FAO
FISHERIES AND
AQUACULTURE
TECHNICAL
PAPER

434

ISSN 2070-7010

Sampling methods applied to fisheries science: a manual

COPEMED

FAO FIAT PANIS

Cover:
Illustration representing port sampling site. Emanuela D’Antoni.

Sampling methods applied to fisheries science: a manual

FAO
FISHERIES
TECHNICAL
PAPER
434

by
Emygdio Landerset Cadima
Consultant
Lisbon, Portugal

Ana Maria Caramelo
Fishery Resources Officer
FAO Fisheries Department
Rome, Italy

Manuel Afonso-Dias
and **Pedro Conte de Barros**
Professors
University of Algarve
Faro, Portugal

Merete O. Tandstad
and **Juan Ignacio de Leiva-Moreno**
Fishery Resources Officers
FAO Fisheries Department
Rome, Italy

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
Rome, 2005

The designations employed and the presentation of material in this information product do not imply the expression of any opinion whatsoever on the part of the Food and Agriculture Organization of the United Nations concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

ISBN 92-5-105397-9

Preparation of this document

This document is based on a series of lectures, seminars and working groups held at different venues, mainly in Europe, Latin America and Africa. The contents of these courses were gradually developed by Professor Emygdio Landerset Cadima and a number of academic contributors.

The project "Advice, technical support and establishment of cooperation networks to facilitate coordination to support fisheries management in the Mediterranean" (COPEMED GCP/REM/057/SPA) promoted two courses on sampling methods applied to fisheries. The contents of this manual were successfully tested in these COPEMED courses held in Malta and Morocco.

This manual deals primarily with the theoretical aspects of the most-used models for collecting data for fish stock assessment. The practical application is considered to be a complementary part to aid understanding of the theory.

The initial version of this document was reworked by Cristina Morgado IPIMAR researcher. Acknowledgements are due to Stephen Cofield and Marie Thérèse Magnan for their help editing this document in English. Merete Tandstad, Pedro Barros and Ana Maria Caramelo reviewed the final copy.

Abstract

The main objective of this manual is to present the basic and standard concepts of sampling methods applied to fisheries science. In order to ensure sound fisheries research, it is essential to have reliable data from landing ports, fishery stocks and research surveys. A rational management of fishing resources can then be established to ensure a sustainable exploitation rate and responsible fisheries management, providing long term benefits for all.

This document is divided into nine chapters. Chapter 1 provides an introduction to sampling theory. Chapter 2 introduces the theory of the three worlds (population, sample and sampling) as well as a short revision of probability concepts.

Chapters 3 to 6 provide an overview of the simple random, random stratified, cluster and two-stage sampling methods. The expressions for estimating the mean and total of the populations, their sampling distributions, the expected values, the sampling variances and their estimates are included and justified for each of the sampling designs.

Chapter 7 presents a case study of biological sampling from landing ports. Chapter 8, an essential part of the manual, provides exercises that should be used to further understand the objectives of sampling and its advantages to fishery resource studies. Finally, Chapter 9 provides possible solutions to those exercises.

Cadima, E.L.; Caramelo, A.M.; Afonso-Dias, M.; Conte de Barros, P.;
Tandstad, M.O.; de Leiva-Moreno, J.I.
Sampling methods applied to fisheries science: a manual.
FAO Fisheries Technical Paper. No. 434. Rome, FAO. 2005. 88p.

Contents

Tables

Figures

1. Introduction

To sample is to collect part of the elements of a set. The elements will be the sampling units, part of the elements will form the sample and the full set is the population. Why is sampling important? Sampling is important because most of the time it is impossible, difficult or expensive to observe all the elements of a population. If the samples are selected with an adequate criterion, it is possible to measure the precision of the conclusions or inferences about that population.

In fisheries research, the objective of sampling the fisheries is to obtain data from the stocks and their exploitation, to analyse the characteristics of the resources, the effects of exploitation on the abundance of these resources and to determine appropriate fishing levels to obtain the best possible catches at present and during future years.

In the sampling process it is convenient to distinguish between population, sample and sampling. Figure 1.1 illustrates the three different "worlds" of the whole sampling process and the relationships among them.

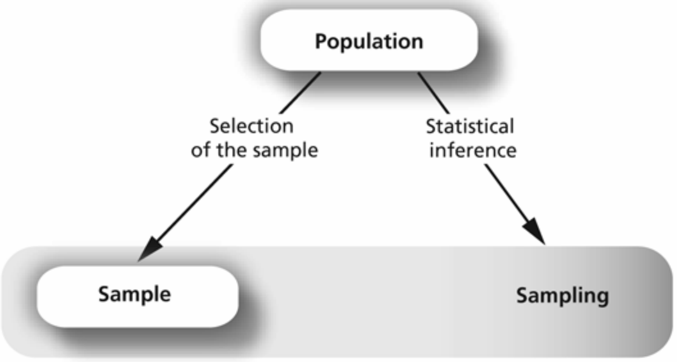
The first "world" of the sampling process is the population. The population "world" is entirely or partly unknown. Its elements can be of various types but they should be well-defined (e.g. the boats of one fleet, sardines landed in a fishing harbour in one day, etc.).

The second "world" of the sampling process is the sample. The sample "world" is completely known (in this aspect the sample is totally different from the population). It is from the sample data that the characteristics of the population will be estimated and, for this purpose, the selection of the sample has to be made with a well-defined criterion.

The third "world" is the sampling "world". The sampling "world" is the set of all samples with the **same size** that could be selected **with the same criterion** from the population. It is from the properties of the sampling "world", and based on the values of the characteristic of interest in the sample selected, that statistical inference can be carried out with pre-defined precision.

Population parameters summarize or characterize the distribution of the population values. The corresponding values in the sample are called statistics.

FIGURE 1.1
The three worlds of the sampling process



Statistical inference is based on the distribution of a given statistic in the sampling. Sampling distributions are also characterized by parameters. It should be noted that often parameters of the population, statistics of the sample and parameters of the sampling distribution share the same name, e.g. the population mean, the sample mean and the sampling mean. However, practitioners should be aware that each of these represents a different quantity with different properties, and the "world" they refer to should always be made explicit.

The sampling "world" of a statistic is formed by calculating the values of this statistic in all possible samples that could be selected from that population, with the same pre-defined criterion. Sampling distributions are probability distributions. The actual sampling distribution of its values is unknown, but often the expected theoretical distribution, and its main properties, can be derived from the knowledge of the sampling process and of the statistic being considered. Sample statistics that are used in the process of estimation of population parameters are called estimators. The sampling distribution of an estimator is the basis for all statistical inference from the sample to the population regarding this estimator. Namely, it is based on these probability distributions of the estimators in the sampling that the precision of the estimation can be evaluated.

There are several different methods for selecting a sample. The most usual are simple random sampling, stratified random sampling and cluster sampling. In fisheries research, multistage sampling, a combination of several of the basic methods, is also commonly used.

In simple random sampling each possible sample has the same probability of being selected. There are two ways of selecting a simple random sample, with or without replacement of the sampling units selected. Simple random sampling is not frequently used in fisheries research, except as part of more complex methods.

In stratified random sampling the whole population is divided into sub-populations, called *strata*. A sample is selected using a random design within each *stratum*. Stratified sampling is usually applied to biological sampling of the landings and in scientific surveys.

In cluster sampling the population is also partitioned into groups, which are designated clusters. Each cluster contains one or more elements, but it is the clusters and not the elements that are the sampling units. Cluster sampling has been used in fisheries to estimate landings per trip from data of artisanal fisheries with many landing sites (beaches) and a small number of vessels operating from each site (beach).

Multistage sampling is a combination of the various methods previously mentioned. At each stage, there is a random selection of the sampling units, which can be elements or clusters. Two-stage sampling and a particular case of three-stage sampling are discussed in this manual.

In systematic sampling all the elements of the population are grouped into classes of the same size. The first element to be sampled is chosen randomly from one class. The remaining sampling units occupy the same relative position in each class. For instance, if the third element of the first class were selected then all the third elements of the other classes would also be chosen. In this method, all classes have one element sampled, and therefore, the size of the sample is equal to the number of classes. This method is not discussed further in this manual.

Fisheries research is most often concerned with the estimation of population mean and totals. The estimation of the proportion of the population that shares some characteristic of interest, e.g the proportion of vessels in a small-scale fishery that make more than two trips per day, is also a common task in fisheries research. The sampling distribution and properties of several estimators, mainly of the estimators of population means and totals, and sometimes of proportions, are discussed in the next chapters of this technical paper.

2. The three worlds of sampling

2.1 THE POPULATION WORLD

Let us consider a population of several elements, where $\mathrm{Y_i}$ is the value in element *i* of a characteristic, represented by the variable *Y*. For example, if the total length of sardines is taken from a landing box, the characteristic, *Y*, can be the total length of a sardine, and the i^{th} measure will have length $Y_{i.}$ In this case *Y* is a continuous variable. Another characteristic could be the age of the sardines, as it is measured in fisheries. In this case age is considered a discrete variable, *Y*, which can take the values 0, 1, 2, ..., i,...

The distribution of the values of a characteristic in the population can be represented in the form of a list, a table, a function, a graph, etc. The distribution may be characterized by parameters, for instance, the mean, the variance, the standard deviation, the quantiles, etc. The population is usually unknown and, therefore, these parameters cannot be calculated.

Greek alphabet letters or Latin alphabet upper case letters will be used to denote the parameters of the population world.

Total and mean values

Populations can be finite or infinite. The total number of elements of a finite population is the size of the population and is denoted by N. When the number of elements of the population is very large, N can be considered as infinite. For example, the population of sardines landed in a country during one year is finite, but for some statistical purposes the population can be considered infinite.

The population mean of a characteristic *Y* is represented by $\overline{Y}$ or μ.

If the population is finite the total value of the characteristic Y will be:

$$Y = \sum_{i=1}^{N} Y_i$$

Then the mean will be:

$$\overline{Y} = \frac{Y}{N} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

In this case the total value can be expressed as:

$$Y = N\overline{Y}$$

Note that *Y* denotes not only the variable, but also the population total value.

Dispersion measures

Several measures of dispersion of the values of the characteristic in the population can be defined. The variance, the standard deviation, the coefficient of variation and the range are the most common ones.

The population variance of the characteristic *Y* is represented by σ^2.

In order to define variance let us consider the deviation of a value Y_i to the mean, that is:

$$Y_i - \overline{Y}$$

For finite populations the sum of squares of the deviations is represented by:

$$SS = \sum_{i=1}^{N} \left(Y_i - \overline{\mathrm{Y}}\right)^2$$

The variance is then defined as:

$$\sigma^2 = \frac{SS}{N}$$

A modified variance, S^2, is introduced in some sampling manuals, with the purpose of simplifying formulas and keeping the parallelism between the formulas in the population and the corresponding formulas in the samples.

S^2 is defined as:

$$S^2 = \frac{SS}{N-1}$$

Note that $(N-1)S^2 = N\sigma^2$. The two variances are practically the same for large population sizes.

The population standard deviation of the characteristic Y is represented by σ (or S) and it is defined as $\sigma = \sqrt{\sigma^2}$ (or $S = \sqrt{S^2}$). The standard deviation is also a measure of dispersion. Compared with the variance, it has the advantage of being expressed in the same units as the variable, but the variance is preferred in most cases for theoretical reasons.

The coefficient of variation is defined as $CV = \frac{S}{\overline{Y}}$ and it is a relative measure of dispersion, making it possible to compare the dispersions of two populations with very different absolute values. For example, the lengths of sardines and the lengths of some tuna species, have standard deviations with different absolute values, but in terms of *CV*s, that is values relative to the means, the dispersions can be comparable.

The range, i.e., the difference between the larger and the smaller value of the population, is also a dispersion measure that can be useful in some cases.

Proportions

Some characteristics can be classified into two categories. For instance fish maturity can be classified into “adults” and “not adults”. In these cases the proportions of the total population elements that belong to one or the other category as well as the number of elements in each category are the parameters to be estimated. This type of characteristic is qualitative, that is, the characteristic of the elements is not measured but its quality (to belong or not to belong to a category) is observed. These characteristics are called attributes.

An attribute can be represented by a variable *Y*, which takes the value 1 if the element belongs to a category and 0 otherwise.

Let *N* be the number of elements in a finite population. The proportion of elements of the population that belongs to the category is represented by *P* and the proportion of the elements that does not belong to the category is *Q=1-P*. The product *NP* is the total number of elements belonging to the category and *NQ* is the total number of elements that do not belong to that category. Then the characteristic *Y* can be represented in the form:

$$Y = \begin{Bmatrix} 1 & P \\ 0 & Q \end{Bmatrix} \text{ with } P + Q = 1$$

The population total value of the attribute Y is:

$$Y = (1 \times NP) + (0 \times NQ) = NP$$

The population mean of the characteristic Y is $\overline{Y}$ and can be calculated by:

$$\overline{Y} = \frac{Y}{N} = P$$

It is important to note that the proportion P of elements belonging to the category of interest is given by the mean of the characteristic Y. This result simplifies greatly the analysis of proportions, as most results can be obtained directly from those for mean values.

The population variance is obtained from:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N} = \frac{(1-P)^2 NP + (0-P)^2 NQ}{N}$$

and hence $\sigma^2 = PQ$

The modified population variance:

$S^2 = \frac{N}{N-1}\sigma^2$ is $S^2 = \frac{N}{N-1}PQ$ therefore

The population standard deviation will be: $\sigma = \sqrt{\sigma^2}$ or $S = \sqrt{S^2}$

Note that as previously mentioned the population is usually unknown and none of these parameters can be calculated.

2.2 THE SAMPLE WORLD

A sample of size n was drawn from a population with a variable Y. The observed value of the characteristic Y of the element i in the sample will be designated as y_i. Therefore, a sample of size n will be formed by the values $y_1, y_2, \ldots, y_n$. The observed values can be sorted by sizes, such as $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(i)} \leq \ldots \leq y_{(n)}$ where the sub-indices indicate the orders of magnitude.

When the sample size is large, the values can be grouped into classes. The classes will be denoted by $j=1, 2, \ldots, k$ where k is the total number of classes. The class interval is the difference between the upper limit y_{j+1} and the lower limit y_j, that is $y_{j+1} - y_j$ for class j.

In fisheries research, the classes should have constant intervals and the total number of classes should not be less than 12. The central value $y_{central\,j}$ of the j^{th} class is:

$$y_{central\,j} = \frac{y_j + y_{j+1}}{2}$$

The number of elements inside each class is the absolute frequency of the class. The quotient of the number of elements in each class by the total number of elements in the sample is the relative frequency, which is often expressed as a percentage (%) and sometimes as per thousand (‰). Frequencies can be accumulated and then they are called cumulative frequencies. For example, if we group the data into k classes (1, 2, 3, …, k), the cumulative frequency of the first class will be the frequency of class 1, the cumulative frequency of the second class will be the sum of the frequency of class 1 plus the frequency of class 2, and so on, up to the frequency of the last class, k, which will be the result of adding the frequencies of all classes 1, 2, ..., k and should be equal to the size of the sample considering absolute frequencies or equal to 1 considering relative frequencies. The absolute, relative or cumulative frequencies can be graphically represented by histograms.

Values calculated from the sample data are called statistics. Latin alphabet lower case letters will be used to denote the statistics of the sample world.

The statistics of location describe the central position of the values of a characteristic, while the statistics of dispersion give an idea of the dispersion of the values in the sample. Examples of statistics of location are the arithmetic mean (commonly called mean or average), the median and the mode. The range, variance, standard deviation and coefficient of variation are examples of statistics of dispersion.

The total of the observed values is designated by y and it is calculated as:

$$y = \sum_{i=1}^{n} y_i$$

Statistics of location

The mean

The arithmetic mean is the most common statistic of location. It is the quotient between the total value, y, and the sample size, n, that is:

$$\bar{y} = \frac{y}{n}$$

When the sample is organized in the form of a frequency table, the mean can be calculated as:

$$\bar{y} = \frac{\sum_{j=1}^{k} f_j y_{centralj}}{\sum_{j=1}^{k} f_j}$$ where f_j is the frequency of class j.

Note that $\sum_{j=1}^{k} f_j = n$ or $\sum_{j=1}^{k} f_j = 1$ depending on whether the f_j represent absolute or relative frequencies.

The median

In an ordered array, the median is defined as the value that separates the set of observations into two parts of equal sizes. When the sample is composed of an odd number of observations the median is the central value, that is, the element of order $\frac{n+1}{2}$. When the sample has an even number of observations the median can be calculated as the midpoint between the $\frac{n}{2}$ and the $\frac{n}{2}+1$ observations.

Quantiles

The median is just one of a family of statistics called quantiles that divide the frequency distribution into several equal parts. The quantiles are designated as *quartiles* when the number of parts is four. The first quartile cuts the frequency distribution at 25% of the total, the second quartile at 50%, this being the median, as seen before, and the third quartile at 75% of the total.

Other quantiles are also used, for instance, *deciles* (division into 10 equal parts), *percentiles* (division into 100 equal parts) and per thousand parts (division into 1000 equal parts). The *percentile* of order p would be the value of the percentile that separates the smallest $p\%$ values of the total number of the frequency distribution. For example the first quartile will be the percentile of order 25%.

The mode

The mode refers to the most frequently observed value of the sample. When the observations are grouped into classes, the class with the highest frequency is called the modal class. In this case, the central value of this class can be taken as the mode.

Some distributions present different local modes, as for instance most of the length compositions of fish landings.

Statistics of dispersion

The range

The range is the difference between the largest and the smallest observed value in the sample, *i.e.* Range = $y_{largest}$ - $y_{smallest}$.

The difference $y_{75\%} - y_{25\%}$ is called the inter-quartile range, which is another useful statistic of dispersion.

The variance

The sample variance, s^2, is the quotient between two quantities, the sum of squares (*ss*) of the deviations of each observation y_i from the arithmetic mean $\bar{y}$ and the size of the sample minus one:

$$ss = \sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ and } s^2 = \frac{ss}{n-1}$$

There are other expressions to calculate the sum of squares (*ss*):

$$ss = \sum_{i=1}^{n} y_i(y_i - \bar{y})$$

$$ss = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

$$ss = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

The standard deviation

The standard deviation is the square root of the variance: $s = \sqrt{s^2}$

The coefficient of variation

The coefficient of variation is the quotient between the standard deviation and the arithmetic mean:

$$cv = \frac{s}{\bar{y}}$$

The parallelism between the sample statistics and the finite population parameters should be noted.

Proportions

As in the population, the proportion of elements of the sample belonging to the category of interest can be calculated for a sample taken from that population. The characteristic of interest, y_i is then such that $y_i = 1$ if element i belongs to the category or $y_i = 0$ if it does not.

Under these conditions, the proportion of the sample elements belonging to the category is defined by the sample mean of y,

$$p = \frac{\sum_{i=1}^{n} y_i}{n}$$

The relation $p + q = 1$ is always valid and can be used to calculate q, which is the proportion of the sample elements that do not belong to the category.

The total value of the variable in the sample is np.

The sample variance can be calculated as:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{(1-p)^2 np + (0-p)^2 nq}{n-1} = \frac{npq}{n-1}$$

and the sample standard deviation is:

$$s = \sqrt{s^2}$$

Sample statistics – example

The original measurements of total lengths (expressed in cm) of a sample of 32 sardines landed from a trip of a purse seiner are presented in the following list, arranged in increasing order:

17.3, 17.7, 17.8, 18.1, 18.1, 18.3, 18.3, 18.7, 18.7, 19.1, 19.3, 19.3, 19.3, 19.4, 19.6, 19.7, 19.7, 19.8, 19.8, 20.1, 20.1, 20.1, 20.1, 20.1, 20.2, 20.2, 20.3, 20.6, 20.6, 20.6, 21.3, 21.7

These measurements were grouped into classes of 0.5 cm interval. The results are shown in Table 2.1.

TABLE 2.1
Distribution of total length measurements (in cm)

Class Interval *(cm)*	Class Central Value *(cm)*	Absolute Frequencies	Relative Frequencies *(%)*	Cumulative Absolute Frequencies
17.0-	17.25	1	3	1
17.5-	17.75	2	6	3
18.0-	18.25	4	13	7
18.5-	18.75	2	6	9
19.0-	19.25	5	16	14
19.5-	19.75	5	16	19
20.0-	20.25	8	25	27
20.5-	20.75	3	9	30
21.0-	21.25	1	3	31
21.5-	21.75	1	3	32
Total		**32**	**100**	–

The following figures represent the histograms of relative frequencies (%) (Figure 2.1) and of the cumulative absolute frequencies (Figure 2.2).

FIGURE 2.1
Histogram of relative frequencies of total lengths

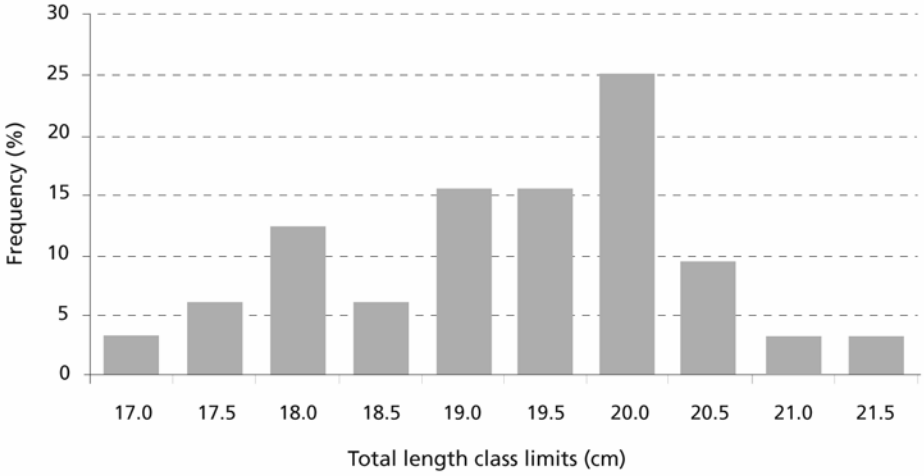
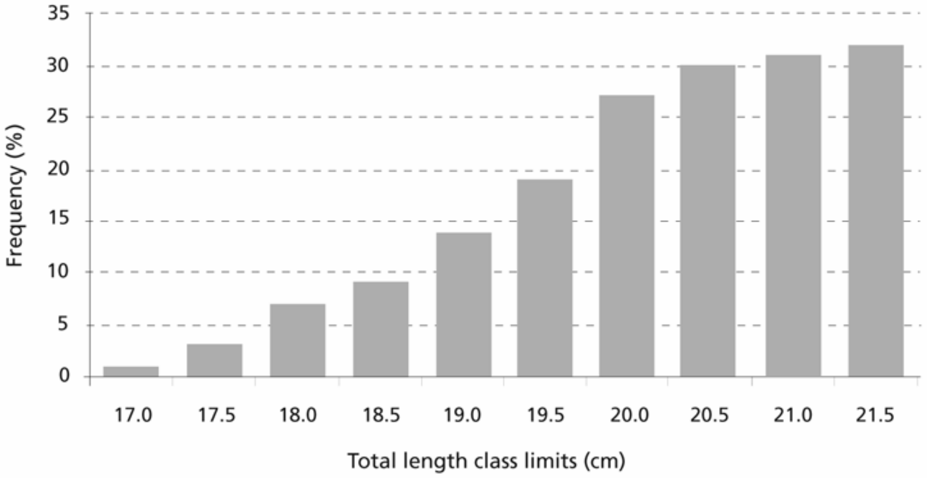


FIGURE 2.2
Histogram of the cumulative absolute frequencies of total lengths



The most common statistics of location and dispersion were calculated from the original values:

Total value: $y = \sum_{i=1}^{n} y_i$ = 624.0 cm

Arithmetic mean: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{624.0}{32}$ = 19.5 cm

Median: 19.7 cm

Modal class: 20.0 cm

Range: $y_{\text{largest}} - y_{\text{smallest}}$ = 4.4 cm

Sum of squares: $ss = \sum_{i=1}^{n} (y_i - \bar{y})^2$ = 34.74 cm^2

Variance: $s^2 = \dfrac{ss}{n-1} = \dfrac{34.74}{31}$ = 1.12 cm^2

Standard deviation: $s = \sqrt{s^2}$ = 1.05 cm

Coefficient of variation: $cv = \dfrac{s}{\bar{y}} = \dfrac{1.05}{19.50} = 0.054 = 5.4\%$

Proportion of fish above 20 cm $p = \dfrac{5}{32} = 0.156 = 15.6\%$

2.3 THE SAMPLING WORLD

Revision of probabilities and some useful distributions

Before starting the discussion of the world of sampling it would be advantageous to review some concepts of probabilities and some probability distributions, as the sampling world is mainly a world of probabilities.

Even if not rigorous, the concept of probability will be presented in a simple and practical way.

In a sample, the relative frequencies are calculated after taking the sample and observing or measuring the characteristics, i.e., the relative frequencies are calculated *a posteriori.*

Before the extraction of the sample *a priori*, the concept of relative frequency should be replaced by a new concept, that of probability. For example, if one randomly selects a vessel of a fleet, there will be a certain probability that this vessel is a purse seiner. If we assume that 19% of the total vessels in a fishing harbour are purse seiners, the probability of randomly selecting a purse seiner will be 19% or 0.19. In that case, the properties of relative frequencies can be transformed into properties of probabilities. In summary:

- the probabilities, *P*, are numbers between 0 and 1 ($0 \le P \le 1$);
- the probabilities can be added (under certain conditions, like percentages);
- the probabilities can be multiplied (under certain conditions, like percentages).

In the theory of probabilities it is convenient to define random variables, that is, variables that have probabilities associated with their values. Mathematically, discrete and continuous variables are studied differently. Thus, the probability that a variable *X* takes a particular value, *x*, can be defined when *X* is a discrete variable. When *X* is a continuous variable, however, this probability is always equal to 0. For continuous variables, what is defined is not the probability that *X* will take the value x, but rather the probability that X will take a value within an interval of two values x_1 and x_2.

An example of a discrete variable is the age of the fishes in an age composition. The values of this variable (age) are 0, 1, 2, 3, etc. years, and are attributed to each fish otolith observed.

The probability of selecting a fish of a certain age from a box is associated with the number of fishes of that age in the box. However, age can also be an example of a continuous variable when it is taken as the time elapsed since birth up to the moment of capture. In this case, only the probability of a fish having an age within an interval of time is considered.

Probability and distribution functions of discrete variables

The probability function, also called probability mass function, *P(x)*, defines the probability that a discrete variable, *X*, takes a value *x*:

$$P(x) = \Pr\{X = x\}$$

The distribution function, also called probability distribution function, *F(x),* gives the probability that the variable *X* will take a value less than or equal to a certain value, *x*:

$$F(x) = \Pr\{X \leq x\}$$

i.e. $F(x) = \sum P(x_i)$, where the summation extends to all values less than or equal to *x*.

Two important parameters of the probability distribution are the mean, *E*, and the variance, *V*. The mean, which in probability theory is also called the expected value of *X*, is the sum of the products of the values, x_i, times their probability, $P(x_i)$:

$$E[X] = \sum [x_i P(x_i)]$$

The variance is defined as the expected value of the square of the deviations of the values of variable *X* relative to its mean, that is:

$$V[X] = E[X - E[X]]^2 = \sum \left[(x_i - E[X])^2 P(x_i)\right]$$

Another expression of the variance is:

$$V[X] = E[X]^2 - E^2[X]$$

The standard deviation is the square root of the variance.

If the variable *X* is continuous, the definition of the parameters *E* and *V* needs differential and integral *calculus*, as is shown below.

Density and distribution functions of continuous variables

In the case of a continuous variable *X*, a function, *f(x),* called a density function, or probability density function, is defined to obtain the probabilities and the distribution function. Among the properties of this function it is useful to mention:

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$

The distribution function, *F(x)*, *i.e.* the probability that the variable *X* takes a value smaller than *x* is therefore defined by:

$$F(x) = \Pr\{X \leq x\} = \int_{-\infty}^{x} f(x)\, dx$$

Note that *x* is used as the reference value and also as the generic value.

The probability that *X* takes a value within the interval limited by the extremes $x=x_1$ and $x=x_2$ is given by:

$$\Pr\{x_1 \leq X \leq x_2\} = F(x_2) - F(x_1)$$

The expected value (or mean) of X is defined as:

$$E[X]=\int_{-\infty}^{+\infty} xf(x)\,dx$$

The variance is defined as:

$$V[X]=\mathrm{E}[\mathrm{X}-\mathrm{E}[\mathrm{X}]]^2=\int_{-\infty}^{+\infty}(x-E[X])^2 f(x)\,dx$$

As for discrete variables, another expression of the variance is:

$$V[X]=\mathrm{E}[\mathrm{X}^2]-\mathrm{E}^2[\mathrm{X}]$$

The standard deviation is the square root of the variance.

Some probability distributions useful for sampling theory

The normal distribution

One of the most important distributions of the theory of probabilities is the normal distribution, also designated as De Moivre distribution (1733), Gauss distribution (1809) or Laplace distribution (1813).

It is a distribution of a continuous variable, X, characterized by the following density function:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Where $-\infty<x<+\infty$ and the parameters μ and σ^2 are the mean and the variance.

The density function is symmetrical relative to the vertical ordinate passing through the mean μ.

The distribution function, F(x), the expected value, E[X], and the variance, V[X] could be calculated (by numerical methods) from the integral of the expression indicated above, but there is no need to present those methods in this manual.

The following notation is usually used to indicate that X follows a normal distribution:

$$X \cap N(\mu,\sigma^2)$$

The median and the mode of this distribution are equal to the mean, μ.

A useful theorem for sampling theory is:

If *A and B are constants and* $X \cap N(\mu, \sigma^2)$

Then $(BX+A) \cap N(B\mu+A\,,\, B^2\sigma^2)$

The standard normal distribution

Applying the previous theorem with $B=\frac{1}{\sigma}$ and $A=-\frac{\mu}{\sigma}$ gives:

$(X-\mu)/\sigma \cap N(0,1)$

The new variable (X-μ)/σ is said to be a variable in standard measures and is, in this case, called the standard normal variable **Z**. Any normal distribution can be reduced to its standard form using the relation:

$$z = \frac{x - \mu}{\sigma}$$

Note that this expression is equivalent to: $x = \mu + z\sigma$

Some particular probability values of the normal distribution should be mentioned:

In terms of *Z*:	In terms of *X*:
$Pr\{-1 < Z < +1\} = 0.68$	$Pr\{\mu - 1\sigma < X < \mu + 1\sigma\} = 0.68$
$Pr\{-1.96 < Z < +1.96\} = 0.95$	$Pr\{\mu - 1.96\sigma < X < \mu + 1.96\sigma\} = 0.95$
$Pr\{-2.58 < Z < +2.58\} = 0.99$	$Pr\{\mu - 2.58\sigma < X < \mu + 2.58\sigma\} = 0.99$

Note that if one is using a normal variable *X* then the probability that *X* takes values between x_1 and x_2 is equal to the probability that the standard normal variable *Z* takes values between the corresponding $z_1 = (x_1 - \mu)/\sigma$ and $z_2 = (x_2 - \mu)/\sigma$.

Figure 2.3 represents the density function of the standard normal distribution between Z=-3 and Z=+3.

In this graphical form, areas represent the probabilities. For instance, the probability that the standard normal variable Z takes values between z_1 and z_2 is the area limited by the curve, by the horizontal axis and by the two vertical ordinates passing through the values $z=z_1$ and $z=z_2$.

Remember that the total area under the density curve is equal to 1.

Figure 2.4 represents the distribution function of the standard normal variable Z.

FIGURE 2.3

Density function of the standard normal distribution

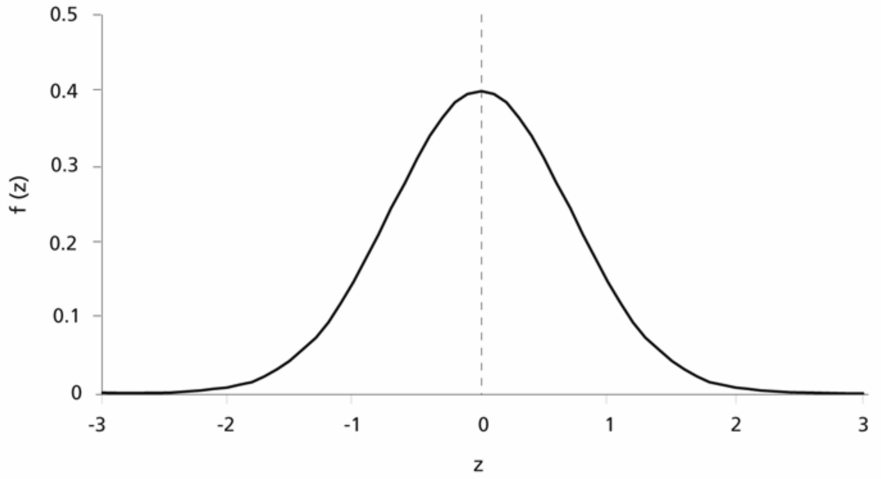
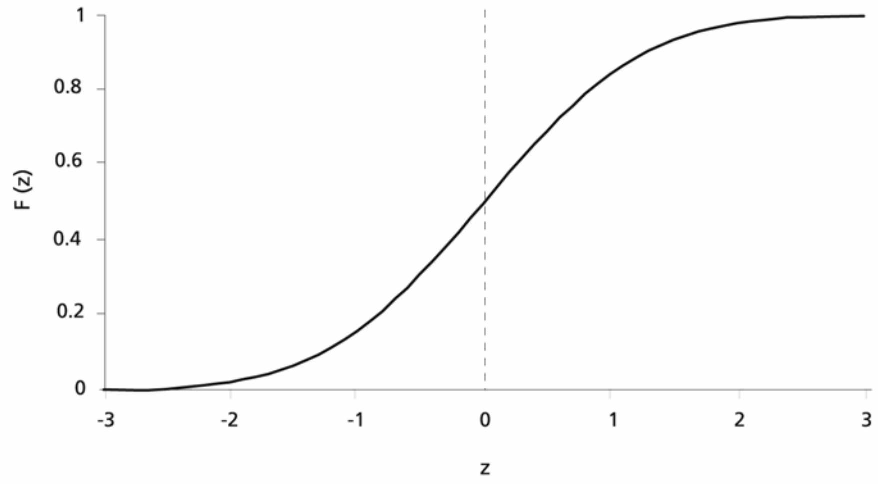


FIGURE 2.4
Distribution function of the standard normal variable, Z



In this graphical form the probability that the variable *Z* is smaller than a value *z* is the ordinate passing through this *z* point.

Note that the probabilities are given by areas when using the density function and by ordinates when using the distribution function.

The t-student distribution

The *t*-student distribution was introduced by Gosset in 1908.

It is a distribution of a continuous variable with one parameter denoted by ν (degrees of freedom) and it is generally designated by *t(* ν) or t_ν .

The mathematical expressions of the density and distribution functions of the *t*-student variable are not discussed in this manual.

The graphical representation of the probability density function is similar to that of the standard normal variable, but more dispersed, with heavier tails. The dispersion depends on the number of degrees of freedom – the fewer degrees, the larger the dispersion. The mean, the median and the mode of this distribution are all equal to zero.

The *t*-student distribution is important in statistical methods and in particular for calculating confidence intervals for parameters of the population.

Associated with the normal and *t*-student distributions are the *chi-square* (χ^2) distribution and the *F*-distribution, which are useful in many statistical methods, but are not dealt with in this manual.

Bernoulli distribution

This distribution is attributed to Bernoulli (1713).

Consider a discrete variable *X* which takes the value 1 with probability *P* and the value 0 with probability *Q=1-P*. In symbolic terms this is:

$$X = \begin{cases} 1 & P \\ 0 & Q \end{cases} \quad P + Q = 1$$

The most important parameters of this distribution are:

Expected value $E[X]$ = Mean $\mu = P$

Variance V[X] = σ^2 = PQ

Standard deviation $\sigma = \sqrt{PQ}$

The binomial distribution, the multinomial distribution, and other probability distributions are associated with the Bernoulli distribution. They are, in certain cases, important to the sampling world. They are combinations of Bernoulli distributions, with the same parameter P in the case of the binomial distribution and with different parameters P in the case of the multinomial distribution.

Introduction to the world of sampling

As previously mentioned an estimator is a statistic used to estimate a parameter.

Let us consider an estimator $\hat{\theta}$ of the population parameter θ. From a sample of size n, taken with a certain criterion, one can calculate a value for this estimator, that is called an estimate of θ.

Sampling distribution of an estimator

The set of estimates that could be calculated from all possible samples (selected with the same criteria) is, by definition, the sampling distribution of the estimator.

This sampling distribution is the basis for measuring the precision and the error of the estimation of the population parameter of interest.

The sampling distribution of an estimator (or in the general case of a statistic) is a probability distribution, because it is the expected distribution of all the possible samples, which could have been selected under the same conditions. Therefore, probability theory can be applied to obtain the properties of the sampling distributions.

The sampling distribution of an estimator is denoted as:

$$\hat{\theta} \cap F(E,V)$$

where $F(E,V)$ indicates a probability or density distribution with expected mean, E, and expected variance, V.

In the case of an approximate distribution, the symbol $\overset{\bullet}{\cap}$ will be used, as for instance:

$$\hat{\theta} \overset{\bullet}{\cap} F(E,V)$$

Expected value of the estimator

Let us consider an estimator $\hat{\theta}$. The expected value or sampling mean of this estimator will be denoted by $E[\hat{\theta}]$. This expected value does not always coincide with the population parameter θ. The difference $E[\hat{\theta}]-\theta$ is called bias. When $E[\hat{\theta}]=\theta$ the estimator, $\hat{\theta}$, is an unbiased estimator of the population parameter, θ.

Sampling variance and error of the estimator

The sampling variance, $V[\hat{\theta}]$ or $\sigma^2_{\hat{\theta}}$, of the estimator $\hat{\theta}$ is the expected value $E\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]^2$. This means that the sampling variance measures the spread of the estimator around its expected value.

Another measure of dispersion is the mean square error, defined as $MSE\left[\hat{\theta}\right]=E\left[\hat{\theta}-\theta\right]^2$. The MSE is a measure of the dispersion of the sampling distribution of the estimator $\hat{\theta}$ around the population parameter θ, while $V[\hat{\theta}]$ is a measure of dispersion around the expected value of the estimator.

It can be proven that $MSE\left[\hat{\theta}\right]=V\left[\hat{\theta}\right]+bias^2$.

For an unbiased estimator the sampling variance is equal to the *MSE*.

The accuracy of an estimator refers to the difference between the estimator and the parameter θ, while the precision refers to the difference from the expected value.

The sampling standard deviation is called the error of the estimator, which is denoted by $\sigma_{\hat{\theta}}$. Estimates of the sampling variance and of the error, denoted by $s^2_{\hat{\theta}}$ and $s_{\hat{\theta}}$ respectively, can be obtained from the size and the variance of the sample.

Confidence intervals

The sampling distribution of an unbiased estimator gives the opportunity to establish the probability that an interval, (l_1, l_2) calculated from the sample values, will contain the population parameter, θ.

The relation $\text{Prob}\{l_1 \le \theta \le l_2\} = C$ allows one to calculate the confidence interval (l_1, l_2) but the solution of this equation is not unique. One should take the smallest of all intervals to which the probability adopted corresponds. This probability is the confidence level, *C*. The interval is the confidence interval, *CI*, and its extremes are the confidence limits.

In some cases, it is preferable to take an interval (not necessarily the smallest) corresponding to the probability *C* but with equal probabilities in both tails of the sampling distribution, which implies:

$\text{Prob}\{\theta \le l_1\} = \dfrac{1-C}{2}$ and

$\text{Prob}\{\theta \ge l_2\} = \dfrac{1-C}{2}$ or, alternatively, in the equivalent form $\text{Prob}\{\theta \le l_2\} = \dfrac{1+C}{2}$

In practical terms, to calculate the limits l_1 and l_2 of a confidence interval one should adopt the desired level of confidence, *C* and a sample size, *n*. Then, based on the sampling distribution of the estimator, the expressions of l_1 and l_2, can be derived. Finally, one selects a sample and estimates l_1 and l_2.

Let us consider the particular case that $\hat{\theta}$ has a normal sampling distribution, that is, $\hat{\theta} \cap N[E,V]$, the population variance, σ^2, is known and the sampling fraction is negligible. Then, according to sampling theory, the confidence limits are:

$$\begin{cases} l_1 = \hat{\theta} - z\,\sigma_{\hat{\theta}} \\ l_2 = \hat{\theta} + z\,\sigma_{\hat{\theta}} \end{cases}$$

where *z* is the value of the standard normal distribution corresponding to the confidence level *C* and $\sigma_{\hat{\theta}}$ is the square root of $\dfrac{\sigma^2}{n}$.

The range of the confidence interval, l_2-l_1 is in this case equal to $2\,z\,\sigma_{\hat{\theta}}$.

If the population variance, σ^2, is unknown, the expression above will be:

$$\begin{cases} l_1 = \hat{\theta} - t_{n-1}\,s_{\hat{\theta}} \\ l_2 = \hat{\theta} + t_{n-1}\,s_{\hat{\theta}} \end{cases}$$

where t_{n-1} is the value of *t* corresponding to the confidence *C*, of the *t*-student distribution, with *n-1* degrees of freedom and $\sigma_{\hat{\theta}}$ is the square root of $\dfrac{s^2}{n}$ designated by $s_{\hat{\theta}}$ and calculated from the sample.

The confidence interval range, l_2-l_1, is equal to $2t_{(n-1)}\,s_{\hat{\theta}}$.

Note that for a large sample size, let us say n larger than 100, there is practically no difference between the t-distribution and the Z-distribution.

Let us take an example in which the estimator $\hat{\theta}$ follows a normal sampling distribution. A sample of 100 elements gave an estimate of $\hat{\theta}$ equal to 13.40 and a variance $s^2 = 30.25$. The error is the square root of, $\frac{s^2}{n}$ *i.e.*, 0.55.

A 95% confidence interval (symmetrical) for the parameter θ can be expressed as:

$$\begin{cases} l_1 = 13.40 - 1.96 \times 0.55 \\ l_2 = 13.40 + 1.96 \times 0.55 \end{cases} \text{or } 13.40 \pm 1.96 \times \text{error}$$

The factor 1.96 corresponds to the 95% confidence level and is obtained from the normal distribution.

The confidence limits can be presented in different ways. Some authors present the limits in absolute values; others prefer to give the limits in relative terms. In the above example the 95% confidence limits in absolute terms would be:

$$\begin{cases} l_1 = 12.30 \\ l_2 = 14.50 \end{cases} \text{or } \begin{cases} l_1 = 13.40 - 1.08 \\ l_2 = 13.40 + 1.08 \end{cases}$$

Note that the interval should never be presented as the mean plus or minus the error, *i.e.* 13.40 ± 0.55 because, according to the sampling distribution of the estimator, the confidence limits are defined as the mean plus or minus (approximately) twice the error.

The limits in relative terms would be:

$$\begin{cases} l_1 = 13.40 - 8.06\% \\ l_2 = 13.40 + 8.06\% \end{cases} \text{where } 8.06\% = \frac{1.08}{13.40}$$

or 13.40 ± 8.06%

or only ± 8.06% of the estimated mean.

Sample size

As mentioned above, to calculate the confidence interval, one needs to establish the confidence level, C, and the sample size, n. Based on these assumptions one calculates the error and the confidence limits.

The same expressions can be used to estimate the sample size, but in this case it is necessary to start by adopting a confidence interval, CI, and the error. From the sampling distribution of the estimator one can then derive the expression that gives the sample size.

In the above example it is easy to see that:

$$CI = l_2 - l_1 = 2 \times z \times error$$

Then considering that the error is $\frac{s}{\sqrt{n}}$, it will be: $\frac{s}{\sqrt{n}} = \frac{CI}{2z}$

and the sample size is:

$$n = \left(\frac{2zs}{CI}\right)^2$$

If the confidence interval were calculated with the t-student distribution, the calculation would be more tedious as the sample size n appears also in the degrees of freedom of t.

An iterative method can be used to solve the equation, which in this case is:

$$n = \left(\frac{2t_{n-1}s}{CI} \right)^2$$

The first trial can be done with the *Z* distribution instead of the *t*-student distribution. The value of *n* obtained will then be the next value to use to obtain the t_{n-1} value corresponding to the confidence level *C*. The process is repeated until arriving at a convergence of the *n* values, with a given approximation.

The following expressions summarize the links between the confidence level, precision, error and confidence limits:

> *"The greater the error, the larger will be the confidence interval and the smaller the precision. The smaller the error, the shorter will be the confidence interval and the greater the precision".*

3. Simple random sampling

3.1 INTRODUCTION

Simple random sampling is the simplest way to sample a population. Its simplicity arises from the way that the sample is selected. In this design, all possible samples have the same probability to be chosen.

Other sampling methods have procedures that include this method for selection of parts of a total sample. For this reason, when describing sampling methods it is convenient to start with the simple random sampling.

3.2 THE POPULATION WORLD

In simple random sampling, the population to be sampled is considered as a simple collection of elements, where no subgroups are considered.

Let Y denote a characteristic of a population. Then Y_i will be the value of the characteristic of the i^{th} element. The main parameters of the population that are more relevant to fisheries research are discussed in Chapter 2. A short summary is presented in Table 3.1.

TABLE 3.1
Main population parameters of interest to fisheries research

N	Population size
Y_i	Value of the characteristic of the i th element (i =1, 2, ..., N)
$Y = \sum_{i=1}^{N} Y_i$	Total value
$\overline{Y} = \frac{Y}{N} = \frac{\sum_{i=1}^{N} Y_i}{N}$	Mean value $\overline{Y}$ or μ (The relation $Y = N\overline{Y}$ is valid)
$\sigma^2 = \frac{SS}{N}$ and $S^2 = \frac{SS}{N-1}$	Variance, σ^2 and modified variance, S^2
$\sigma = \sqrt{\sigma^2}$ and $S = \sqrt{S^2}$	Standard deviation σ, and modified standard deviation S
$CV = \frac{\sigma}{\overline{Y}}$ and $CV = \frac{S}{\overline{Y}}$	Coefficient of variation, CV

3.3 THE SAMPLE WORLD

The main statistics of the sample that are more relevant to fisheries research are discussed in Chapter 2. A short summary of the more common ones is presented in Table 3.2.

TABLE 3.2
Some sample statistics more frequently used in fisheries research

n	Sample size
y_i	Value of the characteristic of the i^{th} element (i =1, 2, ..., n)
$y = \sum_{i=1}^{n} y_i$	Total value of the characteristic in the sample
$\bar{y} = \frac{y}{n}$	Sample mean
$ss = \sum_{i=1}^{n} (y_i - \bar{y})^2$	Sum of squares of the deviations from the sample mean
$s^2 = \frac{ss}{n-1}$	Sample variance
$s = \sqrt{s^2}$	Sample standard deviation
$cv = \frac{s}{\bar{y}}$	Sample coefficient of variation

3.4 THE SAMPLING WORLD

3.4.1 Selection of the sample

Simple random sampling is defined as any sampling system that ensures that all possible samples with a given sample size have the same probability of being selected. Alternatively one could say that every element of the population has the same probability of being selected for inclusion in the sample, in one extraction.

In this sampling method, the probability, P_i, of selecting an element, i, from a finite population of size N, is:

$$P_i = \frac{1}{N}$$

There are two ways to take a simple random sample: either the elements are selected with replacement of the element into the population after each extraction, or without replacement.

When using sampling without replacement from a finite population, it is usual to define the sampling fraction, f, as $f = \frac{n}{N}$. In that case, $(1\text{-}f)$ is called the finite correction factor. When the number of elements of the population is infinite or sampling is with replacement, the sampling fraction is zero, the correction factor is equal to 1 and, therefore, it is not considered.

3.4.2 Estimator of the population mean $\bar{y}$

Several statistics (e.g. the median) can be used to estimate the mean of the population. However, the most frequently used estimator of the population mean is the sample mean:

$$\bar{y} = \frac{y}{n}$$

In simple random sampling, the sampling distribution of the sample mean has some important properties:

1. The expected value of the sampling distribution is equal to the population mean:

   $E[\bar{y}] = \mu$

   Note: the sample mean is an unbiased estimator of the population mean.

2. When the sample size *n* is large (for example, n>100), the sampling distribution of the sample mean tends to be a normal distribution. Using the notation presented in chapter 2, we may write:

   $\bar{y} \overset{\bullet}{\cap} N(E,V)$

   where

   $E = E[\bar{y}] = \bar{Y} = \mu$

   $$V = V[\bar{y}] = \sigma_{\bar{y}}^2 = (1-f)\frac{S^2}{n}$$

   When *n* is very small compared to *N* the correction factor, *(1-f),* can be ignored and, in this case:

   $$V = V[\bar{y}] = \sigma_{\bar{y}}^2 = \frac{S^2}{n}$$

An estimate of the sampling variance of the estimator can be calculated by replacing the population parameter *S,* in the variance expressions, with the sample statistic *s*, that is:

$$v(\bar{y}) = s_{\bar{y}}^2 = (1-f)\frac{s^2}{n}$$

or

$v[\bar{y}] = \dfrac{s^2}{n}$ when *N* is large or sampling is with replacement.

The error is given by: $\sigma_{\bar{y}} = \sqrt{V[\bar{y}]}$

and the estimate of the error is: $s_{\bar{y}} = \sqrt{v\,[\bar{y}]}$

The C confidence interval (*l₁, l₂*)for the estimator of the population mean, can be with:

$$\begin{cases} l_1 = \bar{y} - z\,\sigma_{\bar{y}} \\ l_2 = \bar{y} + z\,\sigma_{\bar{y}} \end{cases}$$

if σ^2 is known (or *n* is large), *z* can be calculated from the probability relation:

$$\text{Prob}\{Z \le z\} = \frac{1+C}{2}$$

When σ^2 is unknown, but *n* is large, the standard error of the mean, $\sigma_{\bar{y}}$, can be replaced with the estimated standard error of the mean, $s_{\bar{y}}$.

If σ^2 is unknown, and *n* is small, the standard error of the mean, $\sigma_{\bar{y}}$, can still be replaced with the estimated standard error of the mean, $s_{\bar{y}}$. In these situations, however, the *Z*-distribution should be replaced with the *t*-student distribution. The confidence interval will then be:

$$\begin{cases} l_1 = \overline{y} - t_{n-1} s_{\overline{y}} \\ l_2 = \overline{y} + t_{n-1} s_{\overline{y}} \end{cases}$$

where t_{n-1} follows the *t*- distribution with *n-1* degrees of freedom.

The value of t_{n-1} is given by the probability relation:

Prob {t outside the interval (-t_{n-1}, t_{n-1})} = (1- C), where C is the desired confidence level.

This value t_{n-1} can be obtained from *t*-student tables.

Note that the limits *(l_1, l_2)* of the confidence interval are derived from the following relation:

$$z = \frac{\overline{y} - \mu}{\sigma_{\overline{y}}} \quad \text{with} \quad Z \overset{\bullet}{\cap} N(0,1)$$

or, when the population variance is unknown and the sample size is small, are derived using the *t*-student distribution as follows:

$$t = \frac{\overline{y} - \mu}{s_{\overline{y}}} \quad \text{with} \quad t \overset{\bullet}{\cap} t(n-1)$$

3.4.3 Estimator of the total value of the population

An unbiased estimator of the total value of the population is:

$$\hat{Y} = N\overline{y}$$

The sampling distribution of this estimator is approximately normal:

$$\hat{Y} \overset{\bullet}{\cap} N(E,V)$$

where

$$E = E\left[\hat{Y}\right] = Y \quad \text{and} \quad V = V\left[\hat{Y}\right] = \sigma_{\hat{Y}}^2 = (1-f)\frac{N^2 S^2}{n}$$

The error of the estimator is the square root of the sampling variance:

$$\sigma_{\overline{y}} = \sqrt{V[\hat{Y}]}$$

Approximations to the sampling variance and to the error of the estimator can be obtained by replacing the population variance, S^2, with the sample variance, s^2, in the respective formulas, that is:

$$v = v\left[\hat{Y}\right] = s_{\hat{Y}}^2 = (1-f)\frac{N^2 s^2}{n} \quad \text{and} \quad s_{\hat{Y}} = \sqrt{v(\hat{Y})}$$

It should be noted that the estimator $\hat{Y} = N\,\overline{y}$ can be written as: $\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$

This last expression shows that the estimator of the total value of the population was obtained by raising, extrapolating or amplifying the total value of the sample by the raising factor *N/n*. This is the most common way of obtaining the estimator total values of the population in fishery research. It is also common in fisheries research to apply, instead of the quotient between the size of the population, *N*, and the size of the sample, *n*, the quotient of the corresponding total weights, *W* and *w* (this implies that the same mean weight in the sample and in the population is assumed).

In these circumstances, the expression given above for the estimated variance can be written in terms of the quotient *N/n* or *W/w*, e.g.:

$$v\left(\hat{Y}\right)=s_{\hat{Y}}^{2}=(1-f)\left(\frac{N}{n}\right)^{2}ns^{2}$$

3.4.4 Estimators of proportions

In the case of proportions, the quantities to be estimated are the proportion *P* of the elements belonging to a category or the population mean and the total number, *NP*, of the elements belonging to the category or the population total of the variable Y_i (in the case of finite populations).

Estimator of the population proportion

In the general situation of simple random sampling, the sample mean, that in this case is the proportion of the sample elements belonging to the category of interest, *p*, is an estimator of the population mean, *P*.

The sampling distribution of this estimator has the following properties:

Expected value: *E[p]=P*

Sampling Variance: *V[p]=* $\left(1-f\right)\frac{S^{2}}{n}$

As previously mentioned, the sampling variance of a mean is $\left(1-f\right)\frac{S^{2}}{n}$, where S^2 is the population variance. In this case S^2 is equal to $\frac{NPQ}{N-1}$ and the sampling variance of *p* will be:

$$V\left[p\right]=(1-f)\frac{N}{N-1}\frac{PQ}{n}$$

An unbiased estimate of *V[p]* can be obtained by replacing the population variance S^2 with the sample variance s^2 in the general expression of the sampling variance of a mean:

$$v\left[p\right]=(1-f)\frac{pq}{n-1}$$

The error S_p is the square root of the sampling variance of *p*: $S_p=\sqrt{V[p]}$

An estimate, s_p, of the sampling error, is given by: $s_p=\sqrt{v\,(p)}$

Estimator $\hat{Y}$ *of the total number of elements in the population, NP*

In the cases of finite populations we are often interested in estimating the total number of individuals belonging to the category. An estimator of the total value can be obtained from the mean, $N\,\bar{y}$ and so: $\hat{Y}=N\,p$

The expected value of *Np* is the total value of the population *NP*.

$$E[Np]=NP$$

The sampling variance of *Np* is:

$$V[Np]=(1-f)N^2\frac{N}{N-1}\frac{PQ}{n}$$

An estimate of the sampling variance $v[Np]$ is:

$$v[Np]=N^2(1-f)\frac{pq}{n-1}$$

The error of *Np* will be:

$$S_{\hat{Y}}=\sqrt{V[\mathrm{Np}]}$$

and an estimate of the sampling error of *Np* is

$$s_{\hat{Y}}=\sqrt{v[\mathrm{Np}]}$$

Comments

The expected value, the variance and the error of the sampling distributions of *p* or *Np* have been presented, but sometimes it is helpful to know other aspects of the sampling distributions.

The sampling distributions of *p* and of *Np* are derived from the binomial distribution.

The variable, Y_i is a Bernoulli variable with probability *P* of being equal to 1 and probability *Q=1-P* otherwise. Therefore the binomial distribution is a combination of *n* independent Bernoulli variables with a common constant parameter *P*.

This distribution is usually denoted as $Y \cap b(n,P)$. The parameters are *n*, the sample size, and *P,* the proportion of elements belonging to the category.

When *P* is close to 0.5 and *n* is large, the binomial distribution approximates the normal distribution. Thus the mean will follow approximately the normal distribution:

$p \overset{\bullet}{\cap} N[E,V]$ with *E* and *V* as previously indicated.

The total value *Np* is also approximately normally distributed:

$Np \overset{\bullet}{\cap} N[E,V]$ with *E* and *V* as previously indicated.

3.4.5 Estimator of several proportions of the population

Some characteristics can be classified into more than two categories. For instance, maturity can be classified into stage I, stage II, stage III, etc.

Consider a population divided into *K* categories or classes and let *h* designate one of these classes. To estimate the proportion of elements belonging to the class *h*, the population can be thought of as divided into only two classes, that is, the class *h* and another class covering all the remaining categories. In this way we can apply the previous conclusions about populations divided into two classes, being $p_h= n_h/n$ the estimator of the proportion of elements belonging to class *h,* where *n* is the size of a simple random sample. The expected value and the sampling variance could be derived for the class *h* as mentioned for the binomial case.

As an example, the expected value and an estimate of the sampling variance of p_h is:

$$E[p_h]=P_h \quad \text{and} \quad v[p_h]=(1-f_h)\frac{p_h(1-p_h)}{n_h-1}$$

The sampling distribution of the sample proportions, when the population is divided into k classes, with different proportions of elements in each class, can be considered as an extension of the binomial distribution, that is, as a combination of n independent Bernoulli distributions with different parameters (n_h, P_h) with $h=1, 2,\ldots, k$. This probability distribution is called the multinomial distribution.

4. Stratified random sampling

4.1 INTRODUCTION

When the population is heterogeneous, dividing the whole population into sub-populations, called *strata*, can increase the precision of the estimates. The *strata* should not overlap and each *stratum* should be sampled following some design. All *strata* must be sampled. The *strata* are sampled separately and the estimates from each *stratum* combined into one estimate for the whole population.

The theory of stratified sampling deals with the properties of the sampling distribution of the estimators and with different types of allocation of the sample sizes to obtain the maximum precision.

The principle of stratification is the partition of the population in such a way that the elements within a *stratum* are as similar as possible and the means of the *strata* are as different as possible.

The design is called stratified random sampling if simple random sampling is applied to each *stratum*.

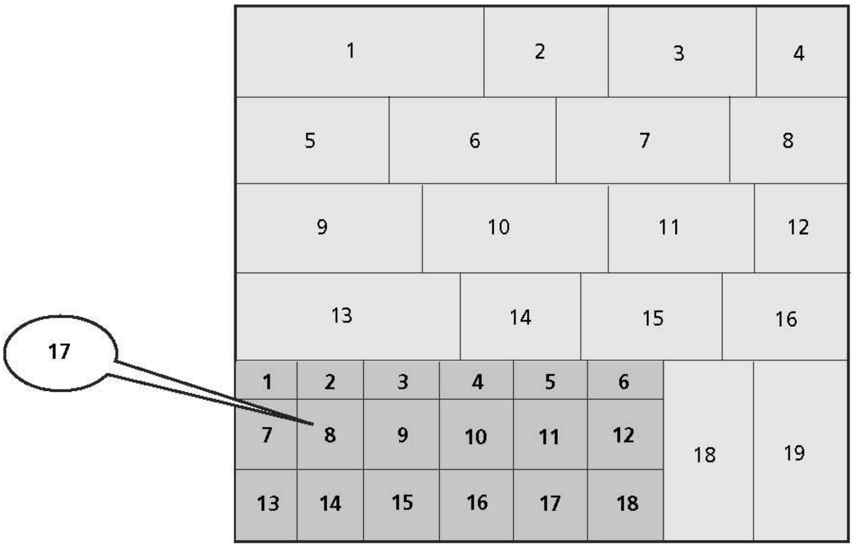
4.2 THE POPULATION

In stratified sampling the population of N elements is divided into k *strata* of sizes:

$N_1, N_2, ..., N_h, \ldots, N_k$ elements, where $N = \sum_{h=1}^{k} N_h$

Every element in the population belongs to at least one *stratum*, and no element of the population belongs to more than one *stratum*. Figure 4.1 shows a stratified sampling scheme for a shrimp fishing ground.

FIGURE 4.1
A stratified sampling scheme for a shrimp fishing ground



The population was divided into 19 *strata*. As an illustration *stratum* 17 shows the 18 trawling unit areas into which the *stratum* was divided. A similar subdivision was used for each of the other *strata*.

4.2.1 *Stratum h*

Let N_h represent the size and Y_{hi} the value of the characteristic Y in the i^{th} element of *stratum h*. The total value of the characteristic Y in *stratum h* is:

$$Y_h = \sum_{i=1}^{N_h} Y_{hi}$$

and the mean value is: $\overline{Y}_h = \dfrac{Y_h}{N_h}$

The modified population variance of *stratum h* is:

$$S_h^2 = \frac{SS_h}{N_h - 1} = \frac{\sum_{i=1}^{N_h}(Y_{hi} - \overline{Y}_h)^2}{N_h - 1}$$

Note that the sum of squares of residuals, SS_h, is divided by *(N_h- 1)*, to obtain S_h^2, and not σ_h^2. The standard deviation is the square root of the variance, $S_h = \sqrt{S_h^2}$.

4.2.2 All *strata*

The total value of the characteristic Y in the population is the sum of the total values of all *strata*:

$$Y = \sum_{h=1}^{k} Y_h$$

and the mean value is a weighted average of the means of all *strata*,

$$\overline{Y} = \frac{Y}{N} = \frac{\sum_{h=1}^{k}\sum_{i=1}^{N_h} Y_{hi}}{N} = \frac{\sum_{h=1}^{k} N_h \overline{Y}_h}{N} = \sum_{h=1}^{k} \frac{N_h}{N}\overline{Y}_h = \sum_{h=1}^{k} W_h \overline{Y}_h$$

where N is the size of the population with k *strata*:

$N = N_1 + N_2 + ... + N_h + N ... + N_k$ and $W_h = \dfrac{N_h}{N}$ is the size of *stratum h*, relative to the total population size, and is used as the weighting factor.

4.3 THE SAMPLE

In stratified sampling, a sample is selected from each *stratum* by simple random sampling. Independent selections are used in each *strata*.

4.3.1 *Stratum h*

Consider a sample of size n_h selected from *stratum h* by simple random sampling without replacement. The value of characteristic Y in the i^{th} element of the sample from the *stratum* is denoted by y_{hi}. Then $y_h = \sum_{i=1}^{n_h} y_{hi}$ is the sample total value and $\overline{y}_h = \dfrac{y_h}{n_h}$ is the sample mean value of characteristic Y in the *stratum*.

The sample variance of characteristic Y in *stratum h* is: $s_h^2 = \frac{\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2}{n_h - 1}$

The sample standard deviation, s_h, is the square root of the variance, $s_h = \sqrt{s_h^2}$ and the coefficient of variation will be $cv = \frac{s_h}{\bar{y}_h}$.

4.3.2 All *strata*

Given independent simple random samples from each *strata*, each of size n_h, the total sample size is $n = \sum_{h=1}^{k} n_h$.

Under these conditions, the total value of characteristic Y in the whole sample is the sum of the sample total values in each *stratum*, $y = \sum_{h=1}^{k} y_h$

The stratified sample mean, $\bar{y}_{st}$, is given by the weighted average of the sample means of the characteristic of interest from each *stratum*,

$$\bar{y}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{k} W_h \bar{y}_h$$

and the stratified sample variance is simply the sum of the variances within each *stratum*. This is achieved because there is no sampling of *strata* (all are observed) and sampling is carried out independently within each of them,

$$s_{st}^2 = \sum_{h=1}^{k} s_h^2$$

The stratified sample standard deviation, s_{st}, is the square root of the variance,

$s_{st} = \sqrt{s_{st}^2}$ and the coefficient of variation will be $cv_{st} = \frac{s_{st}}{\bar{y}_{st}}$.

4.4 THE SAMPLING WORLD

4.4.1 *Stratum h*

Estimator of the mean value

Within each *stratum*, simple random sampling is used. So, the sampling distribution of the estimators of the population parameters of each *stratum* is that given for simple random sampling.

An unbiased estimator $\hat{\bar{Y}}_h$, of the mean of characteristic Y, of the *stratum h*, $\bar{Y}_h$, is $\bar{y}_h$.

The sampling distribution of $\hat{\bar{Y}}_h$ is approximately normal, $\hat{\bar{Y}}_h \overset{\bullet}{\frown} N(E,V)$, where *E* is the expected value and *V* is the sampling variance of the estimator in *stratum h*:

$$E\left[\hat{\bar{Y}}_h\right] = \bar{Y}_h \quad \text{and} \quad V\left[\hat{\bar{Y}}_h\right] = (1 - f_h)\frac{S_h^2}{n_h}$$

where f_h is the sampling fraction defined by $f_h = \frac{n_h}{N_h}$, n_h is the sample size in *stratum h* and N_h is the size of *stratum h*.

An unbiased estimator of S_h^2 is the sample variance s_h^2 :

$$s_h^2 = \frac{\sum_{i=1}^{n_h}(y_{hi} - \overline{y}_h)^2}{n_h - 1}$$

An estimator of the sampling variance of the estimator can thus be obtained by replacing S_h^2 by s_h^2 in the corresponding expression:

$$v\left[\hat{\overline{Y}}_h\right] = (1 - f_h)\frac{s_h^2}{n_h}$$

Estimator of the total value

Let $\hat{Y}_h$ be an estimator of the total value Y_h of *stratum h*, given by: $\hat{Y}_h = N_h\,\overline{y}_h$

where $\overline{y}_h$ is the mean of *stratum h*.

$\hat{Y}_h$ is an unbiased estimator of Y_h with an approximately normal sampling distribution, $\hat{Y}_h \overset{\bullet}{\cap} N[E,V]$, where:

$$E = E[\hat{Y}_h] = Y_h$$

and

$$V\left[\hat{Y}_h\right] = N_h^2\frac{S_h^2}{n_h}(1 - f_h)$$

where f_h is the sampling fraction of *stratum h*, $\frac{n_h}{N_h}$.

The square root of the sampling variance, $\sqrt{V}$ is the error of the estimator.

An unbiased estimator of the sampling variance V is obtained by replacing the population variance by the sample variance in the corresponding expression:

$$v(\hat{Y}_h) = N_h^2\frac{s_h^2}{n_h}(1 - f_h)$$

where s_h^2 is an estimate of S_h^2, given by the sample variance.

4.4.2 All *strata*

Estimator of the mean value

An unbiased estimator of the population stratified mean, for all *strata*, is given by the sample stratified mean,

$$\hat{\overline{Y}} = \overline{y}_{st} = \sum_{h=1}^{k}\frac{N_h}{N}\overline{y}_h$$

The sampling distribution of $\hat{\overline{Y}}$ is approximately normal, $\hat{\overline{Y}} \overset{\bullet}{\cap} N(E,V)$, where E is the expected value and V is the sampling variance of the estimator, given by:

$$E\left[\hat{\overline{Y}}\right] = \overline{Y} \quad \text{and} \quad V\left[\hat{\overline{Y}}\right] = \sum_{h=1}^{k}\frac{N_h^2}{N^2}\cdot V\left[\hat{\overline{Y}}_h\right] = \sum_{h=1}^{k}\frac{N_h^2}{N^2}(1 - f_h)\frac{S_h^2}{n_h}$$

In these expressions, f_h is the sampling fraction defined by $f_h = \frac{n_h}{N_h}$, n_h is the sample size in *stratum h* and N_h is the size of *stratum h*.

An unbiased estimator of the sampling variance of the estimator of the stratified mean value can be obtained by replacing S_h^2 by s_h^2 in the corresponding expression:

$$v\left[\hat{\overline{Y}}\right] = \sum_{h=1}^{k} \frac{N_h^2}{N^2}(1-f_h)\frac{s_h^2}{n_h}$$

Estimator of the total value

As sample selections in different *strata* have been made independently, an estimator of the total value of the population is: $\hat{Y} = N\,\overline{y}_{st}$

where $\overline{y}_{st}$ is the stratified sample mean, given by

$$\overline{y}_{st} = \sum_{h=1}^{k} \frac{N_h}{N}\overline{y}_h = \sum_{h=1}^{k} W_h \overline{y}_h$$

The estimator $\hat{Y}$ has an approximately normal distribution, $\hat{Y}_h \overset{\bullet}{\cap} N[E,V]$, where E and V are the expected value and the sampling variance, respectively, of the estimator, and are given by

$$E = E[\hat{Y}] = \overline{Y} \text{ and } V[\hat{Y}] = N^2 V[\overline{y}_{st}] = \sum_{h=1}^{k} N_h^2 (1-f_h)\frac{S_h^2}{n_h}$$

Like for the mean value, an unbiased estimator of the sampling variance of the estimator of the stratified mean value can be obtained, by replacing the population variance S_h^2 by the corresponding sample variance s_h^2 in its expression:

$$v[\hat{Y}] = \sum_{h=1}^{k} N_h^2 (1-f_h)\frac{s_h^2}{n_h}$$

4.5 ALLOCATION OF THE SAMPLE AMONG THE *STRATA*

In stratified sampling, the size of the sample from each *stratum* is chosen by the sampler, or to put it another way, given a total sample size $n = n_1 + n_2 + \ldots + n_h + \ldots + n_k$, a choice can be made on how to allocate the sample among the *k strata*. There are rules governing how a sample from a given *stratum* should be taken. Sample size should be larger in *strata* that are larger, with greater variability and where sampling has lower cost. If the *strata* are of the same size and there is no information about the variability of the population, a reasonable choice would be to assign equal sample sizes to all *strata*.

4.5.1 Proportional allocation

Let *n* be the total size of the sample to be taken.

If the *strata* sizes are different, proportional allocation could be used to maintain a steady sampling fraction throughout the population. The total sample size, *n*, should be allocated to the *strata* proportionally to their sizes:

$$\frac{n_h}{N_h} = \frac{n}{N} \qquad \text{or} \qquad n_h = n \cdot \frac{N_h}{N} = n \cdot W_h$$

4.5.2 Optimum allocation

Optimum allocation takes into consideration both the sizes of the *strata* and the variability inside the *strata*. In order to obtain the minimum sampling variance the total sample size should be allocated to the *strata* proportionally to their sizes and also to the standard deviation of their values, i.e. to the square root of the variances.

$$n_h = \text{constant} \times N_h \; s_h$$

Given that $n = \sum_{h=1}^{k} n_h$, in this case

$$\text{constant} = \frac{n}{\sum_{h=1}^{k} N_h s_h} \quad \text{so that} \quad \mathrm{n_h} = n \cdot \frac{N_h s_h}{\sum_{h=1}^{k} N_h s_h}$$

where n is total sample size, n_h is the sample size in *stratum h*, N_h is the size of *stratum h* and s_h is the square root of the variance in *stratum h.*

4.5.3 Optimum allocation with variable cost

In some sampling situations, the cost of sampling in terms of time or money is composed of a fixed part and of a variable part depending on the *stratum*.

The sampling cost function is thus of the form:

$$C = c_0 + \sum_{h=1}^{k} c_h n_h$$

where C is the total cost of the sampling, c_0 is an overhead cost and c_h is the cost per sampling unit in stratum *h*, which may vary from *stratum* to *stratum*.

The optimum allocation of the sample to the *strata* in this situation is allocating sample size to the *strata* proportional to the size, and the standard error, and inversely proportional to the cost of sampling in each *stratum*. This gives the following sample size for *stratum h*:

$$n_h = n \cdot \frac{N_h \cdot \dfrac{S_h}{\sqrt{c_h}}}{\sum_{h=1}^{k} \left(N_h \cdot \dfrac{S_h}{\sqrt{c_h}} \right)}$$

Very often, it is the total cost of the sampling, rather than the total sample size, that is fixed. This is usually the case with research vessel surveys, in which the number of days is fixed beforehand. In this case, the optimum allocation of sample size among *strata* is

$$n_h = \frac{(C - c_0) N_h \dfrac{S_h}{\sqrt{c_h}}}{\sum_{h=1}^{k} (N_h S_h \sqrt{c_h})}$$

4.6 COMMENTS

To obtain the full benefits of the stratification technique, the relative sizes of *strata* must be known.

Each *stratum* should be internally homogeneous. If information about heterogeneity is not available then consider all *strata* equally variable. A short stratified pilot survey can sometimes provide useful information about internal dispersion within *strata*.

A small sized sample could be taken from a *stratum* if the variability among their units is small.

Compared with the simple random sample, stratification results almost always in a smaller sampling variance of the mean or total value estimators, when:

- The *strata* are heterogeneous among themselves
- The variance of each *stratum* is small.

A larger sample from a *stratum* should be taken if:

- The *stratum* is larger
- The *stratum* is more heterogeneous
- The cost of sampling the *stratum* is low.

5. Cluster sampling

5.1 INTRODUCTION

In cluster sampling the population is partitioned into groups, called clusters. The clusters, which are composed of elements, are not necessarily of the same size. Each element should belong to one cluster only and none of the elements of the population should be left out.

The clusters, and not the elements, are the units to be sampled. Whenever a cluster is sampled, every element within it is observed.

In cluster sampling, only a few clusters are sampled. Hence, in order to increase the precision of the estimates, the population should be partitioned into clusters in such a way that the clusters will have similar mean values. As the elements inside the clusters are not sampled, the variance within clusters does not contribute to the sampling variance of the estimators. Therefore, in order to decrease the sampling variance of the estimators the variation within the clusters should be as large as possible, while the variation between clusters should be as small as possible.

It should be noted that the partitioning of the population into clusters follows two opposite criteria to the criteria of partitioning a population into *strata*, that is, the heterogeneity within clusters as opposed to the homogeneity within *strata* and the similarity of cluster means as opposed to the differences in *strata* means.

Cluster sampling is often more cost effective than other sampling designs, as one does not have to sample all the clusters. However, if the size of a cluster is large it might not be possible to observe all its elements. The next chapter will show that there are ways to overcome these difficulties.

In fisheries, cluster sampling has been used to estimate landings per trip in artisanal fisheries with a small number of vessels landing at many sites (beaches). Consider, for instance, a fishery with 100 small beaches, where a few vessels land at each beach. One is interested in the total catch per day of the vessels landing at these beaches, but one does not have the possibility to visit all of them. In this case each beach can be a cluster. If a beach is sampled, all its elements (vessel landings) should be observed.

Another example of cluster sampling in fisheries is the sampling of the length composition of an unsorted large catch of a species kept in fish boxes onboard a vessel. Let us assume that the catch in each box is as heterogeneous as possible. The fish boxes can then be looked upon as clusters, and when a box has been selected for sampling, all the elements (fish) inside the box have to be observed.

5.2 THE POPULATION

To be consistent with other methods, the symbol N is used to designate the total number of population sampling units, which in this case are the clusters and not the elements.

The number of elements in a cluster *i* is denoted by M_i.

The total number of elements in the population is:

$$M_o = \sum_{i=1}^{N} M_i$$

and the mean number of elements per cluster is:

$$\overline{M} = \frac{M_o}{N}$$

The value of the characteristic Y in element *j* from cluster *i* is Y_{ij} and Y_i is the total value of the characteristic in cluster *i*:

$$Y_i = \sum_{j=1}^{M_i} Y_{ij}$$

The mean value per element in cluster *i* is:

$$\overline{Y}_i = \frac{Y_i}{M_i}$$

The total value of the characteristic in all the elements of the population is:

$$Y = \sum_{i=1}^{N} Y_i$$

In cluster sampling, two means can be considered:
The mean per cluster

$$\overline{Y} = \frac{Y}{N} \text{ and}$$

the mean per element

$$\overline{\overline{Y}} = \frac{Y}{M_o} = \frac{Y}{N\overline{M}} = \frac{\overline{Y}}{\overline{M}}$$

In cluster sampling, two types of variances can be considered. The first is the variance between clusters, and the second is the variance within clusters, or between elements within clusters.

The variance between cluster totals is:

$$S_1^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}$$

The variance within a cluster i, denoted by S_{2i}^2, is

$$S_{2i}^2 = \frac{\sum_{j=1}^{M_i}(Y_{ij} - \overline{Y}_i)^2}{M_i - 1}$$

Table 5.1 presents a summary of the main parameters of a discrete population divided into clusters that are most used in fisheries research.

5.3 The sample

In cluster sampling, *n* is the number of clusters to be sampled and m_i is the number of elements sampled from cluster *i*. Note that, in this case, $m_i = M_i$, since all elements of every cluster sampled are observed. The total number of elements observed (*i.e* the total number of elements in the sample) is then $m_o = \sum_{i=1}^{n} m_i$ and the mean number of elements per cluster in the sample is $\overline{m} = \frac{m_o}{n}$.

For any sampled cluster *i*, the value of the chosen characteristic of element *j* is y_{ij} and $y_i = Y_i = \sum_{j=1}^{m_i} y_{ij}$ is the total value of the characteristic in cluster *i*. Note that the sampled cluster total y_i is equal to the population cluster total Y_i, since there is no sampling inside the clusters (all elements of the clusters sampled are observed).

TABLE 5.1
Summary of population parameters of most interest in fisheries research

N	Number of clusters in the population
M_i	Number of elements in cluster *i*
$M_o = \sum_{i=1}^{N} M_i$	Total number of elements in the population
$\overline{M} = \frac{M_o}{N}$	Mean number of elements per cluster
Y_{ij}	Value of characteristic Y in element *j* of cluster *i*
$Y_i = \sum_{j=1}^{M_i} Y_{ij}$	Total value of the characteristic Y in cluster *i*
$\overline{Y}_i = \frac{Y_i}{M_i}$	Mean value of the characteristic Y in the elements of cluster *i*
$Y = \sum_{i=1}^{N} Y_i$	Total value of characteristic Y of all the elements in the population
$\overline{Y} = \frac{Y}{N}$	Mean value of characteristic Y per cluster
$\overline{\overline{Y}} = \frac{Y}{M_o} = \frac{Y}{N\overline{M}}$	Mean value of characteristic Y per element
$S_1^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}$	Variance of characteristic Y between cluster totals
$S_{2i}^2 = \frac{\sum_{j=1}^{M_i}(Y_{ij} - \overline{Y}_i)^2}{M_i - 1}$	Variance of characteristic Y within cluster *i*

The mean value of the characteristic Y in all the elements of cluster *i* is:

$$\overline{y}_i = \frac{y_i}{m_i}$$

The total value of the characteristic Y in all the elements of all the clusters sampled is denoted by: $y = \sum_{i=1}^{n} y_i$

The mean value of the characteristic Y per cluster is: $\bar{y} = \frac{y}{n}$

The mean value of the characteristic Y per element is: $\bar{\bar{y}} = \frac{y}{m_o} = \frac{y}{n\bar{m}}$

The variance between total values of the characteristic Y in the clusters sampled is:

$$s_1^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$$

The variance of the values of the characteristic Y within the i^{th} sampled cluster is:

$$s_{2i}^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$$

Table 5.2 presents a summary of the most common sample statistics in cluster sampling applied to fisheries.

TABLE 5.2
Most common sample statistics in cluster sampling

n	Number of clusters sampled
m_i	Number of elements in cluster i (Note that $m_i = M_i$)
$m_o = \sum_{i=1}^{n} m_i$	Total number of elements in sample
$\bar{m} = \frac{m_o}{n}$	Sample mean number of elements per cluster
y_{ij}	Value of the characteristic Y in element j of cluster i
$y_i = Y_i = \sum_{j=1}^{m_i} y_{ij}$	Total value of the characteristic Y in the sampled cluster i
$\bar{y}_i = \frac{y_i}{m_i}$	Mean value of the characteristic Y in the sampled cluster i
$y = \sum_{i=1}^{n} y_i$	Total value of the characteristic Y in the sample
$\bar{y} = \frac{y}{n}$	Mean value of the characteristic Y per cluster sampled
$\bar{\bar{y}} = \frac{y}{m_o} = \frac{y}{n\bar{m}}$	Sample mean value of the characteristic Y per element
$s_1^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$	Variance between total values of the characteristic Y in the clusters sampled
$s_{2i}^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$	Variance of the values of characteristic Y within the sampled cluster i

5.4 THE SAMPLING WORLD

As mentioned in the introduction to this chapter, in cluster sampling the sampling units are the clusters. The selection of the clusters can be made by random sampling with equal probabilities (simple random sampling) or with different probabilities. A particular case of random sampling with different probabilities is when the probabilities are proportional to the sizes of the clusters.

The most important estimators in cluster sampling are the estimators of the total value, of the mean per cluster and of the mean per element.

5.4.1 Selection with equal probabilities

First, let us consider an example where the clusters are selected by simple random sampling without replacement. In this case, the probability of selecting any cluster *i*, in one extraction, is constant and equal to $P_i = \dfrac{1}{N}$.

An unbiased estimator of the population total value, *Y*, is:

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i \quad \text{or} \quad \hat{Y} = N\,\overline{y}$$

The factor $\dfrac{N}{n}$ is a raising factor, which raises the sample total, *y*, to the estimator of the population total $\hat{Y}$.

The sampling distribution of this estimator is approximately normal with expected value *E*, and variance *V*:

$$\hat{Y} \overset{\bullet}{\frown} N(E,V)$$

where

$$E = E\left[\hat{Y}\right] = Y \;\; (\hat{Y} \text{ is an unbiased estimator of } Y)$$

and

$$V = V\left[\hat{Y}\right] = N^2(1-f)\frac{S_1^{\;2}}{n}$$

An estimate of the sampling variance can be obtained by replacing the population variance $S_1^{\;2}$ with the sample variance $s_1^{\;2}$ in the expression of the sampling variance:

$$v\left[\hat{Y}\right] = N^2(1-f)\frac{s_1^{\;2}}{n}$$

5.4.2 Selection with unequal probabilities

Two cases of selecting the sample with unequal probabilities will be considered: selection with replacement and selection without replacement. In the former, the Hansen-Hurwitz estimator will be described. The particular case of selection with probabilities proportional to the sizes of the clusters will be also studied.

Another estimator, the Horvitz-Thompson estimator, is applicable in both cases, *i.e.*, with or without replacement.

Hansen-Hurwitz estimator – selection with replacement

Let P_i be the probability of selecting cluster *i*, with replacement, in one extraction (Note that $\sum_{i=1}^{N} P_i = 1$). In this case an unbiased estimator of the population total value is:

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{P_i}$$

This estimator has an approximately normal distribution with expected value *E*, and variance, *V*:

$$\hat{Y} \overset{\bullet}{\frown} N(E,V)$$

where

$E = E\left[\hat{Y}\right] = Y$ ($\hat{Y}$ is an unbiased estimator of Y)

and

$$V = V\left[\hat{Y}\right] = \frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{y_i}{P_i} - Y\right)^2$$

An estimate of the sampling variance $V\left[\hat{\bar{Y}}\right]$ would be:

$$v\left[\hat{Y}\right] = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{P_i} - \hat{Y}\right)^2$$

Selection with probabilities proportional to cluster sizes

Let us consider the special case where the selection probability is proportional to the size of the clusters, $P_i = \frac{M_i}{M_o}$

In this case, the estimator of the total value, $\hat{Y}$, its sampling variance and all the other expressions can be obtained replacing P_i in the case previously described, by $\frac{M_i}{M_o}$ *i.e.*:

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{M_i/M_0}\right) = \frac{M_o}{n}\sum_{i=1}^{n}\frac{y_i}{M_i} \text{ or } \hat{Y} = M_o\frac{\sum_{i=1}^{n}\bar{y}_i}{n}$$

The sampling variance is:

$$V\left[\hat{Y}\right] = \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M_o}\left(\frac{M_o}{M_i}y_i - Y\right)^2$$

considering that:

$$\bar{\bar{Y}} = \frac{Y}{M_o} \text{ and } \frac{y_i}{M_i} = \bar{y}_i$$

Another expression of this variance can also be obtained:

$$V\left[\hat{Y}\right]=\frac{M_o}{n}\sum_{i=1}^{N}M_i\left(\bar{y}_i-\bar{\bar{Y}}\right)^2$$

An estimate of this variance is:

$$v\left[\hat{Y}\right]=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{M_o}{M_i}y_i-\hat{Y}\right)^2 \text{ or } v\left[\hat{Y}\right]=\frac{M_o^2}{n(n-1)}\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2$$

Note that in order to use this estimate one needs to know the total number of elements in the population, M_o.

An estimator of the mean value per element is: $\hat{\bar{\bar{Y}}}=\frac{\hat{Y}}{M_o}$ or $\hat{\bar{\bar{Y}}}=\frac{\sum_{i=1}^{n}\bar{y}_i}{n}$

which has a sampling variance given by: $V\left[\hat{\bar{\bar{Y}}}\right]=\frac{V\left[\hat{Y}\right]}{M_0^2}$

An estimate of the sampling variance can be calculated replacing the variance of the population total by its sample estimate:

$$v\left[\hat{\bar{\bar{Y}}}\right]=\frac{v\left[\hat{Y}\right]}{M_0^2}$$

resulting in: $v\left[\hat{\bar{\bar{Y}}}\right]=\frac{\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2}{n(n-1)}$

Note that in order to use this estimate one does not need to know the total number of elements in the population, M_o.

The estimator of the mean value per cluster is:

$$\hat{\bar{Y}}=\frac{\hat{Y}}{N}$$

with a sampling variance:

$$V\left[\hat{\bar{Y}}\right]=\frac{V\left[\hat{Y}\right]}{N^2}$$

An estimate of this sampling variance can be obtained from $v\left[\hat{\bar{Y}}\right]=\frac{v\left[\hat{Y}\right]}{N^2}$ and is expressed as:

$$v\left[\hat{\bar{Y}}\right]=\frac{M_o^2}{N^2}\frac{\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2}{n(n-1)}=\bar{M}^2\frac{\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2}{n(n-1)}$$

In order to use this estimator one needs to know M_o, or at least the mean number of elements per cluster, $\bar{M}=\frac{M_0}{N}$.

Selecting clusters with probabilities proportional to their sizes is not always easy. A simple procedure for selecting n clusters with probabilities proportional to their sizes ($P_i = \frac{M_i}{M_o}$), that can be easily used in fisheries research, is given below:

- calculate the cumulative numbers of elements of the population in each cluster;
- assign intervals of "selection numbers" to each cluster, based on these cumulative numbers;
- use the "selection numbers" in order to choose the *n* clusters to be sampled, with a probability proportional to sizes. For this purpose, select (applying a simple random sampling design) one of the total number of the "selection numbers" to get the corresponding cluster;
- repeat the selection of "selection numbers" to obtain the required number of clusters.

A simple example to illustrate the procedure:

Consider a situation where one wishes to select three out of five boats landing fish on a beach. The boats are considered as the clusters to be sampled. Each boat carries a different number of fish boxes to be landed. The percentages of the total number of fish boxes carried by each one of the five boats will be considered as the probabilities proportional to the sizes of the clusters. Table 5.3 shows the original data and how to calculate what can be designated as the "boat selection numbers".

TABLE 5.3
Original data and calculation of "boat selection numbers"

Boat	Number of fish boxes	Cumulative numbers	Boat selection numbers	Selection Probability
1	5	5	1- 5	5/50=0.10
2	10	15	6-15	10/50=0.20
3	7	22	16-22	7/50=0.14
4	13	35	23-35	13/50=0.26
5	15	50	36-50	15/50=0.30

By repeating three times a simple random sampling of one out of fifty numbers, one will get three boat selection numbers corresponding to the boats to be sampled (ignore the selected numbers corresponding to clusters already chosen).

Horvitz-Thompson estimator – selection with or without replacement

It is convenient, before describing this estimator and its sampling characteristics, to show how inclusion probabilities can be calculated.

Let π_i denote the probability of including cluster *i* in a sample of size *n*. The inclusion probability, π_i, is connected with the probability P_i of selecting cluster *i* in one single extraction.

To derive the relation between π_i and P_i it is preferable to use complementary probabilities. In this way, the probability, 1-π_i, of not including cluster *i* in the sample of size *n* can be calculated as the probability of not selecting the cluster *i* in any of the *n* extractions, that is, $1-\pi_i = (1-P_i)^n$. Therefore the inclusion probability π_i will be:

$$\pi_i = 1-(1-P_i)^n$$

Let us now consider the probability, π_{ij}, that both cluster *i* and cluster *j* are included in a sample of size *n*. The probability of extracting either cluster *i* or cluster *j*, in one extraction, is $P_i + P_j$ and so the probability of neither extracting cluster *i* nor cluster *j*, will be $1-(P_i + P_j)$.

In *n* independent extractions the probability of neither extracting cluster *i* nor cluster *j* will be $[1-(P_i+P_j)]^n$. Therefore the probability of extracting either cluster *i* or cluster *j* in *n* extractions will be $1-[1-(P_i+P_j)]^n$.

Alternatively, the same probability - that either cluster *i* or *j* be included in the sample, could also be expressed as the probability of including cluster *i* plus the probability of including cluster *j* minus the probability of including both *i* and *j*, that is, $(\pi_i + \pi_j) - \pi_{ij}$.

The two last expressions are different ways to refer to the same probability, thus:

$$(\pi_i + \pi_j) - \pi_{ij} = 1 - \left[1-\left(P_i + P_j\right)\right]^n$$

Finally the inclusion probability, π_{ij}, can be calculated as:

$$\pi_{ij} = (\pi_i + \pi_j) - \left\{1 - \left[1-\left(P_i + P_j\right)\right]^n\right\}$$

The calculations of the inclusion probabilities, for the example presented previously, are shown below.

Boat, *i*	**Prob. of selection, *Pi***	**Prob. of inclusion, *πi***	**Inclusion probabilities, *πij***			
			1	**2**	**3**	**4**
1	5/50=0.10	π_1 = 0.271				
2	10/50=0.20	π_2 = 0.488	0.102			
3	7/50=0.14	π_3 = 0.364	0.074	0.139		
4	13/50=0.26	π_4 = 0.595	0.128	0.240	0.175	
5	15/50=0.30	π_5 = 0.657	0.144	0.270	0.197	0.337
Total	**1.00**					

Calculations:

$$\pi_1 = 1-(1-0.10)^3 = 0.271$$
$$\pi_2 = 1-(1-0.20)^3 = 0.488$$
$$\pi_3 = 1-(1-0.14)^3 = 0.364$$

$$\pi_{12} = 0.217 + 0.488 - \{1 - [1-(0.10+0.20)]^3\} = 0.102$$
$$\pi_{13} = 0.217 + 0.364 - \{1 - [1-(0.10+0.14)]^3\} = 0.074$$
$$\pi_{23} = 0.488 + 0.364 - \{1 - [1-(0.20+0.14)]^3\} = 0.139$$

etc.

The Horvitz-Thompson estimator of the total value of the population is:

$$\hat{Y} = \sum_{i=1}^{n^*} \frac{y_i}{\pi_i}$$

where y_i is the total value of the variable, in the distinct sampled cluster i; π_i is the probability of inclusion of the cluster i in the sample and n^* is the "effective" size of the sample, that is, the number of distinct clusters, in a sample of size n. Note that clusters sampled repeatedly are eliminated from the calculations.

The estimator is unbiased and its sampling variance can be written as:

$$V\left[\hat{Y}\right] = \sum_{i=1}^{N}\left(\frac{1}{\pi_i} - 1\right)Y_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)Y_iY_j$$

An unbiased estimate of this variance is:

$$v\left[\hat{Y}\right] = \sum_{i=1}^{n^*}\left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right)y_i^2 + 2\sum_{i=1}^{n^*}\sum_{j\neq i}\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}\right)y_iy_j$$

Note that inclusion probabilities should be different from zero.

These estimates of the variances can be negative. A way to avoid this inconvenience is as follows:

Calculate, from each effective sampled cluster i, the following t_i statistics:

$$t_i = n^*\frac{y}{\pi_i}$$

Each of the t_i values calculated can be considered as an estimate of the total value, Y.

The mean of t_i for all clusters effectively sampled is a Horvitz-Thompson estimator of the total value Y:

$$\bar{t} = \frac{1}{n^*}\sum_{i=1}^{n^*} t_i = \sum_{i=1}^{n^*}\frac{y_i}{\pi_i} = \hat{Y}$$

The estimate, $v\left[\hat{Y}\right]$, of the sampling variance of this estimator is:

$$v\left[\hat{Y}\right] = \left(1 - f^*\right)\frac{s_1^2}{n^*} \text{ where } s_1^2 = \frac{\sum_{i=1}^{n^*}\left(t_i - \hat{Y}\right)^2}{n^* - 1} \text{ and } f^* = \frac{n^*}{N}$$

6. Two-stage sampling

6.1 INTRODUCTION

In the two-stage sampling design the population is partitioned into groups, like cluster sampling, but in this design new samples are taken from each cluster sampled. The clusters are the first stage units to be sampled, called primary or first sampling units and denoted by *SU1*. The second-stage units are the elements of those clusters, called sub-units, secondary or second sampling units and will be denoted by *SU2*.

Two-stage sampling is used when the sizes of the clusters are large, making it difficult or expensive to observe all the units inside them. This is, for example, the situation when one wishes to estimate total landing per trip of a fishery with many landing sites and also with a large number of vessels.

Sometimes, in order to decrease the sizes of the primary sampling units, one can previously stratify the population and apply two-stage sampling to each *stratum*.

It is possible to extend the two-stage sampling design to three or more stages. A short reference will be made to a three-stage sampling design, using a case where the procedure to estimate errors is simple.

6.2 THE POPULATION

Most of the population parameters of interest to fisheries research in the two-stage sampling design are the same as in cluster sampling. These are summarised below.

TABLE 6.1
Main population parameters of interest to fisheries research in two-stage sampling design

N	Number of clusters *(SU1)* in the population
M_i	Number of elements *(SU2)* in cluster *(SU1) i*
$M_o = \sum_{i=1}^{N} M_i$	Total number of elements *(SU2)* in the population
$\overline{M} = \frac{M_o}{N}$	Mean number of elements *(SU2)* per cluster *(SU1).* This is useful when a population has clusters *(SU1)* of unequal size.
Y_{ij}	Value of the chosen characteristic of element *(SU2) j* in cluster *(SU1) i*
$Y_i = \sum_{j=1}^{M_i} Y_{ij}$	Total value of the chosen characteristic in cluster *(SU1) i*
$\overline{Y}_i = \frac{Y_i}{M_i}$	Mean value of the characteristic Y in the elements *(SU2)* of cluster *(SU1) i*
$Y = \sum_{i=1}^{N} Y_i$	Total value of the characteristic Y in the population
$\overline{Y} = \frac{Y}{N}$	Mean value of the characteristic Y per cluster (*SU1)*

TABLE 6.1 *(cont.)*

$\overline{\overline{Y}} = \frac{Y}{M_o} = \frac{Y}{N\overline{M}} = \frac{\overline{Y}}{\overline{M}}$	Mean value of the characteristic Y per element *(SU2)*
$\overline{\overline{Y}} = \sum_{i=1}^{N} \frac{\overline{Y}_i}{N}$	Mean value of the characteristic Y per element *(SU2)* if *M_i = constant = M*
Variances	
$S_1^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}$	Variance between total values of the characteristic Y per cluster
$S_1^{*2} = \frac{\sum_{i=1}^{N}(\overline{Y}_i - \overline{\overline{Y}})^2}{N-1}$	Variance between mean values of the characteristic Y per cluster *(SU1)* The asterisk is used in the symbol, S_1^{*2}, to differentiate the variance between mean values of the characteristic per cluster *(SU1)* and the variance S_1^2 between total values of the characteristic per cluster *(SU1)*
$S_{2i}^2 = \frac{\sum_{j=1}^{M_i}(Y_{ij} - \overline{Y}_i)^2}{M_i - 1}$	Variance between values of the characteristic Y in the elements *(SU2)* within cluster *(SU1) i*
$S_2^2 = \frac{\sum_{i=1}^{N} S_{2i}^2}{N}$	Variance between values of the characteristic Y in the elements *(SU2)* within all clusters *(SU1)*

6.3 THE SAMPLE

In this design, as opposed to cluster sampling, the numbers m_i of elements sampled in the second sampling stage are not equal to the sizes of the corresponding clusters. The sample statistics common of this design that are most important to fisheries research are summarised below.

TABLE 6.2
Main sample statistics of interest to fisheries research in two-stage sampling design

n	Number of clusters *(SU1)* sampled
m_i	Number of elements *(SU2)* sampled from *cluster (SU1) i*
$m_o = \sum_{i=1}^{n} m_i$	Total number of elements *(SU2)* sampled
$\overline{m} = \frac{m_o}{n}$	Mean number of elements *(SU2)* sampled per cluster *(SU1)*
y_{ij}	Value of the characteristic Y in element *(SU2) j* of cluster *(SU1) i*
$y_i = \sum_{j=1}^{m_i} y_{ij}$	Total value of the characteristic Y in the elements *(SU2)* sampled from cluster *(SU1) i*
$\overline{y}_i = \frac{y_i}{m_i}$	Mean value of the characteristic Y in the elements *(SU2)* sampled from cluster *(SU1) i*

TABLE 6.2 *(Cont.)*

$y = \sum\limits_{i=1}^{n} y_i$	Total value of the characteristic Y in the clusters *(SU1)* sampled
$\bar{y} = \dfrac{y}{n}$	Sample mean value of the characteristic Y per cluster *(SU1)*
$\bar{\bar{y}} = \dfrac{y}{m_o} = \dfrac{y}{n\bar{m}}$	Sample mean value of the characteristic Y per element *(SU2)*
$\bar{\bar{y}} = \sum\limits_{i=1}^{n} \dfrac{\bar{y}_i}{n}$	Sample mean value of the characteristic Y per element *(SU2)* if *m*m_i *= constant = m*
$s_1^2 = \dfrac{\sum\limits_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}$	Sample variance between total values of the characteristic Y per cluster *(SU1)*
$s_1^{*2} = \dfrac{\sum\limits_{i=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}\right)^2}{n-1}$	Sample variance between mean values of the characteristic Y per cluster *(SU1)*
$s_{2i}^2 = \dfrac{\sum\limits_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2}{m_i - 1}$	Sample variance between values of the characteristic Y in the elements *(SU2)* sampled within cluster *(SU1) i*
$s_2^2 = \dfrac{\sum\limits_{i=1}^{n} s_{2i}^2}{n}$	Sample variance between values of the characteristic Y in the elements *(SU2)* sampled within all clusters *(SU1),* if m_i *= constant = m*

6.4 THE SAMPLING WORLD

In two-stage sampling design, the expected values and the variances of an estimator in the sampling world are calculated taking into consideration the two stages.

Let the sub-indices 1 and 2 refer respectively to the first and to the second sampling stages.

First sampling stage

E_1 refers to the expected value of the estimator among all possible first-stage samples to be selected from the population. V_1 refers to the sampling variance of the estimator among all possible first-stage samples to be selected from the population.

Second sampling stage

E_2 refers to the expected value of the estimator among all possible second-stage samples to be selected from the first-stage clusters already sampled, that is, conditional on the *SU1* sampled.

V_2 refers to the sampling variance of the estimator among all possible second-stage samples to be selected from the first stage clusters already sampled, that is, conditional on the *SU1* sampled.

Using these definitions it can be demonstrated that, if $\hat{\theta}$ is an estimator of the population parameter θ, the expected value of the estimator is:

$$E\left[\hat{\theta}\right]=E_1\left[E_2\left(\hat{\theta}\right)\right]$$

and its sampling variance is:

$$V\left[\hat{\theta}\right]=V_1\left[E_2\left(\hat{\theta}\right)\right]+E_1\left[V_2\left(\hat{\theta}\right)\right]$$

The first term relates to the sampling variance of the estimator between the clusters *(SU1)* and the second term relates to the sampling variance between the elements *(SU2)* within the clusters *(SU1)*.

This basic theorem is valid for the sampling distribution of any estimator and it is valid for any two-stage sampling design. These results can also be extended to sampling designs with more stages.

There are different methods of selecting the sampling units in the two-stage sampling design. The units can be selected with simple random sampling, or with different probabilities in one or both stages. These choices will affect the sampling distribution of the estimators, and correspondingly the choice of estimators to use for any particular purpose.

In this document the following methods will be analysed:

- Simple random sampling at both stages
- Random sampling with different probabilities at the first stage, and simple random sampling at the second stage.

In two-stage sampling applied to fisheries science, the population total value, Y, and the mean per element, $\overline{\overline{Y}}$, are often the parameters to be estimated.

6.4.1 First selection by simple random sampling, without replacement, and second selection by simple random sampling, without replacement

In this two-stage sampling design, an unbiased estimator of the total value of the population is:

$$\hat{Y}=\frac{N}{n}\sum_{i=1}^{n}\hat{Y}_i$$

where $\hat{Y}_i$ is an estimator of the total value of the characteristic in cluster *(SU1) i.* Taking into consideration that simple random sampling is adopted in the second sampling stage, the estimator $\hat{Y}_i$ would be:

$$\hat{Y}_i=M_i\overline{y_i} \text{ or } \hat{Y}_i=\frac{M_i}{m_i}y_i$$

Applying the general theorem from section 6.4.2, it can be proven that the estimator is unbiased with sampling variance equal to:

$$V(\hat{Y})=N^2(1-f_1)\frac{S_1^2}{n}+\frac{N}{n}\sum_{i=1}^{N}M_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}$$

In this expression, the sampling fractions at the first and second stages are $f_1=\frac{n}{N}$

and $f_{2i}=\frac{m_i}{M_i}$ respectively, S_1^2 is the population variance between total values of the characteristic of clusters *(SU1)* and S_{2i}^2 is the population variance between the values of the characteristic of the elements *(SU2)* within cluster *(SU1) i.*

An estimate of the variance, $V(\hat{Y})$, is obtained by replacing the population variance S_1^2 with a sample estimate, $\hat{S}_1^2$ and the population variances S_{2i}^2 with the sample variances s_{2i}^2:

$$v(\hat{Y})=N^2(1-f_1)\frac{\hat{S}_1^2}{n}+\frac{N}{n}\sum_{i=1}^{n}M_i^2(1-f_{2i})\frac{s_{2i}^2}{m_i}$$

Particular case of SU1s with equal sizes

When the *SU1* sampling units have equal sizes and both selections are with equal probabilities ($\frac{1}{N}$ for the first stage and $\frac{1}{M}$ for the second stage), two-stage sampling design becomes a very simple particular case.

Let *M* be the constant number of second-stage sampling units, *SU2*, in any of the *N* clusters *(SU1)* of the population and *m* the constant number of *SU2* sampled from any *SU1* of the sample.

Replacing M_i with *M* and m_i with *m* in the previous case, the above sampling variance will become:

$$V(\hat{Y})=N^2(1-f_1)\frac{S_1^2}{n}+\frac{NM^2}{n}(1-f_2)\frac{\sum_{i=1}^{N}S_{2i}^2}{m}$$

In this case, the estimator of the mean per element, $\bar{\bar{y}}$, is given by $\bar{\bar{y}}=\frac{\hat{Y}}{NM}$. The sampling variance of this estimator is given by $V(\bar{\bar{y}})=\frac{V(\hat{Y})}{(NM)^2}$ and hence, from the previous expression of the sampling variance, this will be:

$$V(\bar{\bar{y}})=(1-f_1)\frac{S_1^2}{M^2n}+\frac{1}{nN}(1-f_2)\frac{\sum_{i=1}^{N}S_{2i}^2}{m}$$

It was earlier seen that, for equal sizes of clusters *SU1*s and equal sizes of the respective *SU2*s samples, the variances between mean values of the characteristic Y in the *SU1*s and the variances between values of the characteristic Y in the *SU2*s could be written as:

$$S_1^{*2}=\frac{S_1^2}{M^2} \text{ and } S_2^2=\frac{\sum_{i=1}^{N}S_{2i}^2}{N} \text{ respectively}$$

Then the sampling variance of $\bar{\bar{y}}$ will take the form:

$$V(\bar{\bar{y}})=(1-f_1)\frac{S_1^{*2}}{n}+(1-f_2)\frac{S_2^2}{nm}$$

An estimate of the sampling variance is:

$$v(\bar{\bar{y}})=(1-f_1)\frac{s_1^{*2}}{n}+f_1(1-f_2)\frac{s_2^2}{nm}$$

The second term of this variance is negligible when f_1 is small.

In the case of three-stage (or more stages) sampling with sampling units of equal sizes, the extension of the expressions above is simple. For example, considering three-stage sampling, the estimator of the population mean per element is:

$$\hat{\overline{\overline{\overline{Y}}}} = \overline{\overline{y}} \text{ where } \overline{\overline{y}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{k} y_{ijl}}{nmk}$$

The sampling variance of this estimator will be:

$$V\left(\overline{\overline{y}}\right) = (1-f_1)\frac{S_1^2}{n} + f_1(1-f_2)\frac{S_2^2}{nm} + f_1 f_2 (1-f_3)\frac{S_3^2}{nmk}$$

*w*here:

$$S_1^2 = \frac{\sum_{i=1}^{N}\left(\overline{\overline{Y}}_i - \overline{\overline{\overline{Y}}}\right)^2}{N-1} \quad S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\overline{Y}_{ij} - \overline{\overline{Y}}_i\right)^2}{N(M-1)} \quad S_3^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{K}\left(Y_{ijl} - \overline{Y}_{ij}\right)^2}{NM(K-1)}$$

with:

$$\overline{Y}_{ij} = \frac{\sum_{l=1}^{K} Y_{ijl}}{K} \quad \overline{\overline{Y}}_i = \frac{\sum_{j=1}^{M}\sum_{l=1}^{K} Y_{ijl}}{MK} \quad \overline{\overline{\overline{Y}}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{K} Y_{ijl}}{NMK}$$

An estimator of the sampling variance can be obtained from:

$$v\left(\overline{\overline{y}}\right) = (1-f_1)\frac{s_1^2}{n} + f_1(1-f_2)\frac{s_2^2}{nm} + f_1 f_2 (1-f_3)\frac{s_3^2}{nmk}$$

where:

$$s_1^2 = \frac{\sum_{i=1}^{n}\left(\overline{\overline{y}}_i - \overline{\overline{\overline{y}}}\right)^2}{n-1} \quad s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\overline{y}_{ij} - \overline{\overline{y}}_i\right)^2}{n(m-1)} \quad s_3^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{k}\left(y_{ijl} - \overline{y}_{ij}\right)^2}{nm(k-1)} \text{ and}$$

$$\overline{y}_{ij} = \frac{\sum_{l=1}^{k} y_{ijl}}{k} \quad \overline{\overline{y}}_i = \frac{\sum_{j=1}^{m}\sum_{l=1}^{k} y_{ijl}}{mk} \quad \overline{\overline{\overline{y}}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{k} y_{ijl}}{nmk}$$

In the case of two-stage sampling with equal-sized *SU1s*, it is also simple to estimate the proportion of elements of the population belonging to one certain category.

A proportion in a sample of size *n* is considered as a mean of *n* Bernoulli variables. Then, the proportion p_i, in the i^{th} sampled cluster i^{th} *SU1*, is:

$$p_i = \overline{y}_i$$

Then the sample mean per element is the overall proportion:

$$\overline{y} = \frac{\sum_{i=1}^{n}\overline{y}_i}{n} = \frac{\sum_{i=1}^{n} p_i}{n} = \overline{p}$$

Therefore, the estimator of the overall proportion of the elements belonging to the category of interest can be the average, $\hat{P}$, of the proportions, p_i, of the clusters sampled:

$$\hat{P} = \hat{\overline{\overline{Y}}} = \overline{p}$$

This estimator is unbiased, and an estimate of its sampling variance is given by:

$$v(\bar{p}) = (1-f_1)\frac{s_1^{*2}}{n} + f_1(1-f_2)\frac{s_2^2}{nm}$$

where

$$s_1^{*2} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})^2}{n-1} \text{ and } s_2^2 = \frac{\sum_{i=1}^{n} s_{2i}^2}{n} \text{ with } s_{2i}^2 = \frac{mp_i q_i}{m-1}$$

6.4.2 First selection with unequal probabilities, with replacement, and second selection with equal probabilities, with replacement

To analyse this design, let P_i be the known probability of selecting the *ith* cluster ($SU1_i$) in one extraction, $\sum_{i=1}^{N} P_i = 1$.

An unbiased estimator of the population total, Y, is:

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{P_i}$$

where $\hat{Y}_i = M_i \bar{y}_i$ is an estimator of the total value, Y_i, of the i^{th} $SU1$.

The sampling variance of this estimator will depend on the sampling design of the first stage. However if independent estimates, $\hat{Y}_i$, of the total values, Y_i, are available, an unbiased estimator of the sampling variance, will be:

$$v\left[\hat{Y}\right] = \frac{\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{P_i} - \hat{Y}\right)^2}{n(n-1)}$$

The estimators presented for multi-stage sampling are in general efficient, but they suffer from the handicap of having complicated expressions. This complication arises from the unequal selection probabilities, requiring the calculation of weights for each cluster. Under some conditions of sample allocation, however, this estimator can be self-weighting.

In fact, the estimator of the total value in the population can be rewritten as:

$$\hat{Y} = \sum_{i=1}^{n}\frac{M_i}{nP_i}\frac{y_i}{m_i} \text{ with } y_i = \sum_{j=1}^{m_i} y_{ij}$$

If the sample allocation is such that $m_i = const \cdot \frac{M_i}{n \cdot P_i}$, then $\frac{nm_i P_i}{M_i} = const = f_0$, and the estimator can be rewritten:

$$\hat{Y} = \frac{1}{f_0}\sum_{i=1}^{n} y_i \text{ showing that the sample is self-weighted.}$$

The importance of self-weighting sampling is to facilitate the calculations, since the constant weight factors simplify the calculations of the estimators and of the estimated sampling variances.

Under these conditions, an estimate of the sampling variance is:

$$v\left[\hat{Y}\right]=\frac{n}{f_0^2}\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{n-1}$$

The population mean per element, $\overline{\overline{Y}}$, can be estimated using the estimator $\overline{\overline{Y}}=\frac{\hat{Y}}{M_0}$.

The sampling variance of this estimator is:

$$v\left[\hat{\overline{\overline{Y}}}\right]=\frac{v\left[\hat{Y}\right]}{M_0^2}=\frac{1}{M_0^2}\frac{\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{P_i}-\hat{Y}\right)^2}{n(n-1)}$$

As in the previous cases, this estimator requires that M_0, the total number of elements of the population, is known.

7. Biological sampling at landing ports for one resource

7.1 INTRODUCTION

The objective of this chapter is to describe a practical sampling system for length composition of the annual total landings of one fishery resource along the coast of a certain area. The reason why only one resource is being used is just to simplify the explanations and calculations, but the sampling system could be developed for several resources. The sampling methods starts with a stratification of the population of landings of sardine, followed by multi-stage sampling applied to each *stratum*. Two of the sampling designs dealt with in previous chapters, stratified sampling and simple random sampling, are the basis for the whole plan.

The sampling considers a region with several landing ports, different components of the fleet and the months of the annual landings. The specific example used in this chapter is an example of how to calculate the annual length composition of the *Sardinella aurita* fishery in a coastal area with several landing ports, where small purse seiners and trawlers fish this species.

7.2 OBJECTIVES

The annual length composition of the landings is important in itself, and equally as a basis for estimating other biological characteristics of the landings. Combining the information on the length composition of the landings with other information, such as:

- average weight by length class;
- age composition in each length group;
- percentage of mature individuals in each length group;
- percentage of sexes in each length group;
- other biological characteristics (stomach contents, gonad somatic index, etc.) by size groups.

The characteristics of the landings as a whole (e.g. length composition of the overall landings) can be estimated.

In order to achieve these objectives, landing statistics (total annual catch) are needed. In this case it is assumed that all the fish caught in one year, in the area and by all fleet components come from a single population of the resource.

If only the annual length composition of the landings is needed, one single sampling scheme will be set up. If on the other hand, other biological characteristics like those mentioned above (percentage of adults, age-length keys by length group, *etc.*) are intended, separate sampling systems should be established.

A crucial issue in all these sampling designs is to ensure that the observation techniques used to determine the characteristics (length, weight, maturation, sex, age, etc.) are well defined. Although this may seem obvious, often sampling systems are invalidated due to inconsistencies during sampling activities. Therefore if, for example, total length is to be measured, every effort should be made to ensure that all the samplers are measuring the total length and not the fork length. Attention should also be paid as to whether the fish are measured to the *cm* below, to the *cm* above or to the nearest *cm*.

7.3 EXAMPLE OF A SAMPLING SCHEME FOR THE ANNUAL LENGTH COMPOSITION OF THE LANDINGS

Definition of the population and sampling scheme

As mentioned above, the population of interest is composed of all fishes caught by the entire fleet fishing the resource that are landed in all the ports of the area being considered, throughout one year.

Total annual landings, which is the population to be sampled, will be divided into *strata* following a separation by month, fleet component and landing port.

The number of *strata* is the product of the number of months (*m*) by the number of fleet components (f_c) by the number of landing ports (*p*):

Number of strata = m × fc × p

Number of strata = 12 months × 2 fleet components × 3 ports = 72 strata

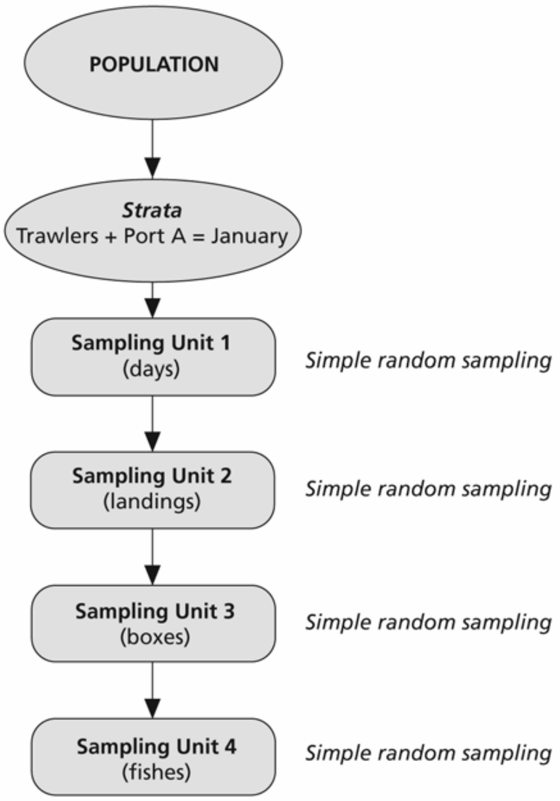
Let us take the example of the *Sardinella aurita* fishery mentioned above and assume that the region has three landing ports. Remember that the *Sardinella aurita* was being fished by two fleet components (small seiners and trawlers) and that the objective is to estimate the annual (twelve months) length composition.

According to sampling theory (see Chapter 4 on stratified sampling) all *strata* must be sampled. This means that a minimum of 72 samples must be taken. But only one sample from each *stratum* (month, port and fleet component) is insufficient for estimating the precision or the error of the annual length composition (at least 2 or preferably 3 samples from each *stratum* should be taken). Let us consider only one *stratum*, represented by the trawlers landing *Sardinella aurita* in Port A during the month of January. Figure 7.1 illustrates the sampling scheme adopted in this situation.

The multistage sampling scheme represented in Figure 7.1 could be explained as follows:

- in a *stratum* (defined by one fleet category, one landing port and one month) several days (selected by simple random sampling) will be sampled (sampling unit 1 – *SU1*);

FIGURE 7.1
Multi-stage sampling scheme for length composition of the landings of *Sardinella aurita* made by trawlers in Port "A" during January



- for each day selected for sampling, out of the total number of trawler landings, several trawler landings are selected (by simple random sampling) (sampling unit 2 – *SU2*);
- for each trawler landing selected in the previous step, several fish boxes are selected (by simple random sampling) (sampling unit 3 – *SU3*);
- finally, from each of these boxes a number of fishes is selected (using simple random sampling without replacement) and measured (sampling unit 4 – *SU4*).

In the last stage the length composition is calculated from the sampling unit 4. Then, by successive back-calculations, the sampled length composition is raised to the length composition of the *stratum*.

Table 7.1 summarizes our sampling scheme. The notation is the same as that used in previous chapters. In order to simplify the explanation, only one *stratum* (one month, one port and one fleet component) has been considered. The procedure is repeated independently, for each stratum.

TABLE 7.1
Summary of a sampling scheme for estimating the length composition of landings

POPULATION	*h* is the index of a *stratum:* one fleet component, one port and one month *h=1…K*	Sample
(N_h) total number of days (*stratum* size) in *stratum h*	*i*, index for days *i=1…N_h*	(n_h) number of days sampled (sample size) in *stratum h* (simple random sampling)
(N_{hi}) total number of landings in day *i* in *stratum h*	*j*, index for landings $j=1\ldots N_{hi}$	(n_{hi}) total number of landings sampled in day i in *stratum h* (simple random sampling)
(N_{hij}) total number of boxes in landing *j* of day *i* in *stratum h*	*k*, index for boxes $k=1\ldots N_{hij}$	(n_{hij}) total number of boxes sampled from landing j of day *i* in *stratum h* (simple random sampling)
(N_{hijk}) total number of fishes in box k of landing *j* from day *i* in *stratum h*	*l*, index for fishes $l=1\ldots N_{hijk}$	(n_{hijk}) total number of fishes sampled from box k of landing j from day *i* in *stratum h* (simple random sampling)

Sampling

The aim of this example was to illustrate how to raise successive estimates of each stage up to the *stratum* level. However to simplify the presentation we have reduced the number of some sampling units to one (as mentioned in the text this is not advisable in practical cases). Another simplification was to eliminate the first sub-index (days).

Let us consider once again the *Sardinella aurita* example in our particular area. The area under consideration has three landing ports (Port A, Port B and Port C). Each port has two fleet components fishing for *Sardinella aurita* (trawlers and small seiners). The objective is to estimate the annual (12 months) length composition of the landings. Remember that according to sampling theory, it is mandatory that all *strata* be sampled. In this particular case study there are hence 72 *strata*.

Only the sampling of one *stratum* (Trawlers, Port A and January) is worked in detail, with the aim of simplifying the calculations. Also for simplicity the sub-index for *stratum* is not indicated in the text and tables. The procedure is carried out for all the 72 *strata* in order to obtain the annual length composition.

One day ("Day 1") is selected (by simple random sampling) from the 31 days of the month of January in Port A. During the selected day, 30 trawlers landed *Sardinella aurita* at Port A. Out of the 30 landings, three landings ("Landing 1", "Landing 2" and "Landing 3") were selected (by simple random sampling).

"Landing 1" was composed of 20 boxes in total, "Landing 2" of 10 boxes in total and "Landing 3" of 8 boxes in total. From "Landing 1" two boxes ("Box 1.1" and "Box 1.2") out of the 20 are selected; from "Landing 2" one box ("Box 2.1") is selected out of 10 and from "Landing 3" one box is also selected ("Box 3.1") out of 8.

All boxes were selected using a simple random sampling method. For simplicity more complicated sampling methods were avoided, e.g. selection with probabilities proportional to sizes of landings measured by the number of boxes.

From each of the boxes selected one bucket ("Bucket 1.1.1", "Bucket 1.2.1", "Bucket 2.1.1", and "Bucket 3.1.1") of fishes is sampled (simple random sampling without replacement). The number of fishes in each bucket is:

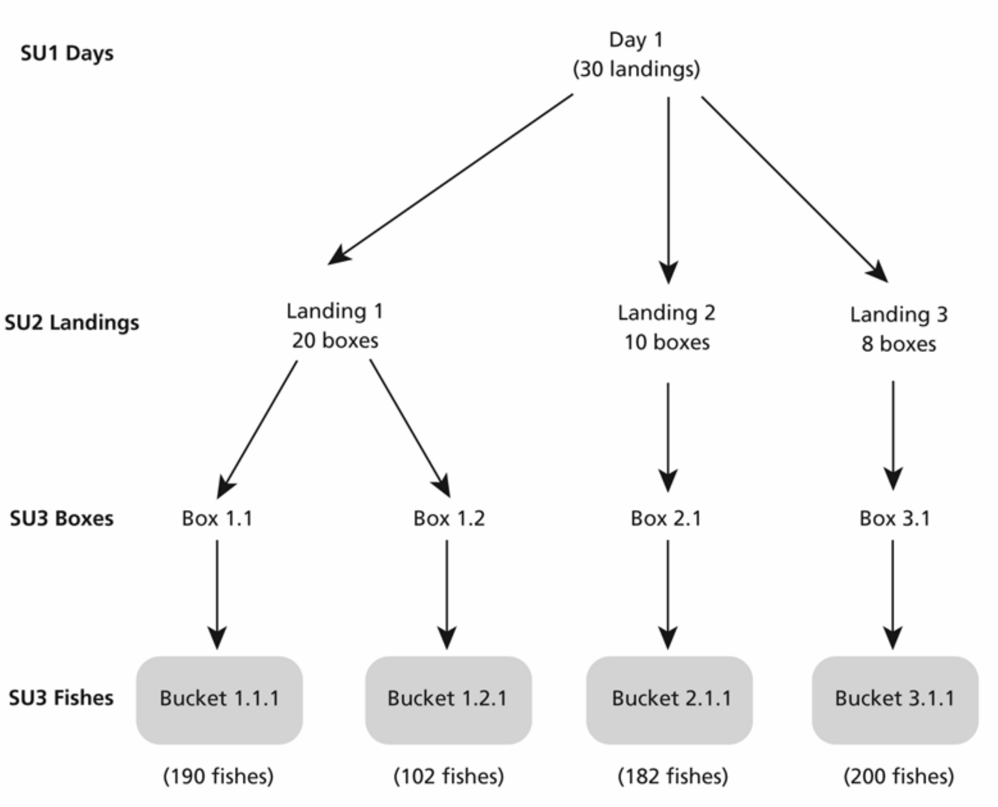
- Bucket 1.1.1 = 190 fishes
- Bucket 1.2.1 = 102 fishes
- Bucket 2.1.1 = 182 fishes
- Bucket 3.1.1 = 200 fishes

Previous calculations estimated that the weight of a box of *Sardinella aurita* was 30 kg and the weight of a bucket of *Sardinella aurita* was 10 kg.

The length of each fish is measured with the same observation technique, and the fishes are grouped into length classes. In the example 10 length classes are considered.

The example could be represented by the following figure (Figure 7.2):

FIGURE 7.2
Scheme of the procedure to obtain the length composition for the landings of *Sardinella aurita* made by trawlers in Port "A" during January



Calculations

The measurements are organized in frequency tables (one table for each bucket).

TABLE 7.2
Frequency table for fish length – fish from buckets

Bucket 1.1.1		Bucket 1.2.1		Bucket 2.1.1		Bucket 3.1.1	
Classes *(cm)*	Frequency	Classes *(cm)*	Frequency	Classes *(cm)*	Frequency	Classes *(cm)*	Frequency
l_1	4	l_1	0	l_1	8	l_1	8
l_2	8	l_2	0	l_2	10	l_2	14
l_3	30	l_3	5	l_3	23	l_3	24
l_4	35	l_4	7	l_4	30	l_4	30
l_5	37	l_5	17	l_5	30	l_5	37
l_6	30	l_6	23	l_6	27	l_6	38
l_7	27	l_7	20	l_7	20	l_7	27
l_8	10	l_8	15	l_8	15	l_8	10
l_9	8	l_9	10	l_9	12	l_9	8
l_{10}	1	l_{10}	5	l_{10}	7	l_{10}	4
Total	**190**	**Total**	**102**	**Total**	**182**	**Total**	**200**

The sample frequencies are raised to the total of each box. In order to do this, a factor, RF, that is the quotient of the box weight to the sample weight (in this particular case the sample weight is the weight of the full bucket of *Sardinella aurita*), is needed. It is known that the weight of each box is 30 kg and the weight of each bucket is 10 kg, therefore the factor would be:

RF = box weight / sample weight = 30/10 = 3

Each frequency is multiplied by the respective RF, in this case equal to 3. Table 7.3 shows the results.

TABLE 7.3
Length composition by box after raising frequencies

Box 1.1		Box 1.2		Box 2.1		Box 3.1	
Classes *(cm)*	Frequency	Classes *(cm)*	Frequency	Classes *(cm)*	Frequency	Classes *(cm)*	Frequency
l_1	12	l_1	0	l_1	24	l_1	24
l_2	24	l_2	0	l_2	30	l_2	42
l_3	90	l_3	15	l_3	69	l_3	72
l_4	105	l_4	21	l_4	90	l_4	90
l_5	111	l_5	51	l_5	90	l_5	111
l_6	90	l_6	69	l_6	81	l_6	114
l_7	81	l_7	60	l_7	60	l_7	81
l_8	30	l_8	45	l_8	45	l_8	30
l_9	24	l_9	30	l_9	36	l_9	24
l_{10}	3	l_{10}	15	l_{10}	21	l_{10}	12

Now the length composition of each landing is calculated. In the case of Landing 1, since two different boxes were sampled, the length compositions of Box 1.1 and Box 1.2 are summed (Table 7.4).

TABLE 7.4
Length composition by landing (boxes 1.1 and 1.2 are summed)

Box 1.1 + Box 1.2		Box 2.1		Box 3.1	
Classes (cm)	Frequency	Classes (cm)	Frequency	Classes (cm)	Frequency
l_1	12	l_1	24	l_1	24
l_2	24	l_2	30	l_2	42
l_3	105	l_3	69	l_3	72
l_4	126	l_4	90	l_4	90
l_5	162	l_5	90	l_5	111
l_6	159	l_6	81	l_6	114
l_7	141	l_7	60	l_7	81
l_8	75	l_8	45	l_8	30
l_9	54	l_9	36	l_9	24
l_{10}	18	l_{10}	21	l_{10}	12

The frequencies of the boxes are raised to the total of each landing. The raising factor, RF, is the quotient of the total number of boxes landed by the total number of boxes sampled. Therefore the factors are:

$$RF_{\text{Landing 1}} = \frac{\text{total number of boxes landed}}{\text{total number of boxes sampled}} = \frac{20}{2} = 10$$

$$RF_{\text{Landing 2}} = \frac{\text{total number of boxes landed}}{\text{total number of boxes sampled}} = \frac{10}{1} = 10$$

$$RF_{\text{Landing 3}} = \frac{\text{total number of boxes landed}}{\text{total number of boxes sampled}} = \frac{8}{1} = 8$$

Table 7.5 shows the results of this raising.

TABLE 7.5
Length composition of the landings sampled

Landing 1		Landing 2		Landing 3	
Classes (cm)	Frequency	Classes (cm)	Frequency	Classes (cm)	Frequency
l_1	120	l_1	240	l_1	192
l_2	240	l_2	300	l_2	336
l_3	1050	l_3	690	l_3	576
l_4	1260	l_4	900	l_4	720
l_5	1620	l_5	900	l_5	888
l_6	1590	l_6	810	l_6	912
l_7	1410	l_7	600	l_7	648
l_8	750	l_8	450	l_8	240
l_9	540	l_9	360	l_9	192
l_{10}	180	l_{10}	210	l_{10}	96

The length compositions of the sampled landings are now summed as shown in Table 7.6.

TABLE 7.6
Sum of the length composition of landings

	Landing 1 + Landing 2 + Landing 3
Classes *(cm)*	**Frequency**
l_1	552
l_2	876
l_3	2316
l_4	2880
l_5	3408
l_6	3312
l_7	2658
l_8	1440
l_9	1092
l_{10}	486

In order to raise the length composition to the total landing of the day, a factor, RF, that is, the quotient between the total number of landings during the day by the total number of landings sampled, is needed. Therefore the raising factor would be:

$$\mathrm{RF} = \frac{\text{total number of landings during the day}}{\text{total number of landings of that day sampled}} = \frac{30}{3} = 10$$

Table 7.7 gives the final result.

TABLE 7.7
Length composition for the total landings of trawlers in Port A on January 1

	Total landings of day
Classes *(cm)*	**Frequency**
l_1	5520
l_2	8760
l_3	23160
l_4	28800
l_5	34080
l_6	33120
l_7	26580
l_8	14400
l_9	10920
l_{10}	4860

The length composition of the total landings from the *stratum* (Trawlers, Port A and January) is obtained applying procedures similar to the ones just described to all the days sampled, summing them, and then raising the total length composition to the whole month. In this last raising, the raising factor is the ratio between the total number of days with landings in the month and the number of days sampled,

$$\mathrm{RF} = \frac{\text{Number of days with landings during the month}}{\text{Number of days sampled}}$$

The total length composition of *Sardinella aurita* landings during a month, and during a year are obtained adding together the raised length compositions for the month or for the year.

7.4 FINAL COMMENTS

- Sampling theory provides an estimator of the length composition of the landings.
- Sampling theory indicates the expected value, the sampling variance and the sampling distribution of the values of the estimator, as well as the relation between these and the parameters of the population.
- Our first objective with this sampling scheme was to obtain the annual length composition of a resource (*Sardinella aurita*) landed in a country. So we presented the calculations only for one day and did not extend the calculations to all *strata*.
- Finally we are not trying to apply any other designs or to calculate sampling variances or errors of the estimators.

8. Exercises

8.1 EXERCISE 1

Group I – Small samples

During the year 2000, a sample of Portuguese landings of hake (*Merluccius merluccius*) were recorded. Catches, in tonnes were the following:

Catches (in tonnes) of a sample of Portuguese landings of hake

48	372	174	165	130	473	148
39	474	155	288	176	277	349
301	508	339	114	211	477	170
304	255	409	267	274	166	211
299	353	192				

1. Calculate:
   a) The sample size.
   b) The range of the sample values.
   c) The median.
   d) The mean.
   e) The total value.
   f) The sample variance.
   g) The sample standard deviation.
   h) The sample coefficient of variation.

Group II – Large samples

A sample of 195 individuals was extracted from the catch of a trawler. The individual total lengths were measured to the cm below. Registered values were the following:

Total length (in cm) of a catch

24	15	16	17	18	19	20	30	26	23	22	24	27	22	21
24	20	21	17	18	19	20	30	26	28	22	24	27	22	21
19	20	21	17	18	19	20	30	26	28	22	24	23	22	21
19	20	21	29	18	19	20	30	27	28	19	24	23	22	21
19	20	21	18	30	19	20	30	27	28	19	24	23	26	21
25	20	21	25	20	19	19	26	27	28	19	24	23	26	21
25	20	21	25	20	21	19	26	27	28	29	24	23	26	21
25	20	21	25	20	21	19	22	27	28	29	24	23	26	18
25	20	23	18	20	21	19	22	31	28	29	24	23	22	18
19	23	23	18	20	21	19	22	31	28	29	24	23	22	18
19	23	23	18	20	21	19	22	26	20	29	26	21	27	22
19	22	23	18	20	21	19	19	22	20	20	20	21	27	22
26	26	23	27	23	27	23	19	21	25	25	25	25	25	22

1. From this data calculate the mean and the variance.
2. Choose an adequate class interval and build up a table with the length frequencies distribution.
3. From the table built in 2., calculate:
   a) The sample mean and sample variance. Compare these results with the ones obtained in 1.
   b) Three statistics of location.
   c) Three statistics of dispersion.
   d) Number of individuals with a length less than 20 cm.
   e) Percentage of individuals with a length equal to or greater than 20 cm.
   f) Percentage of observations between 23 and 25 cm.
   g) The value that corresponds to a length equal to or greater than 45% of all the observations.
   h) The value that corresponds to a length smaller than 21% of all the observations.
   i) The quantile of order 96%.

8.2 EXERCISE 2

Group I – Relative frequencies

In certain ports, the fishing gears used by the vessels were classified into "purse seines", "trawls", "handlines", "longlines" and "trammel nets". We also know the types of the fishing vessels, that is, "small boats without engine", "small boats with engine", "purse seiners" and "stern trawlers". The information about the numbers of vessels according to boat type and gear used is summarized in Table 8.1.

TABLE 8.1
Numbers of vessels according to the type of boat and gear used

Type of vessel	Fishing gear					Total
	Purse seines	Trawls	Handlines	Longlines	Trammel nets	
Small boats without engine	15	0	22	20	8	65
Small boats with engine	15	4	23	40	17	99
Purse seiners	50	0	0	0	0	50
Stern Trawlers	0	51	0	0	0	51
Total	**80**	**55**	**45**	**60**	**25**	**265**

1. Calculate:
   a) Relative frequencies of number of boats by fishing gear.
   b) Percentage of vessels that operate handlines.
   c) Percentage of vessels that operate trammel nets.
2. Calculate:
   a) Relative frequency of boats not operating trammel nets.
   b) Percentage of boats operating purse seines or longlines.
   c) Relative frequency of boats that do not operate with handlines, nor trammel nets, nor longlines.
   d) Proportion of the total fleet that are small boats without engine.
   e) Proportion of the total fleet that are small boats with engine.
   f) Proportion of the total fleet that are small boats.
   g) Check that the proportion of 2.f) is the sum of 2.d) plus 2.e).

3. Calculate:
    a) Proportion of small boats without engine that operate handlines.
    b) Proportion of the total fleet that are small boats without engine.
    c) Proportion of the total fleet that are small boats without engine operating handlines.
    d) Check that the proportion of 3.c) is the product of 3.a) times 3.b).
4. Calculate:
    a) The percentage of small boats without engine operating handlines or longlines.
    b) The percentage of purse seiners operating purse seines.
    c) The relative frequency of vessels that are not purse-seiners.
    d) The percentage of small boats with engine that operate trawls.
    e) The percentage of the fleet that fishes with traps.

Group II – Properties of probabilities

Consider Table 8.1 presented in Group I.

We want to choose one boat, randomly, out of the 265 boats. Every boat has the same probability of being selected.

1. What is the probability that the boat operates:
    a) Handlines.
    b) Trammel nets.
    c) Does not operate trammel nets.
    d) Operates purse-seines or longlines.
    e) Does not operate handlines, nor longlines neither trammel nets.
2. What is the probability that the boat will be:
    a) A small boat without engine.
    b) A small boat with engine.
    c) A small boat.
    d) Show that the probability 2.c) is equal to the sum of probability 2.a) plus the probability 2.b).
3. Calculate the probability of the boat being:
    a) A purse seiner.
    b) A stern trawler.
    c) Neither a purse seiner nor a stern trawler.
    d) Show that probability 3.c) is equal to probability 2.c).
4. Calculate:
    a) If we choose a boat from the small boats without engine, what is the probability that she operates with handline?
    b) If we choose a boat out of the total fleet what is the probability that she is a small boat without engine?
    c) If we choose a boat out of the total fleet what is the probability that she is a small boat without engine operating with handline?
    d) Check that the probability of 6.c) is equal to the product of the probability of 4.a) times the probability of the 4.b).

Group III – Normal distribution

A random variable X is normally distributed with a mean μ = 20.6 and a standard deviation $\sigma = 2$.

1. Calculate:
    a) The probability of X being less than or equal to 18.
    b) The probability of X being greater than 18.
    c) The probability of X being less than 25.
    d) The probability of X being between 18 and 25.

2. Calculate x such that:
    a) $Prob\{X \leq x\} = 0.8413$.
    b) $Prob\{X \geq x\} = 0.9772$.
    c) $Prob\{X < x\} = 0.9986$.
    d) x is the 95% quantile of the distribution of X.
    e) x is the median of the distribution of X.

Group IV – The standard normal distribution

The random variable Z has a standard normal distribution.

1. Calculate:
    a) The probability of the values of Z being between -1 and 1.
    b) The probability of the values of Z being between -2 and 2.
    c) The probability of the values of Z being between -3 and 3.
2. Calculate:
    a) The z_1 value for which the probability that the values of the variable Z will be smaller than z_1 is 2.5% ($Prob\{Z<z_1\}=0.025$).
    b) The z_2 value for which the probability of the variable Z being smaller than z_2 is 97.5% ($Prob\{Z<z_2\}=0.975$).
3. Consider the interval limited by z_1 and z_2 of the previous exercises 2.a) and 2.b).
    a) Compute the probability that the variable Z will be within the interval (z_1, z_2).
    b) Repeat exercise 3.a) but with ($Prob\{Z<z_1\}= 0.004$) and ($Prob\{Z<z_2\}= 0.954$).
    c) Repeat exercise 3.a) but with probabilities 0.012 and 0.962.
    d) Note that the $Prob\{z_1 \leq Z \leq z_2\}$ is equal to 0.950 in the exercises 3.a), 3.b) and 3.c). Verify that the smallest of these intervals is the interval with symmetrical values, z_1 and z_2.

Group V – *t*-student distribution

Using the *t*-student distribution:

1. Calculate:
    a) $Prob\{t(10) >1.812\}$.
    b) $Prob\{t(19) <1.729\}$.
    c) $Prob\{-1.34 < t(15) < +2.602\}$.
2. Calculate the value of a that makes the following expressions true:
    a) $Prob\{t(8) < a\} = 0.95$.
    b) $Prob\{t(26) > a\} = 0.99$.
    c) $Prob\{-a < t(20) < +a\} = 0.95$.

The general result of Group IV 3.d) is also true for the *t*-student distribution.

3. Calculate the a value such that:
    a) $Prob\{t < a\} = 0.95$, with 40, 60, 120 and infinite degrees of freedom.
    b) Compare the values a obtained in 3.a) with the corresponding a values if the probability distribution of 3.a) was replaced with the Z-distribution, *i.e.*, $Prob\{Z < a\} = 0.95$.

Group VI – Bernoulli distribution

Consider the discrete variable X which takes the value 1 with probability $P= 0.18$ and the value 0 with probability $Q= 0.82$.

1. Show that the expected value of the random variable X is equal to P.
2. Show that the variance of the random variable X is equal to PQ.

8.3 EXERCISE 3

An important element in fish stock assessment is the knowledge of the total catch landed for each of the main fisheries and species. The total catch per fishing trip is an element in this assessment, but given the irregular pattern of port calls and the low numbers of port samplers, it is difficult to get many records.

In a given fishing port, the port samplers have been instructed to record the total shrimp catch of at least three landings every week, which they do using a pre-determined and fixed sampling strategy.

Group I – Sample

In the first week of May 2000, they recorded the catch of three shrimp trawlers. The data collected is shown in the following table.

Shrimp landings recorded in three landings (first week of May 2000)

Landing no.	1	2	3
Shrimp landing (Kg)	538	435	1352

1. Calculate, from the sample and the knowledge of total number of landings:
   a) The mean shrimp landing per sampled fishing trip.
   b) The variance of the sampled landings.
   c) The standard deviation of the sampled landings.
2. By the end of the week, and from the port records, the scientist in charge of the sampling programme learned that a total of 10 landings of shrimp were done during that week. Guess the total amount of shrimp landed in that week.

Group II – Population

A few months later, for the purpose of this exercise, an agreement with the fishing companies gave the scientist access to the data from all landings actually done on that week (table below). These data represent thus the population of shrimp landings done during that week.

Population of all shrimp landings (in kg)

Landing no.	1	2	3	4	5	6	7	8	9	10
Shrimp landing	538	0	906	442	598	0	435	859	1352	711

1. Using these data, calculate the following parameters:
   a) The population mean, that is, the average shrimp landing per fishing trip during that week.
   b) The population variance and the modified variance of the landings.
   c) The standard deviation of the landings.
   d) The total amount of shrimp landed.
   e) The proportion of all landings below 400 Kg.
   f) The relative frequency of landings between 400 and 800 Kg.
2. Build at least 10 samples of 3 landings each that could have been selected from that population.
3. Repeat the calculations done on number 1. a) to d), for each of these samples.
4. Compare the values of the statistics obtained in the previous item with the values of the corresponding population parameters.

Group III – Sampling

Table 8.2 is presented at the end of this exercise with data from the 120 different samples of 3 landings that could have been taken from the 10 landings that actually took place during that week.

Using these data, and adopting the sample mean landing as the estimator, $\hat{\bar{Y}}$, of the population mean landing.

1. Plot the histogram of the sampling distribution of the estimator, $\hat{\bar{Y}}$, using an appropriate class interval.
2. Calculate from the data of the sampling distribution of the samples presented in Table 8.2:
   a) The 120 values of the estimator.
   b) The expected value of the estimator, $E\left[\hat{\bar{Y}}\right]$.
   c) The sampling variance of the same estimator, $V\left[\hat{\bar{Y}}\right]$.
   d) The error, $\sigma_{\hat{\bar{Y}}}$ of the estimator.
3. Compare the expected value obtained in 2.b) with the population calculated in Group I – 1.a).
4. Using the results obtained in previous exercises, check the theoretical expression:

$$V\left[\hat{\bar{Y}}\right] = (1 - \frac{n}{N})\frac{S^2}{n}$$

5. Calculate:
   a) The percentiles of the sampling distribution of the estimator with the following orders:
      i) 1.0%.
      ii) 2.5%.
      iii) 3.5%.
      iv) 50.0%.
      v) 95.0%.
      vi) 96.0%.
      vii) 97.5%.
      viii) 98.5%.
   b) Four intervals that encompass 95% of all possible sample means.
   c) The width of the four intervals.
6. What is the shortest of these intervals that holds 95% of all possible sample means.
7. Considering that the port samplers could have taken any of the possible samples, calculate the probability of getting a sample of 3 landings with an average landing.
   a) Below or equal to 600 Kg.
   b) Above 600 Kg.
   c) Between 199 and 953 Kg.
8. Calculate:
   a) The value l such that there is a probability of 95% of getting a sample with an average landing smaller than l.
   b) Two values l_1 and l_2 that there is a probability of 95% of getting a sample with an average landing between l_1 and l_2.

TABLE 8.2

All possible samples of 3 landings that could have been taken from the 10 landings that actually took place in that week

No.	Landing 1	Landing 2	Landing 3	No.	Landing 1	Landing 2	Landing 3	No.	Landing 1	Landing 2	Landing 3
1	538	0	906	**41**	0	906	859	**81**	906	435	1123
2	538	0	230	**42**	0	906	1123	**82**	906	435	711
3	538	0	598	**43**	0	906	711	**83**	906	859	1123
4	538	0	20	**44**	0	230	598	**84**	906	859	711
5	538	0	435	**45**	0	230	20	**85**	906	1123	711
6	538	0	859	**46**	0	230	435	**86**	230	598	20
7	538	0	1123	**47**	0	230	859	**87**	230	598	435
8	538	0	711	**48**	0	230	1123	**88**	230	598	859
9	538	906	230	**49**	0	230	711	**89**	230	598	1123
10	538	906	598	**50**	0	598	20	**90**	230	598	711
11	538	906	20	**51**	0	598	435	**91**	230	20	435
12	538	906	435	**52**	0	598	859	**92**	230	20	859
13	538	906	859	**53**	0	598	1123	**93**	230	20	1123
14	538	906	1123	**54**	0	598	711	**94**	230	20	711
15	538	906	711	**55**	0	20	435	**95**	230	435	859
16	538	230	598	**56**	0	20	859	**96**	230	435	1123
17	538	230	20	**57**	0	20	1123	**97**	230	435	711
18	538	230	435	**58**	0	20	711	**98**	230	859	1123
19	538	230	859	**59**	0	435	859	**99**	230	859	711
20	538	230	1123	**60**	0	435	1123	**100**	230	1123	711
21	538	230	711	**61**	0	435	711	**101**	598	20	435
22	538	598	20	**62**	0	859	1123	**102**	598	20	859
23	538	598	435	**63**	0	859	711	**103**	598	20	1123
24	538	598	859	**64**	0	1123	711	**104**	598	20	711
25	538	598	1123	**65**	906	230	598	**105**	598	435	859
26	538	598	711	**66**	906	230	20	**106**	598	435	1123
27	538	20	435	**67**	906	230	435	**107**	598	435	711
28	538	20	859	**68**	906	230	859	**108**	598	859	1123
29	538	20	1123	**69**	906	230	1123	**109**	598	859	711
30	538	20	711	**70**	906	230	711	**110**	598	1123	711
31	538	435	859	**71**	906	598	20	**111**	20	435	859
32	538	435	1123	**72**	906	598	435	**112**	20	435	1123
33	538	435	711	**73**	906	598	859	**113**	20	435	711
34	538	859	1123	**74**	906	598	1123	**114**	20	859	1123
35	538	859	711	**75**	906	598	711	**115**	20	859	711
36	538	1123	711	**76**	906	20	435	**116**	20	1123	711
37	0	906	230	**77**	906	20	859	**117**	435	859	1123
38	0	906	598	**78**	906	20	1123	**118**	435	859	711
39	0	906	20	**79**	906	20	711	**119**	435	1123	711
40	0	906	435	**80**	906	435	859	**120**	859	1123	711

8.4 EXERCISE 4

Group I – Estimation of the population mean

Consider a Population with $\overline{Y} = 40$, $S^2 = 25$ and $N = 2000$. A sample of 21 elements was selected from the population, with a simple random sampling design, without replacement.

Table below presents the values selected:

Sample data

30	42	38	38	41	42	42	46	36	42	34	35	40	35	39	38	39	40	37	46	45

1. Compute the following statistics from the sample:
   c) The mean $\overline{y}$.
   d) The variance s^2.
   e) The standard deviation s.
2. Adopt $\overline{y}$ as an estimator of the population mean, μ, and estimate:
   a) The population mean.
   b) The estimator is biased?
   c) The sampling variance of $\overline{y}$.
   d) The error of $\overline{y}$.
   e) A 95% confidence interval of μ.

Group II – Estimation of the population total

Consider a population of 20 purse seiners landing their catches at a certain port during one day.

The vessels are numbered from 1 to 20 according to the arrival time.

1. Describe a procedure to select a simple random sample of 4 numbered vessels.
2. Consider the following possible sample:
   Vessels number **3, 15, 10** and **6**
   At the arrival of the vessels it was verified that their corresponding landings were:
   Landings (Kg) **17.9, 2.8, 6.5** and **3.5**
   a) Choose an estimator of the total amount of fish landed in the port during this day.
   b) Indicate the sampling distribution of that estimator and present the formulae to obtain the expected value and the expected sampling variance.
3. Based on the sample presented in 2., estimate:
   a) The total landing.
   b) The sampling variance.
   c) The error of the estimate.
   d) A 95% confidence interval for the population total landings.
4. Estimate the approximate size of the sample necessary if one would like to have an error 10% smaller than the one previously calculated in 3.c).

Group III – Proportions

Consider a population of 100 shrimps in a box.

The aim is to estimate the proportion, P, of females within the box and the total number of females within the box. It was decided to select a simple random sample of size $n = 30$, and to adopt as estimator of P, the proportion, p, of females in the sample. The number of females in the sample was 12.

1. Estimator of the proportions.
    a) Calculate the proportion of females in the sample.
    b) Calculate the sample variance.
    c) Write, according to the sampling theory, the expressions for the expected value of p, and for the sampling variance of p.
    d) Estimate the sampling variance of p.
    e) Estimate the error of p.
    f) Estimate the 95% confidence interval for the proportion P applying the binomial distribution.
    g) Estimate the 95% confidence interval for the proportion P applying the normal approximation to the binomial distribution.
2. Estimator of total number with proportions.
    a) Estimate the total number of females in the population.
    b) Estimate the sampling variance and the error of the total number.
    c) Estimate the 95% confidence interval for the total number of females of the population applying the binomial distribution.
    d) Estimate the 95% confidence interval for the total number of females of the population applying the normal approximation to the binomial distribution.

8.5 EXERCISE 5

Group I – Landing ports

Consider a purse seiner fleet landing sardines in a given fishing port. A stratified sampling design is to be applied in order to estimate the total landings from these vessels. The composition of the fleet is given by Table 8.3.

TABLE 8.3
Number of vessels by power classes of a purse seiner fleet

Power class *(HP)*	Number of vessels
100-	10
200-	50
>300	20
Total	**80**

1. A random stratified sampling design will be applied considering the 3 HP categories as *strata*. The total size of the sample will be 16 vessels allocated proportionally to the number of vessels in each *stratum*. Calculate the number of vessels to be sampled in each *stratum*.
2. The landing of sardines of each sampled vessel was registered. Table 8.4 summarises the sample values of total landings and coefficients of variation

TABLE 8.4
Total landings of sardines and coefficient of variation by vessel power classes

Power class *(HP)*	Total landings *(tonnes)*	CV
100-	4	0.98
200-	60	0.73
>300	20	0.68

a) Calculate the average landing per vessel in each category.
b) Calculate the variance between total landings within each *stratum*.

3. Present estimates for each *stratum* of:
    a) Mean landing.
    b) Expected sampling variance of the estimator of the mean.
    c) Error of the estimator of the mean.
    d) Total landing.
    e) Expected variance of the estimator of total landing.
    f) Error of the estimator of total landing.
4. Present estimates for the total fleet of:
    a) Mean landing.
    b) Expected variance of the estimator of the mean.
    c) Error of the estimator of the mean.
    d) Total landing.
    e) Expected variance of the estimator of total landing.
    f) Error of the estimator of total landing.

Group II – Surveys

A research vessel has carried out a demersal trawl survey on the continental shelf and on the slope off Libya. The goal of the survey was to estimate the biomass of the European hake (*Merluccius merluccius*) in the area.

The survey was designed as a stratified random survey. The study area was divided into 10 *strata*, according to two geographical areas and five depth levels. Each haul was done at a speed of 3 knots, with one-hour duration. The trawl net had a horizontal opening of 50 m. It is assumed that the vertical opening was enough to catch all the hakes that occur in the trawling area.

Table 8.5 presents the two areas and their respective depth zones, the area of each *stratum* in square nautical miles (nm^2), the number of hauls carried out, the average catch and the standard deviation within each *stratum*. The sampling fraction is negligible.

TABLE 8.5
Characteristics of the survey area

Depth *(m)*	**Area** *(mn^2)*	**Number of trawls**	**Average catches** *(kg)*	**Standard deviation of catches** *(Kg)*
		Area 1		
100-200	2085	12	5	1.4
200-300	755	13	28	10.5
300-400	660	9	134	47.8
400-500	540	10	43	14.7
500-600	880	11	13	3.6
		Area 2		
100-200	1252	11	49	14.5
200-300	500	14	122	27.7
300-400	350	8	55	14.0
400-500	445	10	64	15.7
500-600	450	9	57	16.4

1. For each *stratum* estimate:
    a) An index of total biomass of European hake.
    b) The error of the estimator.
    c) The coefficient of variation of the estimator.

2. For the total area estimate:
   a) Total biomass of European hake.
   b) The error of the estimator.
   c) The 95% confidence limits of the total biomass in the area.
3. Consider that only 100 trawls can be carried out during the next year's survey. Under these conditions:
   a) Calculate the proportional allocation of the total 100 trawls to the *strata* areas.
   b) Calculate the *strata* allocation that gives the maximum precision in the estimation of the total abundance.

8.6 EXERCISE 6

Group I - Selection of the clusters

Along the coast of a region, divided into 5 provinces, 35 landing places were identified. The landing places and their number of vessels are presented in Table 8.6. The sizes of the landing places were considered to be the number of vessels in each place.

With the objective of estimating the total landing of the region, it was decided to select 15 landing places.

TABLE 8.6
Number of vessels of each province, by landing place

Landing Place	Number of Vessels	Landing Place	Number of Vessels
- Province 1 -		- Province 4 -	
1	9	19	28
2	30	20	60
3	12	21	16
4	9	22	24
5	9	23	36
6	4	24	20
7	5	25	52
8	10	26	13
		27	35
- Province 2 -		- Province 5 -	
9	30	28	13
10	150	29	48
11	41	30	14
12	18	31	16
13	8	32	12
14	27	33	13
15	4	34	11
		35	38
- Province 3 -			
16	25		
17	5		
18	15		
Total			**860**

1. Select the 15 clusters with equal probabilities.
2. Select the 15 clusters with probabilities proportional to the cluster sizes.
3. Considering the 5 provinces as *strata*, select 3 clusters from each province with probabilities proportional to the sizes of the clusters of each *stratum*.

Group II – Selection with equal probabilities

Consider a population divided into 23 clusters. Aiming at estimating the total value of the population, it was decided to select 5 clusters using the simple random criteria with replacement. Table 8.7 presents a summary of the obtained data.

TABLE 8.7
Sample data

Clusters	Sizes of the clusters	Total values of the clusters
11	50	1244
7	50	1324
2	50	1335
14	50	1300
9	50	1270
Total	**250**	**6473**

1. Indicate:
   a) The number of clusters in the population.
   b) The number of clusters in the sample.
   c) The number of elements in cluster 14.
2. Calculate:
   a) The sample mean value per cluster.
   b) The sample mean value per element.
   c) The sample variance between clusters.
3. Choose an estimator of the total value of the population and estimate:
   a) The expected value of the estimator.
   b) The sampling variance of the estimator.
   c) The error of the estimator.

Group III – Selection with probabilities proportional to sizes, with replacement

Consider a population of fishing vessels divided into 23 clusters with an unequal number of vessels, which are taken as cluster sizes. With the aim of estimating the total landings, Y, of the population, a sample of 5 sites was selected using a random criterion with probabilities proportional to the size of the cluster and with replacement. Table 8.8 summarises the sample data.

TABLE 8.8
Sample data

Clusters sampled	Number of vessels	Mean landings per vessel, $\bar{y}$
1	30	23.78
4	32	24.46
8	20	25.05
13	20	24.15
18	27	23.70
Total fleet	**822**	--

1. Adopting the Hansen-Horowitz estimator of the total landing, estimate:
   a) The total value of the population.
   b) The error of your estimator.
2. Adopting the Horvitz-Thompson estimator of the total landing, estimate:
   a) The total landing of all the vessels.
   b) The sampling variance of your estimator and its error.
   c) An approximate 95% confidence interval of the total landing.

8.7 EXERCISE 7

Group I – Selection with simple random sampling at both stages

A two-stage sampling has been carried out in order to estimate the total landings from the demersal longline fleet. During the first stage 5 vessels out of 58 have been sampled with a simple random criteria without replacement. During the second stage a sample of 50 fish boxes was drawn (by simple random criteria without replacement) from each selected vessel. The sample information of this two-stage sampling is summarized in Table 8.9.

1. Estimate:
   a) The total weight of fish landed.
   b) The error of the estimation.
2. Proportions
   Consider that in the 5 vessels sampled, 10, 15, 7, 5 and 20 boxes of fish were observed among the boxes sampled in vessels 1, 2, 3, 4 and 5, respectively (Table 8.7).
   a) Estimate the proportion of boxes in the total landings.
   b) Estimate the error of the estimation.

TABLE 8.9
Sample data

Vessel	Total number of boxes in the vessels	Number of boxes sampled	Total weight of the sample *(Kg)*	SD of box weight in each vessel *(Kg)*
1	200	50	990	2.02
2	100	50	1405	1.90
3	250	50	1440	2.14
4	90	50	1330	2.21
5	230	50	1105	3.24

Group II – First selection – unequal probabilities with replacement. Second stage – simple random sampling with replacement

A two-stage sampling has been undertaken with the aim of estimating the total weight of shrimp landed.

During the first stage, 5 out of 58 trawlers were randomly sampled with replacement, and unequal probabilities. During the second stage, a sample of 50 boxes was simple randomly drawn from each of the vessels selected in the first stage. The sample information is summarized in Table 8.10.

TABLE 8.10
Sample data

Sampled vessel number	Probability of the vessel being sampled	Total number of boxes	Number of boxes sampled	Total weight of the sample *(Kg)*	SD of box weight in each vessel *(Kg)*
1	0.02	250	50	24.80	1.20
2	0.03	300	50	26.48	1.19
3	0.01	100	50	26.70	1.32
4	0.04	150	50	26.00	1.44
5	0.10	200	50	25.40	2.18
Fleet total		**12000**			

1. Estimate the total weight of shrimps landed.
2. Estimate the error of the estimation.

9. Solutions

Solutions for the exercises in Chapter 8.

Exercise	Group	No.	Question	Answer
8.1	I	1-a)	The sample size	31
			The range of the sample values	469 cm
			The median	267 cm
			The mean	261.9 cm
			The total value	8118 cm
			The sample variance	15636.5
			The sample standard deviation.	125.0 cm
			The sample coefficient of variation.	0.48
8.1	II	1	From this data calculate the mean and the variance.	Mean(x) = 23.2 cm Var(x) = 12.5
		2	Choose an adequate class interval and build up a table with the length frequencies distribution	Class interval chosen – 1 cm Because it is simple, and gives an adequate number of classes (between 15 and 30)
		3-a)	The sample mean and sample variance. Compare this results with the ones obtained in 1	Sample mean = 23.2 cm Sample variance = 12.5 The values of the statistics are equal. Since the class interval used was the same as the resolution of the original measurements, no information was lost.

		3-b)	Three statistics of location	Mean = 23.2 cm
				Median = 22.5 cm
				Mode = 20 cm

		3-c)	Three statistics of dispersion	Inter Quartile Range = 5.5 cm
				Variance = 12.49
				CV = 0.15

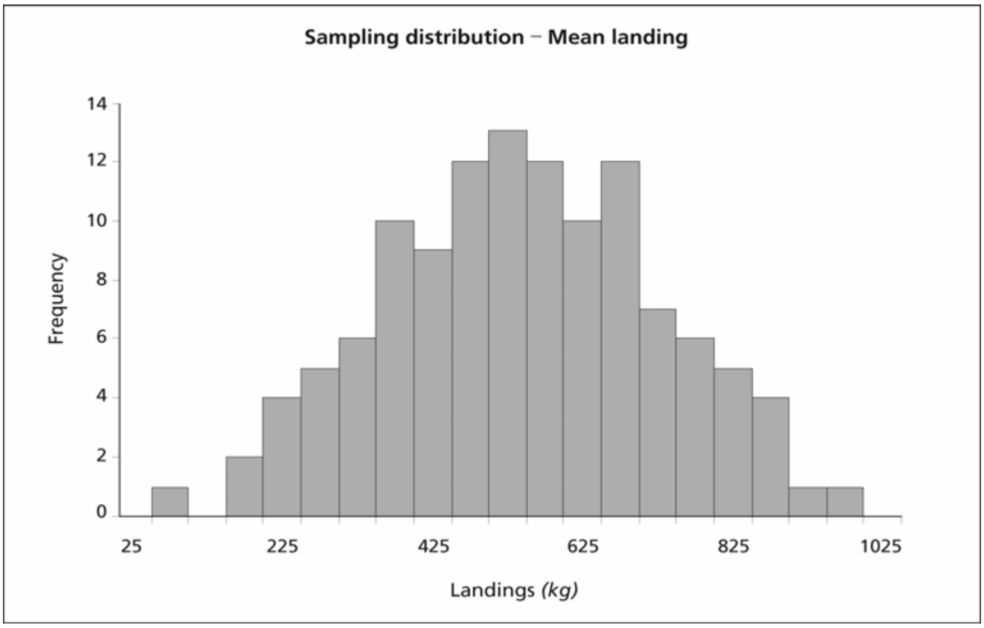
		3-d)	Number of individuals with a length less than 20 cm	41
		3-e)	Percentage of individuals with a length equal to or greater than 20 cm	79%
		3-f)	Percentage of observations between 23 and 25 cm	15%
		3-g)	The value that corresponds to a length equal to or greater than 45% of all the observations	21.5 cm
		3-h)	The value that corresponds to a length less than 21% of all the observations	26.5 cm
		3-i)	The quantile of order 96%	30.5 cm
8.2	I	1-a)	Relative frequencies of number of boats by fishing gear	Purse-seines: 0.3 Trawls: 0.21 Handlines: 0.17 Longlines: 0.23 Trammel nets: 0.09 Purse-seines: 0.3
		1-b)	Percentage of vessels that operate handlines	17%
		1-c)	Percentage of vessels that operate trammel nets	9%

Exercise	Group	No.	Question	Answer
		2-a)	Relative frequency of boats not operating trammel nets	0.91
		2-b)	Percentage of boats operating purse seines or longlines	53%
		2-c)	Relative frequency of boats that do not operate with handlines, nor trammel nets, nor longlines	0.51
		2-d)	Proportion of the total fleet that are small boats without engine	0.25
		2-e)	Proportion of the total fleet that are small boats with engine	0.37
		2-f)	Proportion of the total fleet that are small boats	0.62
		2-g)	Check that the proportion of 2.f) is the sum of 2.d) plus 2.e)	0.62 = 0.25+0.37
		3-a)	Proportion of the small boats without engine that operate handlines	0.34
		3-b)	Proportion of the total fleet that are small boats without engine	0.25
		3-c)	Proportion of the total fleet that are small boats without engine operating handlines	0.08
		3-d)	Check that the proportion of 3.c) is the product of 3.a) times 3.b)	0.08 = 0.34×0.25
		4-a)	Percentage of small boats without engine operating handlines or longlines	65%
		4-b)	Percentage of purse seiners operating purse seines	100%
		4-c)	Relative frequency of vessels that are not purse-seiners.	81%
		4-d)	Percentage of small boats with engine that operate trawls	4%
		4-e)	Percentage of the fleet that fishes with traps	0%
8.2	II	1-a)	Probability that the boat operates handlines	0.17
		1-b)	Probability that the boat operates Trammel nets	0.09
		1-c)	Probability that the boat does not operate trammel nets	0.91
		1-d)	Probability that the boat operates purse-seines or longlines.	0.53
		1-e)	Probability that the boat does not operate handlines, nor longlines nor trammel nets	0.51
		2-a)	Probability that the boat will be a small boat without engine	0.245
		2-b)	Probability that the boat will be a small boat with engine	0.374
		2-c)	Probability that the boat will be a small boat	0.62
		2-d)	Show that the probability 2.c) is equal to the sum of probability 2.a) plus the probability 2.b)	0.619 = 0.245+0.374
		3-a)	Probability of the boat being a purse seiner	0.189
		3-b)	Probability of the boat being a stern trawler	0.192
		3-c)	Probability of the boat not being a purse seiner nor a stern trawler	0.619
		3-d)	Show that probability 3.c) is equal to probability 2.c)	Shown from the comparison of the values
		4-a)	If we choose a boat from the small boats without engine, what is the probability that she operates with handline?	0.338
		4-b)	If we choose a boat out of the total fleet what is the probability that she is a small boat without engine?	0.245

Exercise	Group	No.	Question	Answer
		4-c)	If we choose a boat out of the total fleet what is the probability that she is a small boat without engine operating with handline?	0.083
		4-d)	Check that the probability of 4.c) is equal to the product of the probability of 4.a) times the probability of the 4.b).	0.083 = 0.338×0.245
8.2	III	1-a)	Probability of X being less than or equal to 18	0.097
		1-b)	Probability of X being greater than 18	0.903
		1-c)	Probability of X being less than 25	0.986
		1-d)	Probability of X being between 18 and 25	0.889
		2-a)	x such that Prob $\{X \leq x\} = 0.8413$	22.60
		2-b)	x such that Prob $\{X \geq x\} = 0.9772$	16.60
		2-c)	x such that Prob $\{X < x\} = 0.9986$	26.58
		2-d)	x such that x is the 95% quantile of the distribution of X	23.89
		2-e)	x such that x is the median of the distribution of X	20.60
8.2	IV	1-a)	Probability of the values of *Z* being between -1 and 1	0.683
		1-b)	Probability of the values of *Z* being between -2 and 2	0.954
		1-c)	Probability of the values of *Z* being between -3 and 3	0.997
		2-a)	The z_1 value for which the probability that the values of the variable *Z* will be smaller than z_1 is 2.5% (*Prob*$\{Z<z_1\}$ = 0.025)	-1.96
		2-b)	The z_2 value for which the probability of the variable *Z* being smaller than z_2 is 97.5% (*Prob*$\{Z<z_2\}$ = 0.975)	1.96
		3-a)	Probability that the variable *Z* will be within the interval (z_1, z_2) given by (*Prob*$\{Z<z_1\}$ = 0.025) (*Prob*$\{Z<z_2\}$ = 0.975)	0.950
		3-b)	Probability that the variable *Z* will be within the interval (z_1, z_2) given by (*Prob*$\{Z<z_1\}$ = 0.004) *Prob*$\{Z<z_2\}$ = 0.954).	0.950
		3-c)	Probability that the variable *Z* will be within the interval (z_1, z_2) given by (*Prob*$\{Z<z_1\}$ = 0.012) (*Prob*$\{Z<z_2\}$ = 0.962)	0.950
			Verify that the smallest of these intervals is the interval with symmetrical values, z_1 and z_2	Width 3-a) = 1.96-(-1.96) = 3.92 Width 3-b) =1.68-(-2.65) = 4.33 Width 3-c) = 1.77-(-2.26) = 4.03
8.2	V	1-a)	*Prob* $\{t(10) > 1.812\}$	0.050
		1-b)	*Prob* $\{t(19) < 1.729\}$	0.950
		1-c)	*Prob* $\{-1.34 < t(15) < +2.602\}$	0.890
		2-a)	*a* such that *Prob* $\{t(8) < a\} = 0.95$	1.860
		2-b)	*a* such that *Prob* $\{t(26) > a\} = 0.99$	-2.479
		2-c)	*a* such that *Prob* $\{-a < t(20) < +a\} = 0.95$	2.086
		3-a)	*a* such that *Prob* $\{t < a\} = 0.95$, with 40, 60, 120 and infinite degrees of freedom	40 *df*: a = 1.684 60 *df*: a = 1.671 120 *df*: a = 1.658 ∞ *df*: a = 1.645
		3-b)	Compare *a* obtained in 3.a) with *a* such that *Prob* $\{Z < a\} = 0.95$	$a(Z) = 1.645 = a(t\infty)$

<table>
<tr><th>Exercise</th><th>Group</th><th>No.</th><th>Question</th><th>Answer</th></tr>
<tr><td>8.2</td><td>VI</td><td>1</td><td>Show that the expected value of the random variable X is equal to P</td><td>$$E[X] = \begin{cases} \sum[x_i P(x_i)] \\ 1 \times P + 0 \times Q \\ P \end{cases}$$</td></tr>
<tr><td></td><td></td><td>2</td><td>Show that the variance of the randon variable <i>X</i>
is equal to <i>PQ</i></td><td>$$Var[X] = \begin{cases} \sum[(x_i - E[X])^2 P(x_i)] \\ (1-P)^2 \times P + (0-P)^2 \times Q \\ Q^2 \times P + P^2 \times Q \\ PQ \times Q + P \times PQ \\ PQ \times (P+Q) = PQ \end{cases}$$</td></tr>
<tr><td>8.3</td><td>I</td><td>1-a)</td><td>Mean shrimp landing per sampled fishing trip</td><td>698.7 Kg</td></tr>
<tr><td></td><td></td><td>1-b)</td><td>Variance of the sampled landings</td><td>137 696</td></tr>
<tr><td></td><td></td><td>1-c)</td><td>Standard deviation of the sampled landings</td><td>371.1 Kg</td></tr>
<tr><td></td><td></td><td>2</td><td>Estimate of the total amount of shrimp landed in that week</td><td>6 986.7 Kg</td></tr>
<tr><td>8.3</td><td>II</td><td>1-a)</td><td>Average shrimp landing per fishing trip during that week</td><td>542 Kg</td></tr>
<tr><td></td><td></td><td>1-b)</td><td>Population variance and modified variance of the landings</td><td>σ^2 = 127 730
S^2 = 141 922</td></tr>
<tr><td></td><td></td><td>1-c)</td><td>Standard deviation of the landings</td><td>357.4 Kg</td></tr>
<tr><td></td><td></td><td>1-d)</td><td>Total amount of shrimp landed</td><td>5 420 Kg</td></tr>
<tr><td></td><td></td><td>1-e)</td><td>Proportion of all landings below 400 Kg</td><td>0.300</td></tr>
<tr><td></td><td></td><td>1-f)</td><td>Relative frequency of landings between 400 and 800 Kg</td><td>0.400</td></tr>
<tr><td></td><td></td><td>2</td><td>Build at least 10 samples of 3 landings each that could have been selected from that population.</td><td>
<table>
<tr><th>Sample no.</th><th>Land 1</th><th>Land 2</th><th>Land 3</th></tr>
<tr><td>1</td><td>538</td><td>0</td><td>906</td></tr>
<tr><td>2</td><td>538</td><td>0</td><td>230</td></tr>
<tr><td>3</td><td>538</td><td>0</td><td>598</td></tr>
<tr><td>4</td><td>538</td><td>0</td><td>20</td></tr>
<tr><td>5</td><td>538</td><td>0</td><td>435</td></tr>
<tr><td>6</td><td>230</td><td>598</td><td>0</td></tr>
<tr><td>7</td><td>230</td><td>598</td><td>435</td></tr>
<tr><td>8</td><td>230</td><td>598</td><td>859</td></tr>
<tr><td>9</td><td>230</td><td>598</td><td>1123</td></tr>
<tr><td>10</td><td>230</td><td>598</td><td>711</td></tr>
</table>
</td></tr>
<tr><td></td><td></td><td>3</td><td>Repeat the calculations done on number 1.
a) to d), for each of these samples</td><td>
<table>
<tr><th>Sample. no.</th><th>Average</th><th>Var</th><th>SD</th><th>Total</th></tr>
<tr><td>1</td><td>481</td><td>207617</td><td>456</td><td>4813</td></tr>
<tr><td>2</td><td>256</td><td>72868</td><td>270</td><td>2560</td></tr>
<tr><td>3</td><td>379</td><td>108441</td><td>329</td><td>3787</td></tr>
<tr><td>4</td><td>186</td><td>93028</td><td>305</td><td>1860</td></tr>
<tr><td>5</td><td>324</td><td>81546</td><td>286</td><td>3243</td></tr>
<tr><td>6</td><td>276</td><td>90988</td><td>302</td><td>2760</td></tr>
<tr><td>7</td><td>421</td><td>34003</td><td>184</td><td>4210</td></tr>
<tr><td>8</td><td>562</td><td>99864</td><td>316</td><td>5623</td></tr>
<tr><td>9</td><td>650</td><td>201416</td><td>449</td><td>6503</td></tr>
<tr><td>10</td><td>513</td><td>63259</td><td>252</td><td>5130</td></tr>
</table>
</td></tr>
<tr><td></td><td></td><td>4</td><td>Compare the values of the statistics obtained in the previous item with the values of the corresponding population parameters.</td><td>The values of these statistics both over-estimate and under-estimate the corresponding population parameters</td></tr>
</table>

Exercise	Group	No.	Question	Answer
8.3	III	1	Histogram of the sampling distribution of the estimator, $\hat{\bar{Y}}$, using an appropriate class interval.	Appropriate Class Interval: 50 Kg (approx. 20 classes)



Exercise	Group	No.	Question	Answer			
8.3	III	2-a)	The 120 values of the estimator	No.	Mean Land	No.	Mean Land
				1	1481.3	41	588.3
				2	256.0	42	676.3
				3	378.7	43	539.0
				4	186.0	44	276.0
				5	324.3	45	83.3
				6	465.7	46	221.7
				7	553.7	47	363.0
				8	416.3	48	451.0
				9	558.0	49	313.7
				10	680.7	50	206.0
				11	488.0	51	344.3
				12	626.3	52	485.7
				13	767.7	53	573.7
				14	855.7	54	436.3
				15	718.3	55	151.7
				16	455.3	56	293.0
				17	262.7	57	381.0
				18	401.0	58	243.7
				19	542.3	59	431.3
				20	630.3	60	519.3
				21	493.0	61	382.0
				22	385.3	62	660.7
				23	523.7	63	523.3
				24	665.0	64	611.3
				25	753.0	65	578.0
				26	615.7	66	385.3
				27	331.0	67	523.7
				28	472.3	68	665.0
				29	560.3	69	753.0
				30	423.0	70	615.7
				31	610.7	71	508.0
				32	698.7	72	646.3
				33	561.3	73	787.7
				34	840.0	74	875.7
				35	702.7	75	738.3
				36	790.7	76	453.7
				37	378.7	77	595.0
				38	501.3	78	683.0
				39	308.7	79	545.7
				40	447.0	80	733.3
8.3	III	2-b)	Expected value of the estimator	542.0 Kg			
		2-c)	Sampling variance of the estimator	33 115.2			
		2-d)	Error of the estimator	182.0 Kg			
		3	Compare the expected value obtained in 2.b) with the population mean calculated in Group I – 1.a)	They have the same value			
		4	Check the theoretical expression: $V\left[\hat{\bar{Y}}\right]=(1-\frac{n}{N})\frac{S^2}{n}$	$33115.2=(1-\frac{3}{10})\frac{141922}{3}$			

Exercise	Group	No.	Question	Answer		
		5-a)	Percentiles of the sampling distribution of the estimator with the following orders:			
			i) 1.0%	i) 1.0%: 158.2 Kg		
			ii) 2.5%	ii) 2.5%: 205.5 Kg		
			iii) 3.5%	iii) 3.5%: 222.8 Kg		
			iv) 0.0%	iv) 50.0%: 540.7 Kg		
			v) 95.0%	v) 95.0%: 840.8 Kg		
			vi) 96.0%	vi) 96.0%: 856.7 Kg		
			vii) 97.5%	vii) 97.5%: 876.2 Kg		
			viii) 98.5%	viii) 98.5%: 901.0 Kg		
		5-b)	Four intervals that encompass 95% of all possible sample means	Interval	Lower Limit	Upper Limit
				From minimum value to 95th percentile	83.3	840.8
				From 5th percentile to largest value	242.9	962.7
				From 1st percentile to 96th percentile	158.2	856.7
				From 2.5th percentile to 97.5th percentile	205.5	876.2
8.3	III	5-c)	The width of the four intervals	Interval		Width
				From minimum value to 95th percentile		757.5
				From 5th percentile to largest value		719.8
				From 1st percentile to 96th percentile		698.5
				From 2.5th percentile to 97.5th percentile		670.7
		6	The shortest of these intervals that holds 95% of all possible sample means	The last interval		
		7-a)	Probability of getting a sample of 3 landings with an average landing below or equal to 600 Kg	62.5%		
		7-b)	Probability of getting a sample of 3 landings with an average landing above 600 Kg	37.5%		
		7-c)	Probability of getting a sample of 3 landings with an average landing between 199 and 953 Kg	96.7%		
		8-a)	l such that *Prob* {$\overline{y}$ < l} = 0.95	840.8 Kg		
		8-b)	l_1 and l_2 such that *Prob* {$l_1 < \overline{y} < l_2$} = 0.95	l_1=205.5 Kg l_2=876.2 Kg		
8.4	I	1-a)	Mean	39.29		
		1-b)	Variance	16.41		
		1-c)	Standard deviation	4.05		
		2-a)	Estimate of population mean	39.29		
		2-b)	The estimator is biased?	No		
		2-c)	Sampling variance of $\overline{y}$	4.10		
		2-d)	Error of $\overline{y}$	2.03		
		2-e)	A 95% confidence interval of μ	35.32 – 43.26		

Exercise	Group	No.	Question	Answer
8.4	II	1	A procedure to select a simple random sample of 4 numbered vessels	Select a whole random number between 1 and 20. Include the corresponding vessel in the sample. Repeat 4 times. If at any moment the vessel selected was already included in the sample, repeat the random number selection, until a new vessel is selected.
		2-a)	An estimator of the total amount of fish landed in the port during this day	N x Sample Mean
		2-b)	Sampling distribution of that estimator and the formulae to obtain the expected value and the expected sampling variance	Sampling Distribution: $\hat{Y} \overset{\bullet}{\cap} N\left(E\left[\hat{Y}\right], V\left[\hat{Y}\right]\right)$ $E\left[\hat{Y}\right]=Y$ $V\left[\hat{Y}\right]=(1-f)\dfrac{N^2S^2}{n}$
		3-a)	Estimate of total landing	153.5 Kg
		3-b)	Estimate of sampling variance	3923.4
		3-c)	Estimate of error of the estimate	62.6 Kg
		3-d)	A 95% confidence interval for the population total landings	0 – 352.8
		4	Approximate size of the sample necessary to have an error 10% smaller than the one previously calculated in 3.c)	5
8.4	III	1-a)	Proportion of females in the sample	0.400
		1-b)	Sample variance	0.248
				E[*p*]=P
		1-c)	Expressions for the expected value of *p*, and for the sampling variance of *p*	$V[p]=(1-f)\dfrac{N}{N-1}\dfrac{PQ}{n}$
		1-d)	Estimate of the sampling variance of *p*	0.0058
		1-e)	Estimate of the error of *p*	0.0761
		1-f)	Estimate of the 95% confidence interval for the proportion *P* applying the binomial distribution	0.227 – 0.594
		1-g)	Estimate of the 95% confidence interval for the proportion *P* applying the normal approximation to the binomial distribution	0.251 – 0.549
		2-a)	Estimate of total number of females in the population	40
		2-b)	Estimate of the sampling variance and of the error of the total number	V[N*p*]=57.93 s_{Np}=7.61
		2-c)	Estimate of the 95% confidence interval for the total number of females of the population applying the binomial distribution	23 - 59
		2-d)	Estimate of the 95% confidence interval for the total number of females of the population applying the normal approximation to the binomial distribution	25 - 55

Exercise	Group	No.	Question	Answer	
8.5	I (landing ports)	1	Number of vessels to be sampled in each *stratum*	Class (*stratum*)	Sample Size
				100-	2
				200-	10
				>300	4
				Total	16
		2-a)	Average landing per vessel in each category	Class (*stratum*)	Average Landing
				100-	2
				200-	6
				>300	5
		2-b)	Variance between total landings within each *stratum*	Class (*stratum*)	Variance
				100-	3.84
				200-	19.18
				>300	11.56
		3-a)	Estimates of mean landing for each *stratum*	Class (*stratum*)	Est. Mean Landing
				100-	2
				200-	6
				>300	5
		3-b)	Estimates of expected sampling variance of the estimator of the mean for each *stratum*	Class (*stratum*)	Est. Exp. Var
				100-	1.54
				200-	1.54
				>300	2.31
		3-c)	Estimates of error of the estimator of the mean for each *stratum*	Class (*stratum*)	Est. Error
				100-	1.24
				200-	1.24
				>300	1.52
8.5	I	3-d)	Estimates of total landing for each *stratum*	Class (*stratum*)	Est. Total Landing
				100-	20
				200-	300
				>300	100
		3-e)	Estimates of expected variance of the estimator of total landing for each *stratum*	Class (*stratum*)	Est. Exp. Var. Total
				100-	153.66
				200-	3836.88
				>300	924.80
		3-f)	Estimates of error of the estimator of total landing for each *stratum*	Class (*stratum*)	Est. Exp. Error Total
				100-	12.40
				200-	61.94
				>300	30.41
		4-a)	Estimate of mean landing for total fleet	5.25	
		4-b)	Estimate of expected variance of the estimator of the mean for total fleet	0.77	
		4-c)	Estimate of error of the estimator of the mean for total fleet	0.88	
		4-d)	Estimate of total landing for total fleet	420	
		4-e)	Estimate of expected variance of the estimator of total landing for total fleet	4915.34	
		4-f)	Estimate of error of the estimator of total landing for total fleet		70.11

Exercise	Group	No.	Question	Answer	
8.5	II	1-a)	Estimate of index of total biomass of European hake for each *stratum*	*Stratum*	Est. Index (tonnes)
	(surveys)			A1, 100-200 m	129.8
				A1, 200-300 m	263.3
				A1, 300-400 m	1101.4
				A1, 400-500 m	289.2
				A1, 500-600 m	142.5
				A2, 100-200 m	764.0
				A2, 200-300 m	759.7
				A2, 300-400 m	239.7
				A2, 400-500 m	354.7
				A2, 500-600 m	319.4
		1-b)	Estimate of error of the estimator for each *stratum*	*Stratum*	Est. Error (tonnes)
				A1, 100-200 m	3.0
				A1, 200-300 m	7.6
				A1, 300-400 m	43.7
				A1, 400-500 m	9.9
				A1, 500-600 m	3.6
				A2, 100-200 m	20.6
				A2, 200-300 m	12.3
				A2, 300-400 m	7.6
				A2, 400-500 m	8.7
				A2, 500-600 m	10.2
		1-c)	Estimate of coefficient of variation of the estimator for each *stratum*	*Stratum*	CV Est.
				A1, 100-200m	2.3%
				A1, 200-300 m	2.9%
				A1, 300-400 m	4.0%
				A1, 400-500 m	3.4%
				A1, 500-600 m	2.5%
				A2, 100-200 m	2.7%
				A2, 200-300 m	1.6%
				A2, 300-400 m	3.2%
				A2, 400-500 m	2.5%
				A2, 500-600 m	3.2%
		2-a)	Estimate of total biomass of European hake for the total area	4 634 tonnes	
		2-b)	Estimate of error of the estimator	54 tonnes	

Exercise	Group	No.	Question	Answer	
8.5	II	2-c)	Estimate of 95% confidence limits of the total biomass in the total area	4 258 tonnes - 4 469 tonnes	
		3-a)	Proportional allocation of the total 100 trawls to the *strata* areas	*Stratum*	CV Est.
				A1, 100-200m	26
				A1, 200-300 m	10
				A1, 300-400 m	8
				A1, 400-500 m	7
				A1, 500-600 m	11
				A2, 100-200 m	16
				A2, 200-300 m	6
				A2, 300-400 m	4
				A2, 400-500 m	6
				A2, 500-600 m	6
				Total	100
		3-b)	*Strata* allocation that gives the maximum precision in the estimation of the total abundance	*Stratum*	Num. Hauls
				A1, 100-200m	3
				A1, 200-300 m	8
				A1, 300-400 m	29
				A1, 400-500 m	8
				A1, 500-600 m	3
				A2, 100-200 m	17
				A2, 200-300 m	13
				A2, 300-400 m	5
				A2, 400-500 m	7
				A2, 500-600 m	7
				Total	100
8.6	I (clusters)	1	15 clusters selected with equal probabilities	Element	Cluster N°
				1	20
				2	35
				3	11
				4	25
				5	21
				6	27
				7	5
				8	28
				9	4
				10	13
				11	1
				12	10
				13	30
				14	7
				15	19

Exercise	Group	No.	Question	Answer		
8.6	I	2	15 clusters selected with probabilities proportional to the cluster sizes	Element		Cluster N°
				1		10
				2		11
				3		20
				4		23
				5		9
				6		25
				7		14
				8		35
				9		24
				10		28
				11		12
				12		29
				13		27
				14		34
				15		22
8.6	I	3	15 clusters selected by selecting 3 clusters from each province with probabilities proportional to the sizes of the clusters of each *stratum*	Element	Prov.	Cluster No.
				1	1	3
				2	1	7
				3	1	1
				4	2	10
				5	2	11
				6	2	9
				7	3	17
				8	3	16
				9	3	18
				10	4	20
				11	4	23
				12	4	25
				13	5	35
				14	5	28
				15	5	29
8.6	II	1-a)	Number of clusters in the population	23		
		1-b)	Number of clusters in the sample	5		
		1-c)	Number of elements in cluster 14	50		
		2-a)	Sample mean value per cluster	1294.6		
		2-b)	Sample mean value per element	25.9		
		2-c)	Sample variance between clusters	1422.8		
		3-a)	Estimate of the expected value of the estimator	Estimator: N * Mean Value per Cluster Est. Expected value: 29 775.8		
		3-b)	Estimate of the sampling variance of the estimator	222.7		
		3-c)	Estimate of the error of the estimator	14.9		

Exercise	Group	No.	Question	Answer	
8.6	III	1-a)	Estimate of the total value of the population	19 915.4	
		1-b)	Estimate of the error of your estimator	32.4	
		2-a)	Estimate of the total landing of all the vessels	21 205	
		2-b)	Estimate of the sampling variance of your estimator and its error	Sampling Var Sampling Error	38 482 456 6 203
		2-c)	Estimate of an approximate 95% confidence interval of the total landing	3 981 – 38 428	
8.7	I	1-a)	Estimate of the total weight of fish landed	248 785	
	(two	1-b)	Estimate of the error of the estimation	48 003	
	stages)	2-a)	Estimate of the proportion of boxes of fish in the total landings	0.228	
		2-b)	Estimate of the error of the estimation	0.053	
8.7	II	1	Estimate of the total weight of fish landed	192 090	
		2	Estimate of the error of the estimation	50 285	

Bibliography

Publications related to general sampling methods and survey sampling:

Barnett, V. 1991. *Sample Survey Principles and Methods*. Edward Arnold, London, 173pp.

Chaudhuri, A. & Stenger, H. 1992. *Survey sampling: Theory and methods*. Dekker, 384pp.

Cochran, W.G. 1977. *Sampling techniques*. John Wiley & Sons, Inc., New York, 3rd ed., 428pp.

Efron, B. & Tibshirani, R.J. 1993. *An introduction to the bootstrap*. Chapman & Hall, New York, 436pp.

Green, R.H. 1979. *Sampling design and statistical methods for environmental biologists*. John Wiley & Sons, Inc., New York, 272pp.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. 1993. *Sample survey methods and theory*. John Wiley & Sons, New York, Vol. 1. 664pp.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. 1993. *Sample survey methods and theory*. John Wiley & Sons, New York, Vol. 2. 1016pp.

Levy, P.S. & Lemeshow, S. 1991. *Sampling of populations: Methods and applications*, John Wiley & Sons, Inc., New York, 2nd ed., 420pp.

Som, R.K. 1973. *A manual of sampling techniques*. Heinemann Educational Books Ltd., London, 384 pp.

Thompson, S.K. 1992. *Sampling*. John Wiley & Sons, Inc., New York, 334 pp.

Tryfos, P. 1996. *Sampling methods for applied research: Text and cases*. John Wiley & Sons, Inc., New York, 480pp.

The practical application of sampling theory to estimate fish landings was based on the publications:

Banerji, S.K. 1973. *An assessment of the exploited pelagic fisheries of the Indian Seas*. In *Proceedings of the Symposium on Living Resources*, Seas Around India, Spec. Publ., Cent. Mar. Fish. Res. Inst., p. 114-136, Cochin, India.

Banerji, S.K. 1975. *Fishery statistics needed for development planning*. FAO Fisheries Circulars, No. 630. FAO, Rome. 10pp.

Bazigos, G.P. 1974. *Applied Fisheries Statistics*. FAO Fisheries Technical Paper, No. 135. FAO, Rome. 172pp.

Bazigos, G.P. 1974. *The design of fisheries statistical surveys*. *Inland waters*. FAO Fisheries Technical Paper, FAO, Rome. 133pp.

Bazigos, G.P. 1975. *Applied fishery statistics: vectors and matrices*. FAO Fisheries Technical Paper, No. 135 (suppl. 1), FAO, Rome. 39pp.

Gulland, J.A. 1966. *Manual of sampling and statistical methods for fisheries biology*. Part 1. In: *Sampling methods*. FAO Man. Fish. Sci., 3: 87pp.

Sparre, P.J. 2000. *Manual on sample-based data collection for fisheries assessment*. Examples from Viet Nam. FAO Fisheries Technical Paper, No. 398. FAO, Rome. 171pp.

The main objective of this manual is to present the basic and standard concepts of sampling methods applied to fisheries science. In order to ensure sound fisheries research, it is essential to have reliable data from landing ports, fishery stocks and research surveys. A rational management of fishing resources can then be established to ensure a sustainable exploitation rate and responsible fisheries management, providing long-term benefits for all. This document provides an introduction to sampling theory and introduces the theory of the three worlds (population, sample and sampling), as well as a short revision of probability concepts. It also provides an overview of the simple random, random stratified, cluster and two-stage sampling methods. The expressions for estimating the mean and total of the populations, their sampling distributions, the expected values, the sampling variances and their estimates are included and justified for each of the sampling designs. The document also contains a case study of biological sampling from landing ports and exercises that should be used to further understanding of the objectives of sampling and its advantages for fishery resource studies.

ISBN 92-5-105397-9 ISSN 0429-9345

9 789251 053973

TC/M/A0198E/1/11.05/2000