Jean-Michel Josselin
Benoît Le Maux

Statistical Tools for Program Evaluation

Methods and Applications to Economic Policy, Public Health, and Education

Springer

Statistical Tools for Program Evaluation

Jean-Michel Josselin • Benoît Le Maux

Statistical Tools for Program Evaluation

Methods and Applications to Economic Policy, Public Health, and Education

Springer

Jean-Michel Josselin
Faculty of Economics
University of Rennes 1
Rennes, France

Benoît Le Maux
Faculty of Economics
University of Rennes 1
Rennes, France

ISBN 978-3-319-52826-7 ISBN 978-3-319-52827-4 (eBook)
DOI 10.1007/978-3-319-52827-4

Library of Congress Control Number: 2017940041

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

We would like to express our gratitude to those who helped us and made the completion of this book possible.

First of all, we are deeply indebted to the Springer editorial team and particularly Martina BIHN whose support and encouragement allowed us to finalize this project.

Furthermore, we have benefited from helpful comments by colleagues and we would like to acknowledge the help of Maurice BASLÉ, Arthur CHARPENTIER, Pauline CHAUVIN, Salah GHABRI, and Christophe TAVÉRA. Of course, any mistake that may remain is our entire responsibility.

In addition, we are grateful to our students who have been testing and experimenting our lectures for so many years. Parts of the material provided here have been taught at the Bachelor and Master levels, in France and abroad. Several students and former students have been helping us improve the book. We really appreciated their efforts and are very grateful to them: Erwan AUTIN, Benoît CARRÉ, Aude DAILLÈRE, Kristýna DOSTÁLOVÁ, and Adrien VEZIE.

Finally, we would like to express our sincere gratefulness to our families for their continuous support and encouragement.

Contents

Statistical Tools for Program Evaluation: Introduction and Overview

1

1.1 The Challenge of Program Evaluation

The past 30 years have seen a convergence of management methods and practices between the public sector and the private sector, not only at the central government level (in particular in Western countries) but also at upper levels (European commission, OECD, IMF, World Bank) and local levels (municipalities, cantons, regions). This "new public management" intends to rationalize public spending, boost the performance of services, get closer to citizens' expectations, and contain deficits. A key feature of this evolution is that program evaluation is nowadays part of the policy-making process or, at least, on its way of becoming an important step in the design of public policies. Public programs must show evidence of their relevance, financial sustainability and operationality. Although not yet systematically enacted, program evaluation intends to grasp the impact of public projects on citizens, as comprehensively as possible, from economic to social and environmental consequences on individual and collective welfare. As can be deduced, the task is highly challenging as it is not so easy to put a value on items such as welfare, health, education or changes in environment. The task is all the more demanding that a significant level of expertise is required for measuring those impacts or for comparing different policy options.

The present chapter offers an introduction to the main concepts that will be used throughout the book. First, we shall start with defining the concept of program evaluation itself. Although there is no consensus in this respect, we may refer to the OECD glossary which states that evaluation is the "*process whereby the activities undertaken by ministries and agencies are assessed against a set of objectives or criteria*." According to Michael Quinn Patton, former President of the American Evaluation Association, program evaluation can also be defined as "*the systematic collection of information about the activities, characteristics, and outcomes of programs, for use by people to reduce uncertainties, improve effectiveness, and make decisions*." We may also propose our own definition of the concept: program evaluation is a process that consists in collecting, analyzing, and using information

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_1

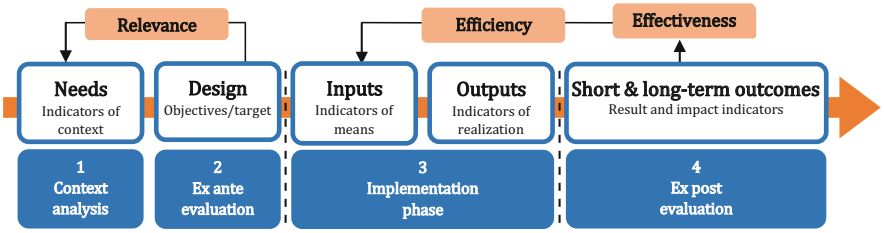

Fig. 1.1 Program evaluation frame

to assess the relevance of a public program, its effectiveness and its efficiency. Those concepts are further detailed below. Note that a distinction will be made throughout the book between a program and its alternative and competing strategies of implementation. By strategies, we mean the range of policy options or public projects that are considered within the framework of the program. The term program, on the other hand, has a broader scope and relates to the whole range of steps that are carried out in order to attain the desired goal.

As shown in Fig. 1.1, a program can be described in terms of needs, design, inputs and outputs, short and long-term outcomes. Needs can be defined as a desire to improve current outcomes or to correct them if they do not reach the required standard. Policy design is about the definition of a course of action intended to meet the needs. The inputs represent the resources or means (human, financial, and material) used by the program to carry out its activities. The outputs stand for what comes out directly from those activities (the intervention) and which are under direct control of the authority concerned. The short-term and long-term outcomes stand for effects that are induced by the program but not directly under the control of the authority. Those include changes in social, economic, environmental and other indicators.

Broadly speaking, the evaluation process can be represented through a linear sequence of four phases (Fig. 1.1). First, a context analysis must gather information and determine needs. For instance, it may evidence a high rate of school dropout among young people in a given area. A program may help teachers, families and children and contribute to prevent or contain dropout. If the authority feels that the consequences on individual and collective welfare are great enough to justify the design of a program, and if such a program falls within their range of competences, then they may wish to put it forward. Context analysis relies on descriptive and inferential statistical tools to point out issues that must be addressed. Then, the assessment of the likely welfare changes that the program would bring in to citizens is a crucial task that uses various techniques of preference revelation and measurement.

Second, ex-ante evaluation is interested in setting up objectives and solutions to address the needs in question. Ensuring the relevance of the program is an essential part of the analysis. Does it make sense within the context of its environment? Coming back to our previous example, the program can for instance consist of

alternative educational strategies of follow-up for targeted schoolchildren, with various projects involving their teachers, families and community. Are those strategies consistent with the overall goal of the program? It is also part of this stage to define the direction of the desired outcome (e.g., dropout reduction) and, sometimes, the desired outcome that should be arrived at, namely the target (e.g., a reduction by half over the project time horizon). Another crucial issue is to select a particular strategy among the competing ones. In this respect, methods of ex-ante evaluation include financial appraisal, budget impact analysis, cost benefit analysis, cost effectiveness analysis and multi-criteria decision analysis. The main concern is to find the most efficient strategy. Efficiency can be defined as the ability of the program to achieve the expected outcomes at reasonable costs (e.g., is the budget burden sustainable? Is the strategy financially and economically profitable? Is it cost-effective?)

Third, during the implementation phase, it is generally advised to design a monitoring system to help the managers follow the implementation and delivery of the program. Typical questions are the following. Are potential beneficiaries aware of the program? Do they have access to it? Is the application and selection procedure appropriate? Indicators of means (operating expenditures, grants received, number of agents) and indicators of realization (number of beneficiaries or users) can be used to measure the inputs and the outputs, respectively. Additionally, a set of management and accounting indicators can be constructed and collected to relate the inputs to the outputs (e.g., operating expenditures per user, number of agents per user). Building a well documented data management system is crucial for two reasons. First, those performance indicators can be used to report progress and alert managers to problems. Second, they can be used subsequently for ex-post evaluation purposes.

Last, the main focus of ex post evaluation is on effectiveness, i.e. the extent to which planned outcomes are achieved as a result of the program, ceteris paribus. Among others, methods include benchmarking, randomized controlled experiments and quasi-experiments. One difficulty is the time frame. For instance, the information needed to assess the program's outcomes is sometimes fully available only several years after the end of the program. For this reason, one generally distinguishes the short-term outcomes, i.e. the immediate effects on individuals' status as measured by a result indicator (e.g., rate of dropout during mandatory school time) from the longer term outcomes, i.e. the environmental, social and economic changes as measured by impact indicators (e.g., the impact of dropout on unemployment). In practice, ex post evaluation focuses mainly on short-term outcomes, with the aim to measure what has happened as a direct consequence of the intervention. The analysis also assesses what the main factors behind success or failure are.

We should come back to this distinction that we already pointed out between efficiency and effectiveness. Effectiveness is about the level of outcome per se and whether the intervention was successful or not in reaching a desired target. Depending on the policy field, the outcome in question may differ greatly. In health, for instance, the outcome can relate to survival. In education, it can be

school completion. Should an environmental program aim at protecting and restoring watersheds, then the outcome would be water quality. An efficiency analysis on the other hand has a broader scope as it relates the outcomes of the intervention to its cost.

Note also that evaluation should not be mistaken for monitoring. Roughly speaking, monitoring refers to the implementation phase and aims to measure progress and achievement all along the program's lifespan by comparing the inputs with the achieved outputs. The approach consists in defining performance indicators, routinely collect data and examine progress through time in order to reduce the likelihood of facing major delays or cost overruns. While it constitutes an important step of the intervention logic of a program, monitoring is not about evaluating outcomes per se and, as such, will be disregarded in the present work.

The remainder of the chapter is as follows. Section 1.2 offers a description of the tools that can be used to assess the context of a public program. Sections 1.3 and 1.4 are about ex-ante and ex-post evaluations respectively. Section 1.5 explains how to use the book.

1.2 Identifying the Context of the Program

The first step of the intervention logic is to describe the social, economic and institutional context in which the program is to be implemented. Identifying needs, determining their extent, and accurately defining the target population are the key issues. The concept of "needs" can be defined as the difference, or gap, between a current situation and a reasonably desired situation. Needs assessment can be based on a cross-sectional study (comparison of several jurisdictions at one specific point in time), a longitudinal study (repeated observations over several periods of time), or a panel data study (both time and individual dimensions are taken into account). Statistical tools which are relevant in this respect are numerous. Figure 1.2 offers an illustration.

First, a distinction is made between descriptive statistics and inferential statistics. Descriptive statistics summarizes data numerically, graphically or with tables. The main goal is the identification of patterns that might emerge in a sample. A sample is a subset of the general population. The process of sampling is far from straightforward and it requires an accurate methodology if the sample is to adequately represent the population of interest. Descriptive statistical tools include measures of central tendency (mean, mode, median) to describe the central position of observations in a group of data, and measures of variability (variance, standard deviation) to summarize how spread out the observations are. Descriptive statistics does not claim to generalize the results to the general population. Inferential statistics on the other hand relies on the concept of confidence interval, a range of values which is likely to include an unknown characteristic of a population. This population parameter and the related confidence interval are estimated from the sample data. The method can also be used to test statistical hypotheses, e.g., whether the population parameter is equal to some given value or not.

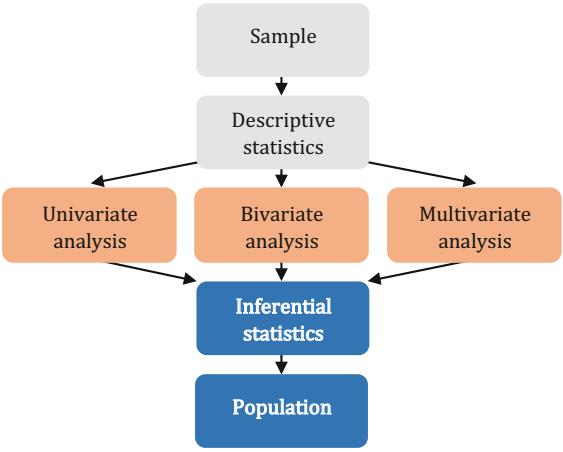

Fig. 1.2 Statistical methods at a glance

Second, depending on the number of variables that are examined, a distinction is made between univariate, bivariate and multivariate analyses. Univariate analysis is the simplest form and it examines one single variable at a time. Bivariate analysis focuses on two variables per observation simultaneously with the goal of identifying and quantifying their relationship using measures of association and making inferences about the population. Last, multivariate analyses are based on more than two variables per observation. More advanced tools, e.g., econometric analysis, must be employed in that context. Broadly speaking, the approach consists in estimating one or several equations that the evaluator think are relevant to explain a phenomenon. A dependent variable (explained or endogenous variable) is then expressed as a function of several independent variables (explanatory or exogenous variables, or regressors).

Third, program evaluation aims at identifying how the population would fare if the identified needs were met. To do so, the evaluator has to assess the indirect costs (negative externalities) as well as benefits (direct utility, positive externalities) to society. When possible, these items are expressed in terms of equivalent money-values and referred to as the willingness to pay for the benefits of the program or the willingness to accept its drawbacks. In other cases, especially in the context of health programs, those items must be expressed in terms of utility levels (e.g., quality adjusted life years lived, also known as QALYs). Several methods exist with their pros and cons (see Fig. 1.3). For instance, stated preference methods (contingent valuation and discrete choice experiment) exploit specially constructed questionnaires to elicit willingness to pay. Their main shortcoming is the failure to properly consider the cognitive constraints and strategic behavior of the agents participating in the experiment, leading to individuals' stated preferences that may not totally reflect their genuine preferences. Revealed preference methods use information from related markets and examine how agents behave in the face of real choices (hedonic-pricing and travel-cost methods). The main advantage of those methods is that they imply real money transactions and, as such, avoid the

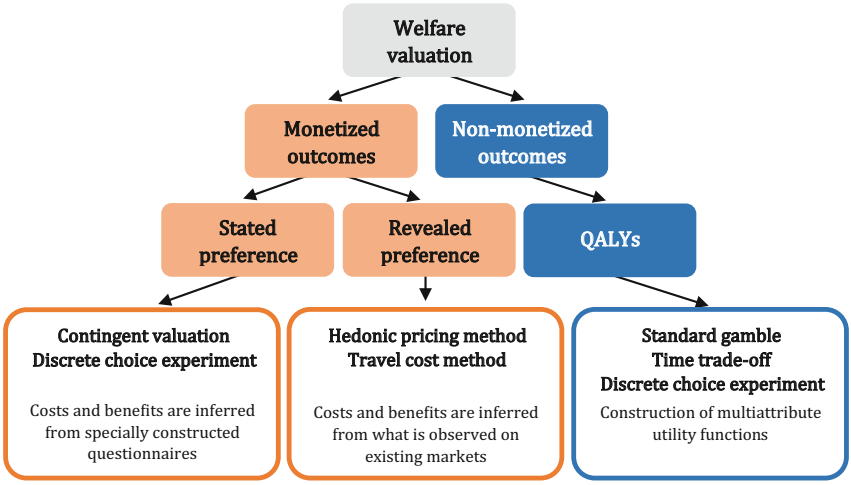

Fig. 1.3 Estimation of welfare changes

potential problems associated with hypothetical responses. They require however a large dataset and are based on sets of assumptions that are controversial. Last, health technology assessment has developed an ambitious framework for evaluating personal perceptions of the health states individuals are in or may fall into. Contrary to revealed or stated preferences, this valuation does not involve any monetization of the consequences of a health program on individual welfare.

Building a reliable and relevant database is a key aspect of context analysis. Often one cannot rely on pre-existing sources of data and a survey must be implemented to collect information from some units of a population. The design of the survey has its importance. It is critical to be clear on the type of information one needs (individuals and organizations involved, time period, geographical area), and on how the results will be used and by whom. The study must not only concern the socio economic conditions of the population (e.g., demographic dynamics, GDP growth, unemployment rate) but must also account for the policy and institutional aspects, the current infrastructure endowment and service provision, the existence of environmental issues, etc. A good description of the context and reliable data are essential, especially if one wants to forecast future trends (e.g., projections on users, benefits and costs) and motivate the assumptions that will be made in the subsequent steps of the program evaluation.

1.3 Ex ante Evaluation Methods

Making decisions in a non-market environment does not mean the absence of budget constraint. In the context of decisions on public projects, there are usually fixed sectoral (healthcare, education, etc.) budgets from which to pick the resources required to fund interventions. Ex ante evaluation is concerned with designing

public programs that achieve some effectiveness, given those budget constraints. Different forms of evaluation can take place depending on the type of outcome that is analyzed. It is therefore crucial to clearly determine the program's goals and objectives before carrying out an evaluation. The goal can be defined as a statement of the desired effect of the program. The objectives on the other hand stand for specific statements that support the accomplishment of the goal.

Different strategies/options can be envisaged to address the objectives of the program. It is important that those alternative strategies are compared on the basis of all relevant dimensions, be it technological, institutional, environmental, financial, social and economic. Among others, most popular methods of comparison include financial analysis, budget impact analysis, cost benefit analysis, cost effectiveness analysis and multi-criteria decision analysis. Each of these methods has its specificities. The key elements of a financial analysis are the cost and revenue forecasts of the program. The development of the financial model must consider how those items interact with each other to ensure both the sustainability (capacity of the project revenues to cover the costs on an annual basis) and profitability (capacity of the project to achieve a satisfactory rate of return) of the program. Budget impact analysis examines the extent to which the introduction of a new strategy in an existing program affects the authority's budget as well as the level and allocation of outcomes amongst the interventions (including the new one). Cost benefit analysis aims to compare cost forecasts with all social, economic and environmental benefits, expressed in monetary terms. Cost effectiveness analysis on the other hand focuses on one single measure of effectiveness and compares the relative costs and outcomes of two or more competing strategies. Last, multi-criteria decision analysis is concerned with the analysis of multiple outcomes that are not monetized but reflect the several dimensions of the pursued objective. Financial flows may be included directly in monetary terms (e.g., a cost, an average wage) but other outcomes are expressed in their natural unit (e.g., success rate, casualty frequency, utility level).

Figure 1.4 underlines roughly the differences between the ex ante evaluation techniques. All approaches account for cost considerations. Their main difference is with respect to the outcome they examine.

Financial Analysis Versus Cost Benefit Analysis A financial appraisal examines the projected revenues with the aim of assessing whether they are sufficient to cover expenditures and to make the investment sufficiently profitable. Cost benefit analysis goes further by considering also the satisfaction derived from the consumption of public services. All effects of the project are taken into account, including social, economic and environmental consequences. The approaches are thereby different, but also complementary, as a project that is financially viable is not necessarily economically relevant and vice versa. In both approaches, discounting can be used to compare flows occurring at different time periods. The idea is based on the principle that, in most cases, citizens prefer to receive goods and services now rather than later.

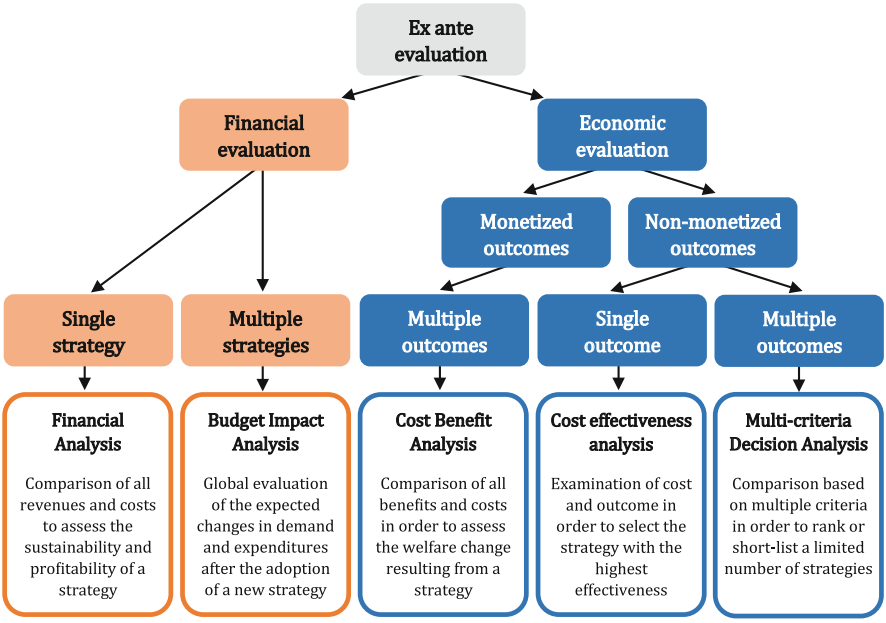

Fig. 1.4 Ex ante evaluation techniques

Budget Impact Versus Cost Effectiveness Analysis Cost effectiveness analysis selects the set of most efficient strategies by comparing their costs and their outcomes. By definition, a strategy is said to be efficient if no other strategy or combination of strategies is as effective at a lower cost. Yet, while efficient, the adoption of a strategy not only modifies the way demand is addressed but may also divert the demand for other types of intervention. The purpose of budget impact analysis is to analyze this change and to evaluate the budget and outcome changes initiated by the introduction of the new strategy. A budget impact analysis measures the evolution of the number of users or patients through time and multiplies this number with the unit cost of the interventions. The aim is to provide the decision-maker with a better understanding of the total budget required to fund the interventions. It is usually performed in parallel to a cost effectiveness analysis. The two approaches are thus complementary.

Cost Benefit Versus Cost Effectiveness Analysis Cost benefit analysis compares strategies based on the net welfare each strategy brings to society. The approach rests on monetary measures to assess those impacts. Cost effectiveness analysis on the other hand is a tool applicable to strategies where benefits can be identified but where it is not possible or relevant to value them in monetary terms (e.g., a survival rate). The approach does not sum the cost with the benefits but, instead, relies on pairwise comparisons by valuing cost and effectiveness differences. A key feature of the approach is that only one benefit can be used as a measure of effectiveness.

For instance, quality adjusted life years (QALYs) are a frequently used measure of outcome. While cost effectiveness analysis has become a common instrument for the assessment of public health decisions, it is far from widely used in other fields of collective decisions (transport, environment, education, security) unlike cost benefit analysis.

Cost Benefit Versus Multi-criteria Decision Analysis Multi-criteria decision analysis is used whenever several outcomes have to be taken into account but yet cannot be easily expressed in monetary terms**.** For instance, a project may have major environmental impacts but it is found difficult to estimate the willingness to pay of agents to avoid ecological and health risks. In that context, it becomes impossible to incorporate these elements into a conventional cost benefit analysis. Multi-criteria decision analysis overcomes this issue by measuring those consequences on numerical scales or by including qualitative descriptions of the effects. In its simplest form, the approach aims to construct a composite indicator that encompasses all those different measurements and allows the stakeholders' opinions to be accounted for. Weights are assigned on the different dimensions by the decision-maker. Cost benefit analysis on the other hand does not need to assign weights. Using a common monetary metric, all effects are summed into a single value, the net benefit of the strategy.

1.4 Ex post Evaluation

Demonstrating that a particular intervention has induced a change in the level of effectiveness is often made difficult by the presence of confounding variables that connect with both the intervention and the outcome variable. It is important to keep in mind that there is a distinction between causation and association. Imagine for instance that we would like to measure the effect of a specific training program, (e.g., evening lectures) on academic success among students at risk of school failure. The characteristics of the students, in particular their motivation and abilities, are likely to affect their grades but also their participation in the program. It is thereby the task of the evaluator to control for those confounding factors and sources of potential bias. As shown in Fig. 1.5., one can distinguish three types of evaluation techniques in this matter: randomized controlled experiment, benchmarking analysis and quasi-experiment.

Basically speaking, a controlled experiment aims to reduce the differences among users before the intervention has taken place by comparing groups of similar characteristics. The subjects are randomly separated into one or more control groups and treatment groups, which allows the effects of the treatment to be isolated. For example, in a clinical trial, one group may receive a drug while another group may receive a placebo. The experimenter then can test whether the differences observed between the groups on average (e.g., health condition) are caused by the intervention or due to other factors. A quasi-experiment on the other hand controls for the differences among units after the intervention has taken place.

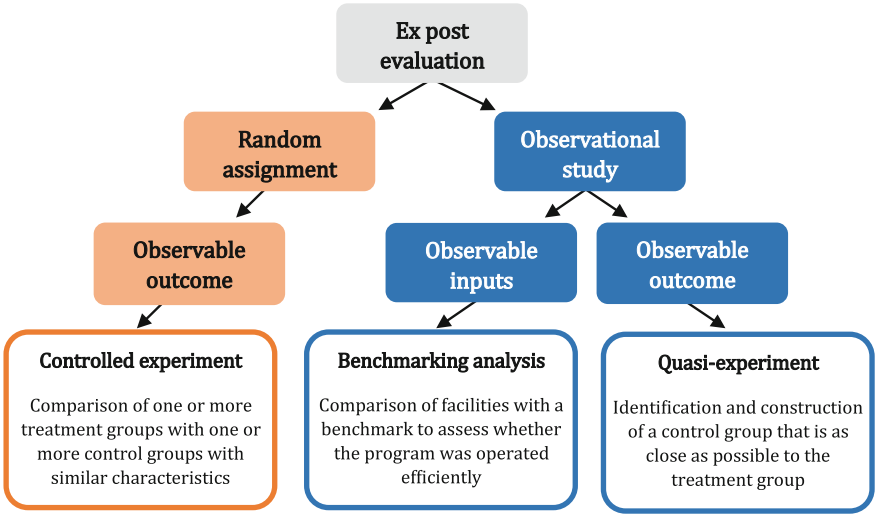

Fig. 1.5 Ex-post evaluation techniques

It does not attempt to manipulate or influence the environment. Data are only observed and collected (observational study). The evaluator then must account for the fact that multiple factors may explain the variations observed in the variable of interest. In both types of study, descriptive and inferential statistics play a determinant role. They can be used to show evidence of a selection bias, for instance when some members of the population are inadequately represented in the sample, or when some individuals select themselves into a group.

The main goal of ex post evaluation is to answer the question of whether the outcome is the result of the intervention or of some other factors. The true challenge here is to obtain a measure of what would have happened if the intervention did not take place, the so-called counterfactual. Different evaluation techniques can be put in place to achieve this goal. As stated above, one way is through a randomized controlled experiment. Other ways include difference-in-differences, propensity score matching, regression discontinuity design, and instrumental variables. All those quasi-experimental techniques aim to prove causality by using an adequate identification strategy to approach a randomized experiment. The idea is to estimate the counterfactual by constructing a control group that is as close as possible to the treatment group.

Another important aspect to account for is whether the program has been operated in the most effectual way in terms of input combination and use. Often, for projects of magnitude, there are several facilities that operate independently in their geographical area. Examples include schools, hospitals, prisons, social centers, fire departments. It is the task of the evaluator to assess whether the provision of services meets with management standards. Yet, the facilities involved in the implementation process may face different constraints, specific demand

settings and may have chosen different organizational patterns. To overcome those issues, one may rely on a benchmarking analysis to compare the cost structure of the facilities with that of a given reference, the benchmark.

Choosing which method to use mainly depends on the context of analysis. For instance, random assignment is not always possible legally, technically or ethically. Another problem with random assignment is that it can demotivate those who have been randomized out, or generate noncompliance among those who have been randomized in. In those cases, running a quasi-experiment is preferable. In other cases, the outcome in question is not easily observable and one may rely instead on a simpler comparison of outputs, and implement a benchmarking analysis. The time horizon and data availability thus also determine the choice of the method.

1.5 How to Use the Book?

The goal of the book is to provide the readers with a practical guide that covers the broad array of methods previously mentioned. The brief description of the methodology, the step by step approach, the systematic use of numerical illustrations allow to become fully operational in handling the statistics of public project evaluation.

The first part of the book is devoted to context analysis. It develops statistical tools that can be used to get a better understanding of problems and needs: Chap. 2 is about sampling methods and the construction of variables; Chap. 3 introduces the basic methods of descriptive statistics and confidence intervals estimation; Chap. 4 explains how to measure and visualize associations among variables; Chap. 5 describes the econometric approach and Chap. 6 is about the estimation of welfare changes.

The second part of the book then presents ex ante evaluation methods: Chap. 7 develops the methodology of financial analysis and details several concepts such as the interest rate, the time value of money or discounting; Chap. 8 includes a detailed description of budget impact analysis and extends the financial methodology to a multiple demand structure; Chaps. 9, 10 and 11 relate to the economic evaluation of the interventions and successively describe the methodology of cost benefit analysis, cost-effectiveness analysis, and multi-criteria decision analysis, respectively. Those economic approaches offer a way to compare alternative courses of action in terms of both their costs and their overall consequences and not on their financial flows only.

Last but not least, the third part of this book is about ex post evaluation, i.e. the assessment of the effects of a strategy after its implementation. The key issue here is to control for all those extra factors that may affect or bias the conclusion of the study. Chapter 12 introduces follow up by benchmarking. Chapter 13 explains the experimental approach. Chapter 14 details the different quasi-experimental techniques (difference-in-differences, propensity score matching, regression discontinuity design, and instrumental variables) that can be used when faced with observational data.

We have tried to make each chapter as independent of the others as possible. The book may therefore be read in any order. Readers can simply refer to the table of contents and select the method they are interested in. Moreover, each chapter contains bibliographical guidelines for readers who wish to explore a statistical tool more deeply. Note that this book assumes at least a basic knowledge of economics, mathematics and statistics. If you are unfamiliar with the concept of inferential statistics, we strongly recommend you to read the first chapters of the book.

Most of the information that is needed to understand a particular technique is contained in the book. Each chapter includes its own material, in particular numerical examples that can be easily reproduced. When possible, formulas in Excel are provided. When Excel is not suitable anymore to address specific statistical issues, we rely instead on R-CRAN, a free software environment for statistical computing and graphics. The software can be easily downloaded from internet. Codes will be provided all along the book with dedicated comments and descriptions. If you have questions about R-CRAN like how to download and install the software, or what the license terms are, please go to https://www.r-project.org/.

Bibliographical Guideline

The book provides a self-contained introduction to the statistical tools required for conducting evaluations of public programs, which are advocated by the World Bank, the European Union, the Organization for Economic Cooperation and Development, as well as many governments. Many other guides exist, most of them being provided by those institutions. We may name in particular the Magenta Book and the Green Book, both published by the HM Treasury in UK. Moreover, the reader can refer to the guidance document on monitoring and evaluation of the European Commission as well as its guide to cost benefit analysis and to the evaluation of socio-economic development. The World Bank also offers an accessible introduction to the topic of impact evaluation and its practice in development. All those guides present the general concepts of program evaluation as well as recommendations. Note that the definition of "program evaluation" used in this book is from Patton (2008, p. 39).

Bibliography

European Commission. (2013). *The resource for the evaluation of socio-economic development*.
European Commission. (2014). *Guide to cost-benefit analysis of investment projects*.
European Commission. (2015). G*uidance document on monitoring and evaluation*.
HM Treasury. (2011a). *The green book. Appraisal and evaluation in Central Government*.
HM Treasury. (2011b). *The magenta book. Guidance for evaluation*.
Patton, M. Q. (2008). *Utilization focused evaluation* (4th ed.). Saint Paul, MN: Sage.
World Bank. (2011). *Impact evaluation in practice*.

Part I

Identifying the Context of the Program

Sampling and Construction of Variables

2

2.1 A Step Not to Be Taken Lightly

Building a reliable and relevant database is a key aspect of any statistical study. Not only can misleading information create bias and mistakes, but it can also seriously affect public decisions if the study is used for guiding policy-makers. The first role of the analyst is therefore to provide a database of good quality. Dealing with this can be a real struggle, and the amount of resources (time, budget, personnel) dedicated to this activity should not be underestimated.

There are two types of sources from which the data can be gathered. On one hand, one may rely on pre-existing sources such as data on privately held companies (employee records, production records, etc.), data from government agencies (ministries, central banks, national institutes of statistics), from international institutions (World Bank, International Monetary Fund, Organization for Economic Co-operation and Development, World Health Organization) or from non-governmental organizations. When such databases are not available, or if information is insufficient or doubtful, the analyst has to rely instead on what we might call a homemade database. In that case, a survey is implemented to collect information from some or all units of a population and to compile the information into a useful summary form. The aim of this chapter is to provide a critical review and analysis of good practices for building such a database.

The primary purpose of a statistical study is to provide an accurate description of a population through the analysis of one or several variables. A variable is a characteristic to be measured for each unit of interest (e.g., individuals, households, local governments, countries). There are two types of design to collect information about those variables: census and sample survey. A census is a study that obtains data from every member of a population of interest. A sample survey is a study that focuses on a subset of a population and estimates population attributes through statistical inference. In both cases, the collected information is used to calculate indicators for the population as a whole.

J.-M. Josselin, B. Le Maux, *Statistical Tools for Program Evaluation*,
DOI 10.1007/978-3-319-52827-4_2

Since the design of information collection may strongly affect the cost of survey administration, as well as the quality of the study, knowing whether the study should be on every member or only on a sample of the population is of high importance. In this respect, the quality of a study can be thought of in terms of two types of error: sampling and non-sampling errors. Sampling errors are inherent to all sample surveys and occur because only a share of the population is examined. Evidently, a census has no sampling error since the whole population is examined. Non-sampling errors consist of a wide variety of inaccuracies or miscalculations that are not related to the sampling process, such as coverage errors, measurement and nonresponse errors, or processing errors. A coverage error arises when there is non-concordance between the study population and the survey frame. Measurement and nonresponse errors occur when the response provided differs from the real value. Such errors may be caused by the respondent, the interviewer, the format of the questionnaire, the data collection method. Last, a processing error is an error arising from data coding, editing or imputation.

Before deciding to collect information, it is important to know whether studies on a similar topic have been implemented before. If this is to be the case, then it may be efficient to review the existing literature and methodologies. It is also critical to be clear on the objectives, especially on the type of information one needs (individuals and organizations involved, time period, geographical area), and on how the results will be used and by whom. Once the process of data collection has been initiated or a fortiori completed, it is usually extremely costly to try and add new variables that were initially overlooked.

The construction of a database includes several steps that can be summarized as follows. Section 2.2 describes how to choose a sample and its size when a census is not carried out. Section 2.3 deals with the various ways of conceiving a questionnaire through different types of questions. Section 2.4 is dedicated to the process of data collection as it details the different types of responding units and the corresponding response rates. Section 2.5 shows how to code data for subsequent statistical analysis.

2.2 Choice of Sample

First of all, it is very important to distinguish between the target population, the sampling frame, the theoretical sample, and the final sample. Figure 2.1 provides a summary description of how these concepts interact and how the sampling process may generate errors.

The target population is the population for which information is desired, it represents the scope of the survey. To identify precisely the target population, there are three main questions that should be answered: who, where and when? The analyst should specify precisely the type of units that is the main focus of the study, their geographical location and the time period of reference. For instance, if the survey aims at evaluating the impact of environmental pollution, the target population would represent those who live within the geographical area over which

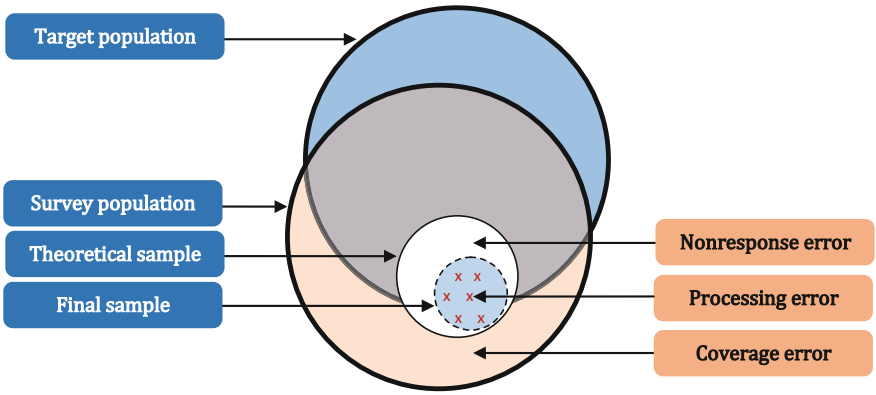

Fig. 2.1 From the target population to the final sample

the pollution is effective or those who may be using the contaminated resource. If the survey is about the provision of a local public good, then the target population may be the local residents or the taxpayers. As to a recreational site, or a better access to that site, the target population consists of all potential users. Even at this stage carefulness is required. For instance, a local public good may generate spill-over effects in neighboring jurisdictions, in which case it may be debated whether the target population should reach beyond local boundaries.

Once the target population has been identified, a sample that best represents it must be obtained. The starting point in defining an appropriate sample is to determine what is called a survey frame, which defines the population to be surveyed (also referred to as survey population, study population or target population). It is a list of all sampling units (list frame), e.g., the members of a population, which is used as a basis for sampling. A distinction is made between identification data (e.g., name, exact address, identification number) and contact data (e.g., mailing address or telephone number). Possible sampling frames include for instance a telephone directory, an electoral register, employment records, school class lists, patient files in a hospital, etc. Since the survey frame is not necessarily under the control of the evaluator, the survey population may end up being quite different from the target population (coverage errors), although ideally the two populations should coincide.

For large populations, because of the costs required for collecting data, a census is not necessarily the most efficient design. In that case, an appropriate sample must be obtained to save the time and, especially, the expense that would otherwise be required to survey the entire population. In practice, if the survey is well-designed, a sample can provide very precise estimates of population parameters. Yet, despite all the efforts made, several errors may remain, in particular nonresponse, if the survey fails to collect complete information on all units in the targeted sample. Thus, depending on survey compliance, there might be a large difference between the theoretical sample that was originally planned and the final sample. In addition to these considerations, several processing errors may finally affect the quality of the database.

A sample is only a portion of the survey population. A distinction is consequently made between the population parameter, which is the true value of the population attribute, and the sample statistic, which is an estimate of the population parameter. Since the value of the sample statistic depends on the selected sample, the approach introduces variability in the estimation results. The computation of a margin of error $\pm e$ is therefore crucial. It yields a confidence interval, i.e. a range of values, which is likely to encompass the true value of the population parameter. It is a proxy for the sampling error and an important issue with sampling design is to minimize this confidence interval.

How large should a sample be? Unfortunately, there is no unique answer to this question since the optimal size can be thought of in terms of a tradeoff between precision requirements ($\pm e$) and operational considerations such as available budget, resources and time. Yet, an indicative formula provides the minimum size of a sample. It is based on the calculation of a confidence interval for a proportion. As an illustration, assume that one wishes to estimate the portion of a population that has a specific characteristic, such as the share of males. The true population proportion is denoted π and the sample proportion is denoted p. Since π is unknown, we can only use the characteristics of the sample to compute a confidence interval. Assume for instance that we find $p = 45\%$ (i.e. 45 percent of the sample units are male) and calculate a margin of error equal to $e = 3\%$. The analyst can specify a range of values $45\% \pm 3\%$ in which the population parameter π is likely to belong, i.e. the confidence interval is $[42\%, 48\%]$. Statistical precision can thus be thought of as how narrow the confidence interval is.

The formula for calculating a margin of error for a proportion is:

$$e = z_\alpha \times \sqrt{\frac{p(1-p)}{n_0}}$$

Three main factors determine the magnitude of the confidence interval. First, the higher is the sample size n_0, the lower is the margin of error e. At first glance, one should then try to maximize the sample size. However, since the margin of error decreases with the square root of the sample size, there is a kind of diminishing returns to increasing sample size. Concurrently, the cost of survey administration is likely to increase linearly with n_0. There is consequently a balance to find between those opposing effects. Second, a sample should be as representative as possible of the population. If the population is highly heterogeneous, the possibility of drawing a non-representative sample is actually high. In contrast, if all members are identical, then the sample characteristics will perfectly match the population, whatever the selected sample is. Imagine for instance that $\pi = 90\%$, i.e. most individuals in the population are males. In that case, if the sample is randomly chosen, the likelihood of selecting a non-representative sample (e.g., only females) is low. On the contrary, if the gender attribute is equally distributed ($\pi = 50\%$), then this likelihood is high. Since the population variance $\pi(1-\pi)$ is unknown, the sample variance $p(1-p)$ will serve as a proxy for measuring the heterogeneity in the

population. The higher is $p(1-p)$, the lower is the precision of the sample estimate. Third, the z_α statistic allows to compute a margin of error with a $(1-\alpha)$ confidence level, which corresponds to the probability that the confidence interval calculated from the sample encompasses the true value of the population parameter. The sampling distribution of p is approximately normally distributed if the population size is sufficiently large. The usually accepted risk is $\alpha = 5\%$ so that the confidence level is 95%. The critical value $z_{5\%} = 1.96$ is computed with a normal distribution calculator.

Let us now consider the formula for the margin of error from a different perspective. Suppose that instead of computing e, we would like to determine the sample size n_0 that achieves a given level of precision, hence keeping the margin of error at the given level e. The equation can be rewritten:

$$n_0 = z_{5\%}^2 \times \frac{p(1-p)}{e^2}$$

Table 2.1 highlights the relationship between the parameters. For instance, when the proportion p is 10% and the margin of error e is set to 5%, the required sample size is $n_0 = 138$. If we want to reach a higher precision, say $e = 1\%$, then we have to survey a substantially higher number of units: $n_0 = 3457$. Of course, the value of p is unknown before the survey has been implemented. Yet, the maximum of the sample variance $p(1-p)$ is obtained for $p = 50\%$. For that value of the proportion, and in order to achieve a level of precision $e = 1\%$, one should survey at least $n_0 = 9604$ units, and $n_0 = 384$ to achieve $e = 5\%$.

The sample size also depends on the size of the target population, denoted N hereafter. Below approximately $N = 200{,}000$, a finite population correction factor has to be used:

Table 2.1 Sample size for an estimated proportion

Proportion		Margin of error			
p (%)	1–p (%)	0.5%	1%	5%	10%
10	90	13,830	3457	138	35
20	80	24,586	6147	246	61
30	70	32,269	8067	323	81
40	60	36,879	9220	369	92
50	50	38,416	9604	384	96
60	40	36,879	9220	369	92
70	30	32,269	8067	323	81
80	20	24,586	6147	246	61
90	10	13,830	3457	138	35

$$e = z_{5\%} \times \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Solving for n yields:

$$n \times \frac{N-1}{N-n} = z_{5\%}^2 \times \frac{p(1-p)}{e^2}$$

$$n \times \frac{N-1}{N-n} = n_{0,}$$

$$n = \frac{n_0 N}{n_0 + N - 1}.$$

For instance, while we were previously suggesting a sample size of $n_0 = 384$ to ensure a margin of error of 5%, now, with the new formula, and if the population size is $N = 500$, we have:

$$n = \frac{384 \times 500}{384 + 500 - 1} \approx 217$$

Table 2.2 provides an overview of the problem. Those figures provide a useful rule of thumb for the analyst. For a desired level of precision e, the lower is the population size N, the lower is the number n of units to survey. Those results, however, have to be taken with caution. What matters at the end is common sense. For instance, according to Table 2.2, if $N = 1000$ the analyst should survey $n = 906$ units to ensure a margin of error of 1%. In that case, sampling would virtually be equivalent to a census, in statistical terms but also in budget and organizational terms. Moving to a less stringent 5% margin of error would provide a much more relevant and tractable number of units to survey.

In practice, most polling companies survey from 400 to 1000 units. For instance, the NBC News/Wall Street Journal conducted in October 2015 a public opinion poll

Table 2.2 Target population and sample size

Population	Margin of error			
N	0.5%	1%	5%	10%
50	50	50	44	33
100	100	99	80	49
500	494	475	217	81
1000	975	906	278	88
2000	1901	1655	322	92
5000	4424	3288	357	94
10,000	7935	4899	370	95
100,000	27,754	8763	383	96

relating to the 2016 United States presidential election (a poll is a type of sample survey dealing mainly with issues of public opinions or elections). A number of 1000 sampling units were interviewed by phone. Most community satisfaction surveys rely on similar sample sizes. For instance, in 2011, the city of Sydney, Australia, focused on a series of $n = 1000$ telephone interviews to obtain a satisfaction score related to community services and facilities. Smaller cities may instead focus on $n = 400$ units. At a national level, sample sizes reach much larger values. To illustrate it, in 2014, the American Community Survey selected a sample of about 207,000 units from an initial frame of 3.5 million addresses. According to our rule of thumb, this would yield a rather high precision, approximately $e = 0.2\%$.

The choice of sample size also depends on the expected in-scope proportion and response rate. First, it is possible that despite all efforts coverage errors exist and that a number of surveyed units do not belong to the target population. On top of these considerations, the survey may fail to reach some sampling units (refusals, noncontacts). To guarantee the desired level of precision, one needs therefore to select a sample larger than predicted by the theory, using information about the expected in-scope and response rates. More specifically, the following adjustment can be implemented:

$$\text{Adjusted sample size} = \frac{n}{\text{Expected response rate} \times \text{Expected in-scope rate}}$$

Suppose for instance that the in-scope rate estimated from similar surveys or pilot tests is 91%. Assume also that the expected response rate is 79%. When $n = 1000$, the adjusted sample size is:

$$\text{Adjusted sample size} = \frac{1000}{0.91 \times 0.79} = 1391$$

A crucial issue here is that once the expected in-scope and response rates have been defined ex ante, their values should serve as a target during the data collection process. A response rate or in-scope rate lower than the desired values will result in a sample size that does not ensure anymore the precision requirement. For instance, in the case of the American Community Survey, if we fictitiously assume an ex-post response rate of 25% and in-scope rate of 85%, which can be realistic in some cases (if not in this particular one), then the margin of error increases from e = 0.2% to 0.5%.

To conclude, whether one chooses a higher or lower sample size (or equivalently, a higher or lower precision) mainly depends on operational constraints such as the budget, but also the time available to conduct the entire survey and the size of the target population. First, there are direct advantages and disadvantages to using a census to study a population. On the one hand, a census provides a true measure of the population but also detailed information about sub-groups within the population, which can be useful if heterogeneity matters. On the other hand, a sample generates lower costs both in staff and monetary terms

and is easier to control and monitor. Second, the time needed to collect and process the data increases with the sample size. Thus, with a sample survey of realistic size, the results are generally available in less time and can still be representative of the population. Third, the population size is also a determinant factor. If the population is small, a census is always preferable. In contrast, for large populations, accurate results can be obtained from reasonably small samples. In any case, the next step now consists in conceiving the questionnaire that will be proposed to respondents.

2.3 Conception of the Questionnaire

A questionnaire is a set of questions designed to elicit information upon a subject, or sequence of subjects, from a respondent. Given its impact on data quality, the questionnaire design plays a central role. The purpose of a survey is to obtain sincere responses from the respondent. One main principle applies in this matter: one should start on the basis that most people do not want to spend time on a survey, and if they do, it could be that they actually are not satisfied with the policy under evaluation, which may be non-representative of the population as a whole. Nonresponses should be minimized as much as possible. This can be done by explaining why the survey is carried out, by keeping it quick and by telling the respondents that the results will be communicated once finalized. Those three rules are even truer nowadays since people are frequently required to participate in surveys in many fields.

An important aspect of questionnaire design is the type of response formats. There are two categories of questions: open-ended versus close-ended. Close-ended questions request the respondent to choose one or several responses among a predetermined set of options. While they limit the range of respondents' answers on the one hand, they require less time and effort for both the interviewer and the participant on the other hand. In contrast, open-ended questions do not give respondents options to choose from. Thereby, they allow them to use their own words and to include more information, including their feelings and understanding of the problem.

Examples of close-ended and open-ended questions are provided in Fig. 2.2. Dichotomous questions (also referred to as two-choice questions) are the simplest version of a close-ended question. They propose only two alternatives to the respondent. Multiple choice questions propose strictly more than two alternatives and ask the respondent to select one single response from the list of possible choices. Checklist questions (or check-all questions) allow a respondent to choose more than one of the alternatives provided. Forced choice questions are similar to checklist questions, although the respondent is required to provide an answer (e.g., yes–no) for every response option individually. Partially closed questions provide an alternative "Other, please specify", followed by an appropriately sized answer box. This type of question is useful when it is difficult to list all possible alternatives or when responses cannot be fully anticipated. Last, open-ended questions can be of two forms, either text or numerical.