Sampling Methods

Exercises and Solutions

Pascal Ardilly
Yves Tillé

Translated from French by Leon Jang

Sampling Methods
Exercises and Solutions

Springer

Pascal Ardilly
INSEE Direction générale
Unité des Méthodes Statistiques,
Timbre F410
18 boulevard Adolphe Pinard
75675 Paris Cedex 14
France
Email: pascal.ardilly@insee.fr

Yves Tillé
Institut de Statistique,
Université de Neuchâtel
Espace de l'Europe 4, CP 805,
2002 Neuchâtel
Switzerland
Email: yves.tille@unine.ch

Library of Congress Control Number: 2005927380

ISBN-10: 0-387-26127-3
ISBN-13: 978-0387-26127-0

Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springeronline.com

Preface

When we agreed to share all of our preparation of exercises in sampling theory to create a book, we were not aware of the scope of the work. It was indeed necessary to compose the information, type out the compilations, standardise the notations and correct the drafts. It is fortunate that we have not yet measured the importance of this project, for this work probably would never have been attempted!

In making available this collection of exercises, we hope to promote the teaching of sampling theory for which we wanted to emphasise its diversity. The exercises are at times purely theoretical while others are originally from real problems, enabling us to approach the sensitive matter of passing from theory to practice that so enriches survey statistics.

The exercises that we present were used as educational material at the *École Nationale de la Statistique et de l'Analyse de l'Information* (ENSAI), where we had successively taught sampling theory. We are not the authors of all the exercises. In fact, some of them are due to Jean-Claude Deville and Laurent Wilms. We thank them for allowing us to reproduce their exercises. It is also possible that certain exercises had been initially conceived by an author that we have not identified. Beyond the contribution of our colleagues, and in all cases, we do not consider ourselves to be the lone authors of these exercises: they actually form part of a common heritage from ENSAI that has been enriched and improved due to questions from students and the work of all the demonstrators of the sampling course at ENSAI.

We would like to thank Laurent Wilms, who is most influential in the organisation of this practical undertaking, and Sylvie Rousseau for her multiple corrections of a preliminary version of this manuscript. Inès Pasini, Yves-Alain Gerber and Anne-Catherine Favre helped us over and over again with typing and composition. We also thank ENSAI, who supported part of the scientific typing. Finally, we particularly express our gratitude to Marjolaine Girin for her meticulous work with typing, layout and composition.

Pascal Ardilly and Yves Tillé

Contents

1

Introduction

1.1 References

This book presents a collection of sampling exercises covering the major chapters of this branch of statistics. We do not have as an objective here to present the necessary theory for solving these exercises. Nevertheless, each chapter contains a brief review that clarifies the notation used. The reader can consult more theoretical works. Let us first of all cite the books that can be considered as classics: Yates (1949), Deming (1950), Hansen et al. (1993a), Hansen et al. (1993b), Deming (1960), Kish (1965), Raj (1968), Sukhatme and Sukhatme (1970), Konijn (1973), Cochran (1977), a simple and clear work that is very often cited as a reference, and Jessen (1978). The *post-mortem* work of Hájek (1981) remains a masterpiece but is unfortunately difficult to understand. Kish (1989) offered a practical and interesting work which largely transcends the agricultural domain. The book by Thompson (1992) is an excellent presentation of spatial sampling. The work devoted to the basics of sampling theory has been recently republished by Cassel et al. (1993). The modern reference book for the past 10 years remains the famous Särndal et al. (1992), even if other interesting works have been published like Hedayat and Sinha (1991), Krishnaiah and Rao (1994), or the book Valliant et al. (2000), dedicated to the model-based approach. The recent book by Lohr (1999) is a very pedagogical work which largely covers the field. We recommend it to discover the subject. We also cite two works exclusively established in sampling with unequal probabilities: Brewer and Hanif (1983) and Gabler (1990), and the book by Wolter (1985) being established in variance estimation.

In French, we can suggest in chronological order the books by Thionet (1953) and by Zarkovich (1966) as well as that by Desabie (1966), which are now classics. Then, we can cite the more recent books by Deroo and Dussaix (1980), Gouriéroux (1981), Grosbras (1987), the collective work edited by Droesbeke et al. (1987), the small book by Morin (1993) and finally the manual of exercises published by Dussaix and Grosbras (1992). The 'Que Sais-je?' by Dussaix and Grosbras (1996) expresses an appreciable translation of the

theory. Obviously, the two theoretical works proposed by the authors Ardilly (1994) and Tillé (2001) are fully adapted to go into detail on the subject. Finally, a very complete work is suggested, in Italian, by Cicchitelli et al. (1992) and, in Chinese, by Ren and Ma (1996).

1.2 Population, variable and function of interest

Consider a finite population composed of N observation units; each of the units can be identified by a *label*, of which the set is denoted

$$U = \{1, ..., N\}.$$

We are interested in a variable y which takes the value y_k on unit k. These values are not random. The objective is to estimate the value of a function of interest

$$\theta = f(y_1, ..., y_k, ..., y_N).$$

The most frequent functions are the total

$$Y = \sum_{k \in U} y_k,$$

the mean

$$\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{Y}{N},$$

the population variance

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} \left(y_k - \overline{Y}\right)^2,$$

and the corrected population variance

$$S_y^2 = \frac{1}{N-1} \sum_{k \in U} \left(y_k - \overline{Y}\right)^2.$$

The size of the population is not necessarily known and can therefore be considered as a total to estimate. In fact, we can write

$$N = \sum_{k \in U} 1.$$

1.3 Sample and sampling design

A sample without replacement s is a subset of U. A sampling design $p(.)$ is a probability distribution for the set of all possible samples such that

$$p(s) \geq 0, \text{ for all } s \subset U \text{ and } \sum_{s \subset U} p(s) = 1.$$

The random sample S is a random set of labels for which the probability distribution is

$$\Pr(S = s) = p(s), \text{ for all } s \subset U.$$

The sample size $n(S)$ can be random. If the sample is of fixed size, we denote the size simply as n. The indicator variable for the presence of units in the sample is defined by

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

The inclusion probability is the probability that unit k is in the sample

$$\pi_k = \Pr(k \in S) = \mathrm{E}(I_k) = \sum_{s \ni k} p(s).$$

This probability can (in theory) be deduced from the sampling design. The second-order inclusion probability is

$$\pi_{k\ell} = \Pr(k \in S \text{ and } \ell \in S) = \mathrm{E}(I_k I_\ell) = \sum_{s \ni k,\ell} p(s).$$

Finally, the covariance of the indicators is

$$\Delta_{k\ell} = \mathrm{cov}(I_k, I_\ell) = \begin{cases} \pi_k(1 - \pi_k) & \text{if } \ell = k \\ \pi_{k\ell} - \pi_k \pi_\ell & \text{if } \ell \neq k. \end{cases} \tag{1.1}$$

If the design is of fixed size n, we have

$$\sum_{k \in U} \pi_k = n, \quad \sum_{k \in U} \pi_{k\ell} = n\pi_\ell, \quad \text{and} \quad \sum_{k \in U} \Delta_{k\ell} = 0.$$

1.4 Horvitz-Thompson estimator

The Horvitz-Thompson estimator of the total is defined by

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

This estimator is unbiased if all the first-order inclusion probabilities are strictly positive. If the population size is known, we can estimate the mean with the Horvitz-Thompson estimator:

$$\widehat{\overline{Y}}_\pi = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}.$$

The variance of $\widehat{Y}_\pi$ is

$$\mathrm{var}(\widehat{Y}_\pi) = \sum_{k\in U}\sum_{\ell\in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

If the sample is of fixed size ($\mathrm{var}(\#S) = 0$), then Sen (1953) and Yates and Grundy (1953) showed that the variance can also be written

$$\mathrm{var}(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{k\in U}\sum_{\ell\in U} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell}.$$

The variance can be estimated by:

$$\widehat{\mathrm{var}}(\widehat{Y}_\pi) = \sum_{k\in S}\sum_{\ell\in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}},$$

where $\pi_{kk} = \pi_k$. If the design is of fixed size, we can construct another estimator from the Sen-Yates-Grundy expression:

$$\widehat{\mathrm{var}}(\widehat{Y}_\pi) = -\frac{1}{2}\sum_{k\in S}\sum_{\substack{\ell\in S,\\ \ell\neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

These two estimators are unbiased if all the second-order inclusion probabilities are strictly positive. When the sample size is 'sufficiently large' (in practice, a few dozen most often suffices), we can construct confidence intervals with a confidence level of $(1-\alpha)$ for Y according to:

$$\mathrm{CI}(1-\alpha) = \left[\widehat{Y}_\pi - z_{1-\alpha/2}\sqrt{\mathrm{var}(\widehat{Y}_\pi)}, \widehat{Y}_\pi + z_{1-\alpha/2}\sqrt{\mathrm{var}(\widehat{Y}_\pi)}\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal random variable (see Tables 10.1, 10.2, and 10.3). These intervals are estimated by replacing $\mathrm{var}(\widehat{Y}_\pi)$ with $\widehat{\mathrm{var}}(\widehat{Y}_\pi)$.

2

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,

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

The Horvitz-Thompson estimator of the total becomes

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

That for the mean is written as

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

The variance of $\widehat{Y}_\pi$ is

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

and its unbiased estimator

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

where

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

The Horvitz-Thompson estimator of the proportion P_D that represents a subpopulation 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

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

is unbiased, and its variance is

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

In a simple design with replacement, the sample variance

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

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

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

is unbiased for the mean and has a smaller variance than that of $\widehat{\overline{Y}}_{WR}$. 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	$\widehat{\overline{Y}} = \frac{1}{n}\sum_{k\in S} y_k$	$\widehat{\overline{Y}}_{WR} = \frac{1}{m}\sum_{i=1}^{m} \tilde{y}_i$
Variance of the mean estimator	$\text{var}\left(\widehat{\overline{Y}}\right) = \frac{(N-n)}{nN} S_y^2$	$\text{var}\left(\widehat{\overline{Y}}_{WR}\right) = \frac{\sigma_y^2}{m}$
Expected sample variance	$\text{E}\left(s_y^2\right) = S_y^2$	$\text{E}\left(\tilde{s}_y^2\right) = \sigma_y^2$
Variance estimator of the mean estimator	$\widehat{\text{var}}\left(\widehat{\overline{Y}}\right) = \frac{(N-n)}{nN} s_y^2$	$\widehat{\text{var}}\left(\widehat{\overline{Y}}_{WR}\right) = \frac{\tilde{s}_y^2}{m}$

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
$$\overline{Y} = \frac{1}{N}\sum_{k\in U} y_k.$$
2. Give a 95 % confidence interval for $\overline{Y}$.

Solution

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

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

The estimator of the dispersion S_y^2 is

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

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

$$\left[\widehat{\overline{Y}} \pm 1.96\sqrt{\frac{N-n}{N}\frac{s_y^2}{n}}\right] = \left[29.07 \pm 1.96\sqrt{\frac{2010-100}{2010} \times \frac{707.45}{100}}\right]$$
$$= [23.99; 34.15]\,.$$

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

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

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

$$\mathrm{CI}(1-\alpha) = \left[\widehat{P}_{WR} - z_{1-\alpha/2}\sqrt{\frac{\widehat{P}_{WR}(1-\widehat{P}_{WR})}{m-1}},\right.$$
$$\left.\widehat{P}_{WR} + z_{1-\alpha/2}\sqrt{\frac{\widehat{P}_{WR}(1-\widehat{P}_{WR})}{m-1}}\right].$$

Indeed, in this case,

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

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

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

By dividing by two and squaring, we get

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

which gives

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

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

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

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

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

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

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

By dividing by two and by squaring, we get

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

which gives

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

or again

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

or

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

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 $\widehat{P}_{WOR}$. 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
$$\Pr(k \in S) = 1 - \left(1 - \frac{1}{N}\right)^m = 1 - \sum_{j=0}^{m} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j}$$
$$= 1 - \left\{ \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} - \frac{m}{N} + 1 \right\} = \frac{m}{N} - \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j}$$
$$= \frac{m}{N} + O\left(\frac{m^2}{N^2}\right).$$
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:

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

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,

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

Therefore,

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

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:

$$P_1 = \text{proportion of individuals owning a dishwasher,}$$
$$P_2 = \text{proportion of individuals owning a laptop computer.}$$

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 ,

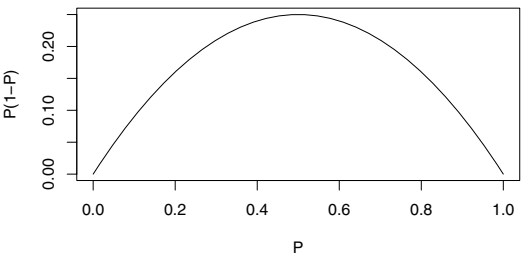
$$\max_{45\,\% \leq P_1 \leq 65\,\%} P_1(1 - P_1) = 0.5(1 - 0.5) = 0.25,$$

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



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

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

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

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:

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

A second manner is:

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

The unbiased estimator of σ_y^2 is

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

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

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

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

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

where

$$\widehat{\overline{Y}} = \frac{1}{n} \sum_{k \in S} y_k.$$

Therefore

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

We get

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

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

2. The same question, if we this time want to estimate an average price during the combined May-June period.
3. 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

1. We denote $\overline{p}_m$ as the simple average of the recorded prices among the n stations for month m (m = May or June).
   We have:
   $$\text{var}(\overline{p}_m) = \frac{1-f}{n}\ S_m^2,$$
   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 $\overline{p}_{\text{June}} - \overline{p}_{\text{May}}$ (the two estimators are calculated on two different *a priori* samples) and
     $$\text{var}_1(\overline{p}_{\text{June}} - \overline{p}_{\text{May}}) = \frac{1-f}{n}\ (S_{\text{May}}^2 + S_{\text{June}}^2).$$
     Indeed, the covariance is null because the two samples (and therefore the two estimators $\overline{p}_{\text{May}}$ and $\overline{p}_{\text{June}}$) are independent.
   - Method 2. We have only one sample (the panel). Still, we estimate the evolution of prices without bias by $\overline{p}_{\text{June}} - \overline{p}_{\text{May}}$, and
     $$\text{var}_2(\overline{p}_{\text{June}} - \overline{p}_{\text{May}}) = \frac{1-f}{n}\ (S_{\text{May}}^2 + S_{\text{June}}^2 - 2S_{\text{May, June}}).$$
     This time, there is a covariance term, with:
     $$\text{cov}\ (\overline{p}_{\text{May}},\ \overline{p}_{\text{June}}) = \frac{1-f}{n}\ S_{\text{May, June}},$$
     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:
     $$\frac{\text{var}_1(\overline{p}_{\text{June}} - \overline{p}_{\text{May}})}{\text{var}_2(\overline{p}_{\text{June}} - \overline{p}_{\text{May}})} = \frac{S_{\text{May}}^2 + S_{\text{June}}^2}{S_{\text{May}}^2 + S_{\text{June}}^2 - 2S_{\text{May, June}}}.$$

After calculating, we find:

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

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:

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

- Method 1:

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

- Method 2:

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

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

In conclusion, we have

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

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 $\widehat{P}_A$ 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:

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

(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 '$\widehat{P}$ greater than a certain level'. We make the hypothesis that $\widehat{P}_A \sim \mathcal{N}(P_A,\ \sigma_A^2)$, 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:

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

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

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

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

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

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 $\ell < k$,

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

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

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

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

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

   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,

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

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