Training Fiche Template

Title	Sampling Theory
Keywords (meta tags)	Data collection, statistical inference, estimation, sample size determination, simple random sampling, stratified sampling
Language	English
Objectives / Goals / Learning outcomes	**The aim of this module is to introduce and explain the basics of sampling theory.** **At the end of this module you will be able to:** - **Understand the differences between population and samples** - **Know the most commonly applied sampling techniques** - **Find optimal sample sizes**

Training course:	
Data Science Literacy	
Data Visualization and Visual Analytics Module	**X**
Introduction to Data science for Human & Social Sciences	
Data Science for good	
Data Journalism and Storytelling	

Description	In this training module you will be introduced to basics of sampling theory. Related to the theory of statistical inference, more specifically to the tools that allow for calculating confidence intervals, we will study the procedures that are used to find optimal sample sizes, depending on the characteristic to be estimated and the sampling technique used. In this module we will study the differences between sample-based data and population-based data and the most commonly applied sampling techniques: simple and statified sampling. Additionally, we will explore the rules for finding optimal sample sizes, conditional on some objectives related to the confidence and the margin of error that we want to have in our inferences.

Co-funded by the
Erasmus+ Programme
of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Contents arranged in 3 levels

1. INTRODUCTION

In statistical analysis, a population is a dataset for which we want to draw some conclusions. A survey is a procedure for which we obtain the data to be analyzed. Surveys can be based on the whole population (census-based or population based) or we might want to select a representative sub-set of this population. This sub-set is defined as a “sample” if its structure reflects the same structure as in the population and the data collected from surveys passed to a sample are called sample-based.

Why collecting datasets in the form of a sample instead of investigating the full population (census-based surveys)? The latter are necessary in counts and exhaustive researches, but they demand using huge resources and this results in high costs. On the contrary, sample-based surveys are appropriate if the population is homogenous, since they will constitute a good representation of the population. Moreover, they are the only option when the population is infinite and in information-destructive process. In any case, samples save time and other costs.

In practical terms, we normally do not have the resources to conduct census-based (population-based) studies, so the alternative is to base our analysis in samples. Basing our conclusions on sample data, implies that there will be an inherent margin of error, on which several factors can impact.

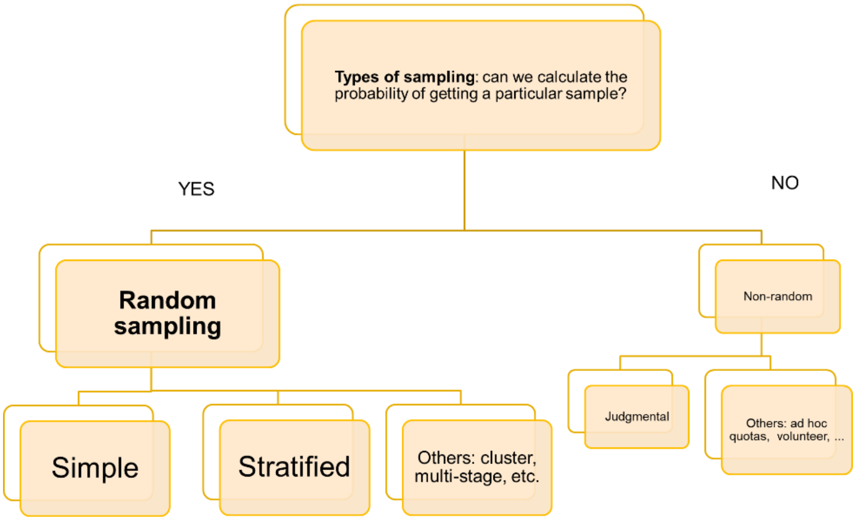
The margin of error will depend, basically on three driving factors:

a. How homogenous the data are in the population: the more heterogenous, all other things being equal, the larger the margin of error.

b. The sample size: the smaller the size, all other things being equal, the larger the margin of error.

c. The sampling technique: depending on the characteristics of your data.

We cannot do much about (a), but there is some room for acting on points (b) and (c). Regarding point (c) about the sampling tecchnique applied, it is important to note that there are a high variey of available sampling techniques that we could apply. The diagram below shows this variety in visual terms:


Types of sampling: can we calculate the probability of getting a particular sample?
YES
NO
Random sampling
Non-random
Simple
Stratified
Others: cluster, multi-stage, etc.
Judgmental
Others: ad hoc quotas, volunteer, ...

We can only control for the margin of error of our conclusions if we work with random samples and the most frequently random sampling techniques are the simple random sample and the stratified random sampling.

2. SAMPLING TECHNIQUES

2.1. Simple random sampling

Simple random sampling is the most elemental sampling technique that relies on random selection of the observations surveyed. It consists on, departing from a listing on the units of the population, selecting randomly n of these units. But even within this simple technique, some specifics of the random selection process can be decided. For example, we can decide if the sampling is going to take place with or without replacement. If the sampling is conducted with replacement, this means that each unit that is randomly selected to be part of the sample is put back into the population after each random selection draw. This obviously implies that one unit can be sampled more than once, but it guarantees that the conditions on which each selection draw take place are equal and constant, and the results of each one of them are independent on each other.

On the contrary, if a simple random sampling without replacement is conducted, each unit is sampled only once, but we cannot guarantee that the conditions are constant along the selection draws. Sampling

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

with and without replacement can produce significantly different results for small populations. They are equivalent only if the populations size (N) is very large.

2.2. Stratified sampling

In may occasions, observations are naturally grouped based on characteristics that they share. For example, data on the distribution of wages are grouped depending on the economic sector of the workers, or they gender, or their region of residence. Strata are defined as parts of the population of interest that present a high internal homogeneity, even when there is a large variability between strata. Stratified sampling takes advantage of these grouping of the observations and selects randomly a number of units on each stratum L (n_L), such us the total sample size is obtained by summing up the elements sampled on each stratum. There are several criteria for allocating the total sample size across strata, being the most common the following:

- Uniform: same sample size on any stratum
- Proportional: proportion of sample members the same as proportion of population members in each stratum
- Optimal: proportional to the size and heterogeneity (variance) on each stratum

Under the same conditions and with the same requirements of precision and confidence, we can affirm that, generally speaking, stratified sampling requires smaller sample size than simple sampling, but issues related to calculating sample sizes will be detailed in the next point.

3. CALCULATING OPTIMAL SAMPLE SIZES

The golden rule in terms of relating the sample size with the precision of our estimates is that the larger the same size, all other things being equal, the smaller the margin of error. However, getting statistical data, even if it is in the form of a sample, can be costly and sometimes we do not have resources to have large samples. As a consequence, there is a compromise solution that sets the optimal (minimum) sample size that we need, given our requirements in terms of precision (margin of error) and confidence of our estimates, and the heterogeneity (variance) of the variable of interest in the population.

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3.1 Solution for simple sampling

Assume first that we want our sample to estimate a population mean for a continuous variable, and our sample is going to be selected applying simple random sampling. The formulas that we need to apply are the following:

$$n^* = k^2 \frac{\sigma^2}{e^2}$$

$$n = \frac{n^*}{1 + \frac{n^*}{N}}$$

Constant k comes from a normal distribution and gets higher if we increase the desired confidence level and the symbol e stands for the margin of error we are willing to assume. Additionally, we need to make and assumption on the homogeneity of the variable in the population. This implies that we need to impose a realistic value (usually coming from some previous study) on the population variance σ^2.

In these equations, n^* is the solution for a simple random sampling with replacement, n is the solution for a simple random sampling without replacement and N is the population size. Generally speaking $n^* \geq n$, and both solutions converge when N is very large.

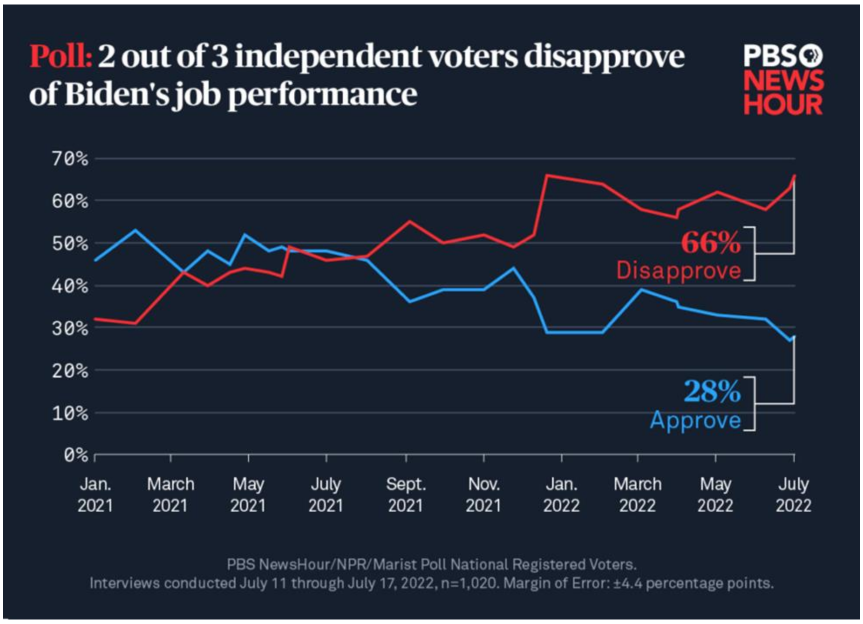
On a similar fashion, if we are interested in estimating the proportion (P) of units in a population that hold a given characteristic, the expressions required to find optimal sample sizes in this sampling technique are:

$$n^* = k^2 \frac{P * (1 - P)}{e^2}$$

$$n = \frac{n^*}{1 + \frac{n^*}{N}}$$

Again, the constant k comes from a normal distribution and gets higher if we increase the desired confidence level, and the term e stands for the margin of error we are willing to assume. In this case, We need to make and assumption on the value of P*(1-P), which is the variance of a binary (yes/no) variable. The usual solution is to suppose P=1-P=0.5, so P*(1-P)=0.25 takes its maximum value.

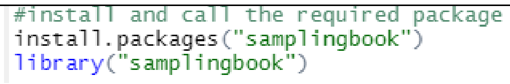
We can illustrate this technique by presenting a practical example on how sample sizes are determined and how applying R can help us on this regard: the Public Broadcasting Service (PBS) in the US regularly estimates the percentage of citizens that approve or disapprove the job of the president. In the case of President Joe Biden, they have been conducting these polls since January 2021. The following chart shows the evolution of their estimates:



In a recent poll along this series, PBS wanted to have estimates with a 99% confidence level, they were willing to have a margin of error of ±4.4% and they assumed the worst-case scenario (usual solution) and suppose that the percentage of people approving (P) is the same as the percentage not approving (1-P). What would be the number of citizens to be sampled with these conditions? The equations displayed above can be implemented in R language to find a solution.

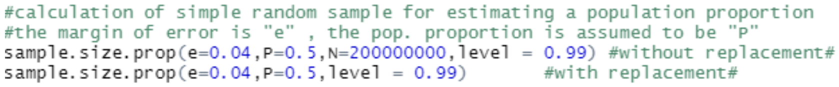
First we need to install and load the required packages:

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



```
#install and call the required package
install.packages("samplingbook")
library("samplingbook")
```

And later, we can find this optimal sample size by calling the function “sample.size.prop” in the package. This function allows for a sampling with or without replacement, although no practical differences will be found between the solution of these two alternatives given the large population size (N) from which the samples are drawn (we can arbitrarily assume that N=200,000,000). The following pieces of code calculate the respective solutions for a sampling without and with replacement:



```
#calculation of simple random sample for estimating a population proportion
#the margin of error is "e" , the pop. proportion is assumed to be "P"
sample.size.prop(e=0.04,P=0.5,N=200000000,level = 0.99) #without replacement#
sample.size.prop(e=0.04,P=0.5,level = 0.99)       #with replacement#
```

Which in both cases finds as solution a sample size of approximately 1,000 units.

3.2. Solution for stratified sampling

In this point the formulas for calculating sample sizes in the case of stratified sampling are detailed. For the shake of simplicity and clarity, we will focus only on the case of estimating a population mean, and we will offer the two most common solutions, which correspond to the cases of proportional (1) and optimal allocation (2):

$$(1) \qquad \mathrm{n} = \frac{\sum_{\mathrm{j}=1}^{\mathrm{L}} \mathrm{N_j}\sigma_{\mathrm{j}}^2}{N\frac{\mathrm{e}^2}{\mathrm{k}^2} + \frac{\sum_{\mathrm{j}=1}^{\mathrm{L}} \mathrm{N_j}\sigma_{\mathrm{j}}^2}{\mathrm{N}}}$$

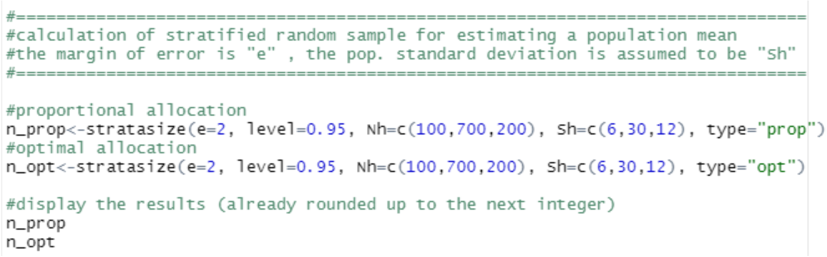
$$(2) \qquad \mathrm{n} = \frac{\frac{1}{N}\left(\sum_{\mathrm{j}=1}^{\mathrm{L}} \mathrm{N_j}\sigma_{\mathrm{j}}\right)^2}{N\frac{\mathrm{e}^2}{\mathrm{k}^2} + \frac{\sum_{\mathrm{j}=1}^{\mathrm{L}} \mathrm{N_j}\sigma_{\mathrm{j}}^2}{\mathrm{N}}}$$

As commented above, in both cases the formula corresponds to the estimation of the population mean for a continuous variable with a stratified sampling without replacement. In these expressions $\mathrm{N_j}$ stands for the size of stratum j and σ_{j}^2 for the variance of the variable on this same stratum.

Similarly to the solutions detailed for simple random sampling, we can illustrate how optimal sample sizes are calculated in stratified sampling by presenting a practical example applying R language.

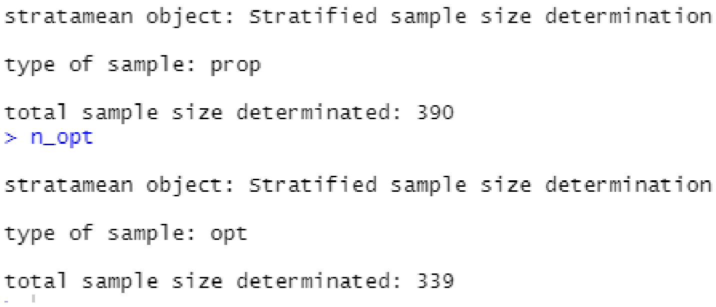
Assume that a charity is conducting a sample survey to study the annual donations made by its members, which are classified in three different groups according to their age with 100, 700 and 200 members each. From a pilot study this charity knows that the respective standard deviations (σ_j) in the annual donations in each group are €6, €30 and €12. We want to find the minimum sample size required to estimate the mean annual donation, setting a margin of error of €2 and a confidence level of 95%.

The following code lines will calculate the optimal sample size, offering the solutions for the case of a proportional and optimal allocation, by calling the function “stratasize” included in package “samplingbook” in R:



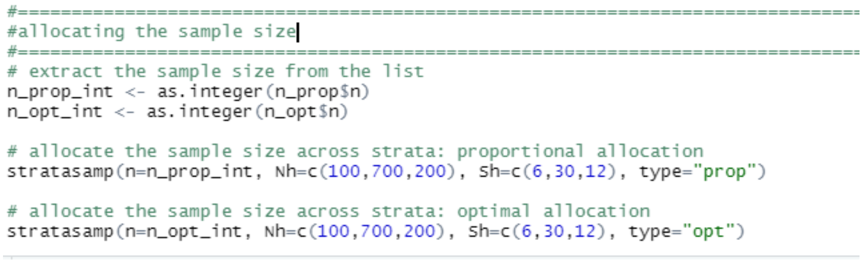
```
#===============================================================================
#calculation of stratified random sample for estimating a population mean
#the margin of error is "e" , the pop. standard deviation is assumed to be "Sh"
#===============================================================================

#proportional allocation
n_prop<-stratasize(e=2, level=0.95, Nh=c(100,700,200), Sh=c(6,30,12), type="prop")
#optimal allocation
n_opt<-stratasize(e=2, level=0.95, Nh=c(100,700,200), Sh=c(6,30,12), type="opt")

#display the results (already rounded up to the next integer)
n_prop
n_opt
```

The respective solutions are 390 and 339 units, as detailed below:



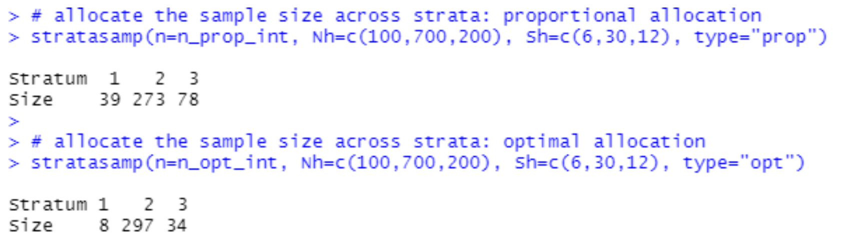
```
stratamean object: Stratified sample size determination

type of sample: prop

total sample size determinated: 390
> n_opt

stratamean object: Stratified sample size determination

type of sample: opt

total sample size determinated: 339
```

Finally, we can wonder of these two sample sizes will be allocated across strata, This can be done by calling the function “stratasamp” in the same package:



```
#===============================================================================
#allocating the sample size
#===============================================================================
# extract the sample size from the list
n_prop_int <- as.integer(n_prop$n)
n_opt_int <- as.integer(n_opt$n)

# allocate the sample size across strata: proportional allocation
stratasamp(n=n_prop_int, Nh=c(100,700,200), Sh=c(6,30,12), type="prop")

# allocate the sample size across strata: optimal allocation
stratasamp(n=n_opt_int, Nh=c(100,700,200), Sh=c(6,30,12), type="opt")
```

Being the solutions:



```
> # allocate the sample size across strata: proportional allocation
> stratasamp(n=n_prop_int, Nh=c(100,700,200), Sh=c(6,30,12), type="prop")

Stratum  1   2  3
Size    39 273 78
>
> # allocate the sample size across strata: optimal allocation
> stratasamp(n=n_opt_int, Nh=c(100,700,200), Sh=c(6,30,12), type="opt")

Stratum 1   2  3
Size    8 297 34
```

Self-assessment (multiple choice queries and answers)

Surveys based on samples:

a) Save resources if compared with a census-based survey
b) Allow for an exhaustive research in a population
c) Both answers are true

Sample size is affected by:

a) The margin of error and the confidence level
b) The sampling technique applied
c) Both answers are true

Proportional allocation distributes sample size across strata basing on:

a) The variance in each stratum
b) The size of each stratum
c) The mean value on each stratum

Resources (videos, reference link)

Related material

Related PPT

Bibliography

NEWBOLD, P. et al. (2008): Statistics for Management and Economics, (6th edition) Ed. Prentice Hall. Chapter 20, pp. 763-784.

Provided by

[Uniovi]

Co-funded by the
Erasmus+ Programme
of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.