International Journal of Zoology Studies
ISSN: 2455-7269; Impact Factor: RJIF 5.14
www.zoologyjournals.com
Volume 2; Issue 2; March 2017; Page No. 43-49

Studies calculating stratified sampling size for eusocial insects

Guedes A.P, Oikawa K.F, Siqueira J.O
Departament of Experimental Psychology, University of São Paulo, São Paulo, São Paulo, Brazil

Abstract
Collecting the whole colony is not always practical or possible to do. Consequently, working with a sample of the colony would be a way to achieve statistical inference. The purpose of this article is to work with a sampling method known as stratified. Thus, both a review and a demonstration of how the stratified sampling for eusocial insects should be performed were conducted.

Keywords: sampling method; eusocial insects; statistical inference

1. Introduction

The purpose of this article is to review a statistical sampling technique known as Stratified Sampling (SS). The stratified sampling is a statistical method confirming that there is a proportional amount of all strata in the same population. This is important to assure that smaller size populations are balanced and keep the critical characteristics [1, 2]. The most proper sample planning for eusocial insects should consider stratified sampling methods, as most colonies are composed of castes and, in accordance with Cochran's studies (1977) , the "stratification may produce a gain in precision in the estimates of characteristics of the whole population. It may possible to divide a heterogeneous population into subpopulations, each of wich is internally homogeneous". In addition, highlights the importance of obtaining the estimates from different strata of the same population [3].

All ant species are social or eusocial, as well as termites [4, 5, 6]. However, not all wasp and bee species are truly social; some of them have solitary habits. An insect might be classified as social or eusocial if overlapping generations, division of labor and cooperative brood care are seen in the colony. The division of labor is associated with the presence of different castes in the colony. There are the worker caste and the reproductive caste, namely, queens and males. Social insects undergo some changes in group formation, but generally all the groups, such as queens, workers and soldiers, the defense group, are found in the colony. Subcastes may be defined by the differentiated morphology or the behavioral polytheism, which might be determined by the age or the individual size factor [7]. Males are not considered as a caste when considering only the biological issue and they are rare, besides having a short life [5, 6]. However, it is noteworthy to mention that, for statistical purposes, males are considered as strata/caste and thus they may participate or not in size calculation, depending on the experiment purpose, and their sampling n will be calculated the same way as the n for the remaining colony. For information purposes, the naked-mole rat (Heterocephalus glaber), an African rodent, is another example of eusocial animals [6]. The individuals of a certain caste or subcaste do not have any difference for statistical purposes, as each one is considered as a stratum. Preferably, the sampling number within an artificial colony (laboratory) should be consistent with the nature, for many reasons, but usually researchers solve this problem by collecting the whole colony when its size is reasonable. When the colony is large, the solution for this problem has been known for more than 200 years. The technique solving this problem is the capture-recapture method. The capture-recapture method aims at estimating the N, the population size, and this is not our objective. However, if estimating N is required, we recommend using this method before applying the stratified sampling method. The capture-recapture method was first used by Laplace in 1783 to estimate the French population [8]. Carl G. J. Pertersen was the first one to use this method in ichthyological studies in order to understand the migration flow in Baltic Sea in 1896. Frederick Lincoln also used this estimation method to assess the size of the wild duck population in North America in 1930 [9]. The capture-recapture method provides the correct estimation on both incidence and prevalence, even though the provided data is incomplete [8]. The reasons for assuring the population representativeness do interfere with the conclusions on the eusocial insect studies. The disproportionate number of a certain caste/subcaste may interfere with the division of labor in the colony [10, 7], and if this is not the purpose of the study, conclusions will have significant bias. The purpose of the stratified sampling is to reach a sample of the sampling unit distribution, represented by castes or subcastes/strata, in order to preserve its representativeness. Thus the number of simulations may be reduced in order to achieve this representation accuracy. The stratified sampling assumes that, when a steady random variable, defined by an actual interval [10], is split in a certain number of separate subintervals which taken together compose the variable domain, for example [2], [4, 6] and [8, 10], the subintervals are designated as strata (castes). Instead of generating n random numbers (according to the variable distribution) on the interval [10], n1, n2 and n3 random numbers would be generated for each subinterval. This statistical idea is relevant for the way colonies are structured.

Size calculating techniques are not applicable only when variables are issues referring to castes/subcastes, thus there is no impediment to take the colonies themselves as the strata. Finally, the proper experimental design is important for the conclusion on the results obtained. The relevance of this study is confirmed and supported by several works developed following the same thinking framework. In the stratified sampling specifically, there are four partitioning techniques used to determine the sampling size n_i, in which $i = 1, \ldots, M$, in each M of the population strata.

To all purposes and intents, the interest level will only rely upon the population mean μ. Additionally, both finite and infinite population circumstances will be taken into account or, in other words, when sampling is conducted without or with replacement, respectively. Finite population means a population in which the number of elements in a group is not that large; in other words, it is countable. The finite population circumstance does not take replacement, therefore, when the sampling unit (ant) is picked for investigation, this experimental unit is no longer part of the population. In this case, we use this kind of sampling for small samples (Bussab and col., 2002; Pestana, 2006) [10, 11]. Infinite population means a population in which the number of elements is too high; in other words, it is uncountable. The infinite population circumstance does take replacement; this means that the sampling unit (ant) picked for evaluation subsequently returns to population after it is drawn (Bussab and col., 2002; Pestana, 2006) [10, 11]. The basic difference between using sampling replacement or not is the possibility of picking or not a member for investigation and then considering the same member again in the investigated sample. In finite sampling, the return of a sampling unit (ant) into the group may interfere with the investigation, as the sample is small and variables are not diluted as in an infinite population.

2. Stratified sampling methods

There are three methods discussed in this work, namely: Equal Partition (or Equipartition), Proportional Partition and eventually Optimal (or Neyman) Partition.

Equipartition is a sampling method in which all subsamples have the same size, that is $n_i = \bar{n}$, for all $i = 1, \ldots, M$. M is the total number of strata in population and sample size is steady $(\bar{n})$.

For Proportional Partition, strata (castes) should be the most homogenous as possible concerning the characteristics relevant for the research (variables strongly correlated to the investigated variable) for the same sample size. The stratified random sampling with proportional partition is more precise, as there is lower variance in estimator. If this is the case, n_i size subsamples are selected in each strata, and they are proportional to N_i population size of each stratum. Concerning the SS size calculation in proportional partition, variance is set as for the Equal Partition.

In Equal Partition, $n_i, i = 1, \ldots, M$ values are steady $(\bar{n})$, while in Proportional Partition they depend on the size of population in the respective N_i strata, through w_i. w_i herein is assumed as the weight of each strata in relation to the population size.

Concerning the optimal (or Neyman) partition, an additional element is added to calculate n_i, namely, the $\mathrm{Var}(\hat{\mu}_e)$, that is, the estimator variance in the stratified population mean, as this method requires the previous knowledge of variances in each stratum.

Essentially, the core problem in Optimal Partition consists of optimizing (that is, minimizing) the size of n_i samples in relation to the variance value established. The solution for this partition technique basically involves Lagrange multiplier method (λ). Just for information, Lagrange multipliers are a method dealing with more than one variable (n1, n2, Nm).

It is noteworthy to observe that, for Optimal Partition, both in SS without or with replacement, n_i values now become dependent on the variability measurement $(S_i \text{ or } \sigma_i)$, standard deviation, and the respective strata, which previously would not happen to Equipartition or Proportional Partition.

Yet following this principle, in both n_i circumstances, it is possible to see that the higher variability in any i stratum $(S_i \text{ or } \sigma_i)$, the higher should be the n_i sampling size for the respective stratum.

After obtaining those values (n_i), we should replace them in their respective variance equations.

2.1 – SS formula applications

Take the following population stratified into three subpopulations $(M = 3)$, composed of 82 individuals, as an example:

Table 1

Stratum	N_i	S_i^2	n_i
1	5	10	
2	31	25	
3	46	20	
	82		N

Each of the 3 population strata will have a sample collected. What should the n size of the sample be, when considering that the variance in this population is $\sigma^2 = 10{,}5$, the maximum acceptable error is $e_A = 1.4$ (mean deviation of the $\overline{X}$ sample in relation to the population mean μ) and the presumed confidence coefficient is $\gamma = 1 - \alpha = 0.95$?

According to the maximum acceptable error criterion, we have:

$$n = z_{\alpha/2}^2 \frac{\sigma^2}{e_A^2}$$

$$n = (1{,}96)^2 \frac{10{,}5}{(1{,}4)^2}$$

$$n = 20.776$$

$$n \approx 21.$$

Based on the size of the total sample $n = 21$, n_1, n_2 and n_3 values are established using the equal partition method, so that $n = n_1 + n_2 + n_3$.

As per the Equal Partition method, every n_i is equal (or steady), that is, $n_i = \bar{n}$. As the sampling is stratified, the total sample (which is already known and equal to 21) should be equally distributed to the 3 strata. Then,

$$n_i = \frac{n}{M}$$

$$n_i = \frac{21}{3}$$

$$n_i = 7.$$

Consequently we have:

Table 2

Stratum	N_i	n_i
1	5	7
2	31	7
3	46	7
Stratum	82	21

The example seems to be improper, but it is fully capable of indicating this technique limitation. This method could be properly used in two circumstances. The first reservation for the use of this method is to obtain $N_i \leq n_i$ in the end of the sampling n calculation. The second issue to be taken into consideration is that this method is more assertive when the experimental unit is

a colony, not a caste/subcaste. This latest reservation though is only an advice, not a key impediment.

Based on the size of the total sample $n = 21$, n_1, n_2 and n_3 values are determined by using the Proportional Partition method, so that $n = n_1 + n_2 + n_3$.

Once proportionality (in comparison with the N size of the population) is the key issue for this partition method, each stratum has a "weight" while determining the n_i sample size of each stratum. These weights are given by:

$$w_1 = \frac{N_1}{N} = \frac{5}{82} = 0.0609.$$
$$w_2 = \frac{N_2}{N} = \frac{31}{82} = 0.3780.$$
$$w_3 = \frac{N_3}{N} = \frac{46}{82} = 0.5609.$$

Thus, by using the proportional partition method, sample sizes n_1, n_2 and n_3 will be given as follows:

$$n_1 = w_1 \cdot n = 0.0609 \cdot 21 = 1.27 \approx 1.$$
$$n_2 = w_2 \cdot n = 0.3780 \cdot 21 = 7.93 \approx 8.$$
$$n_3 = w_3 \cdot n = 0.5609 \cdot 21 = 11.77 \approx 12,$$

and it is consistent, as 1+8+12=21. Therefore,

Stratum	N_i	n_i
1	5	1
2	31	8
3	46	12
	82	21

Based on the size of the total sample $n = 21$, n_1, n_2 and n_3 values are determined by using the Optimal (or Neyman) Partition method, so that $n = n_1 + n_2 + n_3$.

The Optimal (or Neyman) partition requires the (previous) knowledge on the variances of each stratum, that is, the values showed in table provided in the exercise instructions. The referred data is provided in the table below.

Table 3

Strata	N_i	S_i^2	n_i
1	5	10	
2	31	25	
3	46	20	
	82		n

In Optimal partition, n_i values are given by $n_i = \frac{w_i S_i}{\sum_{i=1}^{3} w_i S_i} n$. So, we have to get the standard deviation (S_i), which is obtained by the square root of variance (S_i^2). Example:$S_1 = \sqrt{S_1^2} = \sqrt{10} =$ 3,16.Then,

Table 4

Stratum	N_i	S_i^2	w_i	S_i
1	5	10	0.0609	3.16
2	31	25	0.3780	5.00
3	46	20	0.5609	4.47
	82		1	

In order to calculate n_i, however, it is required to get the $w_i S_i$ column, as well as its sum. Therefore,

Table 5

Stratum	N_i	S_i^2	w_i	S_i	$w_i S_i$
1	5	10	0.0609	3.16	0.1924
2	31	25	0.3780	5.00	1.8900
3	46	20	0.5609	4.47	2.5072
	82		1		4.5896

Then, n_i values will be given by

$$n_1 = \frac{w_1 S_1}{\sum_{i=1}^{3} w_i S_i} n = \frac{0.1924}{4.5896} 21 = 0.88 \approx 1.$$
$$n_2 = \frac{w_2 S_2}{\sum_{i=1}^{3} w_i S_i} n = \frac{1.8900}{4.5896} 21 = 8.64 \approx 9.$$
$$n_3 = \frac{w_3 S_3}{\sum_{i=1}^{3} w_i S_i} n = \frac{2.5072}{4.5896} 21 = 11.47 \approx 11.$$

Stratum	N_i	S_i^2	w_i	S_i	$w_i S_i$	n_i
1	5	10	0,0609	3,16	0,1924	1
2	31	25	0,3780	5,00	1,8900	9
3	46	20	0,5609	4,47	2,5072	11
	82		1		4,5896	21

The Equipartition technique seems to be ideal to address colony-concerning variables, however there is no impediment to calculate the size of castes or subcastes. Variables concerning castes or subcastes have a good response to both Proportional Partition and Optimal Partition and also when colony is its experimental unit.

4. Complements

Mathematic proofs on Equipartition, Proportional Partition and Neyman Partition.

Assuming the idea of building a confidence interval (in relation to the population mean μ) with a confidence coefficient $\gamma = 1 - \alpha$, and also assuming a population with a Normal distribution, we have

$$\mathbb{P}[|\overline{X} - \mu| < e_A] = \gamma.$$

When developing this expression,

$$\mathbb{P}[|\overline{X} - \mu| < e_A] = \mathbb{P}[-e_A < \overline{X} - \mu < e_A].$$

If we divide the three terms by $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, that is, by the standard error provided by the estimator $\overline{X}$ itself, in order to reach the data standardization, we may rewrite it as $Z = \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0.1)$. Where

$$\mathbb{P}\left[\frac{-e_A}{\sigma_{\overline{X}}} < \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} < \frac{e_A}{\sigma_{\overline{X}}}\right] = \mathbb{P}\left[\frac{-e_A}{\sigma_{\overline{X}}} < Z < \frac{e_A}{\sigma_{\overline{X}}}\right],$$

and in both extremities we have the amounts $-z_{\alpha/2} = \frac{-e_A}{\sigma_{\overline{X}}}$ and $z_{\alpha/2} = \frac{e_A}{\sigma_{\overline{X}}}$.

Note that the amount $z_{\alpha/2}$, given by $z_{\alpha/2} = \frac{e_A}{\sigma_{\overline{X}}}$, determines that the maximum acceptable error e_A is given by $z_{\alpha/2} \cdot \sigma_{\overline{X}} = e_A$. Once it aims at obtaining the n value, we should then raise both sides to square, that is

$$z_{\alpha/2}^2 \cdot \mathrm{Var}(\overline{X}) = e_A^2. (2.1)$$

In the event of using replacement (SRSwR),

A known result of the Statistical Inference given by $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$, in the event of independence across consecutive

observations. Thus, the expression (2.1) should be represented by

$$z_{\alpha/2}^2 \frac{\sigma^2}{n} = e_A^2. (2.2)$$

Assuming expression (2.2), and taking into consideration a certain maximum acceptable error e_A, a fixed level of confidence $\gamma = 1 - \alpha$ and an independent sampling, the size of n sample would be represented as

$$z_{\alpha/2}^2 \frac{\sigma^2}{e_A^2} = n$$

or

$$\left(\frac{z_{\alpha/2} \cdot \sigma}{e_A}\right)^2 = n.$$

If sampling is taken without replacement (SRSwoR). If this is the case, a correction factor should be applied into the variance in estimator $\overline{X}$ and its variance would be represented as

$$Var(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right). (2.3)$$

Then, in the event of SRSwoR, taking a certain level of maximum acceptable error e_A and a fixed level of confidence $\gamma = 1 - \alpha$, the size of n sample can be found by replacing expression (2.3) with (2.1), that is,

$$z_{\alpha/2}^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = e_A^2$$
$$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{e_A^2}{z_{\alpha/2}^2}$$
$$\frac{\sigma^2(N-n)}{n(N-1)} = \frac{e_A^2}{z_{\alpha/2}^2}$$
$$z_{\alpha/2}^2\, \sigma^2(N-n) = n(N-1)e_A^2$$
$$Nz_{\alpha/2}^2\, \sigma^2 - nz_{\alpha/2}^2\, \sigma^2 = n(N-1)e_A^2$$
$$Nz_{\alpha/2}^2\, \sigma^2 = n(N-1)e_A^2 + nz_{\alpha/2}^2\, \sigma^2$$
$$Nz_{\alpha/2}^2\, \sigma^2 = n[(N-1)e_A^2 + z_{\alpha/2}^2\, \sigma^2]$$
$$\frac{Nz_{\alpha/2}^2\, \sigma^2}{(N-1)e_A^2 + z_{\alpha/2}^2\, \sigma^2} = n.$$

To sum up, the size of n sample would be

SRSwR (infinite)	SRSwoR (finite)
$\left(\frac{z_{\alpha/2} \cdot \sigma}{e_A}\right)^2 = n$	$\frac{Nz_{\alpha/2}^2\, \sigma^2}{(N-1)e_A^2 + z_{\alpha/2}^2\, \sigma^2} = n$

In the event of infinite sampling, it is not required to know what N (total population size) is.

3.1 Stratified Sampling

The stratified sampling consists of collecting a simple random n_i size sample of each strata.

X_{ij} indicates the j-esim element (sampled) in the i-esim stratum. Concerning the population mean specifically for i-esim stratum, $i = 1, \ldots, M$, this would be represented as

$$\mu_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i},$$

This means that all the elements in this subpopulation (stratum) are taken into account. If the focus was on the whole population mean, the following calculation would be used:

$$\mu = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} X_{ij}}{N},$$

Therefore, we should take into account the sum of all the element values, of every strata. Note that, as the μ_i definition, we could rewrite it as $N_i \cdot \mu_i = \sum_{j=1}^{N_i} X_{ij}$, and, by replacing the general μ expression, we would have

$$\mu = \frac{\sum_{i=1}^{M} N_i \cdot \mu_i}{N}.$$

In the latest expression, the $\frac{N_i}{N}$ ratio can be interpreted as the relative weight of the i-esim stratum in the whole population. When considering $w_i = \frac{N_i}{N}$, we have

$$\mu = \sum_{i=1}^{M} w_i\, \mu_i,$$

Which makes the population mean a certain kind of weighted mean of the population means derived from each stratum, without losing sight of the fact that $\sum_{i=1}^{M} w_i = 1$.

In order to emphasize that this population mean concerns the population mean derived from a stratified population, we will start using μ_e instead of μ, where

$$\mu_e = \sum_{i=1}^{M} w_i\, \mu_i.$$

Thus, when focused on estimating μ_e, or, in other words, obtaining a $\hat{\mu}_e$ value, we will have

$$\hat{\mu}_e = \sum_{i=1}^{M} w_i\, \overline{X}_i,$$

as $\overline{X}$ is a great estimator to get the population mean μ. Furthermore, if we assume that the strata samples are independent of each other, the variance in the estimator for the stratified population mean $\hat{\mu}_e$ will be represented as

$$Var(\hat{\mu}_e) = Var\left(\sum_{i=1}^{M} w_i\, \overline{X}_i\right)$$
$$= \sum_{i=1}^{M} Var(w_i\, \overline{X}_i)$$
$$= \sum_{i=1}^{M} w_i^2\, Var(\overline{X}_i).$$

If this is the case, in the event of a stratified sampling under replacement (SSwR), that is, if $Var(\overline{X}) = \frac{\sigma^2}{n}$, we should have the following:

$$Var(\hat{\mu}_e) = \sum_{i=1}^{M} w_i^2 \frac{\sigma_i^2}{n_i}. (2.4)$$

However, without replacement (SSwoR), that is, $Var(\overline{X}) = \frac{S_i^2}{n_i}\frac{N_i - n_i}{N_i}$, the following should be given:

$$Var(\hat{\mu}_e) = \sum_{i=1}^{M} w_i^2 \frac{S_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right)$$
$$= \sum_{i=1}^{M} w_i^2 \frac{S_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$
$$= \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{n_i} - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}, (2.5)$$

And, in the latest case and by agreement (see Lohr (2009), p. 79), we assumed $\sigma^2 = \frac{N-1}{N} S^2$ and applied the correction factor $\left(\frac{N-n}{N-1}\right)$, which means that

$$\begin{aligned} Var(\bar{X}) &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \\ &= \left(\frac{N-1}{N}\right)\frac{S^2}{n}\left(\frac{N-n}{N-1}\right) \\ &= \frac{S^2}{n}\left(\frac{N-n}{N}\right). \end{aligned}$$

Then, the stratified population mean variance, $Var(\hat{\mu}_e)$, an important result to be used throughout this work, is given by expressions (2.4) and (2.5), depending on SSwR or SSwoR.

Next, we will move on to the analysis of the four partitioning techniques. Initially we will assume that the n value (the size of the SS sample) is unknown, but it may be calculated by following any predetermined criterion (by setting the $Var(\hat{\mu}_e)$ value, the maximum acceptable error or any other criterion).

3.2 Equal Partition (or Equipartition)

By this method, all subsamples should have the same size, that is, $n_i = \bar{n}$, for every $i = 1, \dots, k$. Consider the sample size steady ($\bar{n}$), then the (total) sampling size will be given by the expression

$$n = \bar{n}\, M.$$

Expression on the variance of the stratified population mean estimator (2.5). When using it, we will replaced n_i (provided in the first term denominator) with the value it gets through the equal partition, namely, $n_i = \bar{n}$. However, once $\bar{n} = \frac{n}{M}$, as provided by the equation above, $Var(\hat{\mu}_e)$ expression will be represented as

$$Var_{eq}(\hat{\mu}_e) = \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{n/M} - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}.$$

As $\frac{n}{M}$ is steady (there is no change in i), we might write this as

$$\begin{aligned} Var_{eq}(\hat{\mu}_e) = \frac{M}{n}\sum_{i=1}^{M} w_i^2 S_i^2 \\ - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} \text{ (without replacement).} \end{aligned}$$

But for SSwR, following the expression (2.4), we would have

$$Var_{eq}(\hat{\mu}_e) = \sum_{i=1}^{M} w_i^2 \frac{\sigma_i^2}{n/M},$$

which, by simplifying the expression, we would get to

$$Var_{eq}(\hat{\mu}_e) = \frac{M}{n}\sum_{i=1}^{M} w_i^2\, \sigma_i^2 \text{ (with replacement).}$$

As previously shown, n shall somehow and in both cases depend on $Var_{eq}(\hat{\mu}_e)$. If you set $D^2 = Var_{eq}(\hat{\mu}_e)$, the size calculation, for the SSwoR case, shall be as follows

$$\begin{aligned} D^2 &= \frac{M}{n}\sum_{i=1}^{M} w_i^2 S_i^2 - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} \\ D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} &= \frac{M}{n}\sum_{i=1}^{M} w_i^2 S_i^2 \\ n\left(D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}\right) &= M\sum_{i=1}^{M} w_i^2 S_i^2 \end{aligned}$$

$$n = \frac{M\sum_{i=1}^{M} w_i^2 S_i^2}{D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}} \text{ (without replacement).}$$

However, for SSwR, we would have

$$D^2 = \frac{M}{n}\sum_{i=1}^{M} w_i^2\, \sigma_i^2$$

$$n = \frac{M}{D^2}\sum_{i=1}^{M} w_i^2\, \sigma_i^2 \text{ (with replacement).}$$

3.3 Proportional Partition

Each strata, n_i size subsamples that are proportional to the N_i population size of each strata too. Once $w_i = \frac{N_i}{N}$, we will have

$$n_i = w_i\, n\,, i = 1, \dots, M.$$

In an SSwoR, the $Var_{prop}(\hat{\mu}_e)$ value, properly replacing the n_i value, would be represented by

$$\begin{aligned} Var_{prop}(\hat{\mu}_e) &= \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{w_i\, n} - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} \\ &= \frac{1}{n}\sum_{i=1}^{M} w_i\, S_i^2 - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} \text{ (without replacement).} \end{aligned}$$

For SSwR, we would have

$$\begin{aligned} Var_{prop}(\hat{\mu}_e) &= \sum_{i=1}^{M} w_i^2 \frac{\sigma_i^2}{w_i\, n} \\ &= \frac{1}{n}\sum_{i=1}^{M} w_i\, \sigma_i^2 \text{ (with replacement).} \end{aligned}$$

Regarding the SS size calculation in proportional partitioning, similarly to the equal partitioning, variance is set as $Var_{prop}(\hat{\mu}_e) = D^2$. For the SSwoR case, we have

$$\begin{aligned} D^2 &= \frac{1}{n}\sum_{i=1}^{M} w_i\, S_i^2 - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} \\ D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} &= \frac{1}{n}\sum_{i=1}^{M} w_i\, S_i^2 \\ n\left(D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}\right) &= \sum_{i=1}^{M} w_i\, S_i^2 \end{aligned}$$

$$n = \frac{\sum_{i=1}^{M} w_i\, S_i^2}{D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}} \text{ (without replacement).}$$

Following the same thinking framework, the size calculation in SSwR would be represented by

$$\begin{aligned} D^2 &= \frac{1}{n}\sum_{i=1}^{M} w_i\, \sigma_i^2 \\ n &= \frac{1}{D^2}\sum_{i=1}^{M} w_i\, \sigma_i^2 \text{ (with replacement).} \end{aligned}$$

3.4. Optimal (Neyman) Partition

Regarding the optimal (or Neyman) partition, there is this additional element in order to calculate n_i, namely, the $Var(\hat{\mu}_e)$ value.

Basically, the core issue on optimal partition is to estimate the n_i values minimizing one of their functions (in this case, the $n = \sum_{i=1}^{M} n_i$ sum), as long as $Var(\hat{\mu}_e)$ is set (equally to D^2, for example). The solution for this partitioning technique basically involves Lagrange multiplier method (λ).
Consider the following function to be minimized:

$$\phi_1(n_1, \dots, n_M; \lambda) = n_1 + \cdots + n_M + \lambda[Var(\hat{\mu}_e) - D^2]$$
$$= n_1 + \cdots + n_M + \lambda\left[\sum_{i=1}^{M} \frac{w_i^2 S_i^2}{n_i} - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} - D^2\right].$$

While solving the partial derivative operations, we have:

$$\frac{\partial \phi_1}{\partial n_i} = 1 + \lambda\left[-\frac{w_i^2 S_i^2}{n_i^2}\right], i = 1, \dots, M.$$

As the condition for minimization is $\frac{\partial \phi_1}{\partial n_i} = 0, i = 1, \dots, M$, then

$$1 = \lambda \frac{w_i^2 S_i^2}{n_i^2}, i = 1, \dots, M,$$

Providing

$$n_i^2 = \lambda w_i^2 S_i^2$$
$$n_i = \sqrt{\lambda} w_i S_i. (I)$$

Note that, if we apply the summation to both sides, we have

$$\sum_{i=1}^{M} n_i = \sqrt{\lambda} \sum_{i=1}^{M} w_i S_i$$
$$n = \sqrt{\lambda} \sum_{i=1}^{M} w_i S_i . (II)$$

By dividing (I)/ (II), we will have:

$$\frac{n_i}{n} = \frac{\sqrt{\lambda} w_i S_i}{\sqrt{\lambda} \sum_{i=1}^{M} w_i S_i}$$
$$n_i = \frac{w_i S_i}{\sum_{i=1}^{M} w_i S_i} n . (\text{with replacement})$$

If this is an SSwoR case, the function to be minimized will be given by

$$\phi_2(n_1, \dots, n_M; \lambda) = n_1 + \cdots + n_M + \lambda[Var(\hat{\mu}_e) - D^2]$$
$$= n_1 + \cdots + n_M + \lambda\left[\sum_{i=1}^{M} \frac{w_i^2 \sigma_i^2}{n_i} - D^2\right].$$

Similarly, the partial derivatives will be given by

$$\frac{\partial \phi_2}{\partial n_i} = 1 + \lambda\left[-\frac{w_i^2 \sigma_i^2}{n_i^2}\right], i = 1, \dots, M.$$

By making them equal $\frac{\partial \phi_2}{\partial n_i} = 0, i = 1, \dots, M$, we have

$$1 = \lambda \frac{w_i^2 \sigma_i^2}{n_i^2}, i = 1, \dots, M,$$

then

$$n_i^2 = \lambda w_i^2 \sigma_i^2$$
$$n_i = \sqrt{\lambda} w_i \sigma_i . (I)$$

If, once more, we apply the summation to both sides, we have

$$\sum_{i=1}^{M} n_i = \sqrt{\lambda} \sum_{i=1}^{M} w_i \sigma_i$$
$$n = \sqrt{\lambda} \sum_{i=1}^{M} w_i \sigma_i . (II)$$

By dividing (I)/ (II), for the SSwoR case, we will have:

$$\frac{n_i}{n} = \frac{\sqrt{\lambda} w_i \sigma_i}{\sqrt{\lambda} \sum_{i=1}^{M} w_i \sigma_i}$$
$$n_i = \frac{w_i \sigma_i}{\sum_{i=1}^{M} w_i \sigma_i} n . (\text{with replacement})$$

It is worthy to note that, for optimal partition, both in SSwoR and SSwR, the $n_i = \frac{w_i S_i}{\sum_{i=1}^{M} w_i S_i} n$ and $n_i = \frac{w_i \sigma_i}{\sum_{i=1}^{M} w_i \sigma_i} n$ values now become dependent on a variability measurement (S_i or σ_i) of their respective strata, what once would not happen to both equal and proportional partitions.
Yet following this thinking framework, in both n_i expressions, one can see that the higher the variability in any i strata (S_i or σ_i), the larger should be the n_i sampling size of the respective stratum.
After obtaining those values (n_i), we should replace them in their respective variance equations. For the SSwoR case, we have

$$Var_{ot}(\hat{\mu}_e) = \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{\frac{w_i S_i}{\sum_{i=1}^{M} w_i S_i} n} - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}.$$

In the first term, as n is a constant, it may keep multiplying the sum on the outside of the expression. Yet in the same term, numerator ($w_i^2 S_i^2$) can multiply the inverse of the denominator, that is, $\left(\frac{\sum_{i=1}^{M} w_i S_i}{w_i S_i}\right)$. Then we have the following expression:

$$Var_{ot}(\hat{\mu}_e) = \frac{1}{n} \sum_{i=1}^{M} \frac{w_i^2 S_i^2 \sum_{i=1}^{M} w_i S_i}{w_i S_i} - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}$$
$$= \frac{1}{n}\left(\sum_{i=1}^{M} w_i S_i \sum_{i=1}^{M} w_i S_i\right) - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}$$
$$= \frac{1}{n}\left(\sum_{i=1}^{M} w_i S_i\right)^2 - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}.$$

By setting $Var_{ot}(\hat{\mu}_e) = D^2$, for the SSwoR case, we will have

$$D^2 = \frac{1}{n}\left(\sum_{i=1}^{M} w_i S_i\right)^2 - \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}$$
$$D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i} = \frac{1}{n}\left(\sum_{i=1}^{M} w_i S_i\right)^2$$
$$n\left(D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}\right) = \left(\sum_{i=1}^{M} w_i S_i\right)^2$$
$$n = \frac{\left(\sum_{i=1}^{M} w_i S_i\right)^2}{D^2 + \sum_{i=1}^{M} \frac{w_i^2 S_i^2}{N_i}} . (\text{without replacement})$$

Concerning the size calculation of an SSwR, we would also replace the n_i value in variance equation, that is,

$$Var_{ot}(\hat{\mu}_e) = \sum_{i=1}^{M} \frac{w_i^2 \sigma_i^2}{\frac{w_i \sigma_i}{\sum_{i=1}^{M} w_i \sigma_i} n}.$$

For being a constant (in i), n can multiply the summation, while numerator $(w_i^2\sigma_i^2)$ multiplies the inverse of the denominator $\left(\frac{\sum_{i=1}^{M} w_i\sigma_i}{w_i\sigma_i}\right)$, that is,

$$\begin{aligned}\mathrm{Var}_{ot}(\hat{\mu}_e) &= \frac{1}{n}\sum_{i=1}^{M}\frac{w_i^2\sigma_i^2\sum_{i=1}^{M} w_i\sigma_i}{w_i\sigma_i}\\ &= \frac{1}{n}\sum_{i=1}^{M} w_i\sigma_i\sum_{i=1}^{M} w_i\sigma_i\\ &= \frac{1}{n}\left(\sum_{i=1}^{M} w_i\sigma_i\right)^2.\end{aligned}$$

By setting $\mathrm{Var}_{ot}(\hat{\mu}_e) = D^2$, for the SSwR case, we have

$$D^2 = \frac{1}{n}\left(\sum_{i=1}^{M} w_i\sigma_i\right)^2$$

$n = \frac{1}{D^2}\left(\sum_{i=1}^{M} w_i\sigma_i\right)^2$. (with replacement).

4. Conclusions

The statistical techniques appropriate for eusocial insect studies, and other fields of study working with castes/subgroups, concern the stratified sampling.

The stratified sampling is recommended to decrease spurious associations, strongly driving the validity of the intended assumptions by improving the size of the samples, thus preserving population characteristics. This is because it increases the global estimation precision, maintains population composition according to some basic characteristics and controls the effect of any characteristics on the distribution of the characteristic under investigation (Silva, 2004) [3].

The Equipartition technique proves to be as relevant as the other ones. Its use should be associated with not requiring as much information about the group as the others. However, there are some restrictions associated with the kind of response intended and the $N_i \leq n_i$.

Researchers focused on theoretical issues shall found the mathematic proofs in section Complements. The mathematic proofs are of higher levels of complexity and detailing in order to deepen the subject.

Thanks to financial support CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico.

5. References

1. Moore, David S, George P, McCabe. Introdução à prática da estatística. Trad. FARIAS, AA, 2002.
2. Cochran, William G. Sampling techniques. John Wiley & Sons, 2007.
3. Silva, NN da. Amostragem probabilística: Um Curso Introdutório. 2. ed. 1. reimpr. São Paulo. Edusp, 2004.
4. Grimaldi, David, Michael S. Engel. Evolution of the Insects. Cambridge University Press, 2005.5. Michener, 1974.
5. Wilson EO. *et al.* The insect societies. The insect societies, 1971.
6. Guimarães, Maria Raquel Fellet. Polietismo e expectativa de vida em operárias de Atta laevigata, 2010.
7. Dunn, John, Sérgio Baxter Andreoli. Método de captura e recaptura: nova metodologia para pesquisas epidemiológicas." Revista de Saúde Pública 28.6. 1994, 449-453.
8. Paula, Mde, GO de Almeida, Guedes AC de S. O uso das distribuições Poisson e Gama na estimação do tamanho populacional animal via modelo Bayesiano. Revista Científica da UFPA,[Belém] 7.1. 2009,1-17.
9. Robinson, Gene E. "Regulation of division of labor in insect societies." Annual review of entomology 37.1. 1992, 637-665.
10. BUSSAB WO, MORETIN PA. Métodos uantitativos, Estatística Básica, 5ª edição. São Paulo: Editora Saraiva, 2002.
11. Pestana D. Introdução à probabilidade e estatística. 2ª Edição, Fundação Calouste Gulbenkian, 2006.
12. Lohr S. Sampling: Design and analysis–Second edition. Brooks/Cole Cengage Learning, 2010.