Item ID Number	01700
Author	Carroll, Ray .J.
Corporate Author	
Report/Article Title	Typescript: Report #3, The Effect of Sampling Battalions Rather than Individuals, August 1982
Journal/Book Title	
Year	0000
Month/Day	
Color	□
Number of Images	5
Descripton Notes	

REPORT #3

THE EFFECT OF SAMPLING BATTALIONS RATHER THAN INDIVIDUALS

R.J. Carroll

August 1982

In my letter of August 14 to M.E. LeVois, I raised the possibility that sampling from battalions and then sampling individuals could have different statistical properties from merely taking a simple random sample of individuals. The former method is mentioned by UCLA in their protocol, but it is clear that they intend to use the latter as a basis for analysis. I think it is important to understand the difference between the two and to investigate the effects of this difference. This report is a preliminary analysis of this difference.

While the latter method is called simple random sampling (SRS), the former method might best be called two-stage cluster sampling (TSCS). The two methods are illustrated in Figures #1 and #2.

Suppose there are a total of M battalions and, for simplicity, assume each battalion has m individuals. Let battalion #i (i = 1,2,..., M) have disease rate p_i and suppose the overall disease rate is $\overline{p}$. Suppose we randomly select N battalions and then select n individuals per selected battalion. Then the estimated probability of disease is (for either method) the observed proportion of diseased individuals. If the probabilities of disease are all fairly small (say less than 3% in every battalion), then the variances are approximately

$$\text{Variance(SRS)} \doteq \frac{(1-mn/MN)}{mn}\overline{p}$$

$$\text{Variance(TSCS)} \doteq \frac{(1-n/N)}{mn}\overline{p} + \frac{(1-m/M)}{m}[\frac{1}{M}\sum_{i=1}^{M}(p_i-\overline{p})^2] .$$

Suppose there are M = 500 battalions, of which we sample m = 70. Suppose that each battalion has N = 500 individuals, of whom we sample 100. Further, suppose that the disease probabilities are all less than 3%. Then, to a degree of approximation,

$$\text{Variance (SRS)} \doteq (.0118)^{2}\overline{p}$$

$$\text{Variance(TSCS)} \doteq [.0001149 + .0030715\overline{p}]\overline{p}\ .$$

Table #1 compares the ratio of these two quantities for various values of $\overline{p}$.

The exact numbers in Table #1 are not particularly critical, especially since any real sampling plan will include some stratification. Nonetheless, my calculations indicate the following:

(i) As a general strategy, we should sample as many battalions as possible, with appropriate stratification,

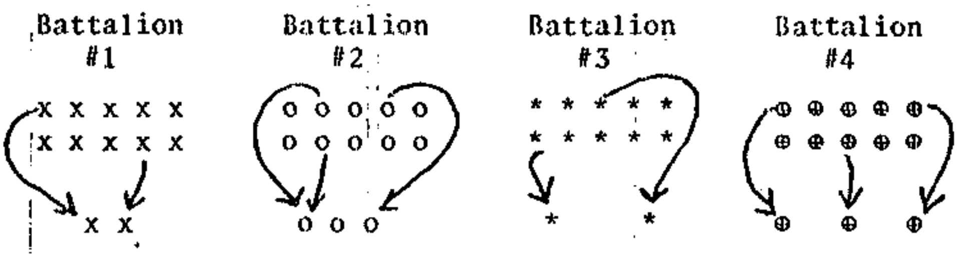
(ii) For larger sample sizes on the order of 6,000 per group, if the event rates are small the effect of TSCS will not be too great,

(iii) If the event rates are large, TSCS will be significantly less efficient than SRS. However, in this instance, we will still have acceptable statistical powers (see Report #2).

This report has not addressed certain problems, such as confounders and misclassification. Also, I have assumed the event rates are fairly homogeneous across battalions; this *might* be a questionable assumption, as it is conceivable that a few battalions had extremely high exposure and event rates. Further study must be guided by the practical nature of the data set.

Figure #1

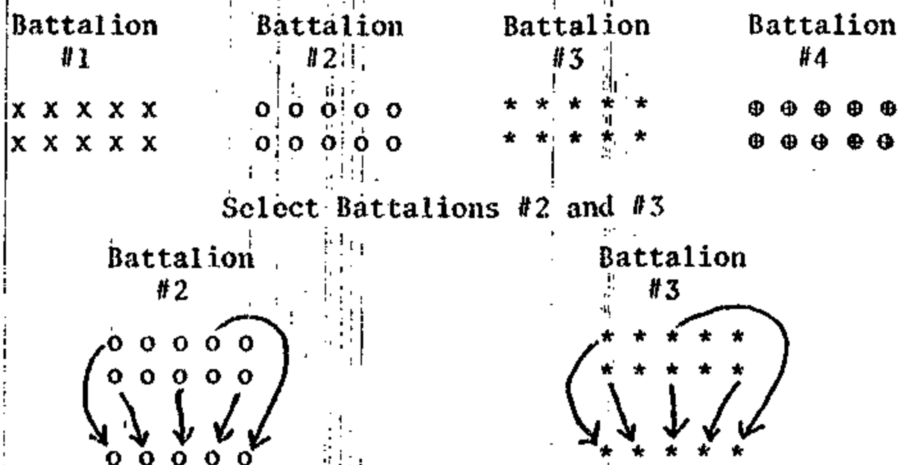
SIMPLE RANDOM SAMPLING OF TEN INDIVIDUALS FROM A TOTAL OF FOUR BATTALIONS.



Note: On average, we will choose someone from every battalion. We could guarantee this by taking a stratified sample.

Figure #2

TWO-STAGE CLUSTER SAMPLING OF TEN INDIVIDUALS, SELECTING AT RANDOM TWO OF FOUR BATTALIONS.



Note: As opposed to simple random or stratified random sampling, in cluster sampling there is no chance of selecting one or more individuals from every battalion.

TABLE #1

$\overline{p}$	Approximate Value of $\frac{\text{Variance (SRS)}}{\text{Variance (TSCS)}}$
.005	.95
.010	1.04
.020	1.26

Appendix, Report #3

$$\text{Variance (SRS)} = \frac{(1-mn/MN)}{mn}\,\bar{p}(1-\bar{p})$$

$$\text{Variance (TSCS)} = \frac{(1-m/M)}{m}S_1^{\,2} + \frac{(1-n/N)}{mn}S_2^{\,2},$$

$$S_1^{\,2} = \frac{1}{n-1}\sum_{i=1}^{M}(p_i-\bar{p})^2$$

$$S_2^{\,2} = \frac{N}{M(N-1)}\sum_{i=1}^{M}p_i(1-p_i)$$

If the event rates are all small,

$$(p_i-\bar{p})^2 \le \frac{1}{4}\,\bar{p}^{2} \quad \text{on average.}$$