I. Simple random sampling

II. Systematic random sampling

I. Simple random sampling

1. Terms / Definitions
2. Simple random sampling
3. Drawing methods
4. Simple random sampling without replacement

4.1 Unbiased estimates for mean values and total values of the population

4.2 Variances of the estimates (error variances)

4.3 Estimating error variances

4.4 Confidence intervals for mean values and total values of the population

5. Simple random sampling with replacement
6. Estimating sizes and shares of subpopulations
7. Estimating ratios
8. Estimating mean values and total values of subpopulations
9. Defining the sample size
10. Particularities in expansion

1. Terms / Definitions

Definitions:

- A *sample* is a subset of a population (total of all statistical units to be covered by a survey) which is to be taken as a basis to estimate specific parameters of the population.

- *Sampling units* are the units taken as a basis for a sampling process.

- A sampling unit which has been included in the sample is referred to as a *sample unit*.

- The *sample size* is the number of units selected for a sample.

Terms:

N **= number of units of the population,**

N_g **= number of units of the population with a specific quality, i.e. size of the so-called subgroup** g**,**

y,x **= survey variables, and**

$Y_1, Y_2, \ldots, Y_N$ **= values of the survey variable** y **of the** N **elements of the population**

Relevant parameters of a population:

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

sum of the variable values of all units of the population (total value),

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

a variable's mean value in the population,

$$Y_g = \sum_{i \in g} Y_i \quad =$$

total value of subgroup *g*,

$$\bar{Y}_g = \frac{Y_g}{N_g} =$$ **mean value of subgroup *g***

and

$$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}} =$$ **ratio between the total values (or mean values) of variables y and x**

Another parameter is the variance of a variable (root mean square deviation of the variable values from their mean value):

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 \qquad \textbf{or} \qquad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$$

It is a major figure in planning sample surveys.

There is:

a) $$S_y^2 = \frac{N}{N-1}\sigma_y^2 \qquad (1.1)$$

b) $$S_y^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} Y_i\right)^2\right) \qquad (1.2)$$

Definition:

A *random sample* (probability sample) is a sample for which the following applies:

- The quantity of possible different samples that can be obtained through a specific sampling method can be determined.
- For each of those possible samples, its probability to be drawn is known.
- For each element of the population, there is a strictly positive probability to become a sample unit.
- There is a calculation rule which, for each of the possible samples, leads to one estimate for the relevant parameter.

For random samples, a sampling theory has been developed. What is important for random samples is the fact that the (frequency) distribution of the estimate can be determined. Ultimately, any conclusions of sampling theory are based on that distribution, whose background is the hypothetical repetition of sampling. For random samples, it is in particular possible to statistically assess the quality of the estimate precision.

Terms:

With

n **= sample size, and**

$y_1, y_2, \ldots, y_n =$ **values of the survey variable *y* of the *n* sample units (generally** $Y_i \neq y_i$ **),**

there is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i =$$ **mean variable value in the sample (sample mean) and**

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 =$$ **variance of the variable in the sample.**

Definitions:

S is the quantity of all possible samples, *s* a specific sample from *S* and *p(s)* the probability with which the sample *s* is drawn from the quantity *S*.

a) A calculation rule $\hat{\theta}$ which is defined for quantity *S*, $s \in S$

i.e. which assigns a value $\hat{\theta}(s)$ to each sample

,

is referred to as (sample) *statistics*.

Examples:

$\bar{y}$ und s_y^2

b) The expected value of a statistics $\hat{\theta}$ is defined by

$$E\left(\hat{\theta}\right) = E_p\left(\hat{\theta}(s)\right) = \sum_{s \in S} p(s)\hat{\theta}(s) .$$

c) The *variance* of a statistics $\hat{\theta}$ is defined by

$$\sigma^2_{\hat{\theta}} = E\left(\left(\hat{\theta} - E\left(\hat{\theta}\right)\right)^2\right) = \sum_{s \in S} p(s)\left(\hat{\theta}(s) - E\left(\hat{\theta}\right)\right)^2 .$$

d) If a statistics $\hat{\theta}$ is used to estimate a parameter θ of the population, $\hat{\theta}$ is referred to as (sample) *estimator*.

e) An estimator $\hat{\theta}$ is referred to as *unbiased* if there is

$$E_p\left(\hat{\theta}(s)\right) = \theta \ .$$

f) The variance of an estimator $\hat{\theta}$ is referred to as

***error variance*.**

2. Simple random sampling

Definition:

A random sampling method in which

- the sampling units are identical with the survey units,
- every possible sample has the same number of sample units, and
- in every sampling process, every sampling unit available has the same probability of being selected,

is referred to as *simple random sampling*.

The background of simple random sampling is a lottery wheel method where the sample units are drawn successively from a lottery wheel through *n* sampling processes. In simple random sampling, a distinction is made between *sampling without replacement* and *sampling with replacement*.

a) Sampling without replacement

By definition, the probability of a specific unit to be included in the sample in the first draw is 1/*N*. The probability of a specific unit to be sampled from the units remaining in the wheel in the second draw is 1/(*N*-1), etc.

Conclusions:

i) The probability of a specific unit to be included in the sample of the size n (inclusion probability) is n/N.

ii) The probability of n specific units to be drawn from the wheel in a specific given sequence (sampling taking account of the sequence) is

$$\frac{1}{N} \times \frac{1}{N-1} \times \ldots \times \frac{1}{N-n+1} \qquad (2.1)$$

iii) The probability of having *n* specific units in a sample of the size *n* (sampling without taking account of the sequence) is

$$\frac{1}{\binom{N}{n}} = \frac{1 \times \ldots \times n}{N \times (N-1) \times \ldots \times (N-n+1)} \quad , \qquad (2.2)$$

wobei

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

and

$$m! = 1 \times \ldots \times m$$

Note: In both cases, all possible samples have the same selection probability.

Notes:

- $\binom{N}{n}$ is the number of possible (different) samples

  if the sequence of sample units is irrelevant.

- If conclusions ii) or iii) apply, it follows conversely that the random sampling method is simple random sampling without replacement.

  Therefore, in literature, you will sometimes find that simple random sampling is defined by conclusions ii) or iii).

b) Sampling with replacement

By definition, the probability of a specific unit to be included in the sample in the first draw is again $1/N$.

As a unit already drawn is replaced into the wheel before the next draw, $1/N$ is the probability of a specific unit to be drawn also in any subsequent draws.

Conclusion:

The probability of n specific units being drawn in a given sequence from the wheel, with multiple selection being allowed, is

$$\frac{1}{N^n} \qquad (2.3)$$

3. Drawing methods

Means used:

a) Lots
b) Random numbers
- genuine random numbers
- pseudo-random numbers

Lots are used for physical drawing processes.

If random numbers are used, the natural numbers from 1 to *N* are arbitrarily assigned to the *N* sampling units, so that the units are listed. Then random or pseudo-random numbers are determined.
The units with the numbers *j* are selected.

$$j \in \{1,\ldots,N\}$$

Sampling without replacement also includes the following option: The values of a random variable that is evenly distributed in an interval (0,1) can successively be assigned to the sampling units. Then, for example, the units with the *n* lowest values or the *n* highest values or the *n* lowest values above a constant $c \in (0,1)$ etc. can be included in the sample.

4. Simple random sampling without replacement

Example:

$$N = 4, \quad (Y_1, Y_2, Y_3, Y_4) = (2,1,5,1) \quad \Rightarrow$$

$$Y = \sum_{i=1}^{4} Y_i = 9 \quad , \quad \bar{Y} = \frac{Y}{4} = 2{,}25$$

$$S_y^2 = \frac{1}{N-1}\left(\sum_{i=1}^{4} Y_i^2 - Y^2 / N \right) = \frac{1}{3}\left(31 - 9^2/4 \right) = \frac{10{,}75}{3} = 3{,}58\overline{3}$$

Objective: Through simple random sampling without replacement with the sample size $n = 2$, $Y, \bar{Y}$ and S_y^2 are to be estimated without bias.

Exercise:

$$N = 4, \quad (Y_1, Y_2, Y_3, Y_4) = (2,1,5,1) \quad \Rightarrow$$

$$Y = \sum_{i=1}^{4} Y_i = 9 \quad , \quad \bar{Y} = \frac{Y}{4} = 2{,}25$$

Determine all possible samples s obtained through simple random sampling without replacement with the size $n = 2$, calculate for each case the sample mean $\bar{y}(s)$ and then calculate the expected value of the sample mean, i. e.

$$E(\bar{y}) = E_p(\bar{y}(s)) = \sum_{s \in S} p(s)\,\bar{y}(s) \; .$$

s	$Y_1=2$	$Y_2=1$	$Y_3=5$	$Y_4=1$	y_1	y_2	$\bar{y}(s)$	$s_y^2(s)$

s^*	$Y_1=2$	$Y_2=1$	$Y_3=5$	$Y_4=1$	y_1	y_2	$\bar{y}(s^*)$	s	$\bar{y}(s)$	$s_y^2(s)$
1	§	§			2	1	1,5	1	1,5	
2	§	§			1	2	1,5			
3	§		§		2	5	3,5	2	3,5	
4	§		§		5	2	3,5			
5	§			§	2	1	1,5	3	1,5	
6	§			§	1	2	1,5			
7		§	§		1	5	3,0	4	3,0	
8		§	§		5	1	3,0			
9		§		§	1	1	1,0	5	1,0	
10		§		§	1	1	1,0			
11			§	§	5	1	3,0	6	3,0	
12			§	§	1	5	3,0			
↙							27,0		13,5	

$(2.1) \Rightarrow$

$$p(s^*) = \frac{1}{N} \times \frac{1}{N-1} \times \ldots \times \frac{1}{N-n+1} = \frac{1}{4 \times 3} = \frac{1}{12}$$ **und**

$$E\left(\bar{y}(s^*)\right) = \sum_{s^* \in S} \frac{1}{12} \bar{y}(s^*) = \frac{1}{12} \sum_{s^* \in S} \bar{y}(s^*) = \frac{27}{12} = 2{,}25$$

As the sequence of sample units is irrelevant in calculating the sample mean, it is not necessary to distinguish between the

s^*samples 1 and 2, 3 and 4, 5 and 6, 7 and 8, 9 and 10 as well as 11 and 12:

$(2.2) \Rightarrow$

$$p(s) = \frac{1 \times \ldots \times n}{N \times (N-1) \times \ldots \times (N-n+1)} = \frac{1 \times 2}{4 \times 3} = \frac{1}{6} \quad \text{und}$$

$$E(\bar{y}(s)) = \sum_{s \in S} \frac{1}{6} \bar{y}(s) = \frac{1}{6} \sum_{s \in S} \bar{y}(s) = \frac{13{,}5}{6} = 2{,}25$$

In this example there is: $E(\bar{y}(s^*)) = E(\bar{y}(s)) = 2{,}25 = \bar{Y}$,

i. e. the sample mean $\bar{y}(s)$ **is an unbiased estimator for the mean value** $\bar{Y}$ **in the population.**

If N – as in this case – is known, the following estimator can also be calculated:

$$\hat{\theta}(s) = N\bar{y}(s)$$

For that estimator there is in this example:

$$E_p\left(N\bar{y}(s)\right) = \sum_{s \in S} p(s) N\bar{y}(s)$$

$$= \frac{1}{6}\sum_{s \in S} N\bar{y}_s = \frac{N}{6}\sum_{s \in S} \bar{y}_s = \frac{4}{6} \times 13{,}5 = 9 = Y$$

i. e. here the sample estimator $\hat{\theta}(s) = N\bar{y}(s)$ is unbiased for the total value Y.

4.1 Unbiased estimates for mean values and total values of the population

In the example, the sample estimators $\hat{\bar{Y}} = \bar{y}$ and $\hat{Y} = N\bar{y}$ are unbiased for $\bar{Y}$ and Y, respectively. This is not a special case because, for simple random sampling without replacement, the following theorem generally applies:

Theorem 4.1.1
$\hat{\bar{Y}} = \bar{y}$ (sample mean) is an unbiased estimator for $\bar{Y}$.

Proof:

$$E(\bar{y}) = \sum_{s \in S} p(s)\,\bar{y}_s = \frac{1}{\binom{N}{n}} \frac{1}{n} \sum_{s \in S} (y_{1s} + \ldots + y_{ns})$$

$$= \frac{1}{\binom{N}{n}} \frac{1}{n} \binom{N-1}{n-1} \sum_{i=1}^{N} Y_i = \frac{n!(N-n)!}{N!} \times \frac{1}{n} \times \frac{(N-1)!}{(n-1)!(N-n)!} Y$$

$$= \frac{1}{N} Y = \bar{Y}$$

Conclusion 4.1.2

$\hat{Y} = N\bar{y}$ is an unbiased estimator for Y because

$$E(N\bar{y}) = NE(\bar{y}) = N\bar{Y} = Y$$

Note: $N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i$, i. e. the total value in the population is estimated by multiplying the sample total value by the factor N/n . As $N/n = 1/(n/N)$ equals the reciprocal *sampling fraction* n/N ,

is referred to as the *expansion factor*. The downsizing process of sampling thus is practically reversed.

Exercise:

For the situation in the above example, determine the

frequency distribution of the sample mean $\bar{y}$ and

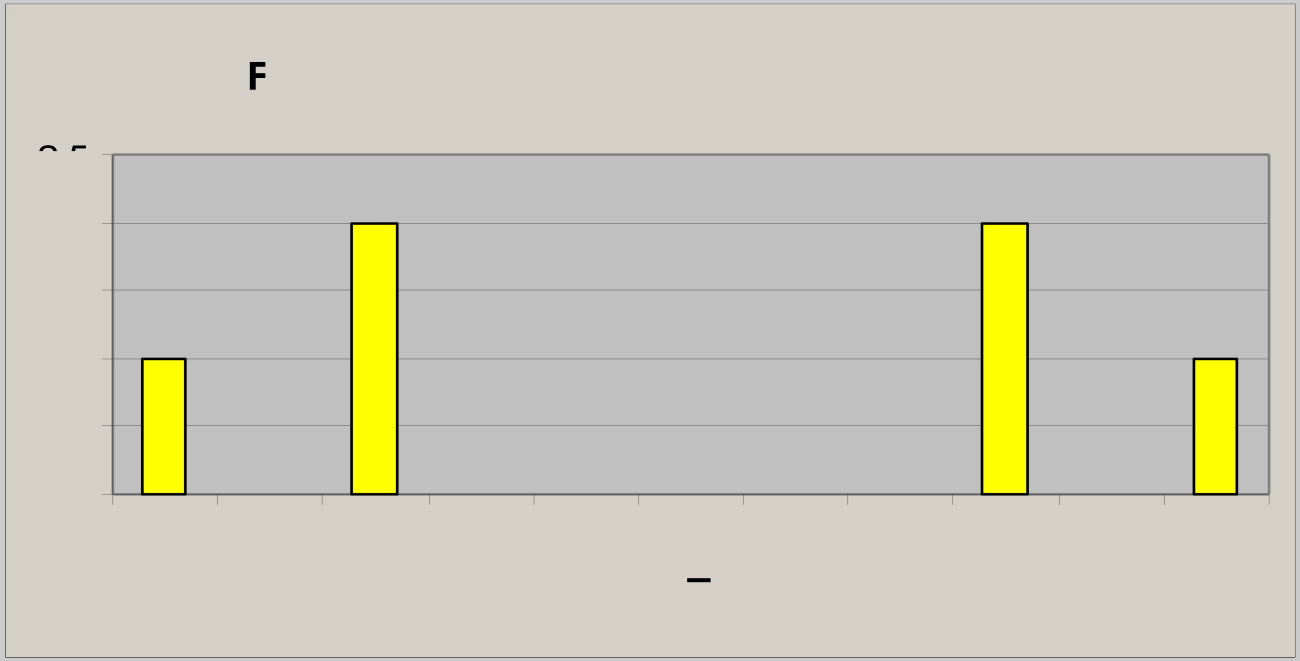
comment the result.

4.2 Variances of the estimates (error variances)

Example (continued):

$$N = 4, \quad (Y_1, Y_2, Y_3, Y_4) = (2,1,5,1) \quad \Rightarrow$$

$$Y = 9, \quad \bar{Y} = \frac{Y}{4} = 2{,}25$$



Theorem 4.2.1

For the error variance of $\hat{\bar{Y}} = \bar{y}$ there is:

$$\sigma^2_{\hat{\bar{Y}}} = \sigma^2_{\bar{y}} = \left(1 - \frac{n}{N}\right)\frac{1}{n} S_y^2$$

Terms

i) The ratio *n*/*N* is referred to as *sampling fraction*.

ii) $1 - \frac{n}{N}$ is referred to as *finite population correction*.

Exercise:

What factors have an impact on the error variance and what is that impact like?