Statistics Taught through Fiction

Statistics Taught through Fiction

By
Krzysztof Z. Górniak

Translated and edited by
Małgorzata Mazurkiewicz

Cambridge
Scholars
Publishing

Statistics Taught through Fiction

By Krzysztof Z. Górniak
Translated and edited by Małgorzata Mazurkiewicz
Illustrated by Katarzyna Mazurkiewicz

This book first published 2017

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-5183-3
ISBN (13): 978-1-4438-5183-1

To our friends: from kindergarten to old age

TABLE OF CONTENTS

There isn't one.

What for? Once you've started reading, you will leave everything behind and simply won't be able to stop.

Moreover, every time you open this book, you will find pleasure in it.

EXPLANATION INSTEAD OF AN INTRODUCTION

Dear Children,

I've used a literary genre that belongs (in general) to the education of adults. I'm very sorry about that.

Why have I done this? My long-standing experience shows that it is one of not so many ways of understanding the nature of statistics. I have no intention of making you remember formulas, which can easily be found on the Internet. I'm not convinced that students should be able to generate the Chi-Square distribution with the first of the ten degrees of freedom and so on off by heart…Less scholasticism, more understanding.

If children can discover more about the world from fairy tales, that means that adults who already know this world have the chance to get to know the Queen of Statistics.

To all teachers giving lectures on statistics – if they're hurt, I'm sorry. For those who are not, I congratulate you on your sense of humour.

You can express your gratitude or deep admiration to:
atocidopiero@zwarszawy.pl

Introduction Instead of Explanation

After reading many times the royal content, the readers can make themselves acquainted with the nightmares which were gnawing at Tosia. This remark is directed to those who are disturbed by their own good mood and to those who desperately need to change this state.

Still Not the Nightmare 1

Tosia wasn't in a good mood whilst going to the university, she really wasn't. Her little girl's head was teeming with troubled thoughts as if they were the snakes. Why did she really need it and what are statistics really?

'The word statistics is an ambiguous term' she heard in a quiet whisper just behind her ear. Tosia was going weak at the knees. Her meniscus creaked. That was the protective Good Sprite who was preparing her for the events which were to come after; even without going into his origins he was carrying on with a story. The word statistics derives from the Medieval Latin word *status*, which meant

state. Originally, (at the end of the 18th century) the description of social-political relations within the country was called statistics.

By its name we understand it to be:

- sets of numerical data (for example: *Statistics of National Income* is the publication of the tables which illustrate the development and the structure of the national income);
- all works connected with collecting and working out mass numerical data (for example: the statistical report in enterprise is the section dealing with the preparation of an overall financial statement on the basis of original information obtained from separate production departments);
- science on methods of the research dedicated to the numerically expressed properties of a statistical population.

In our fairly tale we will use the term *statistics* in the sense of the last of these three meanings which the Good Sprite was casting into Tosia's ear.

Our FAIRY TALE? What does it mean?- Tosia couldn't keep from telling it. She plugged her ears and started to run as if she was crazy.

The research dedicated to numerically expressed properties (mass properties), we will be calling statistical research, and the methods of conducting this research as statistical methods. The Good Sprite was delving further into the subject. The need for conducting statistical research arises due to the fact that this kind of mass research creates the only such opportunity to reveal and establish accurateness in the world of empirically observed phenomena. A whole host of examples from various fields of science can be quoted here.

The same idea is expressed by different authors in various ways. In the literature of statistics, there are a few hundred distinct definitions of statistics as the science about how the methods of conducting mass research are met. Some authors think that the name statistical data should include only the research of social-economic phenomena.

The reason why separate statistical methods and statistics itself as the science about them have been created and developed are certain difficulties unavoidably connected with every single statistical research, and so with each research dedicated to numerically expressed properties of the statistical population. Some of these difficulties are of a technical-organizational nature and directly arise due to the mass character of the statistical research. It's not a simple matter to design and prepare such research, to conduct the observation of the units which belong to the researched statistical population, to collect in an appropriate time and form the essential data and then to work them out and present them if the statistical population counts millions or even only thousands of units. The common census, the research of the size of the country's industrial production, the mass control of the quality of goods, the research of the household budgets, these are examples of

the statistical research, by conducting of which, many complex problems of the technical-organizational nature appear. The branch of statistics called statistical technique deals with such problems.

However, solving the technical-organizational problems does not only mean overcoming all the obstacles connected with the statistical research. The aim of this research is to get to know the properties of all the statistical populations, however it is very difficult or even almost impossible to do it on the basis of the collection of the mass individual data. More synthetic, and at the same time, more specialized and adjusted to the aim of this research forms of the description of these properties are needed. The branch of statistics called statistical description deals with the problem of the selection of the appropriate forms of expressing mass properties.

The next group of difficulties, which can be encountered by conducting statistical research are the difficulties connected with the interpretation of the results. Because of the various results, the scope of the observation very often does not coincide with the scope of the statistical population to which the results of the research need to refer. In such circumstances, while making the decision if it is allowed or not to generalize the results of the observation, a certain risk of making the wrong decision is held. By maintaining certain conditions of conducting the observation and specific form of description of its results, the size of the risk can be assessed by means of the calculus of probability. The branch of statistics called statistical inference deals with the issue of the interpretation of the results based on the calculus of probability.

Our fairy tale is dedicated to the methods of statistical description. The separation of the methods of the description from the methods of statistical inference was dictated by both substantive and didactic reasons. The statistical research of economic phenomena very often (more often than the research of the majority of other phenomena) are based on the total observation, i.e. covering all the units of the researched statistical population. In this case, there is no need to generalize the results, so no probabilistic methods of statistical inference should be applied.

However, if statistical research of the economic phenomena based on the partial observation does not cover all the units of the researched statistical population, then by a generalization of the results, knowledge of the methods of statistical inference is essential. In this case, a point of departure for the generalizations must be the description of the statistical population of the selected units, so this fairy tale must be treated as the first part of the lecture on statistics for the people conducting the research based on the partial observation and for those who make use of the results. From a traditional perspective very often, not only two groups of problems and statistical methods connected with the description and statistical population, but also two separate sets of research have been brought up many times. Recently, this stand has not had too many supporters for what is connected with the wider applications of the research, results of which are

generalized. Statistics is sometimes called the knowledge about taking up the decisions in the conditions of uncertainty. It is good to remember that probabilistic methods of inference can only be applied when the condition of random selection of the observed units is kept. However, because of some technical-organizational difficulties as well as sometimes insufficient knowledge of the principles of selection, such conditions are not always presented in the research of social phenomena, the Good Sprite ended.

He was not discouraged by the lack of interest. He got the hump and muttered something under his breath about the book and about publication of which only he himself knew.

In *Statistics Taught through Fiction* the main emphasis was placed on the problem of the selection of the proper forms of description and the explanation of the meaning of particular numerical descriptive characterizations. The science of statistics should not only lie in assimilating certain terms and methods or formulas and also, or maybe first of all, in understanding the conditions of applying them and the cognitive consequences, to which their usage leads. This fairy tale will not cover the systematic lecture on technique and the organization of statistical research.

The whole course of our 'fairy tale' has been worked out, by and large, on the assumption that the reader has the materialistic knowledge in the scope of secondary school and possesses certain skills in handling the algebraic symbols. Only briefly will we cite the elements of differential and integral calculus; however, these will not be the parts conditioning the assimilation of the whole material. Few theorems will be proved, but the evidence in small print can be omitted while reading the fairy tale.

On account of the academic character of *Statistics Taught through Fiction*, we acknowledge that it is unnecessary to cite extremely extensive literature on the subject in every single part.

Bibliographical positions are only to be found exceptionally, where there is really such a need to show the reader the source of supplementary information.

PART I

NOT GOING OUT IS HARMFUL

Introduction: About Tosia, who earlier didn't have time, and now time is chasing her

It's June. Tosia is at the entrance of her university. The end of the semester is near, and she hasn't had the opportunity to attend the classes on descriptive statistics let alone the lectures.

She heads for the maths department – perhaps she will find details about what she has to do to get credit for the course on the notice board.

What joy! She found it. Tosia is the lucky one!

To pass the course on descriptive statistics the following conditions need to be fulfilled:

1. Statistical research and presentation of the following material need to include:

a) at least 30 objects described by two qualitative variables and two quantitative variables,

b) frequency distribution with class intervals and discrete feature,

c) the correlation table for quantitative feature and qualitative feature,

d) pie chart for qualitative feature,

e) two bar charts for quantitative feature; histogram in it.

2. Calculate and interprete the following parameters of random variable distribution:

a) measures of central tendency: the arithmetic mean and dominant,

b) the average position: median, quartiles, deciles,

c) measures of dispersion: range, quarter deviation, average absolute deviation, variation, standard deviation.

3. Calculate and interprete the parameters of two random variables distribution:

a) Pearson's correlation coefficient and Spearson's rank correlation coefficient for the unsegregat ed data,

b) calculate the regression equations for the correlation table (at least 5×5), calculate the correlation coefficient and correlational relationships,

c) verify for the correlation table of qualitative feature (by means of chi - square test) the null hypothesis of no - reliance features on the level p <0,05 and at least two degrees of freedom.

4. The analysis of dynamics:

a) calculate the linear trend (data from at least 10 years),

b) for the presented data calculate as an example one relative increase, chain increase and index.

Poor Tosia. She didn't understand anything.

She needed some pure black magic. Well, magic anyway, and not black but full of colours, and joyful.

But let us not get ahead of the story. The only solution that Tosia could think of was to go to the library and find the books on statistics.

So that's what she did. She ferreted about for some books, put them in her bag, and then went to the nearby forest to look through them.

She looked through the first book and didn't understand anything.

She looked through the next one. It was much worse: the content was more difficult. Horror of horrors! More symbols. She looked again. It was

only the X repeated many times. She didn't have the heart to even open the last book.

Complete darkness in front of her eyes.

She could not carry out the research.

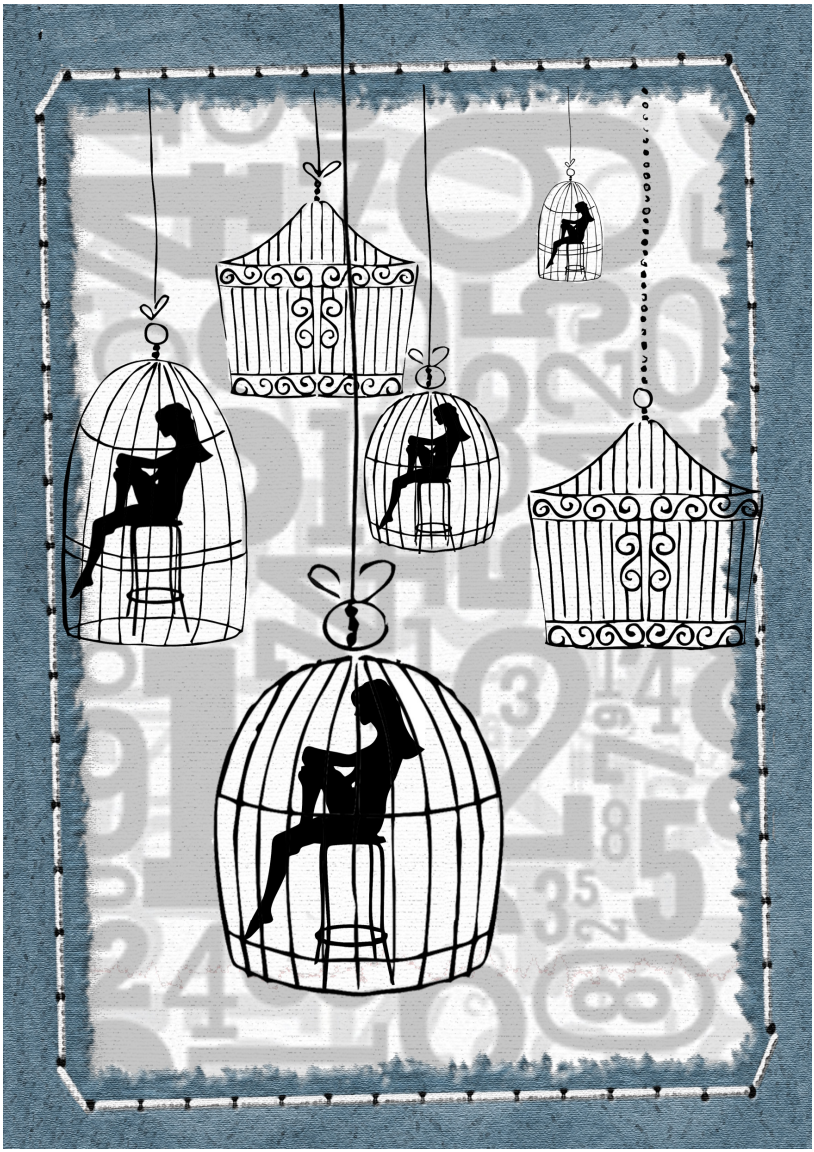
She would not pass the exam!

Good bye, pocket money! Good bye, mobile.

Good bye, Internet; good bye, laptop.

She wailed.

Suddenly the wind blew. The air roared, it started spinning, and…



PART II

INVENTORYING

… a stout gentleman appeared in the middle of the glade. His robes were a little bit worn out but he looked dignified nevertheless. And he said, “Don’t cry, Tosia. I have come to help you. I am the King of Statistics. I will tell you about the Status Kingdom and my wife, the Queen of Statistics. But first, I will tell you about my three daughters: Mean, Median, and Dominant. It all started with them.”


I ♥
Statistics

####

Happiness flourished throughout the Status Kingdom but only up to the time the oldest daughter, Dominant, wanted to get married. Although Median tried to explain to her that it was not necessary to get married immediately, no one listened to Median. It's a pity; maybe some issues would have surfaced earlier and saved a lot of grief.

It was broadcast throughout the kingdom that the winner of the soon-to-be competition would win the hand of the fair Dominant.

Her character was well known so only one candidate entered the competition – Count Pearson's Correlation Coefficient from the Pearson Moment family (the one often mistaken for Spearman's Correlation Coefficient from the Rang Family – he didn't take part in the competition).

The Count was wily. He'd nosed out a good deal. If he was the only one who'd entered the competition, that meant he would win. He quoted a wide range of fairytales in which the following was true: the king, or father-in-law, who wanted his daughter to get married, contributed to her future husband half of his wealth. The Count, with no embarrassment whatsoever, commented, "Next month, I will take Dominant and half of your fortune. It must be ready at this time! You will hold a lavish engagement ceremony, with an orchestra!"

And that was the last they saw of him for quite a while. The King of Statistics was troubled, but for all sorrows the best remedy is simply a piece of good advice so he thought of his wife**.** That is

The Queen of Statistics, who could find regularities in mass phenomena

The King of Statistics anticipated there would be many problems. It's good that he couldn't hear what two Catty People were saying!

He shambled over to see the Queen of Statistics who was in the company of the queen. The King of Statistics hadn't miscalculated the

wisdom of his wife. Immediately, or even much quicker, she understood what to do.

"Do you know, my darling husband, how much half and half makes?" she said.

"Two halves," he answered brilliantly.

He had been good at arithmetic right from his childhood.

"You joker," the queen replied, choking with laughter.

However, she wasn't laughing at the king's joke but because his intellectual level was below the 30th centile. But how could the king know what the 30th centile means? You, my dear reader, know it whether you like it or not.

Statistics said seriously, "Two halves are a whole. You can find out how much it is if you make the Chancellor and the Bursar carry out.

Borowiński said that when the Queen of Statistics was a girl of marriageable age, she wore a bikini that showed quite a lot but hid what was the most important.

And Szulc said that the Queen of Statistics could even open the mouths of numbers.

Statistical Research

No sooner had Statistics finished than the queen threw him off the alcove. She didn't want him to go the whole way, or even two halves of it!

Still Not the Nightmare 2

No one knew, even the King of Statistics, what the Good Sprite was doing at that time. He was musing on the essential information about the basic terms in statistics. He knew them very well but how to hand them over and not scare Tosia? It was a bit of everything.

The basic term in statistics is statistical population that is also called population or mass. Statistical population consists of the individual units that are also called elements of population. If a population is divided into parts, then we call them a partial population, sub-population or, to put it simply, groups. The term population size indicates the number of units in the population or sub-population.

Not every single set of elements is the sensible statistical population. It would not be logical to undertake research of the properties of the whole population if it consisted of accidentally chosen elements that had nothing in common. It would be pointless to research the properties of the whole population if it consisted of the totally identical units, as examining the single element would be enough to find the knowledge of the whole.

However, the nature of the units of a statistical population can be various such as: stars and atoms, the grains of wheat and trees, animals and people, birth and death, deeds of purchase and sales, schools and prisons, children's toys and cars, loaves of bread and bottles of wine, or even literary terms and musical chords. This universality of the term statistical population and the terms and methods connected with it initially cause certain difficulties to Tosia, because it creates the necessity to translate popular or specialized expressions connected with the specific mass phenomena into the universal language of statistics and vice versa. It is important to remember that the very nature of statistics lies in this universality and only because of it the common methods of research of many different mass phenomena can be assimilated instead of rediscovering the methods every time. Tosia will manage with it.

Another basic term in statistics is the feature that means property that allows us to differentiate the units of statistical population. This feature can be qualitative and quantitative. It is qualitative if its varieties are expressed by means of a verbal description, or quantitative if its varieties are expressed in figures. The quantitative feature is, from the mathematical point of view, variable that absorbs various values.

In statistics, we can distinguish between quantitative continuous features and discrete features. This division is of substantial importance from the point of view of the methods that can be applied to statistical research. The feature is called continuous in a certain interval if it can assume the indirect value between every pair of freely chosen values in this interval. That is why a continuous variable accepts infinitely many different varieties, whereas a discrete variable only accepts some (at the same time, in statistical-economical research, as a rule, such discrete features can be met which assume only the integral values, for example: the number of children in a family).

Tosia's attention should be turned to the fact that the division of variables into continuous and discrete, so clearly stated from the theoretical point of view, causes some difficulties in statistical practice. The observation can only provide the completed number of the values, which is why the empirical data about a certain variable can never be perfectly continuous. For example, the age of a man is a continuous feature in the interval from 0 to, let us say, 100 years old, which means that it can endlessly accept many various values in this interval. In this specific empirical research of age, even if we would like to put the whole population of the country or even the world through an observation, we only get the completed number of data, which, in fact, will not create a freely 'dense' set of the age values. Another cause of practical discreteness of many theoretically continuous variables is the imperfection of the measuring devices. The height of a man can be established in practice to an accuracy of 1cm, so because of this reason, we will not get, as the result of observation, indirect values. On the other hand, such features, which by their nature are discrete, can be quantitative, however, they take on so many various values that in statistical practice they are treated as if they were continuous. For example, earnings take on the values of pennies, but in the interval from 1,000 to 2,000, the number of varieties they take on is 100,001 and that is why the salaries are, as a rule, treated as a continuous feature. This incompatibility of theory and practice creates certain difficulties by using some of the methods of description and conclusions, which are adjusted to a continuous variable.

The aim of this statistical research is, as we know, the numerical description of the population's properties. By using the introduced terms, we can specify this aim a little. When talking about researching the properties of the particular population, we, as a rule, mean carrying out one of the following tasks:

1. Determining how particular varieties of a certain feature (or certain features) are spread around the units of this population that means getting to know the distribution of the chosen feature in a given statistical population;
2. Getting to know the development (dynamics) of the given statistical population, which means establishing what changes in time the chosen sizes, which describe the population, are subject to.

Not an easy task waiting for the Good Sprite.

Chancellor and Bursar were summoned to appear before the king. They tumbled down to the floor and shouted in chorus, “Our dear king! We do not steal anymore!”

“I know, I know, that’s not why I called you. Statistics said you need to carry out some research. And now, out of my sight!”

Not to waste any time, after a short pow-wow, they decided to head over to Doctor Methodologist. When they explained the reason for their visit, poor Doctor Methodologist almost suffocated. He laughed so hard, his belly button started to come undone. Doctor Methodologist hadn’t formed a good opinion of these two interlocutors. “Ignoramuses, it’s not a matter of medical investigation, it’s a matter of

Statistical Research

Follow me, I will explain it to you,” he said, still choking with laughter. They went with him to the cellar, where Doctor Methodologist kept his treasures, such as:

The Stages of Statistical Research

1. Planning
2. Observation of particular units
3. Drawing up the material collected
4. Analysis of the results

“And, now, ignoramuses,” he said, “I undertake the quest, which is beyond doubt to briefly clarify what this is all about. The more you understand, the less you will have to work on it. First, go to Wizard Trend and beg him to allow his daughter, Fortuneteller Extrapolation, to transfer you to the year 2012. You will find out what you are to observe at the uni.”

Still Not the Nightmare 3

The Good Sprite was snoozing under the ceiling as if he was the bat. He felt exhilarated only after he heard the word ‘observe’. He knew that this term could bring many dangerous situations. He would have a lot to say to Tosia. In his mind, he was arranging the content of the performance.

‘Just imagine, Tosia, that the statistical population to which all conclusions of statistical research may refer to, we will call the general- or entire population. The entire population relies on collecting the individual information about all the units

included in the general population. Partial observation relies on collecting individual information about only some of the units chosen from the general population in order to deduce the conclusions concerning the general population. We will call the population of the chosen units the statistical sample population or simply, the sample.

There are many premises which induce us to conduct partial observation. It would be absurd to, for example, apply the entire observation in cases when the process of the observation itself is connected with destroying or decreasing the utility of the observed articles. It applies to much research concerning the control of the quality of the products (e.g. research of the period of when potatoes rot; research of the resistance of the glass demijohn etc.). The less labor intensive, the smaller the costs and also the shorter the deadlines of conducting a comparison with the entire observation–it all belongs to the obvious advantages of partial observation. It is not only a matter of shortening the observation process itself, but also the later process of working out the materials collected into a considerably smaller amount. These advantages, in most cases, decide simply about the usefulness of the research.

Let us take as an example the research of a budget of a household, the aim of which is to familiarize us with the factors shaping the structure of the income and expenses of people. If the observation of income coming from various sources and expenses for a particular group of goods and services were conducted with reference to all households in the country, it would accrue enormous costs connected with preparing millions of budgetary books, introductory and current instructions for all families, collecting, transporting and working out the gathered information etc. One should count on a long-standing delay in getting a description of the population of all households in Poland. It is simply clear that even if the entire observation is in this case theoretically possible (we would omit a series of additional obstacles such as, e.g. possible refusals to providing information), from a practical point of view, it could not be accepted: too costly and too long a time in getting the results would blight all the advantages. Therefore, research of the budgets of households are conducted in Poland and other countries on the basis of partial observation; at the same time, the population size of the statistical sample population rarely exceeds a few thousand, and sometimes it amounts to just several hundred families.

When conducting partial observation, we need to take into account the likelihood of making certain errors in the assessment of the general population, which, however, do not disqualify the nature of this kind of observation. It is worth remembering that through the entire observation we will not obtain a 100% accurate picture of the general population. It is difficult to avoid false information, or even omit certain units (we meet with this phenomenon also by analyzing the census' in the most civilized countries). What is more, mass observation sometimes leads to errors, which can be omitted within the smaller scope. So, for example, the list of plenipotentiaries can be professionally better prepared if their

number is smaller; control of the gathered material can be conducted more carefully if it is less extensive; the process of the observation, e.g. control of the quality of the goods, is less tiresome if it lasts for a shorter period of time etc. Moreover, absolute accuracy of the results is by no means needed in statistical research. The degree of the desired accuracy of the results depends, of course, on the subject and the aim of the research. For example, for the needs of economic practice, the knowledge of the average monthly income of Doctors to an accuracy of one penny is not necessary; accuracy of more or less £25 is enough.

Because of this, the efforts of statisticians, when conducting partial observation, go in two directions: in order to be able to establish the amount of possible errors in the assessment of the general population on the basis of an observation of the statistical sample population and to cut the size of possible errors to such limits which will make the assessment useful to the user. Both these problems are directly connected with the matter of choosing the way of selection of the units to the sample.

In principle, we distinguish two ways of selecting the units of the general population from the statistical sample population: the intentional method and the random method. Instantly, the second choice creates the impression of being unreasonable because it links the composition of the sample with the uncontrollable human being factor of chance whereas the intentional method relies on the conscious selection of the units to the sample; it is seemingly more rational. We expect that, in this way, we can get the sample quicker with the proportions in accordance with the proportions of the general population in terms of the researched properties, enabling us to get the representative sample.

However, this kind of conviction relies on the silent assumption, or at least good knowledge of the composition of the general population in terms of the researched properties. If it was like this the ideally representative sample would mean a perfect miniature of the whole could be selected (the one including the units of all variants of the researched features and in the proportions that they could be met in the entire population). In such a case no partial observation would be needed. We would like to conduct it because in terms of the research we do not have sufficient insight with regards to the composition of the general population. This insight cannot be the adequate basis for the intentional selection of the units into the sample.

However, sometimes we know the decomposition in the entire population of the features that remains or will probably remain in the essential relationship with the researched properties. This kind of information can be helpful for selection: we try to choose the units in such a way so that the proportions of the sample are in accordance with the proportions of the general population in terms of the features connected with the research. This method is called the proportional selection method and is often applied in sociological research.