

Simple Random Sampling

2.1 Simple random sampling without replacement

A design is simple without replacement of fixed size n if and only if, for all s ,

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{if } \#s = n \\ 0 & \text{otherwise,} \end{cases}$$

or

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

We can derive the inclusion probabilities

$$\pi_k = \frac{n}{N}, \quad \text{and} \quad \pi_{k\ell} = \frac{n(n-1)}{N(N-1)}.$$

Finally,

$$\Delta_{k\ell} = \frac{n(N-n)}{N^2} \times \begin{cases} 1 & \text{if } k = \ell \\ \frac{-1}{N-1} & \text{if } k \neq \ell. \end{cases}$$

The Horvitz-Thompson estimator of the total becomes

$$\hat{Y}_\pi = \frac{N}{n} \sum_{k \in S} y_k.$$

That for the mean is written as

$$\hat{\bar{Y}}_\pi = \frac{1}{n} \sum_{k \in S} y_k.$$

The variance of \hat{Y}_π is

$$\text{var}(\hat{Y}_\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n},$$

and its unbiased estimator

$$\widehat{\text{var}}(\widehat{Y}_\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n},$$

where

$$s_y^2 = \frac{1}{n-1} \sum_{k \in S} \left(y_k - \widehat{Y}_\pi\right)^2.$$

The Horvitz-Thompson estimator of the proportion P_D that represents a sub-population D in the total population is

$$p = \frac{n_D}{n},$$

where $n_D = \#(S \cap D)$, and p is the proportion of individuals of D in S . We verify:

$$\text{var}(p) = \left(1 - \frac{n}{N}\right) \frac{P_D(1 - P_D)}{n} \frac{N}{N-1},$$

and we estimate without bias this variance by

$$\widehat{\text{var}}(p) = \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}.$$

2.2 Simple random sampling with replacement

If m units are selected with replacement and with equal probabilities at each trial in the population U , then we define \tilde{y}_i as the value of the variable y for the i -th selected unit in the sample. We can select the same unit many times in the sample. The mean estimator

$$\widehat{Y}_{WR} = \frac{1}{m} \sum_{i=1}^m \tilde{y}_i,$$

is unbiased, and its variance is

$$\text{var}(\widehat{Y}_{WR}) = \frac{\sigma_y^2}{m}.$$

In a simple design with replacement, the sample variance

$$\tilde{s}_y^2 = \frac{1}{m-1} \sum_{i=1}^m (\tilde{y}_i - \widehat{Y}_{WR})^2,$$

estimates σ_y^2 without bias. It is possible however to show that if we are interested in n_S units of sample \tilde{S} for distinct units, then the estimator

$$\widehat{Y}_{DU} = \frac{1}{n_S} \sum_{k \in \tilde{S}} y_k,$$

is unbiased for the mean and has a smaller variance than that of \widehat{Y}_{WR} . Table 2.1 presents a summary of the main results under simple designs.

Table 2.1. Simple designs : summary table

Simple sampling design	Without replacement	With replacement
Sample size	n	m
Mean estimator	$\hat{\bar{Y}} = \frac{1}{n} \sum_{k \in S} y_k$	$\hat{\bar{Y}}_{WR} = \frac{1}{m} \sum_{i=1}^m \tilde{y}_i$
Variance of the mean estimator	$\text{var}(\hat{\bar{Y}}) = \frac{(N-n)}{nN} S_y^2$	$\text{var}(\hat{\bar{Y}}_{WR}) = \frac{\sigma_y^2}{m}$
Expected sample variance	$E(s_y^2) = S_y^2$	$E(\tilde{s}_y^2) = \sigma_y^2$
Variance estimator of the mean estimator	$\widehat{\text{var}}(\hat{\bar{Y}}) = \frac{(N-n)}{nN} s_y^2$	$\widehat{\text{var}}(\hat{\bar{Y}}_{WR}) = \frac{\tilde{s}_y^2}{m}$

EXERCISES**Exercise 2.1** *Cultivated surface area*

We want to estimate the surface area cultivated on the farms of a rural township. Of the $N = 2010$ farms that comprise the township, we select 100 using simple random sampling. We measure y_k , the surface area cultivated on the farm k in hectares, and we find

$$\sum_{k \in S} y_k = 2907 \text{ ha and } \sum_{k \in S} y_k^2 = 154593 \text{ ha}^2.$$

1. Give the value of the standard unbiased estimator of the mean

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k.$$

2. Give a 95 % confidence interval for \bar{Y} .

Solution

In a simple design, the unbiased estimator of \bar{Y} is

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{k \in S} y_k = \frac{2907}{100} = 29.07 \text{ ha.}$$

The estimator of the dispersion S_y^2 is

$$s_y^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{k \in S} y_k^2 - \hat{\bar{Y}}^2 \right) = \frac{100}{99} \left(\frac{154593}{100} - 29.07^2 \right) = 707.945.$$

The sample size n being ‘sufficiently large’, the 95% confidence interval is estimated in hectares as follows:

$$\begin{aligned} \left[\hat{\bar{Y}} \pm 1.96 \sqrt{\frac{N-n}{N} \frac{s_y^2}{n}} \right] &= \left[29.07 \pm 1.96 \sqrt{\frac{2010-100}{2010} \times \frac{707.45}{100}} \right] \\ &= [23.99; 34.15]. \end{aligned}$$

Exercise 2.2 Occupational sickness

We are interested in estimating the proportion of men P affected by an occupational sickness in a business of 1500 workers. In addition, we know that three out of 10 workers are usually affected by this sickness in businesses of the same type. We propose to select a sample by means of a simple random sample.

1. What sample size must be selected so that the total length of a confidence interval with a 0.95 confidence level is less than 0.02 for simple designs with replacement and without replacement ?
2. What should we do if we do not know the proportion of men usually affected by the sickness (for the case of a design without replacement) ?

To avoid confusions in notation, we will use the subscript WR for estimators with replacement, and the subscript WOR for estimators without replacement.

Solution

1. a) Design with replacement.

If the design is of size m , the length of the (estimated) confidence interval at a level $(1 - \alpha)$ for a mean is given by

$$CI(1 - \alpha) = \left[\hat{\bar{Y}} - z_{1-\alpha/2} \sqrt{\frac{\hat{s}_y^2}{m}}, \hat{\bar{Y}} + z_{1-\alpha/2} \sqrt{\frac{\hat{s}_y^2}{m}} \right],$$

where $z_{1-\alpha/2}$ is the quantile of order $1 - \alpha/2$ of a random normal standardised variate. If we denote \hat{P}_{WR} as the estimator of the proportion for the design with replacement, we can write

$$\begin{aligned} CI(1 - \alpha) &= \left[\hat{P}_{WR} - z_{1-\alpha/2} \sqrt{\frac{\hat{P}_{WR}(1 - \hat{P}_{WR})}{m - 1}}, \right. \\ &\quad \left. \hat{P}_{WR} + z_{1-\alpha/2} \sqrt{\frac{\hat{P}_{WR}(1 - \hat{P}_{WR})}{m - 1}} \right]. \end{aligned}$$

Indeed, in this case,

$$\widehat{\text{var}}(\hat{P}_{WR}) = \frac{\hat{P}_{WR}(1 - \hat{P}_{WR})}{(m - 1)}.$$

So that the total length of the confidence interval does not exceed 0.02, it is necessary and sufficient that

$$2z_{1-\alpha/2} \sqrt{\frac{\hat{P}_{WR}(1 - \hat{P}_{WR})}{m - 1}} \leq 0.02.$$

By dividing by two and squaring, we get

$$z_{1-\alpha/2}^2 \frac{\hat{P}_{WR}(1 - \hat{P}_{WR})}{m - 1} \leq 0.0001,$$

which gives

$$m - 1 \geq z_{1-\alpha/2}^2 \frac{\hat{P}_{WR}(1 - \hat{P}_{WR})}{0.0001}.$$

For a 95% confidence interval, and with an estimator of P of 0.3 coming from a source external to the survey, we have $z_{1-\alpha/2} = 1.96$, and

$$m = 1 + 1.96^2 \times \frac{0.3 \times 0.7}{0.0001} = 8068.36.$$

The sample size ($m=8069$) is therefore larger than the population size, which is possible (but not prudent) since the sampling is with replacement.

b) Design without replacement.

If the design is of size n , the length of the (estimated) confidence interval at a level $1 - \alpha$ for a mean is given by

$$\text{CI}(1 - \alpha) = \left[\hat{\bar{Y}} - z_{1-\alpha/2} \sqrt{\frac{N - n}{N} \frac{s_y^2}{n}}, \hat{\bar{Y}} + z_{1-\alpha/2} \sqrt{\frac{N - n}{N} \frac{s_y^2}{n}} \right].$$

For a proportion P and denoting \hat{P}_{WOR} as the estimator of the proportion for the design without replacement, we therefore have

$$\text{CI}(1 - \alpha) = \left[\hat{P}_{WOR} - z_{1-\alpha/2} \sqrt{\frac{N - n}{N} \frac{\hat{P}_{WOR}(1 - \hat{P}_{WOR})}{n - 1}}, \right. \\ \left. \hat{P}_{WOR} + z_{1-\alpha/2} \sqrt{\frac{N - n}{N} \frac{\hat{P}_{WOR}(1 - \hat{P}_{WOR})}{n - 1}} \right].$$

So the total length of the confidence interval does not surpass 0.02, it is necessary and sufficient that

$$2z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\hat{P}_{WOR}(1-\hat{P}_{WOR})}{n-1}} \leq 0.02.$$

By dividing by two and by squaring, we get

$$z_{1-\alpha/2}^2 \frac{N-n}{N} \frac{\hat{P}_{WOR}(1-\hat{P}_{WOR})}{n-1} \leq 0.0001,$$

which gives

$$(n-1) \times 0.0001 - z_{1-\alpha/2}^2 \frac{N-n}{N} \hat{P}_{WOR}(1-\hat{P}_{WOR}) \geq 0,$$

or again

$$\begin{aligned} n & \left\{ 0.0001 + z_{1-\alpha/2}^2 \frac{1}{N} \hat{P}_{WOR}(1-\hat{P}_{WOR}) \right\} \\ & \geq 0.0001 + z_{1-\alpha/2}^2 \hat{P}_{WOR}(1-\hat{P}_{WOR}), \end{aligned}$$

or

$$n \geq \frac{0.0001 + z_{1-\alpha/2}^2 \hat{P}_{WOR}(1-\hat{P}_{WOR})}{\left\{ 0.0001 + z_{1-\alpha/2}^2 \frac{1}{N} \hat{P}_{WOR}(1-\hat{P}_{WOR}) \right\}}.$$

For a 95% confidence interval, and with an *a priori* estimator of P of 0.3 coming from a source external to the survey, we have

$$n \geq \frac{0.0001 + 1.96^2 \times 0.30 \times 0.70}{\left\{ 0.0001 + 1.96^2 \times \frac{1}{1500} \times 0.30 \times 0.70 \right\}} = 1264.98.$$

Here, a sample size of 1265 is sufficient. The obtained approximation justifies the hypothesis of a normal distribution for \hat{P}_{WOR} . The impact of the finite population correction $(1 - n/N)$ can therefore be decisive when the population size is small and the desired accuracy is relatively high.

2. If the proportion of affected workers is not estimated *a priori*, we are placed in the most unfavourable situation, that is, one where the variance is greatest: this leads to a likely excessive size n , but ensures that the length of the confidence interval is not longer than the fixed threshold of 0.02. For the design without replacement, this returns to taking a proportion of 50%. In this case, by adapting the calculations from 1-(b), we find $n \geq 1298$. We thus note that a significant variation in the proportion (from 30% to 50%) involves only a minimal variation in the sample size (from 1265 to 1298).

Exercise 2.3 *Probability of inclusion and design with replacement*

In a simple random design with replacement of fixed size m in a population of size N ,

1. Calculate the probability that an individual k is selected at least once in a sample.
2. Show that

$$\Pr(k \in S) = \frac{m}{N} + O\left(\frac{m^2}{N^2}\right),$$

when m/N is small. Recall that a function $f(n)$ of n is of order of magnitude $g(n)$ (noted $f(n) = O(g(n))$) if and only if $f(n)/g(n)$ is limited, that is to say there exists a quantity M such that, for any $n \in \mathbb{N}$, $|f(n)|/g(n) \leq M$.

3. What are the conclusions ?

Solution

1. We obtain this probability from the complementary event:

$$\Pr(k \in S) = 1 - \Pr(k \notin S) = 1 - \left(1 - \frac{1}{N}\right)^m.$$

2. Then, we derive

$$\begin{aligned} \Pr(k \in S) &= 1 - \left(1 - \frac{1}{N}\right)^m = 1 - \sum_{j=0}^m \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\ &= 1 - \left\{ \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} - \frac{m}{N} + 1 \right\} = \frac{m}{N} - \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\ &= \frac{m}{N} + O\left(\frac{m^2}{N^2}\right). \end{aligned}$$

3. We conclude that if the sampling rate m/N is small, $(m/N)^2$ is negligible in relation to m/N . We then again find the probability of inclusion of a sample without replacement, because the two modes of sampling become indistinguishable.

Exercise 2.4 *Sample size*

What sample size is needed if we choose a simple random sample to find, within two percentage points (at least) and with 95 chances out of 100, the proportion of Parisians that wear glasses ?

Solution

There are two reasonable positions from which to deal with these issues:

- The size of Paris is very large: the sampling rate is therefore negligible.
- Obviously not having any *a priori* information on the population sought after, we are placed in a situation which leads to a maximum sample size (strong ‘precautionary’ stance), having $P = 50\%$. If the reality is different (which is almost certain), we have *in fine* a lesser uncertainty than was fixed at the start (2 percentage points).

We set n in a way so that

$$1.96 \times \sqrt{\frac{P(1-P)}{n}} = 0.02, \text{ with } P = 0.5,$$

hence $n = 2\,401$ people.

Exercise 2.5 *Number of clerics*

We want to estimate the number of clerics in the French population. For that, we choose to select n individuals using a simple random sample. If the true proportion (unknown) of clerics in the population is 0.1% , how many people must be selected to obtain a coefficient of variation CV of 5% ?

Solution

By definition:

$$CV = \frac{\sigma(Np)}{NP} = \frac{\sigma(p)}{P},$$

where P is the true proportion to estimate (0.1% here) and p its unbiased estimator, which is the proportion of clerics in the selected sample. A CV of 5% corresponds to a reasonably ‘average’ accuracy. In fact,

$$\text{var}(p) \approx \frac{P(1-P)}{n} \quad (f \text{ a priori negligible compared to } 1).$$

Therefore,

$$CV = \sqrt{\frac{(1-P)}{nP}} \approx \frac{1}{\sqrt{nP}} = 0.05,$$

which gives

$$n = \frac{1}{0.001} \times \frac{1}{0.05^2} = 400\,000.$$

This large size, impossible in practice to obtain, is a direct result of the scarcity of the sub-population studied.

Exercise 2.6 *Size for proportions*

In a population of 4 000 people, we are interested in two proportions:

P_1 = proportion of individuals owning a dishwasher,
 P_2 = proportion of individuals owning a laptop computer.

According to ‘reliable’ information, we know *a priori* that:

$$45 \% \leq P_1 \leq 65 \%, \quad \text{and} \quad 5 \% \leq P_2 \leq 10 \%.$$

What does the sample size n have to be within the framework of a simple random sample if we want to know *at the same time* P_1 near $\pm 2 \%$ and P_2 near $\pm 1 \%$, with a confidence level of 95 % ?

Solution

We estimate without bias $P_i, (i = 1, 2)$ by the proportion p_i calculated in the sample:

$$\text{var}(p_i) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} P_i(1 - P_i).$$

We want

$$1.96 \times \sqrt{\text{var}(p_1)} \leq 0.02, \quad \text{and} \quad 1.96 \times \sqrt{\text{var}(p_2)} \leq 0.01.$$

In fact ,

$$\max_{45 \% \leq P_1 \leq 65 \%} P_1(1 - P_1) = 0.5(1 - 0.5) = 0.25,$$

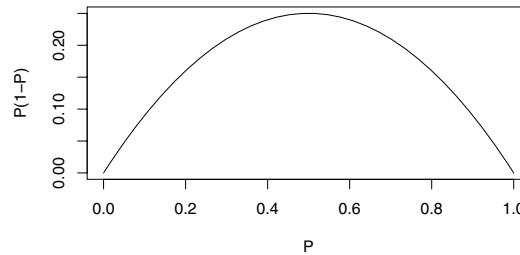
and

$$\max_{5 \% \leq P_2 \leq 10 \%} P_2(1 - P_2) = 0.1(1 - 0.1) = 0.09.$$

The maximum value of $P_i(1 - P_i)$ is 0.25 (see Figure 2.1) and leads to a maximum n (as a security to reach at least the desired accuracy).

It is *jointly* necessary that

Fig. 2.1. Variance according to the proportion: Exercise 2.6



$$\begin{cases} \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} \times 0.25 \leq \left(\frac{0.02}{1.96}\right)^2 \\ \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} \times 0.09 \leq \left(\frac{0.01}{1.96}\right)^2, \end{cases}$$

which implies that

$$\begin{cases} n \geq 1\,500.62 \\ n \geq 1\,854.74. \end{cases}$$

The condition on the accuracy of p_2 being the most demanding, we conclude in choosing: $n = 1\,855$.

Exercise 2.7 *Estimation of the population variance*

Show that

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \bar{Y})^2 = \frac{1}{2N^2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} (y_k - y_\ell)^2. \quad (2.1)$$

Use this equality to (easily) find an unbiased estimator of the population variance S_y^2 in the case of simple random sampling where $S_y^2 = N\sigma_y^2/(N-1)$.

Solution

A first manner of showing this equality is the following:

$$\begin{aligned} \frac{1}{2N^2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} (y_k - y_\ell)^2 &= \frac{1}{2N^2} \sum_{k \in U} \sum_{\ell \in U} (y_k - y_\ell)^2 \\ &= \frac{1}{2N^2} \left(\sum_{k \in U} \sum_{\ell \in U} y_k^2 + \sum_{k \in U} \sum_{\ell \in U} y_\ell^2 - 2 \sum_{k \in U} \sum_{\ell \in U} y_k y_\ell \right) \\ &= \frac{1}{N} \sum_{k \in U} y_k^2 - \frac{1}{N^2} \sum_{k \in U} \sum_{\ell \in U} y_k y_\ell = \frac{1}{N} \sum_{k \in U} y_k^2 - \bar{Y}^2 \\ &= \frac{1}{N} \sum_{k \in U} (y_k - \bar{Y})^2 = \sigma_y^2. \end{aligned}$$

A second manner is:

$$\begin{aligned} \frac{1}{2N^2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} (y_k - y_\ell)^2 &= \frac{1}{2N^2} \sum_{k \in U} \sum_{\ell \in U} (y_k - \bar{Y} - y_\ell + \bar{Y})^2 \\ &= \frac{1}{2N^2} \sum_{k \in U} \sum_{\ell \in U} \{(y_k - \bar{Y})^2 + (y_\ell - \bar{Y})^2 - 2(y_k - \bar{Y})(y_\ell - \bar{Y})\} \\ &= \frac{1}{2N} \sum_{k \in U} (y_k - \bar{Y})^2 + \frac{1}{2N} \sum_{\ell \in U} (y_\ell - \bar{Y})^2 + 0 = \sigma_y^2. \end{aligned}$$

The unbiased estimator of σ_y^2 is

$$\hat{\sigma}_y^2 = \frac{1}{2N^2} \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} \frac{(y_k - y_\ell)^2}{\pi_{k\ell}},$$

where $\pi_{k\ell}$ is the second-order inclusion probability. With a simple design without replacement of fixed sample size,

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)},$$

thus

$$\hat{\sigma}_y^2 = \frac{N(N-1)}{n(n-1)} \frac{1}{2N^2} \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} (y_k - y_\ell)^2.$$

By adapting (2.1) with the sample S (in place of U), we get:

$$\frac{1}{2n^2} \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} (y_k - y_\ell)^2 = \frac{1}{n} \sum_{k \in S} (y_k - \hat{\bar{Y}})^2,$$

where

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{k \in S} y_k.$$

Therefore

$$\hat{\sigma}_y^2 = \frac{(N-1)}{N} \frac{1}{n-1} \sum_{k \in S} (y_k - \hat{\bar{Y}})^2 = \frac{N-1}{N} s_y^2.$$

We get

$$\hat{\sigma}_y^2 = \frac{N-1}{N} s_y^2, \quad \text{and} \quad \hat{S}_y^2 = \frac{N}{N-1} \hat{\sigma}_y^2 = s_y^2.$$

This result is well-known and takes longer to show if we do not use the equality (2.1).

Exercise 2.8 Repeated survey

We consider a population of 10 service-stations and are interested in the price of a litre of high-grade petrol at each station. The prices during two consecutive months, May and June, appears in Table 2.2.

1. We want to estimate the evolution of the average price per litre between May and June. We choose as a parameter the difference in average prices.
Method 1: we sample n stations ($n < 10$) in May and n stations in June, the two samples being completely independent ;
Method 2: we sample n stations in May and we again question these stations in June (*panel* technique).
 Compare the efficiency of the two concurrent methods.

Table 2.2. Price per litre of high-grade petrol: Exercise 2.8

Station	1	2	3	4	5	6	7	8	9	10
May	5.82	5.33	5.76	5.98	6.20	5.89	5.68	5.55	5.69	5.81
June	5.89	5.34	5.92	6.05	6.20	6.00	5.79	5.63	5.78	5.84

- The same question, if we this time want to estimate an average price during the combined May-June period.
- If we are interested in the average price in Question 2, would it not be better to select instead of 10 records twice with Method 1 (10 per month), directly 20 records without worrying about the months (Method 3) ? No calculation is necessary.

N.B.: Question 3 is related to *stratification*.

Solution

- We denote \bar{p}_m as the simple average of the recorded prices among the n stations for month m ($m = \text{May or June}$).

We have:

$$\text{var}(\bar{p}_m) = \frac{1-f}{n} S_m^2,$$

where S_m^2 is the variance of the 10 prices relative to month m .

- Method 1. We estimate without bias the evolution of prices by $\bar{p}_{\text{June}} - \bar{p}_{\text{May}}$ (the two estimators are calculated on two different *a priori* samples) and

$$\text{var}_1(\bar{p}_{\text{June}} - \bar{p}_{\text{May}}) = \frac{1-f}{n} (S_{\text{May}}^2 + S_{\text{June}}^2).$$

Indeed, the covariance is null because the two samples (and therefore the two estimators \bar{p}_{May} and \bar{p}_{June}) are independent.

- Method 2. We have only one sample (the panel). Still, we estimate the evolution of prices without bias by $\bar{p}_{\text{June}} - \bar{p}_{\text{May}}$, and

$$\text{var}_2(\bar{p}_{\text{June}} - \bar{p}_{\text{May}}) = \frac{1-f}{n} (S_{\text{May}}^2 + S_{\text{June}}^2 - 2S_{\text{May, June}}).$$

This time, there is a covariance term, with:

$$\text{cov}(\bar{p}_{\text{May}}, \bar{p}_{\text{June}}) = \frac{1-f}{n} S_{\text{May, June}},$$

where $S_{\text{May, June}}$ represents the true empirical covariance between the 10 records in May and the 10 records in June. We therefore have:

$$\frac{\text{var}_1(\bar{p}_{\text{June}} - \bar{p}_{\text{May}})}{\text{var}_2(\bar{p}_{\text{June}} - \bar{p}_{\text{May}})} = \frac{S_{\text{May}}^2 + S_{\text{June}}^2}{S_{\text{May}}^2 + S_{\text{June}}^2 - 2S_{\text{May, June}}}.$$

After calculating, we find:

$$\left. \begin{aligned} S_{\text{May}}^2 &= 0.05601 \\ S_{\text{June}}^2 &= 0.0564711 \\ S_{\text{May, June}} &= 0.0550289 \end{aligned} \right\} \Rightarrow \frac{\text{var}_1(\bar{p}_{\text{June}} - \bar{p}_{\text{May}})}{\text{var}_2(\bar{p}_{\text{June}} - \bar{p}_{\text{May}})} \approx (6.81)^2.$$

The use of a panel allows for the division of the standard error by 6.81. This enormous gain is due to the very strong correlation between the prices of May and June ($\rho \approx 0.98$): a station where high-grade petrol is expensive in May remains expensive in June compared to other stations (and vice versa). We easily verify this by calculating the true average prices in May (5.77) and June (5.84): if we compare the monthly average prices, only Station 3 changes position between May and June.

2. The average price for the two-month period is estimated without bias, with the two methods, by:

$$\bar{p} = \frac{\bar{p}_{\text{May}} + \bar{p}_{\text{June}}}{2}.$$

- Method 1:

$$\text{var}_1(\bar{p}) = \frac{1}{4} \times \frac{1-f}{n} [S_{\text{May}}^2 + S_{\text{June}}^2].$$

- Method 2:

$$\text{var}_2(\bar{p}) = \frac{1}{4} \times \frac{1-f}{n} [S_{\text{May}}^2 + S_{\text{June}}^2 + 2S_{\text{May, June}}].$$

This time, the covariance is added (due to the '+' sign appearing in \bar{p}).

In conclusion, we have

$$\frac{\text{var}_1(\bar{p})}{\text{var}_2(\bar{p})} = \frac{S_{\text{May}}^2 + S_{\text{June}}^2}{S_{\text{May}}^2 + S_{\text{June}}^2 + 2S_{\text{May, June}}} = (0.71)^2 = 0.50.$$

The use of a panel proves to be ineffective: with equal sample sizes, we lose 29 % of accuracy.

As the variances vary in $1/n$, if we consider that the total cost of a survey is proportional to the sample size, this result amounts to saying that for a given variance, Method 1 allows a saving of 50 % of the budget in comparison to Method 2: this is obviously strongly significant.

3. Method 1 remains the best. Indeed, Method 3 amounts to selecting a simple random sample of size $2n$ in a population of size $2N$, whereas Method 1 amounts to having two strata each of size N and selecting n individuals in each stratum: the latter instead gives a proportional allocation.

In fact, we know that for a fixed total sample ($2n$ here), to estimate a combined average, stratification with proportional allocation is always preferable to simple random sampling.

Exercise 2.9 *Candidates in an election*

In an election, there are two candidates. The day before the election, an opinion poll (simple random sample) is taken among n voters, with n equal to at least 100 voters (the voter population is very large compared to the sample size). The question is to find out the necessary difference in percentage points between the two candidates so that the poll produces the name of the winner (known by census the next day) 95 times out of 100. Perform the numeric application for some values of n .

Hints: Consider that the loser of the election is A and that the percentage of votes he receives on the day of the election is P_A ; the day of the sample, we denote \hat{P}_A as the percentage of votes obtained by this candidate A .

We will convince ourselves of the fact that the problem above posed in ‘common terms’ can be clearly expressed using a statistical point of view: find the critical region so that the probability of declaring A as the winner on the day of the sample (while P_A is in reality less than 50 %) is less than 5 %.

Solution

In adopting the terminology of test theory, we want a ‘critical region’ of the form $]c, +\infty[$, the problem being to find c , with:

$$\Pr[\hat{P}_A > c | P_A < 50\%] \leq 5\%$$

(the event $P_A < 50\%$ is by definition certain; it is presented for reference). Indeed, the rule that will decide on the date of the sample who would win the following day can only be of type ‘ \hat{P} greater than a certain level’. We make the hypothesis that $\hat{P}_A \sim \mathcal{N}(P_A, \sigma_A^2)$, with:

$$\sigma_A^2 = \frac{P_A(1 - P_A)}{n}.$$

This approximation is justified because n is ‘sufficiently large’ ($n \geq 100$). We try to find c such that:

$$\Pr \left[\frac{\hat{P}_A - P_A}{\sigma_A} > \frac{c - P_A}{\sigma_A} \middle| P_A < 50\% \right] \leq 5\%.$$

However, P_A remains unknown. In reality, it is the maximum of these probabilities that must be considered among all P_A possible, meaning all $P_A < 0.5$. Therefore, we try to find c such that:

$$\max_{\{P_A\}} \Pr \left[\mathcal{N}(0.1) > \frac{c - P_A}{\sigma_A} \middle| P_A < 0.5 \right] \leq 0.05.$$

Now, the quantity

$$\frac{c - P_A}{\sqrt{\frac{P_A(1 - P_A)}{n}}}$$

is clearly a decreasing function of P_A (for $P_A < 0.5$). We see that the maximum of the probability is attained for the minimum $(c - P_A)/\sigma_A$, or in other words the maximum P_A (subject to $P_A < 0.5$). Therefore, we have $P_A = 50\%$. We try to find c satisfying:

$$\Pr \left[\mathcal{N}(0, 1) > \frac{c - 0.5}{\sqrt{\frac{0.25}{n}}} \right] \leq 0.05.$$

Consulting a quantile table of the normal distribution shows that it is necessary for:

$$\frac{c - 0.5}{\sqrt{\frac{0.25}{n}}} = 1.65.$$

Conclusion: The critical region is

$$\left\{ \hat{P}_A > \frac{1}{2} + 1.65 \sqrt{\frac{0.25}{n}} \right\}, \quad \text{that is} \quad \left\{ \hat{P}_A > \frac{1}{2} + \frac{1.65}{2\sqrt{n}} \right\}.$$

The difference in percentage points therefore must be at least the following:

$$\hat{P}_A - \hat{P}_B = 2\hat{P}_A - 1 \geq \frac{1.65}{\sqrt{n}}.$$

If the difference in percentage points is *at least* equal to $1.65/\sqrt{n}$, then we have less than a 5 % chance of declaring A the winner on the day of the opinion poll while in reality he will lose on the day of the elections, that is, we have at least a 95 % chance of making the right prediction. Table 2.3 contains several numeric applications. The case $n = 900$ corresponds to the opinion poll sample size traditionally used for elections.

Table 2.3. Numeric applications: Exercise 2.9

n	100	400	900	2000	5000	10000
$1.65/\sqrt{n}$	16.5	8.3	5.5	3.7	2.3	1.7

Exercise 2.10 *Select-reject method*

Select a sample of size 4 in a population of size 10 using a simple random design without replacement with the select-reject method. This method is due to Fan et al. (1962) and is described in detail in Tillé (2001, p. 74). The procedure consists of sequentially reading the frame. At each stage, we decide whether or not to select a unit of observation with the following probability:

$$\frac{\text{number of units remaining to select in the sample}}{\text{number of units remaining to examine in the population}}.$$

Use the following observations of a uniform random variable over $[0, 1]$:

0.375489	0.624004	0.517951	0.0454450	0.632912
0.246090	0.927398	0.32595	0.645951	0.178048

Solution

Noting k as the observation number and j as the number of units already selected at the start of stage k , the algorithm is described in Table 2.4. The sample is composed of units $\{1, 4, 6, 8\}$.

Table 2.4. Select-reject method: Exercise 2.10

k	u_k	j	$\frac{n-j}{N-(k-1)}$	I_k
1	0.375489	0	$4/10 = 0.4000$	1
2	0.624004	1	$3/9 = 0.3333$	0
3	0.517951	1	$3/8 = 0.3750$	0
4	0.045450	1	$3/7 = 0.4286$	1
5	0.632912	2	$2/6 = 0.3333$	0
6	0.246090	2	$2/5 = 0.4000$	1
7	0.927398	3	$1/4 = 0.2500$	0
8	0.325950	3	$1/3 = 0.3333$	1
9	0.645951	4	$0/2 = 0.0000$	0
10	0.178048	4	$0/1 = 0.0000$	0

Exercise 2.11 Sample update method

In selecting a sample according to a simple design without replacement, there exist several algorithms. One method proposed by McLeod and Bellhouse (1983), works in the following manner:

- We select the first n units of the list.
 - We then examine the case of record $(n+1)$. We select unit $n+1$ with a probability $n/(n+1)$. If unit $n+1$ is selected, we remove one unit from the sample that we selected at random and with equal probabilities.
 - For the units k , where $n+1 < k \leq N$, we maintain this rule. Unit k is selected with probability n/k . If unit k is selected, we remove one unit from the sample that we selected at random and with equal probabilities.
1. We denote $\pi_\ell^{(k)}$ as the probability that individual ℓ is in the sample at stage k , where $(\ell \leq k)$, meaning after we have examined the case of record k ($k \geq n$). Show that $\pi_\ell^{(k)} = n/k$. (It can be interesting to proceed in a recursive manner.)
 2. Verify that the final probability of inclusion is indeed that which we obtain for a design with equal probabilities of fixed size.
 3. What is interesting about this method?

Solution

1. • If $k = n$, then $\pi_\ell^{(k)} = 1 = n/n$, for all $\ell \leq n$.
- If $k = n + 1$, then we have directly $\pi_{n+1}^{(n+1)} = n/(n + 1)$. Furthermore, for $\ell < k$,

$$\begin{aligned}
 \pi_\ell^{(n+1)} &= \Pr[\text{unit } \ell \text{ being in the sample at stage } (n + 1)] \\
 &= \Pr[\text{unit } (n + 1) \text{ not being selected at stage } n] \\
 &\quad + \Pr[\text{unit } (n + 1) \text{ being selected at stage } n] \\
 &\quad \times \Pr[\text{unit } \ell \text{ not being removed at stage } n] \\
 &= 1 - \frac{n}{n + 1} + \frac{n}{n + 1} \times \frac{n - 1}{n} = \frac{n}{n + 1}.
 \end{aligned}$$

- If $k > n + 1$, we use a recursive proof. We suppose that, for all $\ell \leq k - 1$,

$$\pi_\ell^{(k-1)} = \frac{n}{k-1}, \quad (2.2)$$

and we are going to show that if (2.2) is true then, for all $\ell \leq k$,

$$\pi_\ell^{(k)} = \frac{n}{k}. \quad (2.3)$$

The initial conditions are confirmed since we have proven (2.3) for $k = n$ and $k = n + 1$. If $\ell = k$, then the algorithm directly gives

$$\pi_k^{(k)} = \frac{n}{k}.$$

- If $\ell < k$, then we calculate in the sample, using Bayes' theorem,

$$\begin{aligned}
 \pi_\ell^k &= \Pr[\text{unit } \ell \text{ being in the sample at stage } k] \\
 &= \Pr[\text{unit } k \text{ not being selected at stage } k] \\
 &\quad \times \Pr[\text{unit } \ell \text{ being in the sample at stage } k - 1] \\
 &\quad + \Pr[\text{unit } k \text{ being selected at stage } k] \\
 &\quad \times \Pr[\text{unit } \ell \text{ being in the sample at stage } k - 1] \\
 &\quad \times \Pr[\text{unit } \ell \text{ not being removed at stage } k] \\
 &= \left(1 - \frac{n}{k}\right) \times \pi_\ell^{(k-1)} + \frac{n}{k} \times \pi_\ell^{(k-1)} \times \frac{n-1}{n} \\
 &= \pi_\ell^{(k-1)} \frac{k-1}{k} = \frac{n}{k}.
 \end{aligned}$$

2. At the end of the algorithm $k = N$ and therefore $\pi_\ell^{(N)} = n/N$, for all $\ell \in U$.

3. What is interesting about this algorithm is that it permits the selection of a sample of fixed size n with equal probabilities without replacement and without having to know *a priori* the size of the population N . For example, we can sample a list that is being filled ‘on the fly’ without needing to wait for everything to be complete before starting the selection procedure. We remark that systematic sampling can be put into place without the population being complete but, in this case, the sample is not necessarily of fixed size.

Exercise 2.12 *Domain estimation*

In a population of size N , we sample n individuals by simple random sampling. We consider a subpopulation D (meaning a ‘domain’) of size N_D , and we denote n_D as the (random) sample size for D . With the selected sample S being decomposable into two parts S_D and $S_{\overline{D}}$, where S_D is the intersection of S and the domain, find the conditional distribution of S_D given n_D (n_D is therefore the cardinality of S_D). What is the practical conclusion?

Solution

$$p(s_D | n_D) = \frac{\Pr(\text{selecting } s_D \text{ and obtaining a size } n_D)}{\Pr(\text{obtaining a size } n_D)}$$

If s_D is indeed of size n_D , the numerator is quite simply $\Pr(\text{selecting } s_D)$. If s_D is not of size n_D , the numerator is null. We are now placed in the first case. In fact:

$$p(s_D) = \sum_{s \supset s_D} p(s) = \frac{\text{Number of } s \text{ containing } s_D}{\binom{N}{n}}.$$

The number of s containing s_D is $\binom{N-N_D}{n-n_D}$ because, in order to go from s_D to s , it is necessary and sufficient to choose $(n-n_D)$ individuals to select outside of the domain D , that is, in a group of size $N - N_D$. Furthermore:

$$\Pr(\text{obtaining a size } n_D) = \sum_{\text{card}(s \cap D) = n_D} p(s) = \frac{\#\{s | \text{card}(s \cap D) = n_D\}}{\binom{N}{n}}.$$

Counting the s such that $\text{card}(s \cap D) = n_D$ brings us back to selecting n_D individuals in D , (there are $\binom{N_D}{n_D}$ possible cases) and $(n - n_D)$ individuals outside of D (there are $\binom{N-N_D}{n-n_D}$ possible cases). Therefore

$$\Pr(\text{obtaining a size } n_D) = \frac{\binom{N_D}{n_D} \binom{N-N_D}{n-n_D}}{\binom{N}{n}},$$

which is a hypergeometric distribution.
Finally, we get:

$$p[s_D | n_D] = \frac{1}{\binom{N_D}{n_D}}.$$

Practical conclusion:

We indeed see that it is the distribution of a simple random sampling of size n_D in a population of size N_D . Thus, all the calculations of bias and variance, if they are conditional on n_D , follow directly from the standard results of simple random sampling, meaning that it is sufficient to continue with the classic formulas in considering that all magnitudes involved are relative to D (we replace n by n_D , N by N_D , S_y^2 by S_{yD}^2 , etc.).

Exercise 2.13 Variance of a domain estimator

Having carried out a simple random sample in a finite population, we are interested in estimating a total Y_0 in a given domain U_0 of the population. We introduce the variable y^* which is

$$y_k^* = \begin{cases} y_k & \text{if } k \in U_0 \\ 0 & \text{otherwise.} \end{cases}$$

1. Throughout this question, the domain size N_0 is unknown and the individuals in the domain are not identifiable *a priori*. The sample size is denoted as n .
 - a) Give the expressions of the unbiased estimator \hat{Y}_0 of the total and its variance.
 - b) Show that

$$(N-1)S_y^{*2} = (N_0-1)S_{y0}^2 + N_0\bar{Y}_0^2 \left(1 - \frac{N_0}{N}\right),$$

where S_{y0}^2 is the population variance of y_k^* (or of y_k) in the domain U_0 and S_y^{*2} is the population variance of y_k^* in the entire population.

- c) Deduce that, when N_0 is very large,

$$\text{var}(\hat{Y}_0) \approx \frac{N^2}{n} \left(1 - \frac{n}{N}\right) (P_0 S_{y0}^2 + P_0 Q_0 \bar{Y}_0^2),$$

where $P_0 = N_0/N$ and $Q_0 = 1 - P_0$.

2. Throughout this question, the domain size N_0 is known, as we henceforth assume that the individuals in the domain are identifiable *a priori* in the survey frame. Recall that the sampling is simple random in the population.
 - a) Give the expressions of the classic unbiased estimator $\hat{\hat{Y}}_0$ of the total and its conditional variance given n_0 . We denote n_0 as the (random) sample size of individuals in the domain U_0 , and we consider that n is sufficiently large so that the probability of obtaining a null n_0 is negligible.

- b) We want to compare the performances of \widehat{Y}_0 and \widehat{Y}_0 . For that, we set $n_0 = nP_0$, and we use this value in the expression $\text{var}(\widehat{Y}_0|n_0)$. Justify this manner of proceeding. Deduce that

$$\text{var}(\widehat{Y}_0|n_0) \approx \text{var}(\widehat{Y}_0) \approx \frac{N^2}{n} \left(1 - \frac{n}{N}\right) P_0 S_{y0}^2.$$

- c) Show that these approximations lead to

$$\frac{\text{var}(\widehat{Y}_0)}{\text{var}(\widehat{Y}_0)} \approx \frac{C_0^2}{C_0^2 + Q_0},$$

where $C_0 = S_{y0}/\overline{Y}_0$ is the coefficient of variation of y_k in the domain U_0 . What do you conclude?

3. In a population of given individuals, we wish to estimate the total number of men in the socio-professional category ‘employees’. We never have at our disposal any information relating to gender except, obviously, in the sample.
- Suppose that we do not know the total number of employees in the population. In what way is this question related to the previous problem (in particular, specify the variable y^* that was used) ?
 - What is the relative gain in accuracy obtained when we suddenly have at our disposal the information ‘total number of employees in the population’ ?
 - How can we estimate this gain? What problem(s) do we face?

Solution

1. a) The estimator is given by

$$\widehat{Y}_0 = N\widehat{\overline{Y}}^* \text{ where } \widehat{\overline{Y}}^* = \frac{1}{n} \sum_{k \in S} y_k^*.$$

We get

$$E(\widehat{Y}_0) = N\overline{Y}^* = \sum_{k \in U} y_k^* = \sum_{k \in U_0} y_k = Y_0$$

(the estimator \widehat{Y}_0 is therefore unbiased), and

$$\text{var}[\widehat{Y}_0] = \text{var}[N\widehat{\overline{Y}}^*] = N^2 \frac{N-n}{Nn} S_y^{*2},$$

where S_y^{*2} is the population variance (unknown) of y_k^* .

b) We have

$$\begin{aligned}
& (N-1)S_y^{*2} \\
&= \sum_{k \in U} (y_k^* - \bar{Y}^*)^2 \\
&= \sum_{k \in U} (y_k^* - \bar{Y}_0 + \bar{Y}_0 - \bar{Y}^*)^2 \\
&= \sum_{k \in U} (y_k^* - \bar{Y}_0)^2 + N(\bar{Y}_0 - \bar{Y}^*)^2 + 2(\bar{Y}_0 - \bar{Y}^*) \sum_{k \in U} (y_k^* - \bar{Y}_0) \\
&= \sum_{k \in U_0} (y_k^* - \bar{Y}_0)^2 + \sum_{k \in U \setminus U_0} (y_k^* - \bar{Y}_0)^2 + N(\bar{Y}_0 - \bar{Y}^*)^2 \\
&\quad + 2(\bar{Y}_0 - \bar{Y}^*)N(\bar{Y}^* - \bar{Y}_0) \\
&= (N_0 - 1)S_{y0}^2 + (N - N_0)\bar{Y}_0^2 - N(\bar{Y}_0 - \bar{Y}^*)^2.
\end{aligned}$$

In fact

$$N(\bar{Y}_0 - \bar{Y}^*)^2 = N \left(\bar{Y}_0 - \bar{Y}_0 \frac{N_0}{N} \right)^2 = N\bar{Y}_0^2 \left(1 - \frac{N_0}{N} \right)^2,$$

which gives

$$(N-1)S_y^{*2} = (N_0-1)S_{y0}^2 + N_0 \left(1 - \frac{N_0}{N} \right) \bar{Y}_0^2.$$

c) If N_0 is very large, then $N_0 \approx (N_0 - 1)$ and $N \approx (N - 1)$:

$$\begin{aligned}
\text{var}(\hat{Y}_0) &\approx N \frac{N-n}{Nn} \left[(N_0-1)S_{y0}^2 + N_0 \left(1 - \frac{N_0}{N} \right) \bar{Y}_0^2 \right] \\
&\approx N^2 \frac{N-n}{Nn} \left(P_0 S_{y0}^2 + P_0 Q_0 \bar{Y}_0^2 \right).
\end{aligned}$$

2. a) We have

$$\hat{\hat{Y}}_0 = N_0 \hat{\hat{Y}}_0,$$

where

$$\begin{aligned}
\hat{\hat{Y}}_0 &= \frac{1}{n_0} \sum_{k \in U_0 \cap S} y_k, \\
n_0 &= \#(U_0 \cap S),
\end{aligned}$$

and

$$\text{var}(\hat{\hat{Y}}_0 | n_0) = N_0^2 \frac{N_0 - n_0}{N_0 n_0} S_{y0}^2.$$

Indeed, in this conditional approach, everything happens as if we had completed a simple random survey of n_0 individuals in U_0 (see Exercise 2.12).

- b) Since n_0 follows a hypergeometric distribution, we have $E(n_0) = nP_0$. The value n_0 does not appear in $\text{var}(\hat{Y}_0)$: to compare similar expressions, it is thus legitimate to substitute $E(n_0)$ with n_0 , which is random. We thus assimilate $\text{var}(\hat{Y}_0|n_0)$ to $\text{var}(\hat{Y}_0)$. Since $N_0 = NP_0$, we get

$$\text{var}(\hat{Y}_0) \approx P_0^2 N^2 \frac{NP_0 - nP_0}{NP_0 n P_0} S_{y0}^2 = P_0 N^2 \frac{N - n}{Nn} S_{y0}^2.$$

Note that we would reach the same expression by starting from the unconditional variance $\text{var}(\hat{Y}_0)$ and by replacing, in the first approximation, the term $E(1/n_0)$ with $1/E(n_0)$.

- c) The relationship between the two variances is:

$$\frac{\text{var}(\hat{Y}_0)}{\text{var}(\hat{Y}_0)} \approx \frac{P_0 S_{y0}^2}{P_0 S_{y0}^2 + P_0 Q_0 \bar{Y}_0^2} = \frac{C_0^2}{C_0^2 + Q_0} < 1.$$

We conclude that the knowledge of N_0 permits having a more efficient estimator. The ‘gain’ is all the more important when C_0 is small, meaning that the domain groups similar individuals (according to y_k), and/or that Q_0 is large, or in other words that the domain is of small size.

3. a) We initially define for the entire population the variable

$$y_k = \begin{cases} 1 & \text{if } k \text{ is male} \\ 0 & \text{otherwise.} \end{cases}$$

Being interested in the domain U_0 of the employees, we will define y_k^* as previously, which comes back to writing:

$$y_k^* = \begin{cases} y_k & \text{if } k \text{ is an employee} \\ 0 & \text{otherwise,} \end{cases}$$

that is to say:

$$y_k^* = \begin{cases} 1 & \text{if } k \text{ is male and an employee} \\ 0 & \text{if } k \text{ is not an employee or not male.} \end{cases}$$

Then, $E(\hat{Y}_0) = N_{h0}$ is the number of male employees in the population.

- b) N_0 is the total number of employees (male + female) henceforth known. The domain U_0 is then defined by the group of employees (male and female). The variable y_k being defined as above, the relative gain from one method to another is

$$\frac{\text{var}(\widehat{\widehat{Y}}_0)}{\text{var}(\widehat{Y}_0)} = \frac{C_0^2}{C_0^2 + Q_0},$$

with

$$C_0 = \frac{S_{y0}}{\overline{Y}_0},$$

and

$$\overline{Y}_0 = \frac{N_{h0}}{N_0} = P_0^h,$$

which is the proportion of men among the employees. As

$$S_{y0}^2 \approx P_0^h(1 - P_0^h),$$

we have

$$C_0^2 = \frac{1 - P_0^h}{P_0^h},$$

and $Q_0 = 1 - P_0$, the proportion of non-employees in the total population (and not only in the domain).

- c) We can estimate without bias (or nearly, because n_0 can be null with a negligible probability) P_0^h by n_{h0}/n_0 and P_0 by n_0/n . However, the gain is a non-linear function of P_0^h and P_0 . The estimator of the gain is therefore biased and the estimation of the associated variance has to rely on a linearisation technique if n is large.

Exercise 2.14 Complementary sampling

Let U be a population of size N . We define the following sampling distribution: we first select a sample S_1 according to a simple design without replacement of fixed size n_1 .

1. We then select a sample S_2 in U outside of S_1 according to a simple random design without replacement of fixed size n_2 . The final sample S consists of S_1 and S_2 . Give the sampling distribution of S . What is interesting about this result?
2. We then select a sample S_3 from S_1 , according to a simple random design without replacement of fixed size n_3 where $(n_3 < n_1)$. Give the sampling distribution of S_3 (in relation to U). What is interesting about this result?
3. Using again the framework from Question 1, we define the estimator of \overline{Y} by:

$$\widehat{\overline{Y}}_\theta = \theta \widehat{\overline{Y}}_1 + (1 - \theta) \widehat{\overline{Y}}_2,$$

with $0 < \theta < 1$,

$$\widehat{Y}_1 = \frac{1}{n_1} \sum_{k \in S_1} y_k \text{ and } \widehat{Y}_2 = \frac{1}{n_2} \sum_{k \in S_2} y_k.$$

Show that, for any θ , \widehat{Y}_θ estimates \bar{Y} without bias.

4. Give the optimal estimator (as θ) in the class of estimators of the form \widehat{Y}_θ .

Solution

1. We have of course $S_1 \subset S, S_2 \subset S$, and $S_1 \cap S_2 = \emptyset$. Therefore, for s of size $n = n_1 + n_2$, we have ($\#S$ indicates the size of the sample S)

$$\begin{aligned} \Pr(S = s) &= \sum_{s_1 \subset s | \#s_1 = n_1} \Pr(S_1 = s_1) \Pr(S_2 = s \setminus s_1 | S_1 = s_1) \\ &= \binom{n_1 + n_2}{n_1} \binom{N}{n_1}^{-1} \binom{N - n_1}{n_2}^{-1} \\ &= \frac{(n_1 + n_2)!}{n_1! n_2!} \times \frac{n_1! (N - n_1)!}{N!} \times \frac{n_2! (N - n_1 - n_2)!}{(N - n_1)!} \\ &= \frac{(n_1 + n_2)! (N - n_1 - n_2)!}{N!} = \binom{N}{n}^{-1}. \end{aligned}$$

The sampling of $S_1 \cup S_2$ is therefore carried out according to a simple random design of fixed size $n = n_1 + n_2$. If we want to increase the sample size already selected using a simple design (for example, to increase the accuracy of an estimator, or because we notice a lower response rate than expected), it is sufficient to reselect a sample according to a simple design among the units that were not selected at the time of the first sampling.

2. The probability of selecting s_3 is calculated as follows using the conditional probabilities.

$$\begin{aligned} \Pr(S_3 = s_3) &= \sum_{s_1 | s_3 \subset s_1} \Pr(S_1 = s_1) \Pr(S_3 = s_3 | S_1 = s_1) \\ &= \binom{N - n_3}{n_1 - n_3} \binom{N}{n_1}^{-1} \binom{n_1}{n_3}^{-1} \\ &= \frac{(N - n_3)!}{(n_1 - n_3)! (N - n_1)!} \times \frac{n_1! (N - n_1)!}{N!} \times \frac{n_3! (n_1 - n_3)!}{n_1!} \\ &= \binom{N}{n_3}^{-1}. \end{aligned}$$

Here once again, we find the distribution characterising the simple random sampling of size n_3 in a population of size N . In practice, to ‘calibrate’ a sample, this property can be used to compete with that shown in 1. We

use *a priori* the sample s_3 , but if its size proves to be insufficient, we call upon s_1 in its group. If we iterate the process, we can set up a group of nested samples, all coming from simple random sampling and using first of all the smallest and then eventually the others as reserve samples, and in relation to the needs as dictated by the field.

3. *Method 1:*

$$E(\widehat{Y}_\theta) = \theta E(\widehat{Y}_1) + (1 - \theta)E(\widehat{Y}_2).$$

The conditional expectation $E(\widehat{Y}_2|S_1)$ is the expectation of a mean in a simple random sample without replacement of fixed size from the population $U \setminus S_1$, which is therefore the true mean of this population, being:

$$E(\widehat{Y}_2|S_1) = \frac{1}{N - n_1} \sum_{U \setminus S_1} y_k = \frac{N\bar{Y} - n_1\widehat{Y}_1}{N - n_1},$$

and therefore

$$E(\widehat{Y}_2) = EE(\widehat{Y}_2|S_1) = \frac{N\bar{Y} - n_1E[\widehat{Y}_1]}{N - n_1} = \frac{N\bar{Y} - n_1\bar{Y}}{N - n_1} = \bar{Y}.$$

Thus

$$E(\widehat{Y}_\theta) = \theta E(\widehat{Y}_1) + (1 - \theta)E(\widehat{Y}_2) = \theta\bar{Y} + (1 - \theta)\bar{Y} = \bar{Y}.$$

Method 2:

We can also use the results from 1., which avoids conditional expectations. Indeed, we can express the simple mean on S of the form

$$\widehat{Y} = \frac{n_1\widehat{Y}_1 + n_2\widehat{Y}_2}{n},$$

thus

$$\widehat{Y}_2 = \frac{n\widehat{Y} - n_1\widehat{Y}_1}{n_2}.$$

We therefore get

$$\widehat{Y}_\theta = \theta\widehat{Y}_1 + (1 - \theta)\frac{n\widehat{Y} - n_1\widehat{Y}_1}{n_2} = \left[\theta - \frac{n_1}{n_2}(1 - \theta) \right] \widehat{Y}_1 + (1 - \theta)\frac{n\widehat{Y}}{n_2}.$$

Since $E(\widehat{Y}_1) = E(\widehat{Y}) = \bar{Y}$,

$$\begin{aligned} E(\widehat{Y}_\theta) &= \left[\theta - \frac{n_1}{n_2}(1 - \theta) \right] E[\widehat{Y}_1] + (1 - \theta)\frac{nE[\widehat{Y}]}{n_2} \\ &= \left[\theta - \frac{n_1}{n_2}(1 - \theta) \right] \bar{Y} + (1 - \theta)\frac{n\bar{Y}}{n_2} \\ &= \bar{Y}. \end{aligned}$$

4. Since \widehat{Y}_θ is unbiased, we find θ that minimises the variance of \widehat{Y}_θ .

Method 1:

$$\text{var}(\widehat{Y}_\theta) = \theta^2 \text{var}(\widehat{Y}_1) + (1 - \theta)^2 \text{var}(\widehat{Y}_2) + 2\theta(1 - \theta) \text{cov}(\widehat{Y}_1, \widehat{Y}_2),$$

$$\text{var}(\widehat{Y}_2) = \text{var}E(\widehat{Y}_2|S_1) + E\text{var}(\widehat{Y}_2|S_1),$$

now

$$\text{var}(\widehat{Y}_2|S_1) = \left(1 - \frac{n_2}{N - n_1}\right) \frac{S_y'^2}{n_2},$$

where $S_y'^2$ is the population variance of y_k in $U \setminus S_1$. Since S_1 is derived from a simple random sample without replacement, it is clear that $U \setminus S_1$ is as well, and $E(S_y'^2) = S_y^2$. Therefore,

$$\begin{aligned} \text{var}(\widehat{Y}_2) &= \text{var}\left(\frac{N\bar{Y} - n_1\widehat{Y}_1}{N - n_1}\right) + E\left[\left(1 - \frac{n_2}{N - n_1}\right) \frac{S_y'^2}{n_2}\right] \\ &= \left(\frac{n_1}{N - n_1}\right)^2 \text{var}(\widehat{Y}_1) + \left(1 - \frac{n_2}{N - n_1}\right) \frac{E[S_y'^2]}{n_2} \\ &= \left(\frac{n_1}{N - n_1}\right)^2 \frac{N - n_1}{Nn_1} S_y^2 + \left(1 - \frac{n_2}{N - n_1}\right) \frac{S_y^2}{n_2} \\ &= \frac{N - n_2}{Nn_2} S_y^2. \end{aligned}$$

We notice that it is the variance of a simple random sample of size n_2 in the complete population. Therefore

$$\text{var}(\widehat{Y}_1) = \frac{N - n_1}{Nn_1} S_y^2 \quad \text{and} \quad \text{var}(\widehat{Y}_2) = \frac{N - n_2}{Nn_2} S_y^2.$$

Furthermore,

$$\text{cov}(\widehat{Y}_1, \widehat{Y}_2) = E[\text{cov}(\widehat{Y}_1, \widehat{Y}_2|S_1)] + \text{cov}[E(\widehat{Y}_1|S_1), E(\widehat{Y}_2|S_1)],$$

where now \widehat{Y}_1 is constant conditionally to S_1 , thus

$$\text{cov}(\widehat{Y}_1, \widehat{Y}_2|S_1) = 0, \text{ and } E(\widehat{Y}_1|S_1) = \widehat{Y}_1,$$

and

$$\begin{aligned} \text{cov}(\widehat{Y}_1, \widehat{Y}_2) &= 0 + \text{cov}\left(\widehat{Y}_1, \frac{N\bar{Y} - n_1\widehat{Y}_1}{N - n_1}\right) = -\frac{n_1}{N - n_1} \frac{N - n_1}{Nn_1} S_y^2 \\ &= -\frac{1}{N} S_y^2. \end{aligned}$$

Therefore,

$$\text{var}(\widehat{Y}_\theta) = \frac{S_y^2}{N} \left[\theta^2 \frac{N-n_1}{n_1} + (1-\theta)^2 \frac{N-n_2}{n_2} - 2\theta(1-\theta) \right].$$

The optimal value of θ is obtained by differentiating $\text{var}(\widehat{Y}_\theta)$ with respect to θ and setting the derivative equal to zero, which gives

$$2\theta^* \frac{N-n_1}{n_1} - 2(1-\theta^*) \frac{N-n_2}{n_2} - 2(1-2\theta^*) = 0,$$

and we get

$$\theta^* = \frac{n_1}{n}.$$

Method 2:

We use the expression \widehat{Y}_θ as a function of \widehat{Y}_1 and \widehat{Y} , which avoids the tedious calculation of the variance of \widehat{Y}_2 . We very easily verify that

$$\widehat{Y}_\theta = \delta \widehat{Y}_1 + (1-\delta) \widehat{Y},$$

with

$$\delta = \frac{n}{n_2} \theta - \frac{n_1}{n_2}.$$

$$\text{var}(\widehat{Y}_\theta) = \delta^2 \text{var}(\widehat{Y}_1) + (1-\delta)^2 \text{var}(\widehat{Y}) + 2\delta(1-\delta) \text{cov}(\widehat{Y}_1, \widehat{Y}).$$

Now,

$$\begin{aligned} \text{cov}(\widehat{Y}_1, \widehat{Y}) &= \text{Ecov}(\widehat{Y}_1, \widehat{Y} | S_1) + \text{cov}[E(\widehat{Y}_1 | S_1), E(\widehat{Y} | S_1)] \\ &= E(0) + \text{cov}\left[\widehat{Y}_1, \frac{n_1}{n} \widehat{Y}_1 + \frac{n_2}{n} E(\widehat{Y}_2 | S_1)\right] \\ &= \text{cov}\left(\widehat{Y}_1, \frac{n_1}{n} \widehat{Y}_1 + \frac{n_2}{n} \frac{N\bar{Y} - n_1 \widehat{Y}_1}{N - n_1}\right) \\ &= \frac{n_1}{n} \frac{N-n}{N-n_1} \text{var}(\widehat{Y}_1) = \frac{n_1}{n} \frac{N-n}{N-n_1} \left(1 - \frac{n_1}{N}\right) \frac{S_y^2}{n_1}. \end{aligned}$$

Finally,

$$\begin{aligned} \text{var}(\widehat{Y}_\theta) &= \delta^2 \left(1 - \frac{n_1}{N}\right) \frac{S_y^2}{n_1} + (1-\delta)^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n} \\ &\quad + 2\delta(1-\delta) \frac{n_1}{n} \frac{N-n}{N-n_1} \left(1 - \frac{n_1}{N}\right) \frac{S_y^2}{n_1} \\ &= \frac{S_y^2}{N} \left[\delta^2 \left(\frac{N-n_1}{n_1}\right) + (1-\delta)^2 \left(\frac{N-n}{n}\right) + 2\delta(1-\delta) \left(\frac{N-n}{n}\right) \right] \\ &= \frac{S_y^2}{N} \left[\left(\frac{N-n}{n}\right) + \frac{N(n-n_1)}{nn_1} \delta^2 \right]. \end{aligned}$$

As $(n - n_1 > 0)$, $\text{var} [\widehat{\bar{Y}}_\theta]$ is manifestly minimal for $\delta^2 = 0$, being

$$\frac{n}{n_2}\theta^* - \frac{n_1}{n_2} = 0,$$

therefore

$$\theta^* = \frac{n_1}{n}.$$

We indeed find again the same θ^* as with Method 1, in a little more ‘elegant’ fashion. No matter the method, the optimal estimator must be the simple mean of the sample S , being $\widehat{\bar{Y}}$. Therefore, when we select samples repeatedly (by simple random sampling each time), the best estimator is still the most simple, meaning that one which we naturally get by combining all the samples *in fine*.

Exercise 2.15 Capture-recapture

In surveys, it sometimes happens that the population size is ignored by the survey taker. One method to remediate this is the following: we identify, among the total population of size N (unknown), M individuals. We then allow these individuals to ‘mix’ with the total population, and we select n individuals by simple random sampling in the total population after mixing. We then pick out from this sample m individuals belonging to the first ‘marked’ population.

1. What is the distribution of m ; what is its expected value and variance?
2. What is the probability that m is equal to zero? We suppose n is small with respect to M and with respect to $N - M$.
3. Considering the expectation of m , give a natural estimator \widehat{N} of N in the case where m is not equal to zero. We verify that in practice this occurs if n and M are ‘sufficiently large’.
4. Calculate $\mathcal{M} = E(m \mid m > 0)$ and $\mathcal{V} = \text{var}(m \mid m > 0)$. In using a Taylor expansion of m around \mathcal{M} , approach $E(\widehat{N} \mid m > 0)$ by considering n ‘large’ (and, consequently, N ‘particularly large’).
5. Conclude about the eventual bias of \widehat{N} .

This method, called ‘capture-recapture’ (see Thompson, 1992), can be used, for example, to estimate the number of wild animals of a certain type in a large forest (we control M , the number of marked animals, and obviously n).

Solution

1. The random variable m follows a hypergeometric distribution with parameters N , M , n :

$$\Pr(m = x) = \frac{\binom{M}{x} \times \binom{N-M}{n-x}}{\binom{N}{n}},$$

for all $x = \max(0, M - N + n), 1, 2, \dots, \min(n, M)$. We can obviously calculate the moments directly by using the previous expression. We can also notice that m/n is the classical unbiased estimator of the true proportion of M/N ‘marked’ individuals. Hence:

$$\mathbb{E}\left(\frac{m}{n}\right) = \frac{M}{N}, \quad \text{and therefore} \quad \mathbb{E}(m) = n \frac{M}{N}.$$

The variance is

$$\text{var}\left(\frac{m}{n}\right) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} \left(\frac{M}{N}\right) \left(1 - \frac{M}{N}\right).$$

If N is large, we then have:

$$\text{var}(m) \approx \left(1 - \frac{n}{N}\right) \left(n \frac{M}{N}\right) \left(1 - \frac{M}{N}\right).$$

2. The probability that m is null is:

$$\Pr(m = 0) = \frac{\binom{M}{0} \binom{N-M}{n}}{\binom{N}{n}} = \frac{\binom{N-M}{n}}{\binom{N}{n}} = \frac{(N-M)^{[n]}}{N^{[n]}} \approx \left(\frac{N-M}{N}\right)^n,$$

where $N^{[n]} = N \times (N-1) \times \dots \times (N-n+1)$. This probability is negligible when M and n are sufficiently large.

3. Since

$$\frac{M}{N} = \mathbb{E}\left(\frac{m}{n}\right),$$

we can use:

$$\hat{N} = M \frac{n}{m},$$

but only if $m > 0$. In practice, this is almost certainly confirmed if M and n are sufficiently large according to Question 2. If $m = 0$, we do not use any estimation (in concrete terms, we continue with the process from the beginning, until $m > 0$).

4. As

$$\mathbb{E}(m) = \mathbb{E}(m \mid m = 0)\Pr(m = 0) + \mathbb{E}(m \mid m > 0)\Pr(m > 0),$$

we have

$$\mathcal{M} = \mathbb{E}(m \mid m > 0) = \frac{\mathbb{E}(m)}{\Pr(m > 0)} = n \frac{M}{N} \frac{1}{1 - \Pr(m = 0)},$$

and

$$\begin{aligned}
\mathcal{V} &= \text{var}(m \mid m > 0) = \text{E}(m^2 \mid m > 0) - [\text{E}(m \mid m > 0)]^2 \\
&= \frac{\text{E}(m^2)}{\Pr(m > 0)} - \mathcal{M}^2 \\
&= \frac{\text{var}(m) + [\text{E}(m)]^2}{\Pr(m > 0)} - \mathcal{M}^2 \\
&= \frac{1}{\Pr(m > 0)} \left[\text{var}(m) - \frac{\Pr(m = 0)}{\Pr(m > 0)} \left(n \frac{M}{N} \right)^2 \right].
\end{aligned}$$

Furthermore, we have:

$$\frac{1}{m} = \frac{1}{\mathcal{M} \left(1 + \frac{m - \mathcal{M}}{\mathcal{M}} \right)}, \quad \text{for all } m > 0 \ (\mathcal{M} > 0).$$

Now, the term

$$\Delta = \frac{m - \mathcal{M}}{\mathcal{M}},$$

conditional on $m > 0$, is of null expectation, by construction. Furthermore:

$$\text{var}(\Delta \mid m > 0) = \frac{1}{\mathcal{M}^2} \text{var}(m \mid m > 0) = \frac{\text{var}(m)}{\Pr(m > 0) \mathcal{M}^2} - \Pr(m = 0).$$

If n and M are large, then $\Pr(m = 0)$ is negligible with respect to the first term of this difference. Since n is large, we can write

$$\left(1 + \frac{m - \mathcal{M}}{\mathcal{M}} \right)^{-1} \approx 1 - \frac{m - \mathcal{M}}{\mathcal{M}} + \left(\frac{m - \mathcal{M}}{\mathcal{M}} \right)^2 + \dots$$

With n large, we neglect the terms of order 3 and above, of order of magnitude by $(nM/N)^{-3/2}$. Thus, for $m > 0$,

$$\hat{N} = \frac{Mn}{m} \approx \frac{Mn}{\mathcal{M}} (1 - \Delta + \Delta^2),$$

and

$$\text{E}(\hat{N} \mid m > 0) \approx \frac{Mn}{\mathcal{M}} (1 + \text{E}(\Delta^2 \mid m > 0)) = N \Pr(m > 0) \left(1 + \frac{\mathcal{V}}{\mathcal{M}^2} \right).$$

5. The estimator is then biased. The bias results from the conjunction of two elements: on the one hand, we are restricted at $m > 0$, and on the other hand the random variable m is in the denominator of the estimator. If n is large, the bias is small because

$$\Pr(m > 0) = 1 - \Pr(m = 0)$$

approaches 1 and that

$$\frac{\mathcal{V}}{\mathcal{M}^2} \text{ varies by } 1/n$$

and therefore approaches zero. The estimator \widehat{N} thus appears as an interesting estimator of N . It would remain to calculate its variance.

Exercise 2.16 *Subsample and covariance*

We consider a simple random sample without replacement of size n in a population U of size N (sample denoted as S). We also consider two individuals k and ℓ *distinct*.

1. Show that:

$$\Pr[k \in S \text{ and } \ell \notin S] = \frac{n(N-n)}{N(N-1)}.$$

2. In the previous sample S , we select by simple random sampling n_1 individuals. We denote S_1 as the sample obtained and S_2 as the complementary sample of S_1 in S . Let k and ℓ be any two distinct individuals belonging to the sample S (we thus work ‘conditionally on S ’). What is $\Pr(k \in S_1 \text{ and } \ell \in S_2 \mid S)$? (Hint: use Question 1.)
3. If k and ℓ are any two elements (but *distinct*) in the population, show that

$$\Pr[k \in S_1 \text{ and } \ell \in S_2] = \frac{n_1(n-n_1)}{N(N-1)}.$$

4. Show, in the conditions of Question 2, that we can consider S_1 as a simple random sample of size n_1 selected from a population of size N . (Hint: calculate $\Pr(S_1 = s_1)$.)
5. By using the results from Question 1, calculate, first of all for k different from ℓ and then for k equal to ℓ , the following:

$$\text{cov}(I\{k \in S_1\}, I\{\ell \in S_2\}),$$

where $I\{A\}$ represents the indicator for event A .

6. Deduce that:

$$\text{cov}(\widehat{Y}_1, \widehat{Y}_2) = -\frac{S_y^2}{N},$$

where \widehat{Y}_ℓ is the simple mean of a real variable y_k calculated in the sample S_ℓ ($\ell = 1, 2$), and S_y^2 is the population variance of y_k .

7. Calculate $\text{cov}(\widehat{Y}, \widehat{Y}_1)$ where \widehat{Y} is the simple mean of y_k calculated in S .

Solution

1. Since

$$\begin{aligned} & \Pr(k \in S \text{ and } \ell \notin S) + \Pr(k \in S \text{ and } \ell \in S) \\ &= \Pr(k \in S \text{ and } (\ell \in S \text{ or } \ell \notin S)) = \Pr(k \in S), \end{aligned}$$

we have

$$\Pr(k \in S \text{ and } \ell \notin S) = \pi_k - \pi_{k\ell} = \frac{n}{N} - \frac{n(n-1)}{N(N-1)} = \frac{n(N-n)}{N(N-1)}.$$

A second method consists of writing:

$$\Pr(k \in S \text{ and } \ell \notin S) = \sum_{\substack{s \ni k \\ s \not\ni \ell}} p(s) = \frac{\#\{s | k \in s \text{ and } \ell \notin s\}}{\binom{N}{n}}.$$

Now the number of samples s containing k but not ℓ is $\binom{N-2}{n-1}$. Indeed, k being in s , there remain $(n-1)$ individuals to select in the population outside of k and of ℓ .

2. The sample S is fixed: we can use the previous result by considering that the population here is the sample S and that the sample is S_1 :

$$\Pr(k \in S_1 \text{ and } \ell \notin S_1 | S) = \frac{n_1(n-n_1)}{n(n-1)}, \text{ with } k \text{ and } \ell \in S.$$

It remains to state that $(\ell \notin S_1 | S)$ is equivalent to $(\ell \in S_2 | S)$, seeing that S_1 and S_2 form a partition of S .

3. As for all S :

$$\Pr[k \in S_1 \text{ and } \ell \in S_2 | S] = \frac{n_1(n-n_1)}{n(n-1)}, \text{ if } k \text{ and } \ell \in S.$$

We have

$$\begin{aligned} & \Pr[k \in S_1 \text{ and } \ell \in S_2] \\ &= \sum_{s | k \in s \text{ and } \ell \in s} \Pr[k \in S_1 \text{ and } \ell \in S_2 | S = s] \Pr[S = s] \\ &= \frac{n_1(n-n_1)}{n(n-1)} \binom{N}{n}^{-1} \binom{N-2}{n-2} \\ &= \frac{n_1(n-n_1)}{n(n-1)} \frac{n(n-1)}{N(N-1)} = \frac{n_1(n-n_1)}{N(N-1)}. \end{aligned}$$

A faster approach, but less natural, consists of stating the result from Question 2 obtained by depending on the lone fact that k and ℓ are in S (the integral composition of S does not provide anything). Also,

$$\Pr(k \in S_1 \text{ and } \ell \in S_2 | k \in S \text{ and } \ell \in S) = \frac{n_1(n - n_1)}{n(n - 1)},$$

which leads to

$$\begin{aligned} & \Pr(k \in S_1 \text{ and } \ell \in S_2) \\ &= \Pr(k \in S_1 \text{ and } \ell \in S_2 | k \in S \text{ and } \ell \in S) \Pr(k \in S \text{ and } \ell \in S) \\ &= \frac{n_1(n - n_1)}{n(n - 1)} \frac{n(n - 1)}{N(N - 1)} = \frac{n_1(n - n_1)}{N(N - 1)}. \end{aligned}$$

4. The probability of selecting s_1 is:

$$\Pr(S_1 = s_1) = \sum_{s \supset s_1} \Pr(S_1 = s_1 | S = s) \Pr(S = s) = \frac{\binom{N-n_1}{n-n_1}}{\binom{n}{n_1} \binom{N}{n}} = \frac{1}{\binom{N}{n_1}}.$$

This result is characteristic of a simple random sampling of size n_1 in a population of size N .

5. If $k \neq \ell$, then

$$\begin{aligned} & \text{cov}(I\{k \in S_1\}, I\{\ell \in S_2\}) \\ &= E(I\{k \in S_1\} I\{\ell \in S_2\}) - (E I\{k \in S_1\})(E I\{\ell \in S_2\}) \\ &= \Pr[k \in S_1 \text{ and } \ell \in S_2] - \Pr[k \in S_1] \Pr[\ell \in S_2] \\ &= \frac{n_1(n - n_1)}{N(N - 1)} - \frac{n_1}{N} \frac{n - n_1}{N} = \frac{n_1(n - n_1)}{N^2(N - 1)}. \end{aligned}$$

If $k = \ell$,

$$\text{cov}(I\{k \in S_1\}, I\{k \in S_2\}) = -\Pr(k \in S_1) \Pr(k \in S_2) = -\frac{n_1(n - n_1)}{N^2}.$$

6. We let $n_2 = n - n_1$:

$$\begin{aligned} & \text{cov}(\widehat{Y}_1, \widehat{Y}_2) \\ &= \text{cov}\left(\sum_{k \in U} \frac{y_k I\{k \in S_1\}}{n_1}, \sum_{\ell \in U} \frac{y_\ell I\{\ell \in S_2\}}{n_2}\right) \\ &= \frac{1}{n_1(n - n_1)} \sum_{k \in U} \sum_{\ell \in U} \text{cov}(I\{k \in S_1\}, I\{\ell \in S_2\}) y_k y_\ell \\ &= \frac{1}{n_1(n - n_1)} \left[-\frac{n_1(n - n_1)}{N^2} \sum_{k \in U} y_k^2 + \frac{n_1(n - n_1)}{N^2(N - 1)} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} y_k y_\ell \right] \\ &= -\frac{1}{N^2} \left(\sum_{k \in U} y_k^2 - \frac{1}{N - 1} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} y_k y_\ell \right) = -\frac{S_y^2}{N}. \end{aligned}$$

This is a second method that depends only on the result from Question 4, meaning that S_1 is a simple random sample of size n_1 in a population of size N and that, by analogy, S_2 is a simple random sample of size n_2 in a population of size $N - n_1$.

$$\text{cov}(\widehat{Y}_1, \widehat{Y}_2) = E_{S_1} \text{cov}(\widehat{Y}_1, \widehat{Y}_2 | S_1) + \text{cov}_{S_1} [E(\widehat{Y}_1 | S_1), E(\widehat{Y}_2 | S_1)].$$

Now, conditionally on S_1 , \widehat{Y}_1 is constant:

$$\text{cov}(\widehat{Y}_1, \widehat{Y}_2) = \text{cov}_{S_1} [\widehat{Y}_1, E(\widehat{Y}_2 | S_1)].$$

We have

$$E(\widehat{Y}_2 | S_1) = \frac{1}{N - n_1} \sum_{k \in U \setminus S_1} y_k = \frac{N\bar{Y} - n_1\widehat{Y}_1}{N - n_1}.$$

Ultimately,

$$\begin{aligned} \text{cov}(\widehat{Y}_1, \widehat{Y}_2) &= \text{cov}_{S_1} \left(\widehat{Y}_1, -\frac{n_1\widehat{Y}_1}{N - n_1} \right) = -\frac{n_1}{N - n_1} \text{var}_{S_1} (\widehat{Y}_1) \\ &= -\frac{n_1}{N - n_1} \left(1 - \frac{n_1}{N} \right) \frac{S_y^2}{n_1} = -\frac{S_y^2}{N}. \end{aligned}$$

7. Finally, the covariance is

$$\begin{aligned} \text{cov}(\widehat{Y}, \widehat{Y}_1) &= \text{cov} \left(\frac{n_1}{n} \widehat{Y}_1 + \frac{n_2}{n} \widehat{Y}_2, \widehat{Y}_1 \right) \quad (\text{with } n_2 = n - n_1) \\ &= \frac{n_1}{n} \text{var}(\widehat{Y}_1) + \frac{n_2}{n} \text{cov}(\widehat{Y}_1, \widehat{Y}_2) \\ &= \frac{n_1}{n} \left(1 - \frac{n_1}{N} \right) \frac{S_y^2}{n_1} - \frac{n_2}{n} \frac{S_y^2}{N} \\ &= \left(1 - \frac{n}{N} \right) \frac{S_y^2}{n} = \text{var}(\widehat{Y}) = \text{cov}(\widehat{Y}, \widehat{Y}). \end{aligned}$$

From this fact, \widehat{Y} and $(\widehat{Y} - \widehat{Y}_1)$ appear to be uncorrelated, which is quite surprising.

Exercise 2.17 Recapture with replacement

The objective is to estimate the number of rats present on an island. We set up a trap which is installed at a location selected at random on the island. When a rat is trapped, it is marked and then released. If, for 50 captured rats, we count 42 distinctly marked rats, estimate using the maximum likelihood method the number of rats living on the island, assuming that the 50 rats were captured at random and with replacement.

Note: the maximum likelihood solution can be obtained through searching using, for example, a spreadsheet.

Solution

In this approach, N is the parameter to estimate and r is the random variable for which it is necessary to express the density and to later maximize. We denote $f_N(r)$ as the probability of obtaining r distinct rats in m trials with replacement (m is a controlled size, known and non-random) in a population of size N . This model is reasonable under the conditions of the process. We note that there are $\binom{N}{r} = N!/r!(N-r)!$ ways of choosing the list of r rats involved. Thus,

$$f_N(r) = \frac{N!}{r!(N-r)!} g_N(r),$$

where $g_N(r)$ is the probability of obtaining r distinct and properly identified rats in m trials with replacement (valid expression because all rats have, for each trial, the same probability of being selected). This list of rats being fixed, the universe Ω of possibilities is formed by the group of mappings of $\{1, \dots, m\}$ to $\{1, \dots, N\}$ (we assume that the r rats listed are identified by the first r integers). We have $m \geq r$, and in fact

$$g_N(r) = \sum_{\omega \in \text{FAV}} p(\omega),$$

where $p(\omega)$ is the probability of obtaining a given mapping ω and FAV is the group of favourable mappings. We have $p(\omega) = N^{-m}$, for all ω . It remains to calculate the total number of favourable cases. It is exactly a question of the number of surjective mappings of $\{1, \dots, m\}$ in $\{1, \dots, r\}$, which is equal to $r!$ multiplied by the Stirling number of second kind $\mathfrak{S}_m^{(r)}$, which is:

$$\mathfrak{S}_m^{(r)} = \frac{1}{r!} \sum_{i=1}^r \binom{r}{i} i^m (-1)^{r-i}.$$

The Stirling number of second kind is equal to the number of ways of finding a group of m elements in r non-empty parts (see Stanley, 1997). However, the calculation of $\mathfrak{S}_m^{(r)}$ does not interest us here. Indeed, $\mathfrak{S}_m^{(r)}$ does not depend on N but only on m and r . Eventually we obtain

$$f_N(r) = \frac{N!}{(N-r)!N^m} \mathfrak{S}_m^{(r)}, r = 1, \dots, \min(m, N). \quad (2.4)$$

We are going to maximize the function $f_N(r)$ for N . Now, maximizing $f_N(r)$ for N comes back to maximizing

$$\frac{N!}{(N-r)!N^m} = \frac{\prod_{i=0}^{r-1} (N-i)}{N^m},$$

as $\mathfrak{S}_m^{(r)}$ does not depend on N . When $m = 50$ and $r = 42$, we find the solution through a search (see Table 2.5). The solution of the maximum likelihood is

Table 2.5. Search for the solution of maximum likelihood: Exercise 2.17

N	$\frac{N!}{(N-r)!N^m} \times 10^{21}$	N	$\frac{N!}{(N-r)!N^m} \times 10^{21}$	N	$\frac{N!}{(N-r)!N^m} \times 10^{21}$
100	3.97038	120	6.49114	140	7.05245
101	4.13269	121	6.56558	141	7.03671
102	4.29281	122	6.63468	142	7.01773
103	4.45037	123	6.69847	143	6.99563
104	4.60503	124	6.75702	144	6.97054
105	4.75645	125	6.81037	145	6.94259
106	4.90434	126	6.85860	146	6.91191
107	5.04842	127	6.90178	147	6.87864
108	5.18844	128	6.94001	148	6.84288
109	5.32416	129	6.97339	149	6.80478
110	5.45539	130	7.00200	150	6.76444
111	5.58193	131	7.02597	151	6.72199
112	5.70364	132	7.04541	152	6.67755
113	5.82038	133	7.06042	153	6.63122
114	5.93203	134	7.07115	154	6.58311
115	6.03850	135	7.07770	155	6.53335
116	6.13971	136	7.08021	156	6.48202
117	6.23561	137	7.07881	157	6.42924
118	6.32616	138	7.07363	158	6.37510
119	6.41134	139	7.06479	159	6.31970

therefore $N = 136$. Another manner of tackling the problem consists of setting the first derivative of the logarithm of the likelihood function equal to zero

$$\frac{d \left(\frac{\log[\prod_{i=0}^{r-1} (N-i)]}{N^m} \right)}{dN} = \frac{d}{dN} \left[\sum_{i=0}^{r-1} \log(N-i) - m \log N \right] = \sum_{i=0}^{r-1} \frac{1}{N-i} - \frac{m}{N} = 0,$$

which gives

$$\sum_{i=0}^{r-1} \frac{N}{N-i} = m.$$

We obtain a non-linear equation that we can also solve by trial and error. Obviously, we obtain the same result.

Exercise 2.18 *Collection*

Your child would like to collect pictures of football players sold in sealed packages. The complete collection consists of 350 distinct pictures. Each package contains one picture ‘at random’ in a totally independent manner from one package to another. Purchasing X packages is similar to taking X samples with replacement and with equal probability in the population of size $N = 350$. To simplify, your child does not trade any pictures.

1. What is the probability distribution of the number of pictures to purchase in order to obtain exactly r different players?
2. How many photos must be purchased on average in order to obtain the complete collection?

Solution

1. In Exercise 2.17, we saw that if n_S represents the number of distinct units obtained by selecting m units with replacement in a population of size N , then

$$p_m(r) = \Pr(n_S = r) = \frac{N!}{(N-r)!N^m} \mathfrak{S}_m^{(r)},$$

where $r = 1, \dots, \min(m, N)$ and $\mathfrak{S}_m^{(r)}$ is a Stirling number of second kind,

$$\mathfrak{S}_m^{(r)} = \frac{1}{r!} \sum_{i=1}^r \binom{r}{i} i^m (-1)^{r-i}.$$

If we let X be the random variable representing the number of drawings necessary to obtain r distinct individuals, then

$$\begin{aligned} \Pr[X = m] &= \Pr[\text{selecting } r-1 \text{ distinct units in } m-1 \text{ samples with replacement}] \\ &\quad \times \Pr[\text{selecting in the } m\text{th sample a unit not yet selected} \\ &\quad \text{knowing that } r-1 \text{ distinct units have already been selected}] \\ &= p_{m-1}(r-1) \times \frac{N-r+1}{N} \\ &= \frac{N!}{(N-r)!N^m} \mathfrak{S}_{m-1}^{(r-1)}, \end{aligned}$$

for $m = r, r+1, \dots$

2. We know the probability distribution of the random variable X . We now wish to calculate its expected value in the case $r = N$, which corresponds to the complete collection. In the case of any r , we have

$$\mathbb{E}(X) = \sum_{m=r}^{\infty} m \frac{N!}{(N-r)!N^m} \mathfrak{S}_{m-1}^{(r-1)}.$$

Since

$$\sum_{m=r}^{\infty} \Pr[X = m] = 1,$$

we have

$$\sum_{m=r}^{\infty} \frac{\mathfrak{s}_{m-1}^{(r-1)}}{N^m} = \frac{(N-r)!}{N!} = \prod_{i=0}^{r-1} \frac{1}{(N-i)}. \quad (2.5)$$

By differentiating Identity (2.5) with respect to N (for the right-hand side, use the logarithmic derivative), we easily obtain

$$\sum_{m=r}^{\infty} m \frac{\mathfrak{s}_{m-1}^{(r-1)}}{N^{m+1}} = \left(\sum_{j=0}^{r-1} \frac{1}{N-j} \right) \frac{(N-r)!}{N!}.$$

We then get

$$E(X) = \frac{N!N}{(N-r)!} \sum_{m=r}^{\infty} m \frac{\mathfrak{s}_{m-1}^{(r-1)}}{N^{m+1}} = \sum_{j=0}^{r-1} \frac{N}{N-j}.$$

For the complete collection:

$$E(X) = \sum_{j=0}^{r-1} \frac{N}{N-j} = N \sum_{j=1}^N \frac{1}{j} = 350 \times \sum_{j=1}^{350} \frac{1}{j} \approx 350 \times (\log 350 + \gamma),$$

where γ is Euler's constant, approximately 0.5772. We get $E(X) \approx 2252$ pictures.

Exercise 2.19 Proportion of students

A sample of 100 students is chosen using a simple random design without replacement from a population of 1000 students. We are then interested in the results obtained by these students in an exam. There are two possible results: success or failure. The outcome is presented in Table 2.6.

Table 2.6. Sample of 100 students: Exercise 2.19

	Men	Women	Total
Success	$n_{11} = 35$	$n_{12} = 25$	$n_{1.} = 60$
Failure	$n_{21} = 20$	$n_{22} = 20$	$n_{2.} = 40$
Total	$n_{.1} = 55$	$n_{.2} = 45$	$n = 100$

1. Estimate the success rate for men and for women.
2. Calculate the approximate bias of the estimated success rates.
3. Estimate the mean square error of these success rates.
4. Give the 95% confidence intervals for the success rate for men R_M and for women R_W . What can we say about their respective positions?

5. What confidence intervals must be considered in order for the true values R_M and R_W to be inside the disjoint confidence intervals? Comment on this.
6. Using the estimation results by domain, find a more simple result for Questions 2 and 3.

Solution

The notation for different proportions in the population U is presented in Table 2.7.

Table 2.7. Notation for different proportions: Exercise 2.19

	Men	Women	Total
Success	P_{11}	P_{12}	$P_{1.}$
Failure	P_{21}	P_{22}	$P_{2.}$
Total	$P_{.1}$	$P_{.2}$	1

1. The success rate for men is naturally estimated by:

$$r_M = \frac{\hat{P}_{11}}{\hat{P}_{.1}} = \frac{n_{11}}{n_{.1}} = \frac{35}{55} \approx 63.6\%.$$

The success rate for women is estimated by:

$$r_W = \frac{\hat{P}_{12}}{\hat{P}_{.2}} = \frac{n_{12}}{n_{.2}} = \frac{25}{45} \approx 55.6\%.$$

These two estimators are ratios. Indeed, the denominators of these estimators are random.

2. Since the sample size n is 100, we can consider without hesitation that n is large. The bias of a ratio $r = \hat{Y}/\hat{X}$ is given by

$$B(r) = E(r) - R \approx R \left(\frac{S_x^2}{\bar{X}^2} - \frac{S_{xy}}{\bar{X}\bar{Y}} \right) \frac{1-f}{n},$$

where

$$x_k = \begin{cases} 1 & \text{if the individual is a man (resp. a woman)} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$y_k = \begin{cases} 1 & \text{if the individual is a man (resp. a woman) who succeeded} \\ 0 & \text{otherwise,} \end{cases}$$

for all $k \in U$. For example, we have for the men:

$$\begin{aligned} S_x^2 &= \frac{1}{N-1} \left(\sum_{k \in U} x_k^2 - N\bar{X}^2 \right) = \frac{1}{N-1} (NP_{.1} - NP_{.1}^2) \\ &= \frac{N}{N-1} P_{.1} (1 - P_{.1}), \end{aligned}$$

and

$$\begin{aligned} S_{xy} &= \frac{1}{N-1} \left(\sum_{k \in U} x_k y_k - N\bar{X}\bar{Y} \right) \\ &= \frac{1}{N-1} (NP_{11} - NP_{.1}P_{11}) \\ &= \frac{N}{N-1} P_{11} (1 - P_{.1}). \end{aligned}$$

We therefore have

$$\frac{S_x^2}{\bar{X}^2} - \frac{S_{xy}}{\bar{X}\bar{Y}} = \frac{1}{P_{.1}^2} \frac{N}{N-1} P_{.1} (1 - P_{.1}) - \frac{1}{P_{.1}P_{11}} \frac{N}{N-1} P_{11} (1 - P_{.1}) = 0.$$

The bias is thus approximately null: $B(r) \approx 0$.

3. Since n is large, the mean square error, similar to the variance, is given by the approximation

$$\text{MSE}(r) \approx \frac{1-f}{n\bar{X}^2} (S_y^2 - 2RS_{xy} + R^2S_x^2),$$

where

$$R = \frac{\bar{Y}}{\bar{X}}.$$

For the men, we get

$$\begin{aligned} \text{MSE}(r_M) &\approx \text{var}(r_M) \\ &\approx \frac{1-f}{nP_{.1}^2} \frac{N}{N-1} \left\{ P_{11}(1 - P_{11}) - 2\frac{P_{11}}{P_{.1}} P_{11}(1 - P_{.1}) + \frac{P_{11}^2}{P_{.1}^2} P_{.1}(1 - P_{.1}) \right\} \\ &= \frac{1-f}{nP_{.1}^2} \frac{N}{N-1} \left\{ P_{11} - 2\frac{P_{11}^2}{P_{.1}} + \frac{P_{11}^2}{P_{.1}} \right\} \\ &= \frac{1-f}{nP_{.1}} \frac{N}{N-1} \frac{P_{11}}{P_{.1}} \left\{ 1 - \frac{P_{11}}{P_{.1}} \right\}. \end{aligned}$$

The estimator (slightly biased) directly becomes

$$\widehat{\text{MSE}}(r_M) = \frac{1-f}{n\hat{P}_{.1}} \frac{N}{N-1} \frac{\hat{P}_{11}}{\hat{P}_{.1}} \left\{ 1 - \frac{\hat{P}_{11}}{\hat{P}_{.1}} \right\}.$$

We get

- For the men:

$$\widehat{\text{MSE}}(r_M) = \frac{1 - \frac{1}{10}}{100 \frac{55}{100}} \frac{1000}{999} \frac{35}{55} \left\{ 1 - \frac{35}{55} \right\} = 0.00379041.$$

- For the women:

$$\widehat{\text{MSE}}(r_W) = \frac{1 - \frac{1}{10}}{100 \frac{45}{100}} \frac{1000}{999} \frac{25}{45} \left\{ 1 - \frac{25}{45} \right\} = 0.0049432148.$$

4. With 95 chances out of 100 (roughly), we have the estimated intervals

$$\begin{aligned} \widehat{\text{CI}}(R_M; 0.95) &= \left[r_M - 1.96 \sqrt{\widehat{\text{MSE}}(r_M)}, r_M + 1.96 \sqrt{\widehat{\text{MSE}}(r_M)} \right] \\ &= [0.636 - 0.121; 0.636 + 0.121] = [0.515; 0.757], \end{aligned}$$

$$\begin{aligned} \widehat{\text{CI}}(R_W; 0.95) &= \left[r_W - 1.96 \sqrt{\widehat{\text{MSE}}(r_W)}, r_W + 1.96 \sqrt{\widehat{\text{MSE}}(r_W)} \right] \\ &= [0.556 - 0.138; 0.556 + 0.138] = [0.418; 0.694]. \end{aligned}$$

The sample size is not very large, but we can consider it to be *a priori* sufficient to approach the distribution of ratios by the normal distribution. Therefore, the two intervals overlap very considerably: we cannot say that the ratios R_M and R_W are significantly different, considering the selected sample size (that is, we do not find two disjoint intervals).

5. With 40 chances out of 100 (roughly), we have the estimated intervals

$$\begin{aligned} \widehat{\text{CI}}(R_M; 0.40) &= \left[r_M - 0.52 \sqrt{\widehat{\text{MSE}}(r_M)}, r_M + 0.52 \sqrt{\widehat{\text{MSE}}(r_M)} \right] \\ &= [0.636 - 0.032; 0.636 + 0.032] = [0.604; 0.668], \end{aligned}$$

$$\begin{aligned} \widehat{\text{CI}}(R_W; 0.40) &= \left[r_W - 0.52 \sqrt{\widehat{\text{MSE}}(r_W)}, r_W + 0.52 \sqrt{\widehat{\text{MSE}}(r_W)} \right] \\ &= [0.556 - 0.037; 0.556 + 0.037] = [0.519; 0.593]. \end{aligned}$$

The establishment of such intervals is an exercise of style which does not represent much in practice. Except for an ‘absolute miracle’, we indeed have $R_M \neq R_W$ (why would we think otherwise?). The question is to find out if the confidence interval actually confirms this evidence or not. If the two intervals do not overlap, the produced statistic could be used as evidence in confirming that $R_M \neq R_W$. If they overlap such as in Question 4, we find the statistic has no usefulness. We can only say that the sample was not large enough to reject the equality hypothesis of the ratios. Obviously, the 40% intervals allow to significantly separate R_M from R_W , but the probability of covering the true values is so poor that we cannot seriously refer to it.

6. The approach (bias and variance) of the previous questions relied upon direct calculations carried out starting from the ratio. However, we can note that we are precisely in the situation of mean estimation in a domain: for example, if we go back to the notation from Question 2, R_M is the mean of y_k for the domain of men ($x_k = 1$). We know that the estimated mean for the domain (here r_M) has, as expected value, the true value R_M as soon as we use a conditional expectation to the sample size matching up with the domain (here $n_{.1}$). A problem occurs when $n_{.1} = 0$, in which case we cannot calculate r_M , but this situation can only occur with a negligible probability (here $n = 100$). Thus, for all $n_{.1} > 0$, we have

$$E[r_M | n_{.1}] = R_M,$$

and therefore

$$E[r_M] = E_{n_{.1}} E[r_M | n_{.1}] = E_{n_{.1}}[R_M] = R_M,$$

where $E_{n_{.1}}[\cdot]$ is the expectation in relation to the hypergeometric distribution of the random variable $n_{.1}$ in a population of size $N_{.1}$ (excluding the case where $n_{.1} = 0$). The bias is approximately null. For the conditional variance, we use the characteristic expression for a simple random sample of size $n_{.1}$:

$$\text{var}[r_M | n_{.1}] = \left(1 - \frac{n_{.1}}{N_{.1}}\right) \frac{S_1}{n_{.1}},$$

where

$$S_1 = \frac{N_{11}}{N_{.1}} \left(1 - \frac{N_{11}}{N_{.1}}\right) = \frac{P_{11}}{P_{.1}} \left(1 - \frac{P_{11}}{P_{.1}}\right),$$

seeing that it is a question of a proportion. The unconditional variance is obtained by

$$\begin{aligned} \text{var}[r_M] &= E_{n_{.1}} \text{var}[r_M | n_{.1}] + \text{var}_{n_{.1}} E[r_M | n_{.1}] \\ &= E_{n_{.1}} \text{var}[r_M | n_{.1}] \\ &= E_{n_{.1}} \left[\left(1 - \frac{n_{.1}}{N_{.1}}\right) \frac{1}{n_{.1}} \right] \frac{P_{11}}{P_{.1}} \left(1 - \frac{P_{11}}{P_{.1}}\right) \\ &= \left(E \left[\frac{1}{n_{.1}} \right] - \frac{1}{N_{.1}} \right) \frac{P_{11}}{P_{.1}} \left(1 - \frac{P_{11}}{P_{.1}}\right). \end{aligned}$$

In the first approximation, as n is large:

$$E \left[\frac{1}{n_{.1}} \right] \approx \frac{1}{E[n_{.1}]} = \frac{1}{nP_{.1}}.$$

Since $N_{.1} = NP_{.1}$, we finally get

$$\text{var}[r_M] = \frac{1 - f}{nP_{.1}} \frac{P_{11}}{P_{.1}} \left(1 - \frac{P_{11}}{P_{.1}}\right),$$

and we indeed find the variance found in Question 3 apart from a factor $N/(N-1)$ (this factor is obviously close to 1).

Exercise 2.20 *Sampling with replacement and estimator improvement*

Consider a population of size N . We perform simple random sampling with replacement of size $m = 3$. We denote \tilde{S} as the random sample selection (with repetitions). For example, with $N = 5$, \tilde{S} can have as values

$$(1, 2, 5), (1, 3, 4), (2, 4, 4), (2, 2, 3), (2, 3, 3), (3, 3, 3).$$

(we consider two samples containing the same units in a different order to be distinct). Consider the reduction function $r(\cdot)$, which suppresses from the sample the information concerning any multiplicity of units. For example:

$$r((2, 2, 3)) = \{2, 3\}, \quad r((2, 3, 3)) = \{2, 3\}, \quad r((3, 3, 3)) = \{3\}.$$

We denote S as the random sample without replacement obtained by suppressing the information concerning the multiplicity of units (in S , the order of individuals does not matter).

1. Calculate the probability R_i that sample \tilde{S} contains exactly i distinct individuals ($i = 1, 2$, or 3).
2. Show that the design of S conditional on its size $\#S$ is a simple random design without replacement of fixed size.
3. Give the sampling design for S , that is, the list of all possible values of S and the probabilities associated with those values.
4. Consider the following two estimators:
The mean with repetition

$$\tilde{Y} = \frac{1}{3} \sum_{k \in \tilde{S}} y_k,$$

the mean calculated on distinct values

$$\hat{Y} = \frac{1}{\#S} \sum_{k \in S} y_k.$$

Calculate the expected values and the variances for these estimators. Make a conclusion.

Solution

1. The probability of having three distinct individuals is

$$R_3 = \frac{N-1}{N} \times \frac{N-2}{N} = \frac{(N-1)(N-2)}{N^2}.$$

In fact, with the first individual (any) being selected, there is a probability $1/N$ that the second individual is identical to the first. Furthermore, with

the first two individuals (distinct) having been selected, there is a probability $2/N$ that the third individual is also one of the first two. Another method consists of counting the number of distinct trios of elements (there are $N(N-1)(N-2)$ combinations) and multiplying this number by the probability of obtaining any given trio, which is $1/N^3$. The probability of getting the same unit three times is

$$R_1 = \frac{1}{N} \times \frac{1}{N} = \frac{1}{N^2}.$$

The probability of obtaining two distinct individuals is obtained by the difference:

$$R_2 = 1 - R_1 - R_3 = \frac{N^2 - (N-1)(N-2) - 1}{N^2} = \frac{3(N-1)}{N^2}.$$

2. For reasons of symmetry between the units, the design of S conditional on $\#S$ had to be simple. However, we are going to calculate this conditional design rigorously. The design of S is obtained from the design of \tilde{S} . Conditional on the size j of S , we have, for $j = 1, 2, 3$:

$$\Pr(S = s | \#S = j) = \begin{cases} \frac{\Pr(S = s)}{R_j} = \frac{\sum_{\tilde{s}|r(\tilde{s})=s} \tilde{p}(\tilde{s})}{R_j} & \text{if } \#s = j \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{p}(\tilde{s})$ is the probability of obtaining an ordered sample with repetition \tilde{s} . Since the sampling is done with replacement, we have $\tilde{p}(\tilde{s}) = 1/N^3$, for all \tilde{s} , which is:

$$\Pr(S = s | \#S = j) = \begin{cases} \frac{1}{R_j} \times \#\{\tilde{s}|r(\tilde{s}) = s\} \times \frac{1}{N^3} & \text{if } \#s = j \\ 0 & \text{otherwise.} \end{cases}$$

- If $j = 1$, and $\#s = 1$, then $\#\{\tilde{s}|r(\tilde{s}) = s\} = 1$.
- If $j = 2$, and $\#s = 2$, then $\#\{\tilde{s}|r(\tilde{s}) = s\} = 6$. In fact, if $s = \{a, b\}$, we can have, for \tilde{S} :

$$(a, a, b) \text{ or } (a, b, a) \text{ or } (b, a, a) \text{ or } (a, b, b) \text{ or } (b, a, b) \text{ or } (b, b, a).$$

- If $j = 3$, and $\#s = 3$, then $\#\{\tilde{s}|r(\tilde{s}) = s\} = 3! = 6$.

We can then calculate the probability $p(s)$ of selecting s , conditional on $\#S$:

- If $j = 1$, and $\#s = 1$, then

$$\Pr(S = s | \#S = 1) = \frac{1/N^3}{1/N^2} = \binom{N}{1}^{-1}.$$

- If $j = 2$, and $\#s = 2$, then

$$\Pr(S = s | \#S = 2) = \frac{\frac{6}{N^3}}{\frac{3(N-1)}{N^2}} = 2 \frac{1}{N(N-1)} = \binom{N}{2}^{-1}.$$

- If $j = 3$, and $\#s = 3$, then

$$\Pr(S = s | \#S = 3) = \frac{\frac{6}{N^3}}{\frac{(N-1)(N-2)}{N^2}} = \frac{6}{N(N-1)(N-2)} = \binom{N}{3}^{-1}.$$

The design conditional on $\#S$ is simple without replacement of fixed size equal to $\#S$.

3. Being conditional on $\#S$, the sample is simple without replacement of fixed size and we have:

$$\begin{aligned} p(s) &= \Pr(S = s) \\ &= \Pr(S = s | \#S = \#s) \Pr(\#S = \#s) \\ &= \begin{cases} R_1 \times \binom{N}{1}^{-1} = \frac{1}{N^3} & \text{if } \#s = 1 \\ R_2 \times \binom{N}{2}^{-1} = \frac{6}{N^3} & \text{if } \#s = 2 \\ R_3 \times \binom{N}{3}^{-1} = \frac{6}{N^3} & \text{if } \#s = 3. \end{cases} \end{aligned}$$

4. The estimator \tilde{Y} is the classical estimator obtained by sampling with replacement of size 3. It is unbiased and

$$\text{var}(\tilde{Y}) = \frac{\sigma_y^2}{3} = \frac{N-1}{3N} S_y^2,$$

where

$$\begin{aligned} \sigma_y^2 &= \frac{1}{N} \sum_{k \in U} (y_k - \bar{Y})^2, \\ \bar{Y} &= \frac{1}{N} \sum_{k \in U} y_k, \end{aligned}$$

and

$$S_y^2 = \frac{N}{N-1} \sigma_y^2.$$

The estimator \hat{Y} is more particular to treat, but we have

$$E(\hat{Y}) = E(\hat{Y} | \#S = 1) R_1 + E(\hat{Y} | \#S = 2) R_2 + E(\hat{Y} | \#S = 3) R_3.$$

Being conditional on the size of S , the design is simple without replacement with fixed size, $E(\hat{Y} | \#S = \alpha) = \bar{Y}$, for $\alpha = 1, 2, 3$, and therefore $E(\hat{Y}) = \bar{Y}$. Moreover,

$$\begin{aligned}
\text{var}(\widehat{\bar{Y}}) &= E\{\text{var}(\widehat{\bar{Y}}|\#S)\} + \underbrace{\text{var}\left\{E(\widehat{\bar{Y}}|\#S)\right\}}_{=\bar{Y}} \\
&= E\{\text{var}(\widehat{\bar{Y}}|\#S)\} \\
&= \text{var}(\widehat{\bar{Y}}|\#S=1)R_1 + \text{var}(\widehat{\bar{Y}}|\#S=2)R_2 + \text{var}(\widehat{\bar{Y}}|\#S=3)R_3 \\
&= \frac{N-1}{N} \frac{S_y^2}{1} R_1 + \frac{N-2}{N} \frac{S_y^2}{2} R_2 + \frac{N-3}{N} \frac{S_y^2}{3} R_3 \\
&= \frac{S_y^2}{N} \left[(N-1)R_1 + \frac{N-2}{2}R_2 + \frac{N-3}{3}R_3 \right] \\
&= \frac{S_y^2(2N-1)(N-1)}{6N^2} \\
&= \left(1 - \frac{1}{2N}\right) \text{var}(\widetilde{\bar{Y}}).
\end{aligned}$$

Thus, $\widehat{\bar{Y}}$ appears to be systematically more efficient than $\widetilde{\bar{Y}}$.

Exercise 2.21 *Variance of the variance*

In a simple random design *without* replacement, give the first- through fourth-order inclusion probabilities. Next, give the variance for the estimator of the sampling variance. Simplify the expression for the case where N is very large, then suppose that y is distributed according to a normal distribution in U . What can we say about the estimator of the variance if n is ‘large’?

Solution

If we denote I_i as the indicator variable for the presence of unit i in sample S , we have

$$I_i = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{if } i \notin S. \end{cases}$$

The first- through fourth-order inclusion probabilities are:

$$\pi_1 = E(I_i) = \frac{n}{N}, i = 1, \dots, N,$$

$$\pi_2 = E(I_i I_j) = \frac{n(n-1)}{N(N-1)}, j \neq i,$$

$$\pi_3 = E(I_i I_j I_k) = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}, j \neq i, k \neq i, k \neq j,$$

and

$$\begin{aligned}
\pi_4 = E(I_i I_j I_k I_\ell) &= \frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)}, \\
&j \neq i, k \neq i, \ell \neq i, k \neq j, \ell \neq j, \ell \neq k.
\end{aligned}$$

The corrected variance in the sample is:

$$s_y^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \widehat{\bar{Y}})^2,$$

where

$$\widehat{\bar{Y}} = \frac{1}{n} \sum_{i \in S} y_i.$$

This estimator is unbiased for the corrected variance in the population

$$\mathbb{E}(s_y^2) = S_y^2, \quad (2.6)$$

where

$$S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})^2 = \frac{N}{N-1} \sigma_y^2,$$

and

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k.$$

In fact, since s_y^2 can also be written (see Exercise 2.7),

$$s_y^2 = \frac{1}{2n(n-1)} \sum_{i \in S} \sum_{j \in S} (y_i - y_j)^2,$$

we get

$$\begin{aligned} \mathbb{E}(s_y^2) &= \frac{1}{2n(n-1)} \sum_{i \in U} \sum_{j \in U} (y_i - y_j)^2 \mathbb{E}(I_i I_j) \\ &= \frac{1}{2N(N-1)} \sum_{i \in U} \sum_{j \in U} (y_i - y_j)^2 = S_y^2. \end{aligned}$$

To calculate the variance of s_y^2 following the sampling, we suppose that the population mean \bar{Y} is null, without sacrificing the general nature of the solution (we can still set $Y_i = Z_i + \bar{Y}$, with $\bar{Z} = 0$). We also denote

$$\mu_4 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{Y})^4.$$

Preliminary calculations

We will subsequently use the following four results:

1. If $\bar{Y} = 0$, then

$$\frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^2 y_j^2 = \sigma_y^4 - \frac{\mu_4}{N}. \quad (2.7)$$

In fact,

$$\frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^2 y_j^2 = \frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} y_i^2 y_j^2 - \frac{1}{N^2} \sum_{i \in U} y_i^4 = \sigma_y^4 - \frac{\mu_4}{N}.$$

2. If $\bar{Y} = 0$, then

$$\frac{1}{N} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^3 y_j = -\mu_4. \quad (2.8)$$

In fact, seeing as $\sum_{j \in U} y_j = 0$,

$$\frac{1}{N} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^3 y_j = \frac{1}{N} \sum_{i \in U} \sum_{j \in U} y_i^3 y_j - \frac{1}{N} \sum_{i \in U} y_i^4 = -\mu_4.$$

3. If $\bar{Y} = 0$, then

$$\frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} y_i^2 y_j y_k = \frac{2\mu_4}{N} - \sigma_y^4. \quad (2.9)$$

Indeed, as

$$\begin{aligned} \frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} \sum_{k \in U} y_i^2 y_j y_k &= 0 = \frac{1}{N^2} \sum_{i \in U} y_i^4 + \frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^2 y_j^2 \\ &\quad + \frac{2}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^3 y_j + \frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} y_i^2 y_j y_k, \end{aligned}$$

with the results from (2.7) and (2.9) we have:

$$0 = \frac{\mu_4}{N} + \sigma_y^4 - \frac{\mu_4}{N} - 2\frac{\mu_4}{N} + \frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} y_i^2 y_j y_k.$$

Therefore,

$$\frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} y_i^2 y_j y_k = \frac{2\mu_4}{N} - \sigma_y^4.$$

4. If $\bar{Y} = 0$, then

$$\frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} \sum_{\substack{\ell \in U \\ \ell \neq i \\ \ell \neq j \\ \ell \neq k}} y_i y_j y_k y_\ell = -3 \left(\frac{2\mu_4}{N} - \sigma_y^4 \right).$$

In fact, since

$$\begin{aligned} & \frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} \sum_{k \in U} \sum_{\ell \in U} y_i y_j y_k y_\ell = 0 \\ &= \frac{1}{N^2} \sum_{i \in U} y_i^4 + \frac{3}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^2 y_j^2 + \frac{4}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^3 y_j \\ &+ \frac{6}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} y_i^2 y_j y_k + \frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} \sum_{\substack{\ell \in U \\ \ell \neq i \\ \ell \neq j \\ \ell \neq k}} y_i y_j y_k y_\ell, \end{aligned}$$

by the results from (2.7), (2.8), and (2.9), we have

$$\begin{aligned} 0 &= \frac{\mu_4}{N} + 3 \left(\sigma_y^4 - \frac{\mu_4}{N} \right) - \frac{4\mu_4}{N} + 6 \left(\frac{2\mu_4}{N} - \sigma_y^4 \right) \\ &+ \frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} \sum_{\substack{\ell \in U \\ \ell \neq i \\ \ell \neq j \\ \ell \neq k}} y_i y_j y_k y_\ell. \end{aligned}$$

Thus

$$\frac{1}{N^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} \sum_{\substack{\ell \in U \\ \ell \neq i \\ \ell \neq j \\ \ell \neq k}} y_i y_j y_k y_\ell = -3 \left(\frac{2\mu_4}{N} - \sigma_y^4 \right).$$

These preliminary calculations will be used to calculate the variance which can be divided into two parts according to:

$$\text{var}(s_y^2) = E(s_y^4) - \{E(s_y^2)\}^2.$$

Since $E(s_y^2)$ is given by (2.6), we must calculate

$$\begin{aligned} E(s_y^4) &= E \left(\frac{1}{n-1} \sum_{i \in S} y_i^2 - \frac{n}{n-1} \widehat{Y}^2 \right)^2 \\ &= \frac{n^2}{(n-1)^2} E \left(\frac{1}{n} \sum_{i \in S} y_i^2 - \widehat{Y}^2 \right)^2 \\ &= \frac{n^2}{(n-1)^2} (A - 2B + C), \end{aligned}$$

where

$$A = E \left(\frac{1}{n} \sum_{i \in S} y_i^2 \right)^2, \quad B = E \left(\frac{1}{n} \sum_{i \in S} y_i^2 \right) \widehat{Y}^2, \quad \text{and} \quad C = E \left(\widehat{Y}^4 \right).$$

Calculation of the 3 terms A, B and C

1. Calculation of A

$$\begin{aligned} A &= E \left(\frac{1}{n} \sum_{i \in S} y_i^2 \right)^2 = E \left(\frac{1}{n^2} \sum_{i \in S} y_i^4 + \frac{1}{n^2} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} y_i^2 y_j^2 \right) \\ &= \frac{1}{n^2} \sum_{i \in U} y_i^4 \pi_1 + \frac{1}{n^2} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^2 y_j^2 \pi_2. \end{aligned}$$

By Result (2.7),

$$A = \frac{N\pi_1\mu_4}{n^2} + \frac{N^2\pi_2}{n^2} \left(\sigma_y^4 - \frac{\mu_4}{N} \right) = \frac{N(\pi_1 - \pi_2)}{n^2} \mu_4 + \frac{N^2\pi_2}{n^2} \sigma_y^4. \quad (2.10)$$

2. Calculation of B

$$\begin{aligned} B &= E \left(\frac{1}{n} \sum_{i \in S} y_i^2 \right) \widehat{Y}^2 = E \left(\frac{1}{n^3} \sum_{i \in S} \sum_{j \in S} \sum_{k \in S} y_i^2 y_j y_k \right) \\ &= E \left(\frac{1}{n^3} \sum_{i \in S} y_i^4 \right) + E \left(\frac{1}{n^3} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} y_i^2 y_j^2 \right) \\ &\quad + E \left(\frac{1}{n^3} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} \sum_{\substack{k \in S \\ k \neq i, k \neq j}} y_i^2 y_j y_k \right) + E \left(\frac{2}{n^3} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} y_i^3 y_j \right) \\ &= \frac{1}{n^3} \sum_{i \in U} y_i^4 \pi_1 + \frac{1}{n^3} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^2 y_j^2 \pi_2 \\ &\quad + \frac{1}{n^3} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i, k \neq j}} y_i^2 y_j y_k \pi_3 + \frac{2}{n^3} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^3 y_j \pi_2. \end{aligned}$$

Through Results (2.7), (2.8), and (2.9), we get:

$$\begin{aligned}
 B &= \frac{A}{n} + \frac{N^2\pi_3}{n^3} \left(\frac{2\mu_4}{N} - \sigma_y^4 \right) - \frac{2N\pi_2}{n^3} \mu_4 \\
 &= \frac{A}{n} + \frac{2N(\pi_3 - \pi_2)}{n^3} \mu_4 - \frac{N^2\pi_3}{n^3} \sigma_y^4 \\
 &= \frac{N(\pi_1 - 3\pi_2 + 2\pi_3)}{n^3} \mu_4 + \frac{N^2(\pi_2 - \pi_3)}{n^3} \sigma_y^4. \quad (2.11)
 \end{aligned}$$

3. Calculation of C

$$\begin{aligned}
 C &= E \left(\widehat{Y}^4 \right) \\
 &= E \left(\frac{1}{n^4} \sum_{i \in S} \sum_{j \in S} \sum_{k \in S} \sum_{\ell \in S} y_i y_j y_k y_\ell \right) \\
 &= E \left(\frac{1}{n^4} \sum_{i \in S} y_i^4 \right) + E \left(\frac{3}{n^4} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} y_i^2 y_j^2 \right) + E \left(\frac{4}{n^4} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} y_i^3 y_j \right) \\
 &\quad + E \left(\frac{6}{n^4} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} \sum_{\substack{k \in S \\ k \neq i \\ k \neq j}} y_i^2 y_j y_k \right) + E \left(\frac{1}{n^4} \sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} \sum_{\substack{k \in S \\ k \neq i \\ k \neq j}} \sum_{\substack{\ell \in S \\ \ell \neq i \\ \ell \neq j \\ \ell \neq k}} y_i y_j y_k y_\ell \right)
 \end{aligned}$$

By calculating the expectations, we have

$$\begin{aligned}
 C &= \frac{1}{n^4} \sum_{i \in U} y_i^4 \pi_1 + \frac{3}{n^4} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^2 y_j^2 \pi_2 + \frac{4}{n^4} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} y_i^3 y_j \pi_2 \\
 &\quad + \frac{6}{n^4} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} y_i^2 y_j y_k \pi_3 + \frac{1}{n^4} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \sum_{\substack{k \in U \\ k \neq i \\ k \neq j}} \sum_{\substack{\ell \in U \\ \ell \neq i \\ \ell \neq j \\ \ell \neq k}} y_i y_j y_k y_\ell \pi_4.
 \end{aligned}$$

Finally, by Results (2.7), (2.8), and (2.9), we get:

$$\begin{aligned}
 C &= \frac{N\pi_1}{n^4} \mu_4 + \frac{3N^2\pi_2}{n^4} \left(\sigma_y^4 - \frac{\mu_4}{N} \right) - \frac{4N\pi_2}{n^4} \mu_4 \\
 &\quad + \frac{6N^2\pi_3}{n^4} \left(\frac{2\mu_4}{N} - \sigma_y^4 \right) - \frac{3N^2\pi_4}{n^4} \left(\frac{2\mu_4}{N} - \sigma_y^4 \right) \\
 &= \frac{N(\pi_1 - 7\pi_2 + 12\pi_3 - 6\pi_4)}{n^4} \mu_4 + \frac{3N^2(\pi_2 - 2\pi_3 + \pi_4)}{n^4} \sigma_y^4. \quad (2.12)
 \end{aligned}$$

From Expressions (2.10), (2.11), (2.12) and (2.6), we finally have the variance of the estimator of the population variance.

$$\begin{aligned}
& \text{var}(s_y^2) \\
&= \frac{n^2}{(n-1)^2} (A - 2B + C) - S_y^4 \\
&= \frac{n^2}{(n-1)^2} \left\{ \frac{N(\pi_1 - \pi_2)}{n^2} \mu_4 + \frac{N^2 \pi_2}{n^2} \sigma_y^4 \right. \\
&\quad \left. - 2 \left[\frac{N(\pi_1 - 3\pi_2 + 2\pi_3)}{n^3} \mu_4 + \frac{N^2(\pi_2 - \pi_3)}{n^3} \sigma_y^4 \right] \right. \\
&\quad \left. + \frac{N(\pi_1 - 7\pi_2 + 12\pi_3 - 6\pi_4)}{n^4} \mu_4 + \frac{3N^2(\pi_2 - 2\pi_3 + \pi_4)}{n^4} \sigma_y^4 \right\} - S_y^4 \\
&= \frac{N(N-n)}{n(n-1)(N-1)^2(N-2)(N-3)} \\
&\quad \times \left\{ \mu_4(N-1)[N(n-1) - (n+1)] - \sigma_y^4 [N^2(n-3) + 6N - 3(n+1)] \right\}.
\end{aligned} \tag{2.13}$$

With simple random sampling, we estimate the sampling variance by:

$$\widehat{\text{var}}(\widehat{Y}) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n},$$

an estimator that has the sampling variance:

$$\text{var}(\widehat{\text{var}}(\widehat{Y})) = \left(1 - \frac{n}{N}\right)^2 \frac{1}{n^2} \text{var}(s_y^2),$$

where $\text{var}(s_y^2)$ is defined in (2.13). So, this expression is surprisingly complex for a problem that *a priori* had appeared to be simple. If N approaches toward infinity (in practice N is ‘very large’), we get the valuable expression for a design with replacement:

$$\text{var}(s_y^2) \approx \frac{1}{n} \left\{ \mu_4 - \frac{n-3}{n-1} \sigma_y^4 \right\}. \tag{2.14}$$

If the variable y has a normal distribution in population U , then we know furthermore that $\mu_4 = 3\sigma_y^4$, and we get

$$\text{var}(s_y^2) \approx \frac{2\sigma_y^4}{n-1}.$$

Finally, in the case:

$$\text{var}(\widehat{\text{var}}(\widehat{Y})) \approx \left(1 - \frac{n}{N}\right)^2 \frac{1}{n^2} \frac{2\sigma_y^4}{n-1}.$$

The standard deviation of $\widehat{\text{var}}(\widehat{Y})$ varies by $1/n^{3/2}$. If n is large, this standard deviation is \sqrt{n} times smaller than $\widehat{\text{var}}(\widehat{Y})$: this is the reason for which in practice we content ourselves with the calculation of $\widehat{\text{var}}(\widehat{Y})$, which we judge to be sufficiently accurate.



<http://www.springer.com/978-0-387-26127-0>

Sampling Methods

Exercises and Solutions

Ardilly, P.; Tillé, Y.

2006, XII, 384 p., Softcover

ISBN: 978-0-387-26127-0