: Let T_{11} be the number of individuals in the population who do not have mutual relationships, so that $2M_{11} + T_{11} = N$ and $y + z_{11} = n$, where n is the number in the binomial sample. We have

$$E\{n\} = Np = E\{y\} + E\{z_{11}\} = 2M_{11}p + T_{11}p.$$

The joint distribution of w_{111}, w_{112}, z_{11} is the following product of a trinomial and binomial:

$$\Pr\{w_{111}, w_{112}, z_{11}\} = (p^2)^{w_{112}}(2pq)^{w_{111}}(q^2)^{M_{11}-x_{11}}p^{z_{11}}q^{T_{11}-z_{11}}K,$$

where K is a product of multinomial coefficients. The distribution of x_{11} is

$$\Pr\{x_{11}\} = (q^2)^{M_{11}-x_{11}}(1 - q^2)^{x_{11}}\binom{M_{11}}{x_{11}}.$$

The conditional distribution of $(w_{111}, w_{112}, z_{11})$, given x_{11} and z_{11}, is

$$\Pr\{w_{111}, w_{112}, z_{11} \mid x_{11}, z_{11}\} = r^{w_{112}}h^{w_{111}}\binom{x_{11}}{w_{111}},$$

where $r = p/(2 - p)$ and $h = 1 - r$. Thus, x_{11} and z_{11} are jointly sufficient for

$(w_{111}, w_{112}, z_{11})$. The joint distribution of x_{11}, z_{11} can be written as

$$\Pr\{x_{11}, z_{11} \mid M_{11}, T_{11}\} = \Pr\{x_{11} \mid M_{11}\} \Pr\{z_{11} \mid T_{11}\}. \tag{1}$$

Since (M_{11}, T_{11}) ranges through a Cartesian product in the case where the population size is completely unknown, equation (1) indicates that z_{11} is irrelevant for the estimation of M_{11}; Blackwell's method [1] can be applied to prove that to any unbiased estimator M^+ of M_{11} based on (x_{11}, z_{11}) there corresponds an unbiased estimator M^{++} based on x_{11} whose variance is no larger than that of M^+ (computed for the true distribution of z_{11}); the fact that, in the case considered here, M^{++} is only known to be within a certain class of estimators (formed by computing the conditional distribution of M^+, given x_{11}, for all admissible distributions of z_{11}) does not weaken the conclusion that if there exists an estimator with minimum variance among all unbiased estimators of M_{11} based on x_{11} (which we shall see below is in fact the case) it will also have minimum variance among all unbiased estimators based on (x_{11}, z_{11}), and it will therefore be sufficient to consider only functions of x_{11} when estimating M_{11}. (More can be said concerning the concept of "irrelevance" presented here, but this would take us too far afield.) We have shown earlier that $\hat{M}_{11} = x_{11}/[p(2 - p)]$ is an unbiased estimator of M_{11}. It is, in fact, the only unbiased estimator of M_{11} that is based on x_{11}, because an unbiased estimator, $g(x_{11})$, must satisfy the system of equations

$$\sum_{x_{11}=0}^{M_{11}} g(x_{11}) \Pr\{x_{11}\} = M_{11}, M_{11} = 0, 1, 2, \cdots,$$

which can be used to define $g(x_{11})$ recursively for $x_{11} = 0, 1, 2, \cdots$. Therefore, $g(x_{11}) = \hat{M}_{11}$ is the unique solution to this system of equations, and is thus the minimum variance unbiased estimator of M_{11} based on the simplified set of data. This concludes the proof. (This theorem could also have been demonstrated by proving that, if the population characteristics are unknown, the only function of (x_{11}, z_{11}) that is an unbiased estimator of M_{11} is $\hat{M}_{11}$; the proof given above of the uniqueness of an unbiased estimator of M_{11} among all functions of x_{11} can be modified in a straightforward manner to prove the uniqueness of an unbiased estimator of M_{11} among all functions of (x_{11}, z_{11}).)

The estimator $\hat{M}_{11}$ will be unbiased whether or not the population size, N, is known. It is however important to note that Theorem 1 deals only with the situation where N is unknown.

In situations where N is known, the statistic x_{11} is not a sufficient statistic for the $(w_{111}, w_{112}, z_{11})$ for the estimation of M_{11} (since $2M_{11} + T_{11} = N$), and so it may be possible to obtain estimators of M_{11} that have a smaller variance than $\hat{M}_{11}$. It is easy to see that, when N is known, the statistic (x_{11}, z_{11}) is a sufficient statistic for the simplified set of data for the estimation of M_{11}. Since the random variable z_{11} has a binomial distribution with expected value $E\{z_{11}\} = T_{11}p$, the estimator $\hat{T}_{11} = z_{11}/p$ is an unbiased estimator of T_{11} and $\tilde{M}_{11} = [N - z_{11}/p]/2$ is an unbiased estimator of M_{11}. The variance $\sigma^2_{\tilde{M}}$ of $\tilde{M}_{11}$ is equal

to $T_{11}q/(4p)$, while the variance $\sigma^2_{\hat{M}}$ of $\hat{M}_{11}$ was equal to $M_{11}q^2/[p(2-p)]$. Since the ratio of these two variance is

$$E = \sigma^2_{\hat{M}}/\sigma^2_{\tilde{M}} = 2M_{11}(1-p)/[T_{11}(1-p/2)],$$

the relative accuracy of $\hat{M}_{11}$ and $\tilde{M}_{11}$ will depend on $2M_{11}/T_{11}$, which is unknown. If $2M_{11}/T_{11}$ is small it will be better to use $\hat{M}_{11}$, while if $2M_{11}/T_{11}$ is large it will be better to use $\tilde{M}_{11}$. Since $\hat{M}_{11}$ and $\tilde{M}_{11}$ are both unbiased, any weighted average $G\hat{M}_{11} + (1-G)\tilde{M}_{11}$ will be unbiased. Since the $\hat{M}_{11}$ and $\tilde{M}_{11}$ are statistically independent, the value of G that will minimize the variance of the weighted average is $G = 1/(1+E)$. Although E is unknown, it may be possible to make a rough guess as to its magnitude, and thus obtain the corresponding value of G to be used in computing the unbiased estimator.

If E and then G are estimated from the same data used to compute $\hat{M}_{11}$ and $\tilde{M}_{11}$, then the weighted average will not in general be unbiased, but this procedure may still be of value. An estimator of E can be based on the following unbiased estimators of $\sigma^2_{\hat{M}}$ and $\sigma^2_{\tilde{M}}$ respectively: $\hat{\sigma}^2_{\hat{M}} = \hat{M}_{11}q^2/[p(2-p)]$ and $\hat{\sigma}^2_{\tilde{M}} = \hat{T}_{11}q/(4p)$. When $N = 2M_{11} + T_{11}$ is known, various other unbiased estimators of $\sigma^2_{\hat{M}}$ and $\sigma^2_{\tilde{M}}$ could be obtained, and various iterative procedures could be suggested for the estimation of M_{11}. We shall not go into these details here, except to mention that, when N is large ($N \to \infty$) and M_{11}/N is a fixed unknown constant, an approximation to the maximum likelihood estimator of M_{11} (based on the simplified set of data) can be obtained by an examination of the roots of a fourth degree equation in M_{11}, where the coefficients of the equation are a particular set of functions of x_{11}, z_{11}, p, and N.

When $s = k = 1$, the expected number of individuals in the population who will be interviewed is the sum of the expected number of individuals in the random sample and the expected number of individuals in the first stage; i.e.,

$$Np + N(1-p)\sum_{i=0}^{N-1}[1-(1-p)^i]b_{11}(i) = N\sum_{i=0}^{N-1}[1-(1-p)^{i+1}]b_{11}(i)$$
$$= N[1-(1-p)\sum_{i=0}^{N-1}(1-p)^i b_{11}(i)],$$

where $b_{11}(i)$ is the proportion of the individuals in the population who are named by i different individuals in the population,

$$\sum_{i=0}^{N-1} b_{11}(i) = 1, \quad \text{and} \quad \sum_{i=0}^{N-1} i b_{11}(i) = 1.$$

We now have the following theorem:

THEOREM 2: *For all $N > 1$, the expected number of individuals interviewed is not greater than $Np(2-p) = N[1-(1-p)^2]$.*

PROOF: We first note that $\sum_{i=0}^{N-1}(1-p)^i b_{11}(i) = E\{(1-p)^i\}$ is greater than or equal to

$$(1-p) = (1-p)^{\sum_{i=0}^{N-1}[i b_{11}(i)]} = (1-p)^{E\{i\}};$$

i.e., that $\log E\{(1-p)^i\} \geqq E\{i\}\log(1-p) = E\{\log(1-p)^i\}$. This fact follows from the convexity of $-\log x$ (see [6], p. 186). The lower bound is attained when $b_{11}(1) = 1$ and $b_{11}(i) = 0$ for $i \neq 1$, which will occur if each individual is named by exactly one individual in the population ($N > 1$).

Theorem 2 indicates that the maximum expected number of individuals interviewed can be computed as a function of p or, on the other hand, the appropriate value of p can be determined when the maximum expected number of individuals interviewed has been specified as Nf_{11}; i.e., then $p = 1 - (1-f_{11})^{\frac{1}{2}}$.

Let us now compare the situation where $s = k = 1$, and the sampling fraction is p, with the situation where $s = 0$, $k = 1$, and the sampling fraction is $f_{11} = 1 - (1-p)^2$. In the former situation, the maximum expected number of individuals interviewed is Nf_{11}; in the latter situation, the expected number of individuals interviewed (which is in this case the expected number of individuals in the random sample) is also Nf_{11}. Although the expected number of individuals interviewed in the former situation will be no more than the expected number of individuals interviewed in the latter situation, we shall see that the variance of $\hat{M}_{11}$ in the former situation ($s = k = 1, p$) is smaller than the variance of the minimum variance unbiased estimator of M_{11} (based on the w_{112}) in the latter situation ($s = 0, k = 1, f_{11}$). In the former situation, w_{111}, w_{112}, and z_{11} can be observed; in the latter situation, w_{111} and z_{11} cannot be observed, but w_{112} can. In the latter situation, w_{112} will have a binomial distribution with expected value $E\{w_{112}\} = M_{11}f_{11}^2$, the unbiased estimator of M_{11} will be $w_{112}/f_{11}^2 = M_{11}^*$, and the variance of M_{11}^* will be $\sigma_{M^*}^2 = M_{11}(1-f_{11}^2)/f_{11}^2$. By an argument similar to that used in the proof for Theorem 1, it can be seen that M_{11}^* is the minimum variance unbiased estimator of M_{11} (based on the w_{112}) in the latter situation when the population characteristics are completely unknown. In the former situation, we have that $\sigma_{\hat{M}}^2 = M_{11}(1-p)^2/p(2-p) = M_{11}(1-f_{11})/f_{11}$. Thus,

$$\sigma_{M^*}^2 - \sigma_{\hat{M}}^2 = M_{11}\{(1-f_{11}^2)/f_{11}^2 - (1-f_{11})/f_{11}\} = M_{11}(1-f_{11})/f_{11}^2,$$

$$[\sigma_{M^*}^2 - \sigma_{\hat{M}}^2]/\sigma_{\hat{M}}^2 = 1/f_{11},$$

and $\sigma_{\hat{M}}^2/\sigma_{M^*}^2 = f_{11}/(1+f_{11})$. This indicates that for estimating M_{11} the former situation ($s = k = 1, p$) is preferable to the latter situation ($s = 0, k = 1, f_{11}$) when $f_{11} < 1$.

The estimator $\hat{M}_{11}$, which we have discussed in this section, is a function of w_{111} and w_{112}; i.e., a function of the number of mutual relationships observed where at least one of the individuals entering the relationship is in the sample. Let us now consider the number, $w_{11\cdot}$, of mutual relationships observed where either none, one, or both of the individuals are in the sample. We have that $w_{11\cdot} = w_{110} + w_{111} + w_{112}$. Let $w_{11\cdot ij}$ be the number of mutual relationships observed where either none, one, or both of the individuals are in the sample and where one of the mutually related individuals is named by i individuals in the sample who do not enter the relationship and the other individual is named by j individuals in the sample ($0 \leqq i \leqq j$) who do not enter the relationship. Let

$w_{11\cdot i\cdot} = \sum_{j \geqq i} w_{11\cdot ij}$ (summed over all j such that $j \geqq i$). Then

$$w_{11\cdot} = \sum_{i \geqq 0} w_{11\cdot i\cdot} = w_{11\cdot 0\cdot} + w_{11\cdot +\cdot},$$

where $w_{11\cdot +\cdot} = \sum_{i \geqq 1} w_{11\cdot i\cdot}$. We note that $w_{11\cdot 0\cdot}$ will include only mutual relationships observed where either one or both of the individuals are in the sample. Let M_{11ij} be the number of mutual relationships in the population where one of the individuals entering the relationship is named by i individuals in the population who do not enter the relationship and the other individual is named by j individuals ($0 \leqq i \leqq j$) who do not enter the relationship. Let $M_{11i\cdot} = \sum_{j \geqq i} M_{11ij}$ (summed over all j such that $j \geqq i$). Then

$$M_{11} = \sum_{i \geqq 0} M_{11i\cdot} = M_{110\cdot} + M_{11+\cdot},$$

where $M_{11+\cdot} = \sum_{i \geqq 1} M_{11i\cdot}$. The expected values of $w_{11\cdot 0\cdot}$ and $w_{11\cdot +\cdot}$ are

$$E\{w_{11\cdot 0\cdot}\} = (1 - q^2) \sum_{i \geqq 0} \sum_{j \geqq i} M_{11ij}[1 - (1 - q^i)(1 - q^j)],$$

and

$$E\{w_{11\cdot +\cdot}\} = \sum_{i \geqq 1} \sum_{j \geqq i} M_{11ij}(1 - q^i)(1 - q^j),$$

respectively. Thus, $\bar{M}_{11} = [w_{11\cdot 0\cdot}/(1 - q^2)] + w_{11\cdot +\cdot}$ is an unbiased estimator of M_{11} since $E\{\bar{M}_{11}\} = M_{110\cdot} + M_{11+\cdot} = M_{11}$. The variance of $w_{11\cdot 0\cdot}$ is

$$\sigma^2\{w_{11\cdot 0\cdot}\} = (1 - q^2) \sum_{i \geqq 0} \sum_{j \geqq i} M_{11ij}(q^i + q^j - q^{i+j}) \cdot[1 - (1 - q^2)(q^i + q^j - q^{i+j})],$$

and the variance of $w_{11\cdot +\cdot}$ is

$$\sigma^2\{w_{11\cdot +\cdot}\} = \sum_{i \geqq 1} \sum_{j \geqq i} M_{11ij}(1 - q^i)(1 - q^j)(q^i + q^j - q^{i+j}).$$

The covariance between $w_{11\cdot 0\cdot}$ and $w_{11\cdot +\cdot}$ is

$$\text{Cov}\{w_{11\cdot 0\cdot}, w_{11\cdot +\cdot}\} = -\sum_{i \geqq 1} \sum_{j \geqq i} M_{11ij}(1 - q^2)(q^i + q^j - q^{i+j})(1 - q^i)(1 - q^j).$$

Thus, the variance of $\bar{M}_{11}$ is

$$\begin{aligned}
\sigma^2_{\bar{M}_{11}} &= \sum_{i \geqq 0} \sum_{j \geqq i} M_{11ij}(q^i + q^j - q^{i+j})[1 - (1 - q^2)(q^i + q^j - q^{i+j})]/(1 - q^2) \\
&\quad - \sum_{i \geqq 1} \sum_{j \geqq i} M_{11ij}(1 - q^i)(1 - q^j)(q^i + q^j - q^{i+j}) \\
&= \sum_{i \geqq 0} \sum_{j \geqq i} M_{11ij}[(q^i + q^j - q^{i+j})q^2/(1 - q^2)] \\
&= \sum_{i \geqq 0} \sum_{j \geqq i} M_{11ij}[1 - (1 - q^i)(1 - q^j)]q^2/(1 - q^2) \\
&= [M_{11} - \sum_{i \geqq 1} \sum_{j \geqq i} M_{11ij}(1 - q^i)(1 - q^j)]q^2/(1 - q^2).
\end{aligned}$$

If $M_{11i;j} = 0$ for all $i \geqq 1$, it is easy to see that $\bar{M}_{11} = \hat{M}_{11}$. If $M_{11i;j} = 0$ for all $i \geqq 1$ except M_{1111}, then $\sigma^2_{\bar{M}_{11}} = (M_{11} - M_{1111}p^2)q^2/(1 - q^2)$. Thus the ratio of the variances in this case is $\sigma^2_{\bar{M}_{11}}/\sigma^2_{\hat{M}_{11}} = 1 - p^2(M_{1111}/M_{11})$; while $\bar{M}_{11}$ has a smaller variance than $\hat{M}_{11}$, the relative decrease in the variance is at most p^2 (which is attained only when $M_{1111} = M_{11}$; i.e., when $M_{110\cdot} = 0$). The relative decrease in the variance in the general case will depend on the unknown parameters $M_{11i;j}$ $(0 < i < j)$. (It is possible to derive an unbiased estimator for each $M_{11i;j}$ $(0 \leqq i \leqq j)$, but we shall not go into these details here.) For some values of the $M_{11i;j}$, the relative decrease in the variance can be large.

The estimator $\hat{M}_{11}$ is generally easier to compute than $\bar{M}_{11}$ and the statistical properties of $\hat{M}_{11}$ are simpler to determine than are the corresponding properties of $\bar{M}_{11}$. On the other hand, the variance of $\hat{M}_{11}$ is greater than or equal to that of $\bar{M}_{11}$. (The variance of $\hat{M}_{11}$ will be greater than that of $\bar{M}_{11}$ whenever $M_{11} > M_{110\cdot}$.) The estimator $\hat{M}_{11}$ does not use information about the number of observed mutual relationships between individuals none of whom are in the sample, while the estimator $\bar{M}_{11}$ does. We have obtained a more accurate estimator, $\bar{M}_{11}$, although one that is not as simple as $\hat{M}_{11}$, by using this information. When accuracy is more important than simplicity, as it is often the case, $\bar{M}_{11}$ should be used rather than $\hat{M}_{11}$. However, in the present paper where we shall show how some of the results obtained concerning the statistical analysis of the data from a one stage one name snowball sample can be generalized to obtain corresponding results concerning the statistical analysis of data from an s stage k name snowball sample where s and k are any specified positive integers, it will be desirable to keep the exposition as simple as possible. For the sake of this simplicity, we shall henceforth ignore the information about the number of observed mutual relationships (and other more general kinds of relationships discussed in subsequent sections) between individuals none of whom are in the sample. (While we have commented upon the effect of ignoring these relationships when $s = k = 1$, it is beyond the scope of this paper to study this effect when s and/or k are greater than 1.)

3. The Case Where s is a Specified Positive Integer and $k = 1$. We shall discuss the estimation of the number, M_{s1}, of $s + 1$ person circular relationships in the population; i.e., the number of combinations of $s + 1$ individuals in the population where the $s + 1$ individuals can be arranged so that the first individual, if asked to name an individual in the population, would name the second individual, the second individual would name the third, $\cdots$, the $(s + 1)$th individual would name the first. A two person circular relationship is a mutual relationship as defined in the preceding section, and the results in the present section are a direct generalization of the prior results. The proofs of these results are similar to the proofs given in the preceding section, and therefore will not be included here.

Let x_{s1} be the number of $s + 1$ person circular relationships observed from the data in an s stage one name snowball sample; i.e., $x_{s1} = w_{s,1,1} + w_{s,1,2} + w_{s,1,3} +$

$\cdots + w_{s,1,s+1}$, where $w_{s,1,j}$ is the number of $s + 1$ person circular relationships observed where j individuals entering the relationship are in the sample and $s + 1 - j$ individuals entering the relationship are in the other observed stages. (For the sake of simplicity, we shall not consider here the number, $w_{s,1,0}$, of $s + 1$ person circular relationships observed where none of the individuals entering the relationship are in the sample; see related comments in Section 2.) The random variable x_{s1} has a binomial distribution with expected value

$$E\{x_{s1}\} = M_{s1}[1 - (1 - p)^{s+1}].$$

Thus, an unbiased estimator of M_{s1} is $x_{s1}/[1 - (1 - p)^{s+1}] = \hat{M}_{s1}$, and the variance of $\hat{M}_{s1}$ is $\sigma^2_{\hat{M}_{s1}} = M_{s1}(1 - p)^{s+1}/[1 - (1 - p)^{s+1}]$. An unbiased estimator of $\sigma^2_{\hat{M}_{s1}}$ is $\hat{\sigma}^2_{\hat{M}_{s1}} = \hat{M}_{s1}(1 - p)^{s+1}/[1 - (1 - p)^{s+1}]$.

Let z_{s1} be the number of individuals in the sample who are not members of $s + 1$ person circular relationships. Because the snowball sampling design has s stages, z_{s1} is observed. We shall refer to the set

$$(w_{s,1,1}, w_{s,1,2}, w_{s,1,3}, \cdots, w_{s,1,s+1}, z_{s1})$$

as the simplified set of data for $s + 1$ person circular relationships when s is a specified integer and $k = 1$. We shall, for the sake of simplicity, now limit our consideration to $(w_{s,1,1}, w_{s,1,2}, \cdots, w_{s,1,s+1}, z_{s1})$, although it may sometimes be worthwhile to make use of additional available data (see related comments in Section 2). By the same method of proof as for Theorem 1, it can be seen that, if the population characteristics (including its size) are unknown, then the estimator $\hat{M}_{s1}$ has minimum variance among all unbiased estimators of M_{s1} based on the simplified set of data. Similarly, $\hat{\sigma}^2_{\hat{M}_{s1}}$ has a minimum variance among all unbiased estimators of $\sigma^2_{\hat{M}_{s1}}$ based on these data.

The estimators $\hat{M}_{s1}$ and $\hat{\sigma}^2_{\hat{M}_{s1}}$ of M_{s1} and $\sigma^2_{\hat{M}_{s1}}$, respectively, will be unbiased whether or not the population size, N, is known. However, when N is known, these estimators need not have minimum variance. When N is known, other unbiased estimators can be based on the relationship $(s + 1)M_{s1} + T_{s1} = N$, where T_{s1} is the number of individuals in the population who do not enter $s + 1$ person circular relationships. Since the random variable z_{s1} has a binomial distribution with expected value $E\{z_{s1}\} = T_{s1}p$, the estimator z_{s1}/p is an unbiased estimator of T_{s1} and $[N - z_{s1}/p]/(s + 1) = \tilde{M}_{s1}$ is an unbiased estimator of M_{s1}. The variance $\sigma^2_{\tilde{M}_{s1}}$ of $\tilde{M}_{s1}$ is $T_{s1}(1 - p)/p(s + 1)^2$. The $\hat{M}_{s1}$ and $\tilde{M}_{s1}$ are statistically independent, and could be combined in various ways, which we shall not discuss here (see related discussion in Section 2).

With an s stage one name snowball sample, the expected number of individuals interviewed is

$$N\{\textstyle\sum_i [1 - (1 - p)^{i+1}]b_{s1}(i)\} = N\{1 - \sum_i (1 - p)^{i+1} b_{s1}(i)\},$$

where $b_{s1}(i)$ is the proportion of the population who are named either directly (in one step) or indirectly in s steps or less by i different individuals in the population; i.e., each individual in the population has an influence score i, where i is

the total number of different individuals who name him (they form step minus one) or who name individuals who in turn name him (they form step minus two) or who name individuals who in turn name individuals who name him (they form step minus three), etc., until step minus s has been considered, and $b_{s1}(i)$ is the proportion of individuals in the population who have the influence score i, for $i = 0, 1, 2, \cdots$. We have that $\sum_i b_{s1}(i) = 1$ and $\sum_i i b_{s1}(i) \leqq s$. As in the proof of Theorem 2, it can be seen that, for all $N > s$, the maximum expected number of individuals interviewed is $N[1 - (1 - p)^{s+1}] = Nf_{s1}$, and the maximum occurs when $b_{s1}(s) = 1$ and $b_{s1}(i) = 0$ for $i \neq s$; this can be attained, for example, if the individuals in the population form an N person circular relationship ($N > s$). Thus, it is possible to determine this maximum as a function of p or, on the other hand, to determine the appropriate value of p when the maximum expected proportion of the population to be interviewed has been specified as f_{s1}; i.e., $p = 1 - (1 - f_{s1})^{1/(s+1)}$.

Let us now compare the situation where an s stage one name snowball sample is drawn and the sampling fraction is p with the situation where an $s - 1$ stage one name snowball sample is drawn and the maximum expected proportion of the population to be interviewed is f_{s1}. In the latter situation, the sampling fraction will be $p_{s-1,1} = 1 - (1 - f_{s1})^{1/s}$. In both situations, the maximum expected proportion of the population to be interviewed is f_{s1}, but we shall see that $\hat{M}_{s1}$ computed in the former situation $(s, 1, p)$ will have a smaller variance than the variance of the minimum variance unbiased estimator of M_{s1} (based on a simplified set of data) in the latter situation $(s - 1, 1, p_{s-1,1})$ when the population characteristics are unknown. Let $w_{s1} = w_{s,1,2} + w_{s,1,3} + \cdots + w_{s,1,s+1}$ be the number of $s + 1$ person circular relationships observed from the data in the latter situation; i.e., w_{s1} is the number of $s + 1$ person circular relationships observed where two or more of the individuals entering the relationship are in the sample obtained. (For the sake of simplicity, we shall not consider here, where an $s - 1$ stage one name sampling procedure is used, the number of $s + 1$ person circular relationships observed where either one or none of the individuals entering the relationship are in the sample; see comments dealing with this point in Section 4.) The random variable w_{s1} will have a binomial distribution with expected value $E\{w_{s1}\}$ equal to $M_{s1}[1 - P_{s-1,1}]$, where

$$P_{s-1,1} = [1 - p_{s-1,1}]^{s+1} + (s + 1)[1 - p_{s-1,1}]^{s} p_{s-1,1}.$$

The unbiased estimator of M_{s1} will be $w_{s1}/[1 - P_{s-1,1}] = M^*_{s1}$ and the variance $\sigma^2_{M^*_{s1}}$ of M^*_{s1} will be $M_{s1}P_{s-1,1}/[1 - P_{s-1,1}]$. Limiting our consideration to the simplified set of data, $(w_{s,1,2}, w_{s,1,3}, \cdots, w_{s,1,s+1})$, when an $s - 1$ stage one name snowball sampling procedure is used to estimate M_{s1}, we find, by an argument similar to that used in the proof of Theorem 1, that M^*_{s1} is the minimum variance unbiased estimator of M_{s1} based on these data in the latter situation $(s - 1, 1, p_{s-1,1})$ when the population characteristics are unknown. We also find that

$$[\sigma^2_{M^*_{s1}} - \sigma^2_{\hat{M}_{s1}}]/\sigma^2_{\hat{M}_{s1}} = sp_{s-1,1}/[1 - P_{s-1,1}],$$

and

$$\sigma^2_{\hat{M}_{s1}}/\sigma^2_{M^*_{s1}} = [1 - P_{s-1,1}]/[1 + sp_{s-1,1} - P_{s-1,1}].$$

This indicates that for estimating M_{s1}, the former situation $(s, 1, p)$ is preferable to the latter situation $(s - 1, 1, p_{s-1,1})$ when $p_{s-1,1} < 1$.

We have discussed the estimation of M_{s1} from an s stage one name snowball sample (and from an $s - 1$ stage one name snowball sample). It is also possible to estimate M_{t1} for any $t \leqq s$ from an s stage one name snowball sample since it contains all the information obtained in a corresponding t stage one name snowball sample. (See comments dealing with this point in Section 4.) From an s stage one name snowball sample it is also possible, using an approach similar to that described in the present section, to estimate M_{t1} for any $t > s$.

In closing this section, we note that an $s + 1$ person circular relationship is an $s + 1$ person closed system in the sense that each of the $s + 1$ individuals entering the relationship names an individual from among the $s + 1$ individuals, and that it is an irreducible system in the sense that no proper subset of the $s + 1$ individuals will form a closed system. Any $s + 1$ individuals forming an $s + 1$ person closed irreducible system, in the sense defined here, will form an $s + 1$ person circular relationship. We also note that an $s + 1$ person circular relationship is an $s + 1$ person s step one direction relationship in the sense that, starting with any given individual entering the relationship, if we include him, the individual he names (this individual is said to be on step one), the individual named by the individual on step one (this individual is said to be on step two), $\cdots$, the individual named by the individual on step $s - 1$ (this individual is said to be on step s), we will have included all $s + 1$ individuals forming the relationship (and no others). Any $s + 1$ individuals forming an $s + 1$ person s step one direction relationship will form an $s + 1$ person circular relationship.

4. The Case Where s and k are Specified Positive Integers. Let M_{sk} be the number of $s + k$ person s step k direction relationships in the population; i.e., the number of combinations of $s + k$ individuals in the population where, starting with any given individual in the combination, if we include him, the k individuals he would name (they are said to be on step one), the k individuals who would be named by each of the individuals on step one (they are said to be on step two), $\cdots$, the k individuals who would be named by each of the individuals on step $s - 1$ (they are said to be on step s), we will have included all $s + k$ indidivuals in the combination (and no others). From the comments in the preceding section, we see that the number, M_{s1}, of $s + 1$ person s step one direction relationships is equal to the number of $s + 1$ person circular relationships. The number, M_{1k}, of $1 + k$ person one step k direction relationships is equal to the number of $k + 1$ person cliques; i.e., the number of combinations of $k + 1$ individuals where each individual in the combination would name the other k individuals in the combination. The results presented in the present section are direct generalizations of the prior results; the proofs of these results are similar to the

proofs presented earlier, and therefore will not be included, except at certain points where they may not be directly evident.

Let x_{sk} be the number of $s + k$ person s step k direction relationships observed from the data in an s stage k name snowball sample; i.e.,

$$x_{sk} = w_{s,k,1} + w_{s,k,2} + w_{s,k,3} + \cdots + w_{s,k,s+k},$$

where $w_{s,k,j}$ is the number of $s + k$ person s step k direction relationships observed where j individuals entering the relationship are in the sample and $s + k - j$ individuals entering the relationship are in the other observed stages. (For the sake of simplicity, we shall not consider here the number, $w_{s,k,0}$, of $s + k$ person s step k direction relationships observed where none of the individuals in the relationship are in the sample.) The random variable x_{sk} has a binomial distribution with expected value $E\{x_{sk}\} = M_{sk}[1 - (1 - p)^{s+k}]$. Thus, an unbiased estimator of M_{sk} is $x_{sk}/[1 - (1 - p)^{s+k}] = \hat{M}_{sk}$, and the variance of $\hat{M}_{sk}$ is $\sigma^2_{\hat{M}_{sk}} = M_{sk}(1 - p)^{s+k}/[1 - (1 - p)^{s+k}]$. An unbiased estimator of $\sigma^2_{\hat{M}_{sk}}$ is $\hat{\sigma}^2_{\hat{M}_{sk}} = \hat{M}_{sk}(1 - p)^{s+k}/[1 - (1 - p)^{s+k}]$.

Let z_{sk} be the number of individuals in the sample who are not members of $s + k$ person s step k direction relationships. Because the snowball sampling design has s stages, z_{sk} is observed. We shall refer to the set of

$$(w_{s,k,1}, w_{s,k,2}, w_{s,k,3}, \cdots, w_{s,k,s+k}, z_{sk})$$

as the simplified set of data for $s + k$ person s step k direction relationships. As in the earlier sections, we shall, for the sake of simplicity, limit our consideration to $(w_{s,k,1}, w_{s,k,2}, \cdots, w_{s,k,s+k}, z_{sk})$, when an s stage k name snowball sample is used to estimate M_{sk}. By the same method of proof as for Theorem 1, it can be seen that, if the population characteristics are unknown, then the estimator $\hat{M}_{sk}$ has minimum variance among all unbiased estimators of M_{sk} based on the simplified set of data when s and k are specified integers. Similarly, the estimator $\hat{\sigma}^2_{\hat{M}_{sk}}$ has minimum variance among all unbiased estimators of $\sigma^2_{\hat{M}_{sk}}$ based on these data.

Although $\hat{M}_{sk}$ and $\hat{\sigma}^2_{\hat{M}_{sk}}$ will be unbiased estimators of M_{sk} and $\sigma^2_{\hat{M}_{sk}}$, respectively, whether or not the population size, N, is known, these estimators need not have minimum variance when N is known. When N is known, unbiased estimators can be based on the fact that $(s + k)M_{sk} + T_{sk} = N$, where T_{sk} is the number of individuals in the population who are not members of $s + k$ person s step k direction relationships. The details in this case are very similar to those appearing earlier (see related comments in Sections 2 and 3).

With an s stage k name snowball sample, the expected number of individuals interviewed is

$$N\{\textstyle\sum_i [1 - (1 - p)^{i+1}]b_{sk}(i)\} = N\{1 - \sum_i (1 - p)^{i+1} b_{sk}(i)\},$$

where $b_{sk}(i)$ is the proportion of the population who are named either directly (in one step) or indirectly in s steps or less by i different individuals in the population; i.e., each individual in the population has an influence score i, where i is the total number of different individuals who name him (they form step minus

one) or who name individuals who in turn name him (they form step minus two) or who name individuals who in turn name individuals who name him (they form step minus three), etc., until step minus s has been considered, and $b_{sk}(i)$ is the proportion of the individuals in the population who have influence score i, for $i = 0, 1, 2, \cdots$. We have that $\sum_i b_{sk}(i) = 1$ and $\sum_i i b_{sk}(i) \leqq c_{sk}$, where $c_{sk} = k + k^2 + \cdots + k^s$. The following theorem is a generalization of Theorem 2:

THEOREM 3: *When s and k are specified integers, the maximum expected number of individuals interviewed is $N[1 - (1 - p)^{c_{sk}+1}]$, for N sufficiently large.*

PROOF: The fact that $N[1 - (1 - p)^{c_{sk}+1}] \geqq N\{1 - \sum_i (1 - p)^{i+1} b_{sk}(i)\}$ can be proved using the same method as for Theorem 2. The bound is attained whenever $b_{sk}(c_{sk}) = 1$ and $b_{sk}(i) = 0$ for $i \neq c_{sk}$. It is possible to prove that, for N sufficiently large, the bound can be attained. The detailed calculations will not be given here.

Theorem 3 indicates that it is possible to determine the maximum expected number of individuals interviewed as a function of p or, on the other hand, to determine the appropriate value of p when the maximum expected proportion of the population to be interviewed has been specified as f_{sk}; i.e.,

$$p = 1 - (1 - f_{sk})^{1/(c_{sk}+1)}.$$

Let us now compare the situation where an s stage k name snowball sample is drawn and the sampling fraction is p with the situation where an $s - 1$ stage k name snowball sample is drawn and the maximum expected proportion of the population interviewed is f_{sk}. In the latter situation, the sampling fraction will be $p_{s-1,k} = 1 - (1 - f_{sk})^{1/(c_{s-1,k}+1)}$. In both situations, the maximum expected proportion of the population to be interviewed is f_{sk}. In the latter situation $(s - 1, k, p_{s-1,k})$, an estimator of M_{sk} will be based (for reasons made clear below) on the number $w_{sk} = w_{s,k,k+1} + w_{s,k,k+2} + \cdots + w_{s,k,s+k}$ of $s + k$ person s step k direction relationships observed, where $k + 1$ or more individuals entering the relationship are in the sample and $s - 1$ or fewer individuals entering the relationship are in the other $s - 1$ stages. (If $k + 1$ (or more) individuals entering an $s + k$ person s step k direction relationship are observed in the sample, then the relationship will be detected in an $s - 1$ stage k name snowball sample since the remaining $s - 1$ (or fewer) individuals will be observed in the other $s - 1$ (or fewer) stages. If one (or more) individuals in an $s + k$ person s step k direction relationship is observed in the sample, then the relationship will be detected in an s stage k name snowball sample since the remaining $k + s - 1$ (or fewer) individuals will be observed in the other stages; k individuals will be observed in the first stage, when one individual is in the sample, and the remaining $s - 1$ individuals will be observed in stages $2, 3, \cdots, s$.) The random variable w_{sk} will have a binomial distribution with expected value $E\{w_{sk}\}$ equal to $M_{sk}[1 - P_{s-1,k}]$, where

$$P_{s-1,k} = \sum_{i=0}^{k} \binom{s+k}{i} p_{s-1,k}^{i} (1 - p_{s-1,k})^{s+k-i}.$$

The unbiased estimator of M_{sk} is $M^*_{sk} = w_{sk}/[1 - P_{s-1,k}]$, and the variance of M^*_{sk} is $\sigma^2_{M^*_{sk}} = M_{sk}P_{s-1,k}/[1 - P_{s-1,k}]$. Limiting our consideration to the simplified set of data, $(w_{s,k,k+1}, w_{s,k,k+2}, \cdots, w_{s,k,s+k})$, when an $s - 1$ stage k name snowball sampling procedure is used to estimate M_{sk}, we find by an argument similar to that used in the proof of Theorem 1 that M^*_{sk} is the minimum variance unbiased estimator of M_{sk} based on these data in the latter situation $(s - 1, k, p_{s-1,k})$ when the population characteristics are unknown. We also find that, in the former situation (s, k, p),

$$\begin{aligned}\sigma^2_{\hat{M}_{sk}} &= M_{sk}(1 - p)^{s+k}/[1 - (1 - p)^{s+k}] \\ &= M_{sk}(1 - f_{sk})^{e_{sk}}/[1 - (1 - f_{sk})^{e_{sk}}] \\ &= M_{sk}(1 - p_{s-1,k})^{g_{sk}}/[1 - (1 - p_{s-1,k})^{g_{sk}}],\end{aligned}$$

where $e_{sk} = (s + k)/(c_{sk} + 1)$, $g_{sk} = (s + k)(c_{s-1,k} + 1)/(c_{s,k} + 1)$, $c_{0,k} = 0$, so that

$$\begin{aligned}\sigma^2_{M^*_{sk}} - \sigma^2_{\hat{M}_{sk}} &= M_{sk}\{P_{s-1,k}/[1 - P_{s-1,k}] - (1 - p_{s-1,k})^{g_{sk}}/[1 - (1 - p_{s-1,k})^{g_{sk}}]\} \\ &= M_{sk}\{P_{s-1,k} - (1 - p_{s-1,k})^{g_{sk}}\}/\{[1 - P_{s-1,k}][1 - (1 - p_{s-1,k})^{g_{sk}}]\}.\end{aligned}$$

In the preceding section, we saw that this difference was positive when $k = 1$. When $s = 1$, this difference is also positive since

$$\left[\sum_{i=0}^{k}\binom{k+1}{i}p^i_{0,k}(1 - p_{0,k})^{k+1-i} - (1 - p_{0,k})\right]\Big/(1 - p_{0,k})$$

$$> \left[\sum_{i=0}^{k}\binom{k}{i}p^i_{0,k}(1 - p_{0,k})^{k-i} - 1\right] = 0,$$

when $p_{0,k} < 1$. Thus, when $s = 1$, the former situation $(1, k, p)$ is preferable to the latter situation with regard to the estimation of M_{1k}, the number of $k + 1$ person cliques. This generalizes the result presented in Section 2 indicating the preferability of the snowball sample with $s = 1$ to the case where $s = 0$. It is interesting to note that, while an s stage one name snowball sample is preferable, with regard to the estimation of M_{s1}, to a comparable $s - 1$ stage one name snowball sample (see Section 3), it is not always the case that an s stage k name snowball sample is preferable, with regard to the estimation of M_{sk}, to a comparable $s - 1$ stage k name snowball sample, except when either $k = 1$ or $s = 1$. This follows from the fact that the difference $\sigma^2_{M^*_{sk}} - \sigma^2_{\hat{M}_{sk}}$ can be negative for certain values of s, k, and $p_{s-1,k}$; e.g., $s = 2$, $k = 2$, and $p_{1,2}$ very close to one.

We have discussed the estimation of M_{sk} from an s stage k name snowball sample (and from an $s - 1$ stage k name snowball sample). It is also possible to estimate M_{tk} for any $t \leqq s$ from an s stage k name snowball sample since it contains all the information obtained in a corresponding t stage k name snowball sample. Thus, by using data for $t + k$ person t step k direction relationships obtained from the first t stages (and from the random sample) of an s stage k name snowball sample $(t \leqq s)$, the methods presented earlier in this section can be applied in order to estimate M_{tk}. These methods will lead to simple estima-

tors; when $t < s$, they will of course not make use of all of the available data. For example, when $t = 1$, $s = 2$, $k = 1$, these methods will not make use of the information available about the number of mutual relationships observed where one of the individuals entering the relationship is in the second stage. Using this information, it is possible, to improve upon the estimator $\hat{M}_{11}$ of M_{11} in, say, the special case where it is assumed that each individual in the population entering a mutual relationship is named by exactly two individuals in the population. In this case, this information could be used in a way similar to that described in Section 2 where the estimator $\hat{M}_{11}$ was improved upon by using the information about the number of observed mutual relationships between individuals both of whom are in the first stage.

When the individuals in the sample are asked to list in a specified order k different individuals in the population (for example, each individual may be asked to name his "k best friends" and to rank them with regard to some specified criterion) and the individuals forming the various stages are asked to do likewise, the methods developed in this paper can be used to estimate M_{th}, for $t = 1, 2, \cdots, s$ and $h = 1, 2, \cdots, k$, from an s stage k name snowball sample, where M_{th} is understood to be the number of $t + h$ person t step h direction relationships obtained when considering the first h individuals listed by each individual in the population (or, more generally, when considering any specified subset of h individuals listed by each individual).

In this section, we have been concerned in the main with the estimation of the number, M_{sk}, of $s + k$ person s step k direction relationships in the population. Let us now consider briefly the number, M_{skg}, of g person s step k direction relationships; i.e., the number of combinations of g individuals in the population where, starting with any given individual in the combination, if we include him, the k individuals he would name (they are said to be on step one), the k individuals who would be named by each of the individuals on step one (they are said to be step two), $\cdots$, the k individuals who would be named by each of the individuals on step $s - 1$ (they are said to be on step s), we will have included all g individuals in the combination (and no others). Obviously, $M_{skg} = M_{sk}$ for $g = s + k$. For $g \leqq s + k$, $M_{skg} = M_{g-k,k}$ (where $g \geqq 1 + k$). This follows from the fact that if we start with any given individual in an g person s step k direction relationship, where $g \leqq s + k$, if we include him and the individuals on step one, two, $\cdots$, $g - k$, we will have included $1 + k + (g - k - 1) = g$ individuals. Thus, for $g \leqq s + k$ the results presented in this section can be applied directly to estimate $M_{skg} = M_{g-k,k}$. Since an upper bound for g is $1 + c_{sk}$ (this bound is not attainable for some values of s and k), we have that $g = 1 + k$ for $s = 1$, and $g \leqq 1 + s$ for $k = 1$. Thus for $s = 1$ or for $k = 1$, $M_{skg} = M_{g-k,k}$ (for $1 + k \leqq g \leqq 1 + c_{sk}$). We note therefore that for $s = 1$ or for $k = 1$ the results presented earlier can be applied directly to estimate M_{skg}. For $g > s + k$ ($s > 1$ and $k > 1$), the methods developed in the present section for the estimation of M_{sk} from an s stage k name snowball sample (and from an $s - 1$ stage k name snowball sample) can be generalized in a straightforward manner in order to obtain similar methods for the estimation of M_{skg}. It is possible to

estimate M_{tkg} for any $t \leqq s$ from an s stage k name snowball sample (since it contains all the information obtained in a corresponding t stage k name snowball sample) and for any $t \geqq s$ using an approach similar to that described earlier in the present section. (Here the range of possible values of g is of course a function of t and k.) The modification of the s stage k name snowball sampling procedure described in the preceding paragraph could also be used to estimate M_{thg} for $h \leqq k$. (Here the range of g is a function of t and h.)

Other kinds of g person relationships can be defined and in some cases the s stage k name snowball sampling procedure can be used to estimate the number of such relationships present in the population. This will be the case if the definition of the g person relationship is such that (i) an individual can belong to at most one such relationship, (ii) the data obtained by the s stage k name snowball sampling procedure can be used to determine whether or not any given individual appearing in the initial sample (i.e., in the zero stage) belongs to such a relationship, (iii) the data obtained can be used to determine whether any two individuals appearing in the initial sample belong to the very same g person relationship or not. These three conditions can be modified in various ways. For example, even if (i) is not satisfied an unbiased estimator is still available for the number, N_g, of such g person relationships present in the population in the case where (ii) and (iii) are satisfied and where the data obtained can also be used to determine the number of such relationships to which each individual in the initial sample belongs; but the formula for the variance of this estimator will not be as simple as the corresponding variance formulas presented earlier herein. Even if (ii) and (iii) are not satisfied an unbiased estimator for N_g is still available in the case where the data obtained can be used to determine whether any set of d (or more) individuals appearing in the initial sample belong to the very same g person relationship or not (d is a specified integer; $1 \leqq d \leqq g$). Even if (iii) is not satisfied (i.e., if the data required under (iii) are not available), an unbiased estimator for N_g is still available when (i) and (ii) hold true. In these cases, and in some other cases too, where modified forms of these conditions hold true, the methods developed herein can be generalized in a straightforward manner. It will therefore not be necessary to include the details here.

The snowball sampling procedure can be used for purposes other than those presented here. It is, however, beyond the scope of the present paper to study the other possible uses of snowball sampling.

5. Random Choices. In the preceding sections, we discussed the situation where each individual, if asked, would name k different individuals from the given finite population according to some specified criterion; e.g., his "k best friends". In the present section, we shall discuss briefly the situation where each individual, would name k other individuals chosen at random (without replacement) from this population. In this situation, the expected number of $k + 1$ person cliques (i.e., $1 + k$ person one step k direction relationships) will be

$$N\left[\binom{N-1}{k}\right]^{-k} \Big/ (k+1) = \binom{N}{k+1}\binom{N-1}{k}^{-(k+1)},$$

and the expected number of individuals who will be members of $k+1$ person cliques will be $N\left[\binom{N-1}{k}\right]^{-k}$. The expected number of $s+1$ person circular relationships (i.e., $s+1$ person s step one direction relationships) will be

$$N[(N-2)^{[s-1]}/(N-1)^s]/(s+1) = \binom{N}{s+1}[s^{[s]}/(N-1)^{s+1}]$$
$$= N^{[s+1]}/[(N-1)^{s+1}(s+1)],$$

where $x^{[s]} = x!/(x-s)!$, and the expected number of individuals who will be members of $s+1$ person circular relationships will be $N^{[s+1]}/(N-1)^{s+1}$. When $N \to \infty$, the expected number of $s+1$ person circular relationships approaches $1/(s+1)$, while the expected number of $k+1$ person cliques approaches zero when $k > 1$; the expected number of two person cliques (i.e., two person circular relationships) will approach $\frac{1}{2}$. The expected number of $s+k$ person s step k direction relationships will be less than or equal to

$$\binom{N}{s+k}\left[\binom{s+k-1}{k}\Big/\binom{N-1}{k}\right]^{s+k},$$

which approaches zero when $k > 1$; for $k = 1$, the expected number of $s+1$ person s step one direction relationships was given above (it approached $1/(s+1)$).

If each individual names k different individuals at random, the proportion $b_{1k}(i)$ of individuals in the population who are named by i different individuals is a random variable with expected value

$$E\{b_{1k}(i)\} = \binom{N-1}{i}\left(\frac{k}{N-1}\right)^i\left(1-\frac{k}{N-1}\right)^{N-1-i} = \Pr_{1k}\{i\}.$$

Thus, the expected proportion of the population interviewed, when a one stage k name snowball sample is drawn, is

$$1 - \sum_{i=0}^{N-1}(1-p)^{i+1}\Pr_{1k}\{i\}$$
$$= 1 - (1-p)\left[(1-p)\left(\frac{k}{N-1}\right) + 1 - \frac{k}{N-1}\right]^{N-1}$$
$$= 1 - (1-p)\left[1 - \frac{pk}{N-1}\right]^{N-1}$$
$$= 1 - (1-p)\sum_{i=0}^{N-1}\left(1-\frac{k}{N-1}\right)^i\binom{N-1}{i}p^i(1-p)^{N-1-i},$$

which approaches $1-(1-p)e^{-pk}$ as $N \to \infty$. The expected proportion interviewed, when an s stage k name snowball sample is drawn, can be written as

$$1 - (1-p)\sum_{i_1,i_2,\cdots,i_s}(1-p)^{i_1+i_2+\cdots+i_s}E\{b_k(i_1, i_2, i_3, \cdots, i_s)\},$$

where $b_k(i_1, i_2, i_3, \cdots, i_s)$ is the proportion of individuals in the population whose step minus one consists of i_1 individuals, whose step minus two consists of i_2 additional individuals (an individual appearing on both steps minus one and two is included in the i_1 count but not in the i_2 count; the individual himself, if he reappears on step minus two, is not counted), whose step minus three consists of i_3 additional individuals, $\cdots$, whose step minus s consists of i_s additional individuals. It is clear that

$$\sum_{i_1+i_2+\cdots+i_s=i} b_k(i_1, i_2, i_3, \cdots, i_s) = b_{sk}(i),$$

where summation is over all values of $i_1, i_2, \cdots, i_s$ such that $i_1 + i_2 + \cdots + i_s = i$. The expected value $E\{b_k(i_1, i_2, \cdots, i_s)\}$ of $b_k(i_1, i_2, \cdots, i_s)$ will be equal to

$$\begin{aligned}
\Pr_k\{i_1, i_2, \cdots, i_s\} = {} & \binom{N-1}{i_1}\left(\frac{k}{N-1}\right)^{i_1}\left(1-\frac{k}{N-1}\right)^{N-1-i_1} \\
& \cdot \binom{N-1-i_1}{i_2}\left[1-\binom{N-2-i_1}{k}\Big/\binom{N-2}{k}\right]^{i_2} \\
& \cdot \left[\binom{N-2-i_1}{k}\Big/\binom{N-2}{k}\right]^{N-1-i_1-i_2} \\
& \cdot \binom{N-1-i_1-i_2}{i_3}\left[1-\binom{N-2-i_1-i_2}{k}\Big/\binom{N-2-i_1}{k}\right]^{i_3} \\
& \cdot \left[\binom{N-2-i_1-i_2}{k}\Big/\binom{N-2-i_1}{k}\right]^{N-1-i_1-i_2-i_3} \\
& \cdots \binom{N-1-i_1-\cdots-i_{s-1}}{i_s} \\
& \cdot \left[1-\binom{N-2-i_1-\cdots-i_{s-1}}{k}\Big/\binom{N-2-i_1-\cdots-i_{s-2}}{k}\right]^{i_s} \\
& \cdot \left[\binom{N-2-i_1-\cdots-i_{s-1}}{k}\Big/\binom{N-2-i_1-\cdots-i_{s-2}}{k}\right]^{N-1-i},
\end{aligned}$$

where $i = i_1 + i_2 + \cdots + i_s$. When $s = 1$, $\Pr_k\{i_1\} = \Pr_{1k}\{i_1\}$ approaches $k^{i_1}e^{-k}/i_1!$ as $N \to \infty$. In other words, the random variable i_1 (the number of individuals forming step minus one for an individual drawn at random from the population) has a Poisson distribution with mean value k, when $N \to \infty$. When $s = 2$,

$$\begin{aligned}
\Pr_k\{i_1, i_2\} = {} & \Pr_k\{i_1\}\binom{N-1-i_1}{i_2} \\
& \cdot \left[1-\frac{(N-2-k)^{[i_1]}}{(N-2)^{[i_1]}}\right]^{i_2}\left[\frac{(N-2-k)^{[i_1]}}{(N-2)^{[i_1]}}\right]^{N-1-i_1-i_2}
\end{aligned}$$

approaches $[k^{i_1}e^{-k}/i_1!][(ki_1)^{i_2}e^{-ki_1}/i_2!]$ as $N \to \infty$. When $s = 3$,

$$\Pr_k\{i_1, i_2, i_3\} = \Pr_k\{i_1, i_2\}\binom{N-1-i_1-i_2}{i_3} \cdot \left[1 - \frac{(N-2-i_1-k)^{[i_2]}}{(N-2-i_1)^{[i_2]}}\right]^{i_3}\left[\frac{(N-2-i_1-k)^{[i_2]}}{(N-2-i_1)^{[i_2]}}\right]^{N-1-i_1-i_2-i_3}$$

approaches $[k^{i_1}e^{-k}/i_1!][(ki_1)^{i_2}e^{-ki_1}/i_2!][(ki_2)^{i_3}e^{-ki_2}/i_3!]$ as $N \to \infty$. More generally, $\Pr_k\{i_1, i_2, \cdots, i_s\}$ will approach

$$\frac{e^{-k(1+i^*)}k^{i_1}(ki_1)^{i_2}(ki_2)^{i_3}\cdots(ki_{s-1})^{i_s}}{i_1!i_2!i_3!\cdots i_s!} = \Pr_k^{\infty}\{i_1, i_2, \cdots, i_s\}$$

as $N \to \infty$, where $i^* = i_1 + i_2 + \cdots + i_{s-1}$. It also can be seen that $b_k(i_1, i_2, \cdots, i_s)$ converges in probability to $\Pr_k^{\infty}\{i_1, i_2, \cdots, i_s\}$, and that

$$\sum_{i_1,i_2,\cdots,i_s}(i_1 + i_2 + \cdots + i_s)b_k(i_1, i_2, \cdots, i_s) = \sum_i ib_{sk}(i)$$

converges in probability to $c_{sk} = k + k^2 + \cdots + k^s$, as $N \to \infty$. This fact is of particular interest since (as observed in the preceding section) c_{sk} is an upper bound for $\sum_i i\,b_{sk}(i)$. Furthermore, the proportion of the population interviewed converges in probability to

$$1 - (1-p)\sum_{i_1,i_2,\cdots,i_s}(1-p)^{i_1+i_2+\cdots+i_s}\Pr_k^{\infty}\{i_1, i_2, \cdots, i_s\}.$$

Thus, for $s = 1$, the proportion interviewed converges in probability to $1 - (1-p)\sum_i(1-p)^i k^i e^{-k}/i! = 1 - (1-p)e^{-kp}$; for $s = 2$, the proportion interviewed converges in probability to

$$1 - (1-p)\sum_{i_1,i_2}(1-p)^{i_1+i_2}\left[\frac{e^{-k(1+i_1)}k^{i_1}(ki_1)^{i_2}}{i_1!i_2!}\right] = 1 - (1-p)e^{-k[1-(1-p)e^{-pk}]}.$$

More generally, the proportion interviewed in an s stage k name snowball sample converges in probability to I_s, where $I_0 = p$ and $I_s = 1 - (1-p)e^{-kI_{s-1}}$. The fact that $I_s \leqq 1 - (1-p)^{c_{sk}+1}$ follows from Theorem 3, or it can be proved directly by induction on s.

6. Binomial Sampling. In [4], the binomial sampling model was used to derive exact formulas for certain sampling procedures and also to obtain approximate formulas for other sampling procedures (where, for example, the sample size-population size ratio is a constant, rather than a random variable). Since binomial sampling does differ from the more usual sampling models, it is of course possible to construct examples where the mathematical results obtained with binomial sampling do not lead to satisfactory approximations for the results obtained with the usual sampling models. Caution must naturally be exercised. Although the statistical problems studied in the present article are different from those presented in [4], we shall observe as in [4] that formulas derived

(earlier herein) for binomial sampling are simpler than related formulas derived with the usual sampling models and that the former formulas lead to good approximations for the latter ones under certain circumstances.

Suppose that a random sample (in the usual sense) of size n is drawn without replacement from a given population of size N (n denotes a fixed positive integer, rather than a random variable, in this section), where each individual in the sample names one other individual, and where there is just one stage beyond the initial sample. In other words, consider the case where $s = k = 1$ with binomial sampling replaced by the more usual sampling model for drawing a sample of fixed size from a finite population. In this case, the random variable y has a hypergeometric distribution with expected value $E\{y\} = n\,2M_{11}/N$ and variance $\sigma_y^2 = n(2M_{11}/N)[1 - (2M_{11}/N)][(N - n)/(N - 1)]$. For binomial sampling (see Section 2), the expected value and the variance of y are $E\{y\} = 2M_{11}p$ and $\sigma_y^2 = 2M_{11}pq$, respectively. Thus, if n/N is set equal to p, the expected value formulas are identical; the variance formula derived for binomial sampling will serve as an approximation to the variance formula derived with the more usual sampling model when $N \to \infty$ (M_{11} is fixed).

The random variables y_1, y_2, and x_{11}, under the usual sampling model, have the expected values

$$E\{y_1\} = n(2M_{11}/N)[(N - n)/(N - 1)],$$

$$E\{y_2\} = n(2M_{11}/N)[(n - 1)/(N - 1)],$$

$$E\{x_{11}\} = n(M_{11}/N)[(2N - n - 1)/(N - 1)],$$

respectively; while for binomial sampling these expected values were $E\{y_1\} = M_{11}2pq$, $E\{y_2\} = M_{11}2p^2$, $E\{x_{11}\} = M_{11}p(2 - p)$, respectively. Again, we note that when n/N is set equal to p ($N \to \infty$), the formulas obtained for binomial sampling will lead to approximations for the formulas derived with the usual sampling model. In addition, the variance formulas for y_1, y_2, and x_{11}, derived with the usual sampling model, will approach the corresponding variance formulas derived for binomial sampling when $n/N = p$ and $N \to \infty$. More generally, the probability distributions of y_1, y_2, and x_{11}, and all of the moments of these statistics, will approach the corresponding probability distributions and moments derived for binomial sampling, when $n/N = p$ and $N \to \infty$. Even more general results concerning the relationship between formulas derived with the usual sampling model and corresponding formulas derived for binomial sampling could be presented (when s and k are any positive integers), but we shall not go into these details.

REFERENCES

[1] David Blackwell, "Conditional expectation and unbiased sequential estimation," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 105–110.

[2] James S. Coleman, "Relational analysis: The study of social organizations with survey methods," *Human Organization*, Vol. 17 (1958–59), pp. 28–36.

[3] W. EDWARDS DEMING AND GERALD J. GLASSER, "On the problem of matching lists by samples," *J. Amer. Stat. Assn.*, Vol. 54 (1959), pp. 403–415.

[4] LEO A. GOODMAN, "On the estimation of the number of classes in a population," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 572–579.

[5] LEO A. GOODMAN, "On the analysis of samples from k lists," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 632–639.

[6] J. L. HODGES, JR. AND E. L. LEHMANN, "Some problems in minimax point estimation," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 182–197.

[7] FREDERICK MOSTELLER, "Questions and Answers," *Amer. Statistician*, Vol. 3 (1949), No. 3, pp. 12–13.

[8] MARTIN A. TROW, "Right wing radicalism and political intolerance: A study of support for McCarthy in a New England town," Ph.D. dissertation, Columbia University, 1957.