Variance Formula for an Estimate from Cluster Sample

From Cochran WG, *Sampling Techniques*

Example 2. A simple random sample of 30 households was drawn from a census taken in 1947 in wards 6 and 7 of the Eastern Health District of Baltimore. The population contains about 15,000 households. In Table 3.5 the persons in each household are classified (a) according to whether they had consulted a doctor in the last 12 months, (b) according to sex.

TABLE 3.5

DATA FOR A SIMPLE RANDOM SAMPLE OF 30 HOUSEHOLDS

Household Number	Number of Persons m_i	Number of: Males a_i	Number of: Females	Doctor Seen in Last Year: Yes a_i	Doctor Seen in Last Year: No
1	5	1	4	5	0
2	6	3	3	0	6
3	3	1	2	2	1
4	3	1	2	3	0
5	2	1	1	0	2
6	3	1	2	0	3
7	3	1	2	0	3
8	3	1	2	0	3
9	4	2	2	0	4
10	4	3	1	0	4
11	3	2	1	0	3
12	2	1	1	0	2
13	7	3	4	0	7
14	4	3	1	4	0
15	3	2	1	1	2
16	5	3	2	2	3
17	4	3	1	0	4
18	4	3	1	0	4
19	3	2	1	1	2
20	3	1	2	3	0
21	4	1	3	2	2
22	3	2	1	0	3
23	3	2	1	0	3
24	1	0	1	0	1
25	2	1	1	2	0
26	4	3	1	2	2
27	3	1	2	0	3
28	4	2	2	2	2
29	2	1	1	0	2
30	4	2	2	1	3
Totals	104	53	51	30	74

Our purpose is to contrast the ratio formula (for variance) with the **inappropriate binomial formula**. Consider first the proportion of people who had consulted a doctor. For the binomial formula, we would take

$$n = 104, \qquad p = \frac{30}{104} = 0.2885$$

Hence

$$v_{bin}(p) = \frac{pq}{n} = \frac{(0.2885) \times (0.7l15)}{104} = \underline{\underline{0.00197}}$$

For the ratio formula, we note that there are 30 clusters and take

$n = 30$

m_i = total number in ith household

a_i = number in ith household who had seen a doctor

$p = 0.2885$, as before

$\bar{m} = 104/30 = 3.4667$

$\Sigma a_i^2 = 86; \sim \Sigma m_i^2 = 404; \Sigma a_i m_i = 113$

The fpc may be ignored. Hence, from (3.26),

$$v(p) = \frac{(86) - 2(0.2885)(113) + (0.2885)^2(404)}{(30)(29)(3.4667)^2} = \underline{\underline{0.00520}}$$

$\frac{0.00520}{0.00197} = 2.63$ = "**Design Effect**":- **104 gives same precision as a SRS of 104/2.63 = 39.5**

The variance given by the ratio method, 0.00520, is <u>much larger than that given by the binomial formula,</u> 0.00197. For various reasons, families differ in the frequency with which their members consult a doctor. For the sample as a whole, the proportion who consult a doctor is only a little more than one in four, but there are several families in which every member has seen a doctor. Similar results would be obtained for any characteristic in which the members of the same family tend to act in the same way.

In estimating the proportion of males in the population, the results are different. By the same type of calculation, we find

binomial formula:	v(p) = 0.00240
ratio formula:	v(p) = 0.00114

Here the binomial formula overestimates the variance. The reason is interesting. Most households are set up as a result of a marriage, hence contain at least one male and one female. Consequently the proportion of males per family varies less from 1/2 than would be expected from the binomial formula. None of the 30 families, except one with only one member, is composed entirely of males, or entirely of females. If the binomial distribution were applicable, with a true P of approximately 1/2, households with all members of the same sex would constitute one quarter of the households of size 3 and one eighth of the households of size 4. This property of the sex ratio has been discussed by Hansen and Hurwitz (1942). Other illustrations of the error committed by improper use of the binomial formula in sociological investigations have been given by Kish (1957).