GENERALIZATIONS FOR COMPLEX PROBABILITY SAMPLING*

Leslie Kish, Survey Research Center, The University of Michigan

Summary

Most statistical theory is based on assuming simple random distribution of the sample elements. However, that assumption is only justified when either the population distribution is random or the sample design is simple random. Generalizations are needed for broader classes of samples in practical use. A broad class of design is epsem: each population element Y_i, (i = 1, 2, ..., N), has the same selection probability f. A broader class of designs is probability sampling: each population element has a known selection probability fP_i--with f and P_i known and positive. We show that for all epsem $E(y) = fY$, where $y = \Sigma\, y_j$, (j = 1, 2, ..., n) is the sample total. Similarly $E(\Sigma\, y_j/P_j) = fY$ for all probability samples. We also have $E(y/n) = Y/N$ for all epsem, if n is a fixed constant. If n is a variate, with $E(n) = fN$, y/n is a ratio mean; similarly for $E[(\Sigma\, y_j/P_j)/(\Sigma\, 1/P_j)]$. These results also apply to powers $Y_i = X_i^k$, and other functions $Y_i = g(X_i)$ of the population variables, and to vectors $Y_i = g(X_i, \ldots, Z_i)$. For example, we have $E(\Sigma\, y_j^2) = f\Sigma\, Y_i^2$. From this we also obtain $E[\Sigma\, y_j^2/n - (y/n)^2] = \sigma_y^2 - \mathrm{Var}(\bar{y})$ for all epsem (exactly if n is fixed).

1. General Statement

Consider a population of N elements, and a variable Y_i associated with each element (i = 1, 2, ..., N). The population total is $Y = \Sigma\, Y_i$, and the population mean is $\bar{Y} = Y/N$. Probability sampling is a selection method, an operation which insures that the expected appearance in the sample for the ith element is P_i', a known positive number. When sampling without replacement of elements, the ith element appears either once or not at all; then P_i' also represents the probability of selection of the ith element.

It is convenient to represent P_i' by its equivalent fP_i, where f is a known positive constant, a selection factor common to all elements. The P_i are positive values known for each element in the population. Thus probability sampling without replacement of elements is an operation which assigns a known positive probability fP_i to each element. Often the P_i are simple numbers, perhaps integers, associated with the elements.

For example, British households may be selected by applying the f = 1/1000 to the electoral list; the P_i denote the number of electors from the ith household on the electoral list. Or, we may select area segments with 1/1000, then subselect large, medium, and small farms with 1/1, 1/5, and 1/20; here f = 1/20,000, and the P_i take values of 20, 4, and 1. The selection factor f may be applied in different ways; the definitions and their consequences will hold. One may select every Fth listing, where f = 1/F; one may select n listings at random, where n = fN; or one may assign the probability f independently to each of the N listings. The selection is often more complex; multistage and multiphase designs may be used, with stratification and other techniques introduced at each stage. Nevertheless, the overall selection factor f is carefully controlled in probability samples.

A selection is epsem (equal probability selection method) when the selection probability is the same known constant f for all elements; that is, $P_i = 1$ for all i. When sampling elements with replacement, an element may appear more than once in the sample, and we substitute expected appearance for selection probability in the definition.

The number of sample elements is n, and the simple total of a variable present in a sample is $y_c = \sum_j y_j$ in epsem (j = 1, 2, ..., n). The analogous simple total for other probability samples is $y_c = \sum_j y_j/P_j$; here y_j and p_j denote the values Y_i and P_i where the jth sample element is the ith population element. This simple total is not the only possible estimator, but it is the one most frequently used in practice. It might be called the simple or symmetrical estimator, and it has theoretical justifications beyond its simplicity and naturalness. For all probability samples we have directly the simple relation

$$E(y_c) = fY. \qquad (1)$$

This relationship follows directly from the definition of probability sampling, and from the basic relationship for a sum of random variables: $E(\Sigma\, y_j) = \sum_j E(y_j)$, (see section 5). Most statistical theory also assumes independence between the j selections, but that independence is lacking in complex selections. Contrariwise our aim is to find some useful results based on (1) alone, without assuming the independence of observations.

*This research was supported by grant G-7571 of the National Science Foundation. Another version appears in section 2.8 of my book, [Kish, 1965]. I am grateful for the suggestions of Morris H. Hansen, James G. Wendel, Bruce M. Hill, and William Ericson.

This and related results have wide utility. It is usually easy to state for a sample design if it is an epsem or a probability sample. The above relations, and others similar to them, then follow immediately, without having to derive them separately for the great variety of specific designs that are widely utilized for selecting samples. Many of these designs are complex, involving several stages with stratification at each stage, with random or systematic selection; multiphase sampling or controlled selection may also be used. With all their complexity, the desired overall selection probability of elements is maintained at fP_i in probability samples, with operations based typically on tables of random numbers. These operations are readily specified with practical office and field procedures.

Several other relations can be also derived easily from the above. When the sample size $\underline{n}$ is a fixed constant, the sample mean $\bar{y} = y/n$ is an unbiased estimate of $\bar{Y}$, because $n = fN$, and

$$E(y/n) = E(y)/n = fY/fN = \bar{Y}. \qquad (2)$$

Epsem with fixed $\underline{n}$ occurs in many varieties of element sampling, and in the sampling and subsampling of equal clusters. For all of these varieties of sampling the unbiasedness of y/f and y/n follows immediately without having to derive them laboriously and separately for the many types current in theory and practice.

Epsem with variable $\underline{n}$ occurs with unequal sized clusters; also when dealing with subclasses. For these, y/f is still unbiased, but y/n is generally not. With a fixed sampling fraction $\underline{f}$, we have $E(n) = fN$, and $E(y/f)/E(y/f) = Y/N = \bar{Y}$. However, $\bar{y} = y/n$ is a ratio mean (the ratio of two random variables) and $E(\bar{y}) \neq \bar{Y}$ in general. Nevertheless, the ratio mean is widely employed and preferred.

The situation is similar for probability samples which are not epsem; when the selection probabilities are fP_i. Note that $E(\Sigma\, 1/p_j) = fN$; this is but a special case of (1), when $Y_i = 1$ for all $\underline{i}$. The commonly preferred mean is a ratio mean similar to the mean above

$$\bar{y} = \frac{\Sigma\, y_j/p_j}{\Sigma\, 1/p_j} \text{ and } \frac{E(\Sigma\, y_j/p_j)}{E(\Sigma\, 1/p_j)} = \frac{fY}{fN} = \bar{Y}, \qquad (3)$$

but $E(\bar{y}) \neq \bar{Y}$ generally. The ratio mean y/n in epsem with variable $\underline{n}$ may be considered a special case of (3). Ratio means generally are accepted either as adequate approximations or as preferred statistics. My concentration on unbiased estimates and expectations is forced by the limitations of my capabilities and of the development of survey sampling literature. Although unbiasedness is given a prominent role, it is usually abandoned for the most important designs of survey sampling, such as unequal clusters. It is possible to make the analysis conditional on the denominator (n or $\Sigma\, 1/p_j$) found in the sample. This may also be done within the frame of likelihood functions instead of sampling distributions [Birnbaum, 1962; Raiffa, and Schlaifer, 1961].

2. Some Applications

Other important relations, similar to $E(y) = fY$ may be derived for all probability samples. The nature of Y_i was not and need not be restricted. It may represent $Y_i = X_i^2$, or $Y_i = X_i^k$, or $Y_i = X_i^k Z_i^m$. It may represent some function of the element values, or a function of several variables: $Y_i = g(Y_{1i}, Y_{2i}, \ldots, Y_{ki})$; this should be confined, I suppose, to real, finite, single-valued functions of the vector of variables defined on individual elements. Since Y_i may also represent the binomial variable $X_i < K$, where $\underline{K}$ is any fixed constant, it follows that the cumulative distribution function of the sample maintains the proportionality $\underline{f}$ to that of the population.

The importance of the principle can be illustrated by obtaining a much-needed result: estimates of the population variance σ_y^2 from any epsem or other probability sample. From the sample we construct $\underline{n}$, $y = \Sigma\, y_j$, and $\Sigma\, y_j^2$, either self-weighted or properly weighted, ($\Sigma\, 1/p_j$, $\Sigma\, y_j/p_j$, and $\Sigma\, y_j^2/p_j$). Since $E(n) = fN$, and $E(y) = fY$, and $E(\Sigma\, y_j^2) = f\Sigma\, Y_i^2$, we get

$$E(\sum_j^n y_j^2 - \frac{y^2}{n}) = E[(\sum_j^n y_j^2 - \frac{fY^2}{N}) - (\frac{y^2}{n} - \frac{f^2Y^2}{fN})]$$

$$= f(\sum_i^N Y_i^2 - \frac{Y^2}{N}) - E(\frac{y^2}{n} - \frac{f^2Y^2}{fN}).$$

Thus

$$E(nv_y^2) = fN\sigma_y^2 - E(\frac{Y}{n} \cdot y - \frac{f^2Y^2}{fN}), \qquad (4)$$

where $v_y^2 = (\Sigma\, y_j^2/n - \bar{y}^2) = (n-1)s_y^2/n$. We should like also to express the expectation of this element variance in the sample. When $\underline{n}$ is fixed at $\underline{fN}$ for the sample, we have directly that

$$E(v_y^2) = \sigma_y^2 - E\frac{(y^2 - f^2Y^2)}{(fN)^2} = \sigma_y^2 - \mathrm{Var}(\bar{y}),$$

and

$$E[v_y^2 + \mathrm{var}(\bar{y})] = \sigma_y^2, \text{ when } E[\mathrm{var}(\bar{y})] = \mathrm{Var}(\bar{y}). \qquad (5)$$

When $\underline{n}$ is not actually fixed, the analysis may be made conditional on a fixed $\underline{n}$, and arrive at essentially the same result. Furthermore, it can be shown that the bias is bound to be usually small for $\underline{nv_y^2/n}$ considered as a ratio mean.

Hence,

$$E(v_y^2) \doteq \frac{E(nv_y^2)}{E(n)} = \sigma_y^2 - E\left(\frac{Y}{n} \cdot \frac{Y}{fN} - \frac{Y}{N} \cdot \frac{Y}{N}\right)$$

$$\doteq \sigma_y^2 - R_{\frac{Y}{n}\frac{Y}{fN}} \sigma_{\frac{Y}{n}} \sigma_{\frac{Y}{fN}} . \qquad (6)$$

The second term becomes $Var(y/n)$ for fixed n, and it should approach the mean-square-error of $\bar{y}$ when n is not fixed. Generally then $v_y^2 \doteq mse(\bar{y})$ computed from the sample will be a good estimate of σ_y^2 (or S_y^2) among the population elements.

$Var(\bar{y})$ is roughly σ_y^2/n for many designs, and then $s_y^2 = v_y^2 n/(n-1) = (\Sigma y_j^2 - y^2/n)/(n-1)$ may be employed to estimate σ_y^2. For the special case for simple random sampling, when $Var(\bar{y}) = (1-f)S_y^2/n$, $E(s_y^2) = S_y^2$ follows immediately.

This result has great practical utility--and provided my chief motivation for this effort. Survey samplers find it useful to compute the ratio of the actual variance of a complex sample to the variance that a simple random sample based on the same number n of elements would have had. I called [Kish, 1965, 8.2] this the "design effect":

$$deff = \frac{var(\bar{y})}{(1-f)\hat{S}_y^2/n} . \qquad (7)$$

Here then we may estimate $\hat{S}_y^2 = v_y^2 + var(\bar{y})$. Often $s_y^2 = nv_y^2/(n-1)$ will serve well enough. Errors in estimating the second term of the denominator are smaller by the factor $1/n$ than the errors in estimating the numerator.

Methods for estimating the population covariance σ_{yx} between two variables Y_i and X_i are similar to those for estimating the population variance σ_y^2. Hence $[v_{yx} + cov(\bar{y}, \bar{x})]$ should be a good estimate, and s_{yx} often an acceptable approximation.

Then it should follow that the correlation coefficient $R_{yx} = \sigma_{yx}/\sigma_y\sigma_x$ can be well estimated from similar statistics. We may use $[v_{yx} + cov(\bar{y}, \bar{x})]/\sqrt{v_y^2 + var(\bar{y})}\ \sqrt{v_x^2 + var(\bar{x})}$ and often merely $v_{yx}/v_y v_x$, computed from any probability sample. These kinds of analytical statistics are frequently computed from complex probability samples, but without adequate (or any) justification--so far as I know.

3. Some Questions

Justification can be found, I believe, in the symmetry and proportionality of the sampling distribution of any epsem selection to the distribution of population elements. Those symmetries also hold for probability samples properly weighted with the $1/p_j$ values. I hope that others will obtain derivations of needed statistics for probability samples which are now available only for simple random samples.

Simple random sampling, with or without replacement, is often called "random sampling," or merely "sampling" in the statistical literature. "Complex" in the title refers to other types of probability sampling. Other epsem methods represent suppression of most of the $\binom{N}{n}$ combinations equally probably in simple random sampling. Yet they preserve the symmetry of equal selection probability of n/N for each element through equal numbers or expectations of the combinations in which it appears. Analogous properties of probability sampling, when not epsem, are more complex because they require the weights P_i. This symmetry is well known and exploited for simple random sampling, but not for other selection methods.

The symmetries of probability samples resemble those of simple random samples, and the sample moments will be similar to those of the population. What is missing from nonrandom samples is the independence of individual observations of random samples. In complex samples the observations are not independent, and the correlation between sample values may have serious effects. Hence, although $\bar{y}$ is a good estimate of $\bar{Y}$, and v_y^2 of σ_y^2, the variance of $\bar{y}$ may be much greater than v_y^2/n. Similarly, the sampling variabilities of other statistics computed from complex samples may differ greatly from those of simple random selections.

I believe these results are important for three reasons. First, to prove that $E(y_c) = fY$ consumes a half-page or page, for each of several types of the selection designs described in sampling textbooks. Second, these proofs imply that the reader must behave similarly when faced with other, perhaps more complicated designs. For example, a multistage design with stratification and systematic sampling at each stage may look formidable to the unwary; a "controlled selection" even worse. Instead one may say directly that "$E(y_c) = fY$, because y_c was based on a probability sample." Third, results are needed and can be had for other valuable statistics--similar to s_y^2 as estimator of S_y^2.

If the results are so important, and the method is so basic and simple, why has it been missed by mathematical statisticians? Because in their derivations they assume the independence of sample observations needlessly and too early. They ignore other methods of probability sampling, and assume simple random sampling without further thought. For example, instead of deriving a result in terms of $Var(\bar{y})$, they write it in terms of $\sigma_y/\sqrt{n}$. When simple random selection

(independence of selection or observation) is unstated, a necessary condition has been omitted. When that assumption is stated, although it is sufficient, it may be an unnecessarily narrow restriction on the results.

For example, the variance of the function of several random normal variates is stated in terms of the covariance matrix $\sum_i \sum_j \frac{\sigma_{ij}}{n} \frac{\partial \bar{y}_i}{\partial \bar{Y}_i} \frac{\partial \bar{y}_j}{\partial \bar{Y}_j}$ in the best books [Rao, 1952]. This result follows also for complex samples, if stated in terms of $\mathrm{Cov}(\bar{y}_i, \bar{y}_j)$ instead of $\sigma_{ij}/\sqrt{n}$ [see Kendall and Stuart, 1958, 10.6].

To assume that the sample observations are random when the selection is complex, amounts to assuming that the population distribution is random. This is never exactly true, and--more important--it is often far from true.

4. Some Extensions

For an extension of probability sampling we coin the term randomized sampling: when the P_i are known, but the selection constant may be either known f or unknown f'. The selection probabilities of elements may be known (P_i) proportionately to each other, but with the common unknown factor f'. If $P_i = 1$ we have an equal choice selection method (escem), an extension of epsem. Many results for probability (and epsem) sampling can be extended to randomized (and escem) sampling.

The selection constant f' may be unknown in two important classes of problems: "urns" of unknown sizes and hypothetical universes. First, consider a list containing N + B listings; the presence of a number B of "blank" (empty) listings may render the number of elements unknown even if the list total N + B is known. For example, on the British electoral list families can be associated with uniquely defined family heads, but their numbers will remain unknown. Most lists assume this aspect when we analyse a subclass whose size N is unknown. If f is fixed and known for the entire list, it will be also for the subclass, but the subclass size n becomes a random variable. On the other hand, if n is fixed (with N unknown) f' becomes unknown. Of the three quantities involved in $f = n/N$, if two are fixed, so is the third, but a pair may remain not fixed together. Second, f' becomes unknown when inference from the results of a sampled population are extended to a larger, hypothetical, and indefinite universe with an assumed similar distribution of the variable Y_i. When the unknown f' may be considered an unknown constant, and when it must be treated as a random variable, we shall not consider here.

Note the vital fact that subclasses inherit from the entire sample the four broad classes of selection we discussed: epsem, equal chance, probability or randomized sampling. Therefore, the results above for an entire sample also holds for its subclasses. Fixed sample size, however, is not inherited generally by subclasses. But for this exception, simple random sampling is also inherited by the subclasses. On the other hand, other, complex, selection types--such as stratified, systematic selection, or equal sized clusters--are not generally inherited by subclasses.

5. Derivation of $E(y_c) = fY$.

Some colleagues, skeptical about the direct validity of (1) for all probability samples, demand some proof. Several others have pointed to as many distinct proofs, each with some claim for preference and reference. Still others say that no proof is necessary, since the basic rule about the expectation of a sum of random variables established $E(y_c) = fY$ for probability sampling directly after its definition. The following brief derivation may clarify the situation, and distinguish the established truth from other things which only resemble it.

We define C, the set of all possible samples c under a specified sampling plan applied to a population, and assume that C is finite. W_c is the probability of obtaining sample c; hence $\sum_{c \in C} W_c = 1$, summed over the entire distribution.

Let δ_{ic} represent the number of times the ith element appears in sample c; $\delta_{ic} = k$ when the ith element appears k times in sample c. When sampling with replacement, k may be 1, 2, ..., n. When sampling elements without replacement $\delta_{ic} = 1$ or $\delta_{ic} = 0$, if the sample does or does not contain the ith element. Thus δ_{ic} is a random variable associated with the ith population element representing its number of appearances. The expected number of appearances of the ith element is

$$E(\delta_{ic}) = \sum_c W_c \delta_{ic} = fP_i, \text{ for all probability samples} \quad (8)$$

and

$$E(\delta_{ic}) = f, \text{ for all epsem.}$$

Best known of epsem methods is simple random sampling: when each of the $\binom{N}{n}$ possible combinations receive the same selection probability, $W_c = 1/\binom{N}{n}$. Each of the N elements appears in $\binom{N-1}{n-1}$ combinations; thus the selection probability of each element is perceived as $\binom{N-1}{n-1}/\binom{N}{n} = n/N$. There are $\binom{N-1}{n-1}$ combinations which contain the ith element, when $\delta_{ic} = 1$. In the other possible combinations $\delta_{ic} = 0$, and their number is $\binom{N}{n} - \binom{N-1}{n-1} = \left(\frac{N}{n} - 1\right)\binom{N-1}{n-1}$. The

expectation of δ_{ic} is $E(\delta_{ic}) = [\binom{N-1}{n-1} \cdot 1 + (\frac{N}{n} - 1)\binom{N-1}{n-1} \cdot 0]/\binom{N}{n} = \binom{N-1}{n-1}/\binom{N}{n} = n/N$, for all i.

The random variable associated with the ith element is the number δ_{ic} of its appearances in the sample; whereas its value remains constant at Y_i. The contribution of the ith population element to the sample is $\delta_{ic}Y_i$, the product of its constant value Y_i with the random variable δ_{ic} that represents its appearance in the sample. The expected contribution of the ith element is

$$E(\delta_{ic}Y_i) = \sum_c Y_i\delta_{ic}W_c = Y_i \sum_c \delta_{ic}W_c$$

$$= Y_i\, E(\delta_{ic}) = Y_i f P_i, \text{ for all probability sampling,} \quad (9)$$

and

$$E(\delta_{ic}Y_i) = Y_i f \text{ for all epsem.}$$

The sample total represents the sum of contributions for all those population elements which appear in the sample:

$$y_c = \sum_j y_j/p_j = \sum_i (Y_i/P_i)\delta_{ic}. \quad (10)$$

Its expected value is

$$E(y_c) = \sum_i (Y_i/P_i)E(\delta_{ic})$$

$$= \sum_i (Y_i/P_i)(P_i f) = fY. \quad (11)$$

References

[1] Birnbaum, A. [1962], "On the foundations of statistical inferences," JASA, 57, 297-306.

[2] Kendall, M. G. and Stuart, A. [1958], The Advanced Theory of Statistics, Vol. II., London: Griffin and Company.

[3] Kish, L. [1965], Survey Sampling, New York: John Wiley and Sons.

[4] Raiffa, H. and Schlaifer, R. [1961], Applied Statistical Decision Theory, Boston: Graduate School of Business Administration, Harvard University.

[5] Rao, C. R. [1952], Advanced Statistical Methods in Biometric Research, New York: John Wiley and Sons.