Przegląd Statystyczny. Statistical Review, 2020, vol. 67, 3, 195–211 DOI: 10.5604/01.3001.0014.7109

Philosophical foundations of statistical research

Józef Pociecha[a]

Abstract. Every researcher desires to uncover the truth about the object of the undertaken study. When conducting statistical research, however, scientists frequently give no deeper thought as to their motivation underlying the choice of the particular purpose and scope of the study, or the choice of analytical tools. The aim of this paper is to provide a reflection on the philosophical foundations of statistical research. The three basic understandings of the term 'statistics' are outlined, followed by a synthetic overview of the understanding of the concept of truth in the key branches of philosophy, with particular attention devoted to the understanding of truth in probabilistic terms. Subsequently, a short discussion is presented on the philosophical bases of statistics, touching upon such topics as determinism and indeterminism, chance and chaos, deductive and inductive reasoning, randomness and uncertainty, and the impact of the information revolution on the development of statistical methods, especially in the context of socio-economic research. The article concludes with the formulation of key questions regarding the future development of statistics.

Keywords: philosophy of science, philosophy of truth, theory of probability, statistical learning, socio-economic investigations

JEL: C00, C10, C60

1. Introduction

The term 'statistics' relates to a variety of concepts, with at least three separate dimensions. Firstly, statistics can be understood as a part of mathematics, commonly referred to as 'mathematical statistics'. This is a branch of mathematics based primarily on the theory of probability, but also involving other areas of mathematics, such as algebra or calculus. The substantial use of inductive reasoning is a specific feature of mathematical statistics, which stands in contrast to deductive reasoning, widely applied in other branches of mathematics.

Statistics as 'the science about the condition of a state' is another interpretation of the concept in question, and in fact the one closest to the original understanding of the term (Pociecha, 2016). 'Statistics' traditionally referred to a set of data presented in a table form, describing the condition of a given state. Nowadays, when we mean such a 'quantitative description of the state of things in a state' we use the term 'public statistics'. Contemporary public statistics is a system of gathering statistical data, involving the collection, storage, processing and publishing of such data, as well as making accessible or distributing the results of statistical research. Public statistics is an essential element of the information system of a democratic society, providing state authorities, national and local administration bodies, the economic sector, and society at large with official statistical data on the economic, demographic, social and environmental situation in a given country (Oleński, 2006).

[a] Cracow University of Economics, Department of Statistics, ul. Rakowicka 27, 31-510 Kraków, Poland, e-mail: pociecha@uek.krakow.pl, ORCID: https://orcid.org/0000-0003-3140-481X.

Statistics may also be perceived as the science of quantitative methods of studying mass processes. This most general and common conception of statistics includes the application of mathematical statistics and other quantitative methods as the methodological basis for research. Moreover, it views public statistics as the basic source of statistical data in social and economic research. At its core, this understanding of statistics considers it the science of uncovering the truth of the surrounding reality based on data which describe this reality. It is an empirical science – the one which makes use of inductive methods. In consequence, the determination of the degree to which an analysis or forecast corresponds to the studied reality (i.e. the degree of veracity of the results of statistical research) is of key importance here. It is crucial then in statistical analysis to be always aware that by using statistical methods we attempt to uncover the truth about the reality under study. In light of the above, the perception of truth in scientific research must be considered.

2. Philosophical notion of truth

Truth is a philosophical notion. It is the main object of the study of epistemology (the study of the nature of knowledge) and one of the branches of philosophy which examines the relationship between knowledge and reality. Scientific knowledge is a specific type of knowledge which requires the fulfilment of a selection of defined rules. The nature of this knowledge is the subject of the philosophy of science (Woleński, 2014). The aim of science, as Strawiński (2011) stresses, is the explanation of the world formulated in scientific theories. Other, concurrent functions of science, such as education, innovation, or 'emancipation', are the derivatives of the cognitive function.

One of the primary concerns of epistemology is the search for an adequate definition of truth and the criteria for suppositions to be true. This problem is addressed by the theory of truth, a subset of epistemology. The key issues that this science investigates is whether truth exists, whether it can be determined, or in what way we can come to know the truth. A variety of answers have been formulated to the abovementioned questions by a number of schools of philosophy, jointly with a wide range of definitions of truth. Below is a concise overview of the most commonly accepted definitions of truth.

2.1. The classical (correspondence) definition of truth

Truth is a supposition which is in accordance with the state of things this supposition concerns (Tatarkiewicz, 1978a). This definition was formulated by Aristotle (4th century BCE) as: 'To say of what is that it is not, or of what is not that it is, is false, while

to say of what is that it is, and of what is not that it is not, is true'. This statement was written in 'Metaphysics', Aristotle's most important work on philosophy, handed down through the ages via the publication by Andronicus of Rhodes (1st century BCE). This classical definition of truth is known mainly in the form worded by Thomas Aquinas (13th century AD) 'Veritas est adaequatio intellectus et rei' ('The truth consists of an adequation between the intellect and a thing'). Truth exists, therefore, if what is in our minds corresponds to reality. This classical definition of truth has been subject to critique of many types (Woleński, 2014), including the absence of a set of universal criteria for the 'adequation' and the problem of 'the replication of reality via language' (Pruś, 2018). These shortcomings in the classical definition of truth were rectified by a Polish logician, Alfred Tarski, who formulated a semantic definition of truth (Tarski, 1995).

Despite many attempts to devalue the Aristotelean conception of truth, his definition, formed within the classical Greek philosophy, still presents a challenge to empirical scientific study, as everyone wishes to discover the surrounding reality. The consequence of following the Aristotelean definition of truth is accepting that the world (reality) exists in an objective sense (outside our minds) and that it is knowable. Of course, knowing reality is difficult, but possible, making scientific inquiry relevant. It also means that it is possible and appropriate to conduct scientific research using statistical methods.

2.2. Neopositivist notions of truth

Among many well-known trends of the 19th century philosophy, a significant influence was exerted by the views of the neopositivist school, also known as the Vienna Circle, representing logical positivism. The third wave of the positivistic ideas, called also the 'Third Positivism' (Tatarkiewicz, 1978b), expressed theses of a minimalistic philosophy combining three complementary theories: empiricism, positivism and physicalism. Empiricism presupposes the establishment of all knowledge on the basis of empirical data and the rejection of anything that does not find support in empirical facts, while accepting that experience is the source of all knowledge in the real world. Positivism, on the other hand, assumes that the objects of study may only be facts. It rejects metaphysics and states that only scientific knowledge is certain. Still, neopositivists differentiated between knowledge of the real world which is of an empirical nature, and formal knowledge which is of an axiomatic nature, such as logic and mathematics. This trend was called logical positivism (logical empiricism). Science establishes facts in the form of theorems or formulates tautologies concerning a coherent system of logic and mathematics. Neopositivists thought that the

logical-mathematical language is the only language of science. They believed in physicalism, meaning the reduction of all sciences to the expression of physics, or at least the application of research techniques and mathematical descriptions drawn from physics in all branches of science, including social sciences, such as psychology or economics.

Two members of the Vienna Circle, Rudolf Carnap and Otto Neurath, were supporters of the coherence-based definition of truth, according to which 'that which is true is internally coherent', and is 'true on the basis of certain statements of experience', in other words based on acquired experience. This means that truth is determined by the absence of a logical contradiction between these statements (Tatarkiewicz, 1978b). This view is the basis for the logical positivism which was propagated by neopositivists.

The legacy of neopositivism in social and economic sciences includes the appreciation of statistical data as a source of knowledge about the real world, and of quantitative studies based on mathematical modelling of social and economic processes. The recognition of the primacy of physics over other branches of science has become the basis for the methodology of contemporary 'econophysics', a tool used to describe these processes.

2.3. Popper's critical rationalism

Karl Raimund Popper was an active member of the Vienna Circle of neopositivists, but neverthless he expressed theses which were not always in line with the group's views. Popper's discussions held during seminars of the Vienna Circle led him to writing a book entitled 'The Logic of Scientific Discovery', which encapsulated his main philosophical views (Popper, 1977), shaped both under the influence of, and in contrast to, logical positivism.

Popper called his philosophy 'critical rationalism'. It assumed that everything which was proved at a certain moment might at some other point become doubtful. The key points in Popper's views were as follows:

- the perception of thought as the act of solving problems using a deductive method, in which the mind constructs notions, hypotheses and theories which subsequently become subject to falsification;
- the rejection of all *a priori* notions and primary elements of cognition, assuming that the work of the mind is of a temporary nature and does not uncover any unchangeable laws;
- the aim of science is the creation of new, increasingly bold theories which describe an increasingly broad class of phenomena.

Popper introduced the notion of falsificationism, which involves a set of methodological procedures a researcher must apply if his/her goal is to advance scientific knowledge. The approach does not advocate searching for the confirmation of scientific theories (verification), but rather investigating contrary cases that could prove the falsehood of the studied theory. A scientist's aim, then, sholud be the attempt to falsify a theory (to demonstrate that it does not correspond with experience), and if this attempt is unsuccessful, the theory should be temporarily accepted until one of the subsequent attempts at falsification results in the refutation of the theory.

For Popper, the truth needs to correspond to facts, and only such an understanding makes rational critique possible. He proposes treating the notion of 'truth' as a synonym of the notion 'corresponds with the facts'. The truth is a mountain peak enveloped in clouds. An experienced alpinist may not have difficulty reaching it, but may not know when the peak is reached as it may be indistinguishable among other nearby peaks, obscured by clouds. This fact, however, does not in any objective way affect the existence of the peak; an authentic idea of error or doubt entails the idea of an absolute objective truth which can never be attained (Popper, 1977). Popper's critical rationalism constitutes the basis for the testing of statistical hypotheses which aim to refute the null hypothesis. According to Popper's philosophy, we are never able to verify the null hypothesis, but only to refute it. The absence of a basis for the rejection of the null hypothesis is synonymous with its temporary acceptance, until it is falsified.

2.4. Thomas Kuhn's philosophy of science

An American physicist, historian, and philosopher of science Thomas Samuel Kuhn was the creator of the idea of the scientific paradigm. His most important work in the area of the philosophy of science is 'The Structure of Scientific Revolutions' (Kuhn, 1962). In it, Kuhn provides a critique of Popper's falsification, casting doubt on its assumptions. First and foremost, he draws attention to the fact that research hypotheses are verified in the context of a set of generally accepted scientific knowledge, and that this set of knowledge itself is not subject to verification. Kuhn strongly rejects Popper's thesis that in science we learn from our own mistakes and replace erroneous theories with better ones. The aim of science according to Popper is to constantly substantiate its theses, which is a sign of scientific progress. According to Kuhn, however, the problem with the validity of scientific theories being the sign of scientific progress is the conformity of thought within the scientific community, which leads to the rapid development of research in a given field.

Kuhn's primary achievement as a philosopher of science was the introduction of a paradigm as a set of notions and theories constituting the basis of a given science. These notions and theories are seldom questioned as long as the paradigm is cognitively creative, meaning that by making use of it, detailed theories may be formed in accordance with experimental (historical) data assigned to the given branch of science. The most general paradigm is the paradigm of the scientific method, the criterion for recognition of a given scientific operation.

Kuhn argued that science is not a monolithic, cumulative process of acquiring knowledge. Instead, he believed that science is a series of periods of calm interspersed with sudden intellectual revolutions, leading to the replacement of one conceptual worldview with another.

He also analyses the relationship between the philosophy of science and its history. He considers two approaches towards the philosophy of science: one represented by the formal methodology of science, and the other by a historically-oriented theory of science. Kuhn is a representative of cognitive scepticism, holding the belief that although scientific inquiry leads to the discovery of fundamental truths, the shifting of paradigms does not necessarily bring scientists closer to the attainment of truth.

In science, particularly in social science, various paradigms may exist simultaneously, for example the paradigms of classical economics and Keynesian economics. In statistics, the paradigms of mathematical statistics, Bayesian statistics and statistical learning are currently the prevailing ones.

Having provided this brief overview of philosophical notions, the author wishes to emphasise that it is important to be aware of how we wish to discover an understanding of the truth of the studied reality in the course of conducting all statistical research.

3. Determinism and indeterminism

Statistical research is by nature empirical, relating to the surrounding reality. Thus, it is crucial to take a philosophical stand regarding the nature of the relationships between things, properties, quantities and events which constitute this reality. In philosophy, two main opposing views have been formed regarding the functioning of the world: determinism and indeterminism.

Determinism is a philosophical conception which assumes that all events are related through the notion of cause and effect: every event and every state is determined by previously existing causes, consisting of other events and states. Everything which occurs in the world, including human actions, is conditioned in advance, outlined, defined and must take place within a cause and effect series of events.

Determinism in the history and prehistory of human thought is the primaeval philosophical standpoint. For ages, the prevailing belief was that all events in nature were predetermined. In ancient times, the most well-known proponent of deterministic causation was Democritus (Tatarkiewicz, 1978a). In the history of philosophy, the most extreme deterministic view is represented by Laplace's 'mathematical daemon', a spirit with an unlimited capacity for mathematical deduction, who would be able to predict all future events if only it knew all the quantities which characterised the present state (van Strien, 2014).

Determinism proclaims unconditional faith in the power and omnipotence of formal logic, which is a tool enabling the discovery and description of the world. It rejects the idea of chance as an objective phenomenon, claiming that the impression of randomness is strictly a subjective state resulting from insufficient information.

Indeterminism, on the other hand, is a philosophical conception assuming that the relationship between cause and effect in nature is not absolute, it presupposes the existence of chance, and rejects the possibility of predicting subsequent events based on previous ones, as the same causes need not necessarily lead to the same effects. In its extreme form, indeterminism totally rejects the existence of (or the possibility of knowing) any conditions. Indeterminism also exists in moderate forms which accept the presence of objective regularities (laws), but only in certain fields and conditions of reality. Indeterminism has become the contemporary scientific viewpoint to the extent to which determinism was the original philosophical standpoint. Indeterminism started as a result of 200 years of discoveries in physics, when it became apparent that in the world of atoms and quants there is no place for determinism, and that regularities occur only in mass events.

Indeterminism negates determinism. The conflict between indeterminism and determinism, ongoing in philosophical debate since ancient times, relates in particular to the issue of the free will, man's responsibility, the aim of nature, causation in nature, necessity and chance.

Statistical research is based on an indeterministic understanding of the world. Statistical regularities are of a stochastic nature; they appear in mass phenomena, and individual cases may differ from general regularities.

4. Chance and chaos

As pointed out in the previous section, determinism was the original philosophical standpoint on nature, which assumed the existence of an eternal order, and the aim of science was its discovery. Yet, philosophers ever since Aristotle's times have

recognised the role of chance, considering it as something which violated that eternal order, was beyond human understanding, and thus was impenetrable by science. In the mid-19th century, philosophers realised that the search for the deterministic laws of nature is hampered by logical and practical difficulties, and subsequently they initiated research into models of the laws of nature based on the mechanisms of chance. The key inspiration in the search for such models were Adolph Quetelet's achievements in the field of statistics, involving the application of the notion of probability for the description of social and biological phenomena, including posing 'Quetelet's question' (Ostasiewicz, 2012). Additionally, the formulation of the laws of inheritance by Gregor Mendel, which laid foundation for the science of genetics, stimulated progress in seeking these new models. In terms of physical sciences, inspiration was provided by the statistical interpretation of the fundamental theorem of theoretical physics, namely the second law of thermodynamics, as formulated by Ludwig Boltzmann. These achievements, as well as many others, brought forth a revolution in the understanding of nature. Over time, the roles of order and chance in science were reversed: chance became the primary notion.

A conventional way of thinking would suggest that chance causes chaos, which further leads to the question about the relationship between the two notions. The word 'chance' is used to describe random phenomena, such as drawing a number in a lottery in which the numbers are in a random sequence. A sufficiently long series of random occurrences reveals a certain order which can be discovered using calculations of probability. On the other hand, numbers generated in a deterministic process may express random behaviours which we call chaos. It was from this ground that the theory of deterministic chaos arose, dealing with the irregular, unordered behaviour of deterministic systems which are practically unpredictable over lengthy periods of time (Schuster, 1995). Deterministic chaos, as per the chaos theory, is the property of an equation solution being highly sensitive to any minor disturbances of its parameters. This usually concerns nonlinear finite and non-finite differential equations describing dynamical systems.

Radhakrishna Rao (1994) defines the notions of chance and chaos as follows: 'Chance deals with order in disorder while chaos deals with disorder in order'. Both chance and chaos may be observed and modelled. Chance is modelled using the tools of probability and mathematical statistics. Chaos, which is of a mathematical nature, is described using deterministic models.

Without evaluating the usefulness of the analytical tools for social and economic research provided by the theory of deterministic chaos, it should be noted that contemporary statistical research is based on an indeterministic understanding of the

nature of the relationships between the components of the world which surrounds us, modelled by methods of probability and mathematical statistics, and that chance is an inevitable element of this reality.

5. Deductive and inductive reasoning

The assumption of a deterministic or indeterministic approach to understanding the world is linked with the way in which we reason about it. Reasoning is the method by which we come to accept a previously unaccepted position on the basis of previously accepted positions (Ajdukiewicz, 2006). In philosophy and logic, two basic and opposing types of reasoning are recognised: deductive and inductive.

Deduction is a type of logical thinking which aims to arrive at a defined conclusion based on a previously established set of premises. Deductive reasoning is entirely self-contained within its assumptions, meaning that it does not require the creation of new theorems or notions, but is simply a process of drawing conclusions. If it is carried out correctly, meaning that the set of premises does not include false statements, then the conclusions drawn as a result of deductive reasoning are irrefutably true and cannot be validly questioned.

Using deductive reasoning, no new knowledge going beyond the premises is created, as all theses generated are contained within axioms. It is not claimed that either the axioms or the theses generated from them are linked to reality. Logic and mathematics make use of deductive reasoning, including such branches of mathematics as probability or mathematical statistics, e.g. 'Let (Ω, σ, P) be a probabilistic space...', Kolmogorov's axiomatic probability definition, or the definition of a stochastic process.

Deductive reasoning was introduced in the earliest stages of the development of the Greek philosophy, with the greatest role in its development played by Parmenides (Tatarkiewicz, 1978a).

Inductive reasoning involves starting from detail and arriving at the general idea, i.e. the correctness of the statement (conclusion) stems from the correctness of its consequences (premises). In inductive reasoning, we decide on the premises when we have certain data on their consequences. Using this type of reasoning, decisions in the real world are made based on incomplete or faulty information. Inductive reasoning is a logical process by which a hypothesis is selected to fit the data and generalisation is made from an individual case. In consequence, new knowledge is created, however it is uncertain due to the lack of bilateral, unambiguous conformity between the data and the hypothesis. For the human mind, accustomed to deductive

logic, this lack of precision in reasoning based on existing data, in contrast to reasoning based on axioms, has resulted in a reluctance towards the application of the rules of inductive reasoning. Knowledge gained by generalisation from details, although initially uncertain, becomes certain if its associated uncertainty is expressed quantitatively.

Induction in mathematics can be complete or incomplete. Complete induction is reasoning about a general regularity based on statements covering all possible cases of the occurrence of this regularity. Complete induction is a method of proving statements about natural numbers. Incomplete induction involves reasoning about a general regularity based on a finite number of statements which cover some occurrences of this regularity. It is the basic tool for discovering the truth in experimental science. A problems arises, however, when selecting a criterion for the purpose of distinguishing between valuable and worthless results of research obtained through incomplete induction. In many cases, one has to simply use common sense to do that.

Statistics is based on the principle of inductive reasoning in its incomplete form. This branch of science discovers the surrounding reality on the basis of a finite number of statistical data to which the theory of probability is applied.

6. Randomness and uncertainty

Randomness, denoting the absence of purpose, cause, or predictable behaviour, is inextricably linked with indeterminism. Randomness is understood as a random process whose results cannot be exactly predicted, but can be presented as a distribution instead. A tool for describing random processes is provided by probability, formally defining a random event as follows: 'Let (Ω, σ, P) be a probabilistic space...'. Intuitively, a random event is the one whose outcome cannot be predicted with certainty. Another notion is the random variable, defined as a function which reflects a probabilistic space in the world of numbers. Subsequently, the distribution of the random variable and its endless analytical forms become defined. This way, the idea of randomness of a philosophical nature was formalised in the basic mathematical notion of probability and mathematical statistics, as shown by several authors, including Richard von Mises (1957).

Randomness occurs in nature, thus it is necessary for the laws of nature to be expressed in probabilistic categories. This constitutes the basis not only for contemporary physics and biology, but also for psychology and social sciences. Random behaviours are also considered as an inherent aspect of the functioning of many classes of objects and their manner of existence. Radhakrishna Rao (1994) poses the

following question: 'Does randomness play any role in the development of new ideas and can creative capacity be explained using random processes?' Creative capacity is understood as the source of a new idea or theory which does not align with or cannot be drawn from the existing paradigm, and which explains a broader set of phenomena than any other existing theory. A good example of creative capacity could be Albert Einstein's creation of his theory of relativity. Writer Arthur Koestler, describing the act of creation, said: 'At the decisive stage of discovery, the codes of disciplined reasoning are suspended, as they are in the dream, the reverie, the manic flight of thought, when the stream of ideation is free to drift by its own emotional gravity, as it were in an apparently 'lawless' fashion'. Writer Douglas Hofstadter notes: 'It is a commonly held view that randomness is an indispensable element of the creative arts'. R. Rao claims that random thinking is an important component of creative ability. Thus, the role of randomness is in fact considerably broader than it can appear at first. Carl Gauss once said: 'I've had my results for a long time, but I do not yet know how I am to arrive at them'.

A further philosophical notion which has its own mathematical expression is uncertainty. Within the theory of decision making (Szapiro, 1993), it refers to a situation where defined decisions may cause various effects depending on which of the sets of possible states occurs, with the caveat that the probability of the occurrence of individual states is not known. Formalised principles for decision making are outlined in the theory of mathematical programming (Trzaskalik, 1990).

Uncertainty is an integral part of nature and society. It manifests itself in the behaviour of elementary particles in physics, of genes and chromosomes in biology and of individuals in society, when acting in situations of stress and tension. This makes it necessary to develop theories based on stochastic laws within the natural and social sciences which utilise the notion of random events. The feeling of uncertainty is heightened by such factors as the lack of information, an unknown degree of inaccuracy in the available information, the absence of technical possibilities to obtain the necessary information, and the inability to conduct relevant measurements.

When aiming to reduce uncertainty, it is crucial to express it in quantitative terms. The first attempts at the quantitative expression of uncertainty were made by Thomas Bayes (in the 18th century), who introduced the notion of an *a priori* distribution of a set of possible hypotheses indicating the degree of confidence in them before observing the data. If this distribution is given, then together with the knowledge of the distribution of probability resulting from the data, in the conditions of a given hypothesis, total probability is obtained. The conditional distribution

of probability in a specific hypothesis is calculated this way. The Bayes Theorem is an example of the application of the probability theory as a tool in inductive reasoning. It constitutes one of the foundations of Bayesian statistics (Osiewalski, 1991) as a means to taming uncertainty, in the situation where classical statistical methods fail to do so. This theory makes use of the notion of subjective probability. In classical mathematical statistics, probability is an objective quantity understood as a family of measurements serving to describe the certainty of a random event. Subjective probability is defined by subjective opinions of individuals based on the available data. Bayesian statistics combines a deductive mathematical approach with an inductive empirical approach to statistical research, including decision-making under uncertainty.

7. The probabilistic notion of truth

A crucial question thus arises: Can we discover truth using statistical methods? Can we, on this basis, define truth? Statisticians believe we can. In statistical terms, truth is a belief with an acceptable level of probability (error) which corresponds to reality. Reaching the truth in statistical terms means making a point or interval estimation. Radhakrishna Rao (1994) ends his book by citing an ancient piece of Eastern wisdom:

The road to wisdom?
Well it is plain and simple to express,
Err,
And err
And err again,
But LESS,
And LESS,
And LESS.

A probabilistic understanding of truth on a philosophical basis is represented by probabilism. This is a contemporary variant of scepticism which assumes that our knowledge is only probable knowledge. It is not possible to demonstrate what truth is and what falsehood is, but only to recognise those theorems or suppositions with a high degree of probability. The roots of probabilism date back to the views of the ancient sceptics, who claimed that in practical life it is not necessary to have certainty – a reasonable level of probability is sufficient. The main representative of such ancient probabilism was Carneades of Cyrene (Tatarkiewicz, 1978a).

The ideas of probabilism were reborn on the foundation of neopositivism. Despite the fact that scientific theories, in neopositivists' view, are unprovable, they can be assigned various degrees of probability, estimated using statistical methods, based on the available empirical material. Major advances in this field were made by Rudolf Carnap and Imre Lakatos. Probabilism proposes the replacement of the idea of verification through facts with the idea of probability based on induction. Carnap was the co-founder of the Vienna Circle and a proponent of a radical version of neopositivism (Tatarkiewicz, 1978b), while simultaneously conducting research into the foundations of mathematics and issues of probability (Carnap, 1950). The approach of Imre Lakatos to the philosophy of science was an attempt to reach a compromise between the falsificationism of Popper and the theory of scientific revolutions expressed by Kuhn. Popper's theory requires scientists to abandon their theories the moment they encounter an observation which falsifies them, and form 'bold hypotheses' in their place. On the other hand, Kuhn believed science is a series of interwoven eras of 'normal science', during which academics hold to their pet theories in spite of accumulating observations contradicting them, and 'scientific revolutions' which bring about changes in the ways of thinking; in Kuhn's opinion, nevertheless, the causes of these revolutions are often devoid of a rational basis. Lakatos (1995) searched for a methodological approach which would make the reconciliation of these contradictory standpoints, and at the same time would provide a rational view corresponding to historical facts of the progress taking place in science.

The tool of probabilism is statistical inference. By taking as a given a sceptical understanding of truth as preached by probabilism, we may, however, come to an approximation of truth at a satisfactory level (with the level of error that we find acceptable).

8. The information revolution and statistics

The aim of statistics is to draw information from data. The processing of these data in order to obtain useful information and formulating conclusions on their basis is the subject of data analysis. In the field of statistical data analysis, it is often said: 'Let the data speak for themselves'. This is known as 'learning from data' or 'statistical learning'. The latter is a set of statistical tools for the modelling and analysis of complex data sets. Among the foundational research papers, covering in detail the theory and applications of the methods of statistical learning, are books written by Hastie et al. (2009), as well as by James et al. (2013).

Modern data analysis is based on the application of methods which automatically search for procedures allowing an optimum data analysis. They are part of the field of machine learning, involving the creation of systems which perfect their own operations based on experiences from the past. Two basic types of machine learning can be distinguished: supervised learning and unsupervised learning. The former assumes human supervision in the process of creating a function mapping the system input to its output. This supervision involves providing a program with a set of input-output pairs in order to teach it to take decisions in the future. Unsupervised learning, on the other hand, is a kind of machine learning which assumes the absence of an exact or even approximate output in the training data. The task of learning without supervision involves the determination of interdependencies among various features or the discovery of an internal structure within the data set. Examples of unsupervised learning include cluster analysis and correspondence analysis. Taxonomic methods are also unsupervised learning methods. The literature on machine learning is very extensive and consists of papers which synthesise IT, mathematical, engineering and statistical issues, including work by Cichosz (2000), Koronacki and Ćwik (2005), or Krzyśko et al. (2008).

Machine learning is a scientific discipline connected with the issue of artificial intelligence. This subject involves knowledge from the field of mathematics, statistics, engineering, and IT. It has emerged as a result of the attempts to mathematically model the processes which take place within the human body. At the same time, artificial intelligence is an IT subfield which concerns the creation of models of intelligent behaviours and of computer programmes which simulate these behaviours. It can also be defined as an IT subfield which deals with solving problems that cannot be effectively algorithmised (Rutkowski, 2009). The concept of artificial intelligence has two basic meanings: one involving hypothetical modelling of intelligence, and the other of a technology in service of scientific research. The main task of research in the field of artificial intelligence is to construct machines and computer programmes which are capable of carrying out selected functions of the human mind and senses which cannot be easily reproduced with a numerical algorithm. Thus, AI-related issues are connected with the field of IT, but involve a number of other disciplines, including neurology, psychology, cognitive studies, systematics, as well as contemporary philosophy.

The development of telecommunications and information technologies, the Internet and IT, occurring along with the decrease in the unit costs of gathering and storing data, have caused significant quantitative and qualitative changes in the approach to data itself and to the possibilities of analysing them. This dense, constant,

and unstructured stream of data is known as Big Data. Enormous amounts of information are generated not only in many fields involving the study of the natural world, but also in social and economic fields. The role of statistics in this context is to find some meaning and purpose within these data.

Since the introduction of high-performance computers, which signalled the beginning of the 'information era', a dramatic increase has occurred in terms of the possibilities of effectively analysing large and complex statistical problems. This growth in both the technological capacity to store, organise, and search data, as well as in the available methods of analysing data, has led to the emergence of a new field of statistical research, called Data Mining. However, very large data sets also pose various challenges regarding the reliability of inferences drawn from the reality studied through the analysis of the said data. In Big Data sets, apart from information characterised by a sufficient degree of clarity (Clear Data), there is also a significant amount of false, outdated, fuzzy, duplicated, incomplete, and erroneous data (Dirty Data), as well as those data whose quality or usefulness is unknown (Dark Data) (Migdał-Najman & Najman, 2017). Can such data, subject to no filtering based on appropriate methodological assumptions, be used to uncover the truth of the studied reality? Here again we have to pose questions regarding the philosophical and methodological nature of scientific research.

9. Questions

In relation to the new possibilities of conducting statistical research supported by contemporary IT tools, as outlined in the previous chapter, the following questions emerge, jointly with their potential answers.

Is data analysis based solely on empirical foundations expressed by the assertions: 'Let the data speak for themselves' or 'In data analysis no assumptions are necessary'? It can be proven that data analysis procedures, including those regarding Big Data, are also based on assumptions which, nevertheless, are in most cases concealed and more flexible than those observed in classical statistical analysis.

Can large sets of data be analysed at a chosen degree of precision, thus bringing us closer to the truth about the reality we study? Or is it simply a matter of the amount of time dedicated to the execution of computer calculations and the related costs of a given study? Can the future be predicted at the assumed level of accuracy based on such studies? The answers to such questions are inconsistent: IT specialists believe that the computing capacity of modern computers has no limits, and yet, e.g. problems relating to projecting the course of the Covid-19 pandemic indicate that data

analysis procedures do in fact have limitations, despite the volume of the tools and resources involved.

Can an AI-based computer justify the relevance of conducting a given piece of scientific research? Here, the answer is rather negative. It can indicate the analytical or prognostic efficiency of the applied, specific research procedures, yet it is unable to formulate general research aims.

Is an AI-based computer capable of philosophical thought? Rather not, but certainly human beings can form a philosophy of artificial intelligence; such a branch of philosophy has already been created (Russell & Norvig, 2003).

And finally, is the development of science even possible without any philosophical underpinnings in the era of high-performance computers? The entire development of epistemology and the philosophy of science from ancient times to this day suggests that science cannot be conducted without a philosophical basis (Heller, 2011). In terms of statistics constituting a basic tool of empirical study, this statement also proves true.

Nevertheless, any attempt to provide full answers to the above-mentioned questions require a separate study.

References

Ajdukiewicz, K. (2006). *Język i poznanie*. Warszawa: Wydawnictwo Naukowe PWN.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: The University of Chicago Press.

Cichosz, P. (2000). *Systemy uczące się*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd edition). New York: Springer Science+Business Media. https://doi.org/10.1007/978-0-387-84858-7.

Heller, M. (2011). *Filozofia nauki: Wprowadzenie*. Kraków: Petrus.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer Science+Business Media. https://doi.org/10.1007/978-1-4614-7138-7.

Koronacki, J., Ćwik, J. (2005). *Statystyczne systemy uczące się*. Warszawa: Akademicka Oficyna Wydawnicza Exit.

Krzyśko, M., Wołyński, W., Górecki, T. & Skorzybut, M. (2008). *Systemy uczące się: rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*. Warszawa: Wydawnictwo Naukowo-Techniczne.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Lakatos, I. (1995). *Pisma z filozofii nauk empirycznych*. Warszawa: Wydawnictwo Naukowe PWN.

Migdał-Najman, K., Najman, K. (2017). Big Data = Clear + Dirty + Dark Data. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu / Research Papers of Wrocław University of Economics*, (469), 131–139. https://doi.org/10.15611/pn.2017.469.13.

von Mises, R. (1957). *Probability, Statistics and Truth* (2nd edition). London: George Allen & Unwin.

Oleński, J. (2006). *Misja polskiej statystyki publicznej w latach 2006–2011 oraz strategia jej realizacji*. Warszawa: Główny Urząd Statystyczny.

Osiewalski, J. (1991). *Bayesowska estymacja i predykcja dla jednorównaniowych modeli ekonometrycznych*. Kraków: Akademia Ekonomiczna w Krakowie.

Ostasiewicz, W. (2012). *Myślenie statystyczne*. Warszawa: Wolters Kluwer Polska.

Pociecha, J. (2016). Korzenie polskiej statystyki publicznej – statystyka i statystycy galicyjscy na początku XX wieku. *Folia Oeconomica Cracoviensia*, *57*, 5–18.

Popper, K. R. (1977). *Logika odkrycia naukowego*. Warszawa: Państwowe Wydawnictwo Naukowe.

Pruś, J. (2018). Teorie prawdy: klasyczna, korespondencyjna i semantyczna – próba uściślenia relacji. *Rocznik Filozoficzny Ignatianum, 24*(2), 57–83. https://doi.org/10.5281/zenodo.2542205.

Rao, C. R. (1994). *Statystyka i prawda*. Warszawa: Wydawnictwo Naukowe PWN.

Russel, S., Norvig, P. (2003). *Artificial Inteligence: A Modern Approach* (2nd edition). New Jersey: Pearson Education.

Rutkowski, L. (2009). *Metody i techniki sztucznej inteligencji*. Warszawa: Wydawnictwo Naukowe PWN.

Schuster, H. G. (1995). *Chaos deterministyczny: Wprowadzenie*. Warszawa: Wydawnictwo Naukowe PWN.

Strawiński, W. (2011). Funkcja i cele nauki – zarys problematyki metodologicznej. *Zagadnienia naukoznawstwa*, (3), 323–335. http://journals.pan.pl/dlibra/publication/108246/edition/93890/content.

van Strien, M. (2014). On the origins and foundations of Laplacian determinism. *Studies in History and Philosophy of Science*, *45*(1), 24–31. https://doi.org/10.1016/j.shpsa.2013.12.003.

Szapiro, T. (1993). *Co decyduje o decyzji*. Warszawa: Wydawnictwo Naukowe PWN.

Tarski, A. (1995). *Pisma logiczno-filozoficzne*. Warszawa: Wydawnictwo Naukowe PWN.

Tatarkiewicz, W. (1978a). *Historia filozofii* (t. 1). Warszawa: Państwowe Wydawnictwo Naukowe.

Tatarkiewicz, W. (1978b). *Historia filozofii* (t. 3). Warszawa: Państwowe Wydawnictwo Naukowe.

Trzaskalik, T. (1990). *Wielokryterialne, dyskretne programowania dynamiczne: Teoria i zastosowania w praktyce gospodarczej*. Katowice: Wydawnictwo Akademii Ekonomicznej.

Woleński, J. (2014). *Epistemologia: Poznanie, prawda, wiedza, realizm*. Warszawa: Wydawnictwo Naukowe PWN.