ALLOCATION REVISITED

Dhiren Ghosh, Synectics for Mngmt Decisions & Andrew Vogt, Georgetown U.
Dhiren Ghosh, SMD, 3030 Clarendon Blvd, # 305, Arlington, VA 22201

Keywords: configuration, proportional, optimal, pseudoproportional, duality

0. INTRODUCTION It is well known that in stratified and poststratified random sampling the most desirable configuration, the optimal configuration, is given by:

$$f_i^{opt} = \frac{N_i S_i}{\sum_j N_j S_j},$$

where f_i^{opt} is the fraction of sample units in the i-th stratum, N_i the number of population units in the i-th stratum, and S_i the standard deviation of the variable under study on the population in the i-th stratum. Since the normalization in the denominator so that the sample fractions will add to 1 is easily supplied, for simplicity we write: $f_i^{opt} \propto N_i S_i$ to describe the allocation.

Other commonly considered allocations are

$$\text{proportional: } f_i \propto N_i,$$
$$\text{pseudoproportional: } f_i \propto N_i S_i^2,$$
$$\text{equal: } f_i \propto 1.$$

These are special cases of the two families:

$$\text{k-proportional: } f_i \propto N_i^k,$$
$$\text{k-pseudo: } f_i \propto N_i S_i^k$$

where k denotes an arbitrary real number.

In this paper we demonstrate some theoretical relations between these and other sample configurations and the optimal configuration. We propose a measure of distance from the optimal, a notion of dual configuration, and criteria for the stratum fractions to be larger or smaller than the optimal fractions.

Consider a population of size N divided into m strata ($m \geq 2$) of sizes N_1, ..., N_m. A variable X is defined on this population. As usual, a sample of size n is drawn from the population, consisting of n_1 units, ..., n_m units from the respective strata ($n_i \geq 1$ for each i), and the value of X is computed on each unit. Two estimators of the population mean of X, $\overline{X}$, are of interest: the stratified mean $\overline{x}_{st}$ and the poststratified mean $\overline{x}_{pst}$, both given by the same formula:

$$\sum_{j=1}^{m} \frac{N_j}{N} \overline{x}_j.$$

1. DUALITY Two allocations $\{f_i\}$ and $\{g_i\}$ are said to be <u>dual</u> provided

$$f_i g_i \propto (f_i^{opt})^2 \text{ for all } i.$$

If the first allocation $\{f_i\}$ is given, and $N_i S_i$ is known for each i (up to a common scale factor), then $\{g_i\}$ is completely determined. It is given by:

$$g_i = \frac{\frac{(f_i^{opt})^2}{f_i}}{\sum_j \frac{(f_j^{opt})^2}{f_j}}$$

for each i. We sometimes indicate the configuration dual to $\{f_i\}$ by $\{f_i^d\}$.

The optimal configuration is of course proportional to the stratum-by-stratum geometric mean of a configuration and its dual. In this sense it is a configuration intermediate between $\{f_i\}$ and $\{f_i^d\}$.

PROPOSITION 1:

i) The dual of a dual configuration is the original.

ii) The optimal configuration is self-dual.

iii) The variance of the stratified mean for fixed sample size is the same for a configuration and its dual.

iv) The conditional variance of the poststratified mean, where the condition is a given configuration and sample size, is the same for a configuration and its dual.

PROOF: We omit proof of i) and ii). Note that

$$V(\overline{x}_{st}) = V(\sum_j \frac{N_j}{N}\overline{x}_j) = \sum_j \frac{N_j^2}{N^2}(1-\frac{n_j}{N_j})\frac{S_j^2}{n_j}$$

$$= \frac{1}{n}(\sum_j \frac{N_j}{N}S_j)^2(\sum_j \frac{(f_j^{opt})^2}{f_j}) - \frac{1}{N}(\sum_j \frac{N_j}{N}S_j^2)$$

where $N_jS_j = f_j^{opt}(\sum_j N_jS_j)$. The part that depends on the configuration is:

$$\sum_j \frac{(f_j^{opt})^2}{f_j}.$$

Replacing f_j by f_j^d in this, we get:

$$\sum_j \frac{(f_j^{opt})^2}{f_j^d} = (\sum_j \frac{(f_j^{opt})^2}{\frac{(f_j^{opt})^2}{f_j}})(\sum_j \frac{(f_j^{opt})^2}{f_j})$$

$$= \sum_j \frac{(f_j^{opt})^2}{f_j}.$$

Since the conditional variance of $\overline{x}_{pst}$ equals the variance of $\overline{x}_{st}$, this establishes both iii) and iv). Q. E. D

By Proposition 1 dual configurations give the same variances in stratified and poststratified sampling. Thus, for first-order statistical purposes (when higher moments are neglected as they usually are), in the settings where $\overline{x}_{st}$ and $\overline{x}_{pst}$ are unbiased estimators, dual configurations having the same sample size are equivalent.

PROPOSITION 2: If $\{f_i\}$ and $\{f_i^d\}$ are dual allocations, then

$$f_i^{opt} \geq \min(f_i, f_i^d) \text{ for every } i;$$

and there exist at least two strata in which

$$\max(f_i, f_i^d) \geq f_i^{opt} \geq \min(f_i, f_i^d).$$

PROOF: Suppose $f_i^{opt} < \min(f_i, f_i^d)$ for some index i. Then $(f_i^{opt})^2 < f_if_i^d = f_i(\frac{\frac{(f_i^{opt})^2}{f_i}}{\sum_j \frac{(f_j^{opt})^2}{f_j}})$, and hence $\sum_j \frac{(f_j^{opt})^2}{f_j} < 1$. By the Cauchy-Schwarz inequality, however, $\sum_j \frac{(f_j^{opt})^2}{f_j} = (\sum_j \frac{(f_j^{opt})^2}{f_j})(\sum_j f_j) \geq \sum_j(\frac{f_j^{opt}}{\sqrt{f_j}})\sqrt{f_j} = 1$.

If $f_i^{opt} > \max(f_i, f_i^d)$ for all i, then $1 = \sum_i f_i^{opt} > \sum_i f_i = 1$. So at least one f_i^{opt} is intermediate. If only one is, then $1 - f_i^{opt} = \sum_{j\neq i} f_j^{opt} > \sum_{j\neq i} f_j = 1 - f_i$, and $f_i > f_i^{opt}$. But a similar argument shows that $f_i^d > f_i^{opt}$, and this contradicts the argument of the first paragraph. Q. E. D.

2. AN EXAMPLE: THE k-PSEUDO ALLOCATIONS

For a real number k let $\{f_i^k\}$ denote the k-pseudo allocation: $f_i^k \propto N_iS_i^k$. Then the dual of $\{f_i^k\}$ is evidently $\{f_i^{2-k}\}$. In this case we can refine Proposition 2.

PROPOSITION 3: Let $-\infty < k \leq 1$.

i) If i denotes the stratum of largest variance (or one of sufficiently large variance), then

$$f_i^{2-k} \geq f_i^{opt} \geq f_i^k;$$

ii) If i denotes the stratum of smallest variance (or one of sufficiently small variance), then

$$f_i^k \geq f_i^{opt} \geq f_i^{2-k}.$$

PROOF: The statement is clearly true when $k = 1$ since $\{f_i^1\} = \{f_i^{opt}\}$. For real numbers p and q with $q \neq 0$, we can define an average $A_{p,q}$ of the stratum deviations $S_1, ..., S_m$ by:

$$A_{p,q} = \{\frac{\sum_j N_jS_j^pS_j^q}{\sum_j N_jS_j^p}\}^{\frac{1}{q}}.$$

Since $A_{p,q}$ is an average - the q-th root of a convex combination of $S_1^q, ..., S_m^q$, it is always intermediate between the smallest and the largest of the numbers $S_1, ..., S_m$.

The inequalities $f_i^{2-k} \geq f_i^{opt} \geq f_i^k$ can be rewritten as:

$$\frac{N_i S_i^{2-k}}{\sum_j N_j S_j^{2-k}} \geq \frac{N_i S_i}{\sum_j N_j S_j} \geq \frac{N_i S_i^k}{\sum_j N_j S_j^k}.$$

Rearranging terms and taking a $(k-1)$-st root, we find that these inequalities hold if and only if

$$S_i \geq \max(A_{1,1-k}, A_{k,1-k}).$$

A similar argument shows that $f_i^k \geq f_i^{opt} \geq f_i^{2-k}$ if and only if

$$S_i \leq \min(A_{1,1-k}, A_{k,1-k}).$$

Since the averages are intermediate, i) and ii) follow. Q. E. D.

On the other hand, if $\min(A_{1,1-k}, A_{k,1-k}) < S_i < \max(A_{1,1-k}, A_{k,1-k})$, then $f_i^{opt} > \max(f_i^k, f_i^{2-k})$. Examples exist for any fixed $k \neq 1$ and any fixed $m \geq 3$ with $f_i^{opt} > \max(f_i^k, f_i^{2-k})$ for all but two strata.

Note that for $0 \leq k < 1$, $A_{k,1-k} \leq A_{1,1-k}$.

3. THE DISTANCE FUNCTION Two desirable properties in a function $D(\{f_i\})$ that would measure the distance from an allocation to the optimal allocation are that the function be nonnegative and that it have the same value on configurations that yield the same variance. From the proof of Proposition 1, it follows that

$$D(\{f_i\}) = G(\sum_i \frac{(f_i^{opt})^2}{f_i})$$

where G is a function taking $[1, \infty)$ into $[0, \infty)$. It seems natural also to require continuity, strict monotonicity, and surjectivity. Thus candidates for G would include logarithms and functions of the form $G(x) = a(x-1)^b$ where a and b are positive constants. Here we take the simplest choice:

$$D(\{f_i\}) = (\sum_i \frac{(f_i^{opt})^2}{f_i}) - 1.$$

PROPOSITION 4: The function just defined has the following properties:

i) $D(\{f_i\}) \geq 0$ and equals 0 if and only if $\{f_i\} = \{f_i^{opt}\}$;

ii) $D(\{f_i\}) = D(\{g_i\})$ whenever the configurations $\{f_i\}$ and $\{g_i\}$ yield the same variance and, in particular, for dual configurations;

iii) for a given population and sample size, the variance of the stratified mean on the configuration $\{f_i\}$ (or the conditional variance of the poststratified mean) is a strictly increasing and continuous function of $D(\{f_i\})$, given in fact by:

$$V(\overline{x}_{st}) = \frac{1}{n}(\sum_i \frac{N_i}{N} S_i)^2 (1 + D(\{f_i\})) - \frac{1}{N}(\sum_i \frac{N_i}{N} S_i^2);$$

and iv) D is a convex function of $\{f_i\}$, that is,

$$D(\{\lambda f_i + (1-\lambda) g_i\}) \leq \lambda D(\{f_i\}) + (1-\lambda) D(\{g_i\})$$

for arbitrary configurations $\{f_i\}$ and $\{g_i\}$ and $0 \leq \lambda \leq 1$.

PROOF: We prove only iv). Regarded as a function defined on points $(f_1, ..., f_m)$ in the first orthant of R^m, the Hessian matrix of second partial derivatives of $D(\{f_i\})$ is diagonal and positive-definite and this implies convexity. Q. E. D.

An immediate consequence of iv) is the inequality:

$$D(\{\lambda f_i + (1-\lambda) g_i\}) \leq \max(D(\{f_i\}), D(\{g_i\}).$$

If two experts disagree as to which configuration is best, choose a configuration that is a convex combination of the two proposed configurations and be guaranteed to do as well as or better than one of the two!

Indeed, a natural path in allocation space - i.e., in the set $\mathcal{F} = \{(f_1, ..., f_m) : 0 < f_i \text{ for all } i, \sum_j f_j = 1\}$ - is the map:

$$s \longmapsto \{f_i(s)\} = \{(\frac{(f_i^{opt})(\frac{f_i}{f_i^{opt}})^s}{\sum_j (f_j^{opt})(\frac{f_j}{f_j^{opt}})^s}\}.$$

This is a smooth curve with parameter s varying from $-\infty$ to ∞ and satisfies:

$$f_i(1) = f_i, f_i(0) = f_i^{opt}, f_i(-1) = f_i^d$$

and

$$f_i(-s) = f_i^d(s)$$

for all real numbers s and $i = 1, ..., m$, where $\{f_i\}$ is an arbitrarily prescribed initial allocation and $\{f_i^d\}$ is its dual. This trajectory satisfies:

$$f_i(s) \propto (f_i^{opt})(\frac{f_i}{f_i^{opt}})^s,$$

and has the distinctive property that

$$D(\{f_i(s)\}) = M(s)M(-s) - 1$$

where $M(s) = \sum_i f_i^{opt}(\frac{f_i}{f_i^{opt}})^s$. The function $D(\{f_i(s)\})$ by inspection is an even function of s and is strictly increasing for $s \geq 0$. So the trajectory for $-1 < s < 1$ improves the variance over that at $s = \pm 1$. An intermediate path between two configurations can be an improvement over both. One may be able to do better than both experts!

5. THE k-PROPORTIONAL ALLOCATIONS

In general, the stratum variances may not be known to the sampling statistician, even if the stratum fractions $\frac{N_i}{N}$ are known. Nonetheless, rough information about these variances may enable one to choose a configuration that improves over the proportional. The situation presents both opportunities and dangers.

PROPOSITION 5: i) If $N_i \propto S_i^k$ for some $k > 0$, then an allocation $\{f_i\}$ with $f_i \propto N_i^j$ is superior to proportional allocation so long as $1 < j < 1 + \frac{2}{k}$;

ii) If $N_i \propto S_i^{-k}$ for some $k > 0$, then an allocation $\{f_i\}$ with $f_i \propto N_i^j$ is superior to proportional allocation so long as $1 > j > 1 - \frac{2}{k}$.

PROOF: Using the notation of the last section for case i), we consider $f_i(s)$ where $f_i \propto N_i$ and $f_i^{opt} \propto N_iS_i = N_i^{1+\frac{1}{k}}$. Then

$$f_i(s) \propto N_i^{1+\frac{1}{k}}(\frac{N_i}{N_i^{1+\frac{1}{k}}})^s = N_i^{1+\frac{1-s}{k}},$$

and this represents an improvement over proportional ($s = 1$) for all values of s between 1 and -1.

A similar argument applies in case ii). Q. E. D.

PROPOSITION 6: An improvement in variance relative to proportional results from an allocation $f_i \propto N_i^k$ with k somewhat larger than 1 provided

$$\sum_i \frac{N_i}{N}\log(N_i) < \sum_i \frac{N_iS_i^2}{\sum_j N_jS_j^2}\log(N_i).$$

PROOF: Consider the function of k given by $D(\{f_i(k)\})$ where $f_i(k) \propto N_i^k$ is k-proportional. If we differentiate this function with respect to k and evaluate the derivative at $k = 1$, it is easily seen that the derivative is negative if and only if the inequality above is satisfied. In this case increasing k beyond the value 1 will result in a lowering of variance provided the increase is not too large. Q. E. D.

If one has enough partial knowledge of the stratum variances to be confident that the inequality in Proposition 6 holds, one may safely choose a k-proportional allocation with k larger than 1. Likewise, if the reverse inequality is known to hold strictly, one may safely choose a k smaller than 1. However, one may not vary k too far from 1 without a risk of disimprovement. Furthermore if equality or near-equality holds, then any change in k will immediately or shortly take one outside the convex region of reduced variance. In this case one is better off not to deviate from proportional.