utb.

Andreas Behr

Theory of Sample Surveys with R

utb 4328

Eine Arbeitsgemeinschaft der Verlage

Böhlau Verlag · Wien · Köln · Weimar
Verlag Barbara Budrich · Opladen · Toronto
facultas · Wien
Wilhelm Fink · Paderborn
A. Francke Verlag · Tübingen
Haupt Verlag · Bern
Verlag Julius Klinkhardt · Bad Heilbrunn
Mohr Siebeck · Tübingen
Nomos Verlagsgesellschaft · Baden-Baden
Ernst Reinhardt Verlag · München · Basel
Ferdinand Schöningh · Paderborn
Eugen Ulmer Verlag · Stuttgart
UVK Verlagsgesellschaft · Konstanz, mit UVK/Lucius · München
Vandenhoeck & Ruprecht · Göttingen · Bristol
Waxmann · Münster · New York

Andreas Behr

Theory of Sample Surveys with R

UVK Verlagsgesellschaft mbH · Konstanz
mit UVK/Lucius · München

Prof. Dr. Andreas Behr lehrt Statistik an der Universität Duisburg-Essen.

Lob und Kritik

Wir freuen uns darüber, dass Sie sich für ein UTB-Lehrbuch entschieden haben und hoffen, dass Sie dieses Buch in Ihrem Studium sinnvoll unterstützt.
Für Lob und Kritik haben wir stets ein offenes Ohr: Schreiben Sie uns einfach eine E-Mail an das Lektorat (wirtschaft@uvk.de).

Online-Angebote oder elektronische Ausgaben sind erhältlich unter www.utb-shop.de.

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Lektorat: Rainer Berger
Einbandgestaltung: Atelier Reichert, Stuttgart
Umschlagmotiv: © bluedesign - Fotolia.com
Druck und Bindung: fgb · freiburger graphische betriebe, Freiburg

UVK Verlagsgesellschaft mbH
Schützenstr. 24 · 78462 Konstanz
Tel. 07531/9053-0 · Fax 07531/9053-98
www.uvk.de

UTB-Nr. 4328
ISBN 978-3-8252-4328-9

Preface

The book is based upon my lecture notes developed for the course in sampling theory at Münster University and University Essen-Duisburg. Some basic knowledge of statistical methods, regression analysis and the R programming environment would be helpful for understanding the text.

I am grateful to Ullrich Rendtel, Götz Rohwer and Ulrich Pötter helping me to understand the basics of sampling theory and the feedback from many students of my courses. I would like to thank Christoph Schiwy for providing the layout of the manuscript using Latex and knitr and Katja Theune and Jurij Weinblat for reading the manuscript.

Münster, December 2014 *Andreas Behr*

Some hints for solutions of the exercises which accompany all chapters and the data files used in the text can be downloaded at `www.uvk-lucius.de/behr`.

Contents

List of Figures

1
Introduction

Survey samples provide the most important sources of information in the social sciences. Basic to all statistical considerations is the random selection of elements of the population into the sample. The sampling design specifies the specific procedure of sampling. While in a strict sense our knowledge will be confined to the elements in the sample, knowledge of the properties of specific sampling designs is indispensable to plan, carry out and analyse survey samples.

1.1 Sources of randomness

Introductory courses in probability calculus discuss properties of random variates. Basic to the idea of randomness is the random generator $\mathcal{G}$. In most applications of probability calculus in the social sciences, characteristics of individuals, e.g. the working status or the income, are regarded as realizations of random variates. A stochastic model can be regarded as a detailed description of a complicated random generator.

In the design approach applied in survey sampling the random process is strictly confined to the random sampling of elements from the population. The characteristics of the elements (e.g. their income) are treated as fixed. The difference of the modelling and the design approach can be illustrated by means of simple random experiments, casting dices and dimes and drawing balls from urns.

1.1.1 Stochastic model

Stochastic models are specific random generators that can repeatedly be used to generate realizations of random variates. A very specific understanding of social reality, most common in economics, regards this reality as being the product of the application of random generators. Assume we have a fair dice and will provide each person an amount of money equal to the number obtained from casting the dice (in 1000 euro) and call that amount income. Therefore, the income of a person is a random variate having a specific probability distribution. E.g., the expected income is 3500 euro and so is the average of two generated incomes. If we have the impression that this model is not a realistic description of the social process in which incomes are determined, we can improve (complicate) the random process. E.g., we can additionally cast a dime if the person is male. If head comes up, the income is increased by 1000 euro, if tail comes up by 2000 euro. Being still not satisfied with the model, we can throw additionally a dime if the person has an academic degree and increase the income by 1000 euro if head comes and by 2000 euro if tail. Obviously, we can proceed in complicating the model but the main point is that income is generated by means of a (perhaps rather complicated) random number generator and therefore a random variate. Moreover, we can generate as many incomes as we want making use of the random generator repeatedly. Alternatively, as is most common in econometrics, one can imagine the income of a person being a linear combination of her characteristics, e.g. age, years of education and so on, to which a realization of a normally distributed random variate is added.

1.1.2 Design approach

In the design approach, the characteristics of the individuals, e.g. their working status or their income, are treated as fixed. Assume a population of six individuals having incomes of 1000, 2000, ..., 6000 euro. We do not speculate why person number six earns 6000 whereas person number one only earns 1000 euro. Now assume that we do not know the income of the six persons but that we can sample randomly two persons and get information about their income. To obtain a sample, we number six otherwise identical balls, put them in an urn, and draw blindly two balls. The two numbers obtained refer to two of the six persons and we will be informed about their incomes. Obviously, the incomes we observe depend on the sample we happen to draw. The expected average income of the two persons sampled is 3500 euro just as in the example of stochastic modelling. However, the difference is, in the sampling example the incomes are treated as fixed. Therefore, the income is not regarded as a random variate but as a fixed property of the persons. It is only random, which specific persons will be included in the sample.

1.2 Surveys

Surveys based on random selection are important sources of information about conditions and changes in society. In Germany, e.g. the micro census (Mikrozensus) carried out every year by the German Federal Statistical Office (Statistisches Bundesamt) includes about 1 million individuals. The German Socio Economic Panel (GSOP) carried out by the German Institute for Economic Research (DIW Berlin) is an important source of information about social and economic conditions in Germany and is used in a large number of scientific analyses.

1.2.1 Characteristics of surveys

A survey sample denotes a statistical inquiry and analysis that meets several important requirements.

1. The interest is confined to a well-defined population denoted by U. A census would include all elements of U but is very rarely carried out because of cost and time considerations.

2. Instead of questioning all elements of U, a sample $s \subset U$ is sampled and information for the elements in s is obtained.

3. To apply random calculus, the individuals of the sample have to be selected by making use of (pseudo) random numbers, usually generated by means of implemented random generators. We denote a random generator by $\mathcal{G}$.

4. A most simple random generator can be seen in an urn filled with different balls, which are drawn blindly, that is independently of what is written on the balls or their colour. Therefore, the ideal vision of obtaining a random sample is an urn with a ball for each element of the population marked with a non-ambiguous identifier and the blind draw of a specified number of balls.

5. The sample obtained making use of a random generator is called a random sample. We will restrict the discussion in this text towards random samples.

6. The sampling frame contains information to identify all elements in the population. Ideally, we can think of the sampling frame being a file, which uniquely assigns a non-ambiguous identifier to each element of the population. We will abstract from the fact that, in practice, it is often difficult to obtain a complete list for a defined population.

7. The variables of interest are often quite numerous but we denote a representative variable of interest by Y.

8. The distribution of variable Y in the population can be characterized by different parameters, e.g. total (sum), mean, standard deviation, and so on. As we do not observe Y for all members of the population but only for the members of the sample, we cannot calculate the true parameters for the population. Instead, we try to estimate the population parameters based on the information provided by the sample.

9. We try to grasp the extent of the expected estimation error by providing estimates of the variance of the estimation functions.

1.2.2 Sampling frame

The sampling frame ideally is a list containing a non-ambiguous identifier for all elements of the population. Furthermore, information on how the elements can be contacted must be included in the frame. As an example one can think of a complete and up to date register maintained by the city's registration office including e.g. name, date of birth and address of all inhabitants. Note that in practice a complete and up to date register of the target population is seldom available due to non-registered inhabitants, incomplete registration of persons moving in or out of the relevant area and so on. Furthermore, even if a potential useful register exists, data privacy or prohibitive costs may prevent its use.

1.2.3 Probability sampling

Throughout this text, we focus on probability sampling. The sampling space $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ consists of the M different samples that can possibly be drawn from the population U. The sampling design associates a probability $\mathrm{P}(S = s) = p(s)$ to each of the M possible samples.

1.2.4 Sampling and inference

People often associate some miraculous capabilities with survey sampling. However, from the N elements in the population we will know the specific values y only for the n elements contained in the sample, and therefore our knowledge will be restricted to these sampled elements. About all the $N - n$ elements of the population, which have not been sampled, we cannot say anything specific.

Therefore, the relevant question is not what can be said about the elements which have not been sampled but rather in what specific way we should carry out the sampling, which estimating functions with their specific characteristics we should apply, and what we can expect to happen in doing so. Hence, we focus on the procedure of sampling and on the general characteristics of estimating functions.

A simple trick will be helpful to learn about the properties of sampling designs and estimating functions: We counterfactually consider a population as completely known and observe the outcomes when drawing samples and applying estimating functions to these samples.