Przegląd Statystyczny. Statistical Review, 2021, vol. 68, 1, 1–16 DOI: 10.5604/01.3001.0015.0375
Received: 18.05.2021, revised: 30.06.2021, accepted: 7.07.2021

The paradigm of statistical inference and the paradigm of statistical learning

Józef Pociecha[a]

Abstract. The starting point for the presentation of the similarities and differences between the principles of conducting statistical research according to the rules of both statistical inference and statistical learning is the paradigm theory, formulated by Thomas Kuhn. In the first section of this paper, the essential features of the statistical inference paradigm are characterised, with particular attention devoted to its limitations in contemporary statistical research. Subsequently, the article presents the challenges faced by this research jointly with the expanding opportunities for their effective reduction. The essence of learning from data is discussed and the principles of statistical learning are defined. Moreover, significant features of the statistical learning paradigm are formulated in the context of the differences between the statistical inference paradigm and the statistical learning paradigm. It is emphasised that the statistical learning paradigm, as the more universal one of the two discussed, broadens the possibilities of conducting statistical research, especially in socio-economic sciences.

Keywords: paradigm theory, learning from data, statistical inference, statistical learning
JEL: C000, C180, C83

1. The statistical inference paradigm

Thomas S. Kuhn, an American physicist, historian and philosopher of science, is the founder of the concept of the scientific paradigm. In his basic work on the philosophy of science entitled *The Structure of Scientific Revolutions* (Kuhn, 1962),[1] he introduced into the philosophy of science the idea of a paradigm as a set of concepts and theories that form the basis of a given science. These theories and concepts are not questioned, at least as long as the paradigm is cognitively creative, i.e. it can be used to create specific theories consistent with the experimental or historical data which the science concerns. The most general paradigm is that of the scientific method, which formulates the criteria for recognising an activity as scientific. The paradigm guides the research effort of scientific communities and is the basic criterion for identifying areas of individual sciences. Kuhn's fundamental claim is that a transition from the old to the new paradigm takes place in the process of scientific revolutions. When a paradigm shift occurs, the scientific world changes qualitatively by enriching it with new facts and theories. Thus, according to Kuhn, the development of scientific theories continues.

In science, and especially in social sciences, different paradigms can occur simultaneously, which can even lead to scientific paradigm wars, involving scientists

[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.
[1] Available in Polish, Kuhn (2020).

from different camps to contest each other and deny others their scientific character. Examples of different paradigms in economic sciences are those of classical and Keynesian economics. Parallel paradigms exist in statistics, including descriptive statistics, mathematical statistics, Bayesian statistics or the statistical learning paradigm (Pociecha, 2020).

The paradigm of mathematical statistics (statistical inference) is based on the notions of a (general) population, i.e. a statistical population which we do not study, and a sample, i.e. a subset of the population we are investigating. The paradigm assumes that – on the basis of the sample – we can infer about the population, as long as the sample is representative of the population. A necessary condition for the representativeness of a sample is its random selection. Sample representativeness is a gradual concept; the degree of sample representativeness depends on the sampling method and, for a given sampling scheme, the size of the sample. Quite complicated, layered sampling schemes are generally used in statistical survey practice, especially when the population is very numerous and clearly structured. The sampling theory deals with sampling procedures (see e.g. Bracha, 1996; Steczkowski, 1995; Tillé, 2006).

As part of mathematical statistics, two theories constituting its methodological basis were formulated – the statistical estimation theory and the statistical hypotheses testing theory. The theory of estimation lays down the rules for estimating the parameters of the distribution in the population, based on the obtained sample, with a fixed type of distribution. The theory of statistical hypotheses testing formulates rules for verifying the truthfulness of judgements regarding the parameters of the distribution in a certain population by comparing it with a determined type of distribution, or judgements relating to the compliance of the empirical distribution with a selected theoretical distribution. Moreover, the theory involves examining hypotheses regarding the randomness of a sample or other related judgements about one or more populations.

Statistical inference is of a probabilistic nature and uses two concepts in particular: the concept of a confidence level, defined as a confidence coefficient or interval, understood as a probabilistic measure of an estimation error in parameter estimation, and the concept of a significance level, defined as a predetermined probability (risk) of committing a first type error.

The paradigm of mathematical statistics is based on the stochastic nature of statistical regularities resulting from the indeterministic understanding of the connections between the components of the world around us, on the principles of inductive inference in the incomplete version and the frequency definition of probability. The assumptions of the Aristotelian realist philosophy, assuming that the world (reality) exists objectively (outside our mind) and is knowable constitute

the fundamental philosophical basis of the paradigm of statistical inference as a tool for studying mass processes. This philosophy shows that learning about the reality that surrounds us, although difficult, is possible and the scientific effort it entails is deliberate. This justifies the possibility and purposefulness of conducting scientific research also with the use of statistical methods. An important premise of the statistical inference paradigm is Karl Popper's critical rationalism, which forms the philosophical basis for testing scientific hypotheses. The immediate philosophical foundation of this paradigm is probabilism, which originated from ancient sceptics and was developed by neo-positivists. Popper adopts a sceptical understanding of the truth, which we can approach only at a distance acceptable to us (with an error that we accept), and which allows making statistical inferences (Pociecha, 2020).

Conducting statistical research in accordance with the presented paradigm of mathematical statistics is, however, subject to certain limitations and its correct performance – in both theoretical and practical terms – faces a number of difficulties. In particular, it is challenging for socio-economic sciences to put into practice the theoretical requirements for sampling (Cassel et al., 1977). This relates to defining the substantive, spatial and temporal scope of the population, determining the sampling frame, sampling scheme, sampling procedure (quota, group, systematic, stratified sampling), the multi-stage sampling scheme or determining the sample size. Each decision in any of the mentioned areas affects the obtained sample representativeness. However, in the practice of statistical analyses, often no attention is devoted as to how the data set, which we consider a random sample, was obtained and no tests verifying the randomness of a sample are conducted. In effect, the correctness of the obtained results of statistical investigations is often questioned, and the validity of using statistical methods in socio-economic research becomes uncertain.

The statistical estimation theory and the theory of parametric statistical hypotheses verification requires assuming a specific analytical form of distribution in the population. The vast majority of estimation procedures and the verification of parametric hypotheses require the assumption of normal distribution in the population. While in physico-biological studies the assumption of the normality of distribution in the population is in most cases satisfied, in socio-economic studies it is usually not. However, procedures aiming to test the normality of a population distribution are rarely used. There are, of course, also procedures assuming a different than normal analytical form of distribution in the population, but statistical procedures based on such distributions have not been developed and are theoretically complex.

It should also be noted that parametric hypothesis estimation and testing procedures are limited to solving problems which can be effectively parameterised,

but there are numerous empirical research problems that cannot. The existing non-parametric tests alleviate the problem of analysing non-parameterised issues only to a certain extent.

Another limitation of the mathematical statistics paradigm is the adoption of an axiomatic in theory and frequentative in practice definition of probability. The Bayesian statistic paradigm extends the understanding of probability to an *a priori probability* and a *probability a posteriori,* which broadens the understanding of this key concept to include its subjective aspect and allows for a clear connection between statistical theory and empirical research. The limitation of classical statistical inference is that the testing of statistical hypotheses is based on minimising the risk of committing the first type error, i.e. rejecting the null hypothesis when it is actually true, which occurs in significance tests. In a large number of cases, it is more important to minimise the risk of making a second type error, i.e. accepting the null hypothesis when it is false. These situations arise, for example, in the process of testing the correctness of financial statements when their audit is performed (Hołda & Pociecha, 2009).

The limitations resulting from the failure of empirical data to meet the theoretical assumptions underlying the methods of the estimation and verification of statistical hypotheses are also highlighted by Wiesław Szymczak in his book on the practice of statistical inference (Szymczak, 2018). In his work, the author critically assesses the role of the paradigm in statistics. Summing up, it should be emphasised that the commonly functioning paradigm of statistical inference does not provide a universal basis for empirical statistical research. It is subject to significant limitations and creates a number of difficulties for the correct implementation of empirical research according to this paradigm.

2. Challenges facing modern statistical research

The rapid development of information technology (IT), encompassing more and more efficient computer hardware and involving an increasingly higher quality and reliability computer software, enables processing great amounts of information, offering new analytical possibilities for contemporary statistical research, unlike ever before. Modern computers have an unimaginable computing power. Currently, the most powerful computer in the world, designated as 1/10 1 – Summit – IBM Power System AC922 has 2,801,644 GB of memory and 2,414,592 cores. Its computing power is at the level of 148,600 teraflops per second, and may even exceed the value of 200,795 teraflops (Onet, n.d.). Thus, it can be concluded that the current computational possibilities for statistical analyses face no technical barriers.

The increased ability to collect, process and store data is now leading to the creation of extremely large data sets for which the term Big Data has been adopted.

Big Data is defined as a dense, continuous and unstructured data stream resulting from interpersonal interactions, interactions between devices being part of the infrastructure of the global computer network, and all other instruments through which this data stream is registered and transmitted (Migdał-Najman & Najman, 2017). The most important source of Big Data, however, are the interactions resulting from humans' connections with IT devices, giving people access to numerous services such as transaction systems, online stores, financial services, mobile services, systems monitoring health, emotions, location, and physical activity. Big Data sets are characterised by the amount of data they contain (*volume*), data processing speed (*velocity*) and data diversity (*variety*). The above-mentioned features include the degree of their reliability (*veracity*), their value for the user (*value*) and the possibility of their visualisation (*visualisation*) (Tabakow et al., 2014).

Even if a Big Data set displays the above-mentioned features, it does not necessarily mean that it is directly useful for conducting statistical analyses. Big Data, in addition to providing useful, up-to-date, accessible, comparable, consistent and accurate data, called *clear data*, consists of inaccurate, repeated, incomplete, wrongly named or non-integrated data, referred to as *dirty data*, as well as *dark data*, whose author, place and time of creation, content and connection with other data remains unidentified (Migdał-Najman & Najman, 2017).

Thus, Big Data contains not only the clear data desired by the analyst, but also dirty data, whose removal often involves complicated cleaning procedures (Kim et al., 2003), and dark data, towards which the analyst should make a decision whether to eliminate them from the given data set. It is difficult to clearly state in what proportions the above-mentioned Big Data components occur, but IT specialists and analysts claim that clear data is a substantial minority within Big Data, which is also reflected in the obvious disproportion between the amount of data collected and the amount of relevant data providing valuable information. IT specialists say that dark data can account for up to 90% of the entire volume of Big Data – it is then this percentage of Big Data that is not fit for analytical purposes. Thus, the technical and IT-related capacity for collecting and storing data are much higher than the ability to analyse and draw conclusions from these data; moreover, this disproportion is growing rapidly (Migdał-Najman & Najman, 2017).

The changes in the acquisition, storage, processing and analysis of data presented above pose new challenges for modern statistical research. It is not the issue of limited data availability or restricted computational possibilities that constitute a barrier to the development and application of statistical methods. On the contrary, today's excess of data and the enormous computing power of computers pose a challenge for the rational application of statistical methods in socio-economic

analyses. In consequence, the classical paradigm of mathematical statistics is often insufficiently effective for modern statistical research. This made it necessary to search for a new paradigm of empirical research using statistical tools.

3. Learning from data

The purpose of statistical analysis is to extract information from data, while data analysis involves the processing of data in order to obtain useful information and draw conclusions. Sometimes the term 'data analysis' is understood as a field of knowledge covering the issues of acquiring, storing and processing data, building data warehouses, databases and algorithms; the term may also relate to the knowledge of IT tools such as Excel, Python, R, SQL environment, etc. Data analysis understood in this way is described in many works, e.g. in Alexander & Kusleika (2019). It should be emphasised, however, that the scope of data analysis is significantly broader and involves not only obtaining or processing data with the use of appropriate IT tools, but, above all, inferring about the socio-economic reality which these data come from.

In data analysis, the saying 'let the data speak for itself' means learning from data. The idea is to acquire knowledge not only useful for performing current activities, but also to improve future performance. Learning from data is an inductive inference made on the basis of available observations. Learning outcomes are influenced by three main factors: the components of learning from data, the type of feedback on the basis of which the learning process takes place, and the method of presenting the acquired information (Russel & Norvig, 2003).

In the era of the rapid development of information technology, computers are designed to learn from data. In result, machine learning was created, i.e. self-learning systems based on algorithms which automatically improve through experience (Cichosz, 2000). Machine learning should therefore be understood as the ability of computers to automatically learn from data and transfer this knowledge to the recipient.

Machine learning from data can be realised in three forms: supervised learning, unsupervised learning and reinforcement learning. Supervised learning consists in approximating unknown function f by mapping the input data with the output data by providing individual input data (x_i) and knowing the output data (y_i). The essence of supervised learning is to provide the algorithm with a set of input-output pairs, i.e. (x_i, y_i) pairs in order to approximate unknown function f. According to the supervised learning theory, input-output is entered to find function f that maps the values of x to the value of y. The set of all possible functions that can describe this mapping is called hypothetical space H. Next, function h is selected. This

function belongs to hypothetical space H, which, in the author's opinion, approximates unknown function f well and provides the possibility of making rational future decisions. Function h is a hypothesis of the actual course of function f (Russel & Norvig, 2003). In statistical literature, input data are called predictors, or classically – independent (explanatory) variables. In the machine learning terminology output data are called response variables, and classically – dependent variables (Hastie et al., 2009).

Unsupervised learning is a type of machine learning which assumes that there is no exact or even approximate output in the training data. So unsupervised learning involves learning patterns on the basis of the given data when only the input data are known. The aim of unsupervised learning is to either identify the interdependencies between features or to discover the internal structure of a data set. Examples of unsupervised learning include cluster analysis and correspondence analysis. Unsupervised learning methods are taxonomic methods used to classify objects in a multidimensional space of features according to the adopted measure of their similarity or distance (Pociecha et al., 1988).

Reinforcement learning does not use input or output data. It consists in observing the environment by the learning system and selecting activities in order to maximise the rewards and avoid the punishments. The learning system learns on its own the best strategy, called politics, to collectively obtain the highest reward (Géron, 2018).

Function f, connecting the input data with the output data in supervised learning, can be deterministic or indeterministic; consequently, learning can also be deterministic or indeterministic. In the study of socio-economic phenomena, usually indeterministic relationships are observed, therefore supervised learning should be understood as indeterministic learning. Unknown function f is approximated by hypothetical function h. In theory, the more complex the function, the better chance of an exact approximation of function f. The indeterministic learning process involves an inevitable compromise between the complexity of hypothetical function h and the degree of dispersion of the input data. The learning problem is feasible if hypothesis space H contains the actual function f. Unfortunately, it is not always possible to assess whether a given learning problem is feasible because the true function is unknown. One way to bypass this barrier is to use the previously gained knowledge to derive hypothesis h from space H, when it is certain that the actual function f is contained in this space (Russel & Norvig, 2003).

If the actual function f is of a stochastic nature, supervised learning is understood as statistical learning. Bearing in mind that in the vast majority of cases function f, which assigns input data to output data, is defined in a stochastic manner, machine learning is in fact almost entirely statistical learning. However, due to the fact that it was introduced into the literature by computer scientists, the term 'statistical

learning' has been dominated by the term 'machine learning'. It is only thanks to the fundamental works of Hastie et al. (2009) and James et al. (2013) that statistical learning begins to occupy its rightful place in the world literature.

The formal definition of statistical learning is presented in numerous studies, including that of Vapnik (2000). According to a popular operational definition, statistical learning is a collection of descriptive statistics, mathematical statistics, and non-parametric and non-algorithmisable procedures for modelling and understanding complex data sets. Statistical learning is a new field of knowledge that has been developing since the turn of the 20th and 21st century, being the product of the development of statistics and computer science. It combines the principles of machine learning with statistical methods (James et al., 2013).

4. Principles of statistical learning

In the general approach to statistical learning, we have dependent variable Y, understood as the *response variable* and k explanatory variables (predictors) $X_1, X_2, \ldots, X_k$. We assume that there is a relationship between Y and $X = (X_1, X_2, \ldots, X_k)$ which we can generally define as

$$Y = f(X) + \xi, \tag{1}$$

where
f – an unknown function associating Y with X;
ξ – a random component.

The essence of statistical learning is to guess function f using function h, which is one of the hypotheses belonging to hypothesis space H, concerning unknown function f (Hastie et al., 2009).

There are two primary reasons for attempting to guess function f. The first one, of a practical nature, is the prediction of Y based on the knowledge of X. The other reason, of a more cognitive nature, is the inference about Y on the basis of X. The prediction task is when a set of X predictors is available, but the corresponding values of response variable Y are unknown. In this case, we predict Y using the following equation:

$$\hat{Y} = \hat{f}(\mathrm{X}), \tag{2}$$

where $\hat{f}$ is one of the functions falling in hypothetical space H. In this approach, $\hat{f}$ is treated as a black box, in the sense that it is not usually a specific analytical form of $\hat{f}$, on condition that it provides accurate possible forecasts of Y.

If we want to use statistical learning methods for the purpose of inferring about the relationship between the dependent variable and the explanatory variables, then we cannot treat $\hat{f}$ as a black box, but we have to take the specific form of function $\widehat{f}$. Subsequently, we attempt to answer the following questions:

- Which pre-adopted explanatory variables actually affect the response variable?
- What is the direction of the relationship between the response variable and individual explanatory variables?
- What is the appropriate analytical form for $\hat{f}$?
- Is the linear form sufficient?

Statistical learning methods are designed to answer these types of questions (James et al., 2013).

The approximation of the actual f function is the key statistical learning problem. Its estimation is based on a data set, called the training data or training set, containing input and output information (x_{ij}, y_i). In other words, we seek such a function $\hat{f}$, for which

$$Y \approx \hat{f}(X) \tag{3}$$

for any pair of observations from set (X, Y). A parametric or non-parametric approach can be applied here.

The parametric method of statistical learning involves specifying the analytical form of function $\hat{f}$. In the simplest and most common case, we assume it as a linear multivariate model. Then, using the data from the training set, the partial regression coefficients of this model are estimated, most often by means of the least squares method. Of course, in the case of parametric statistical learning there are many options for both the selection of the vector of explanatory variables and the analytical form of the regression function.

Non-parametric statistical learning methods do not make explicit assumptions about the analytical form of the functions for f. Instead, they look for a form of function f which fits as closely as possible to the data from the training set. The non-parametric approach can have a great advantage over the parametric approach, because by avoiding the assumption of a specific analytical form of function f, it can fit the empirical data more accurately. The parametric approach involves the risk that the analytical form of function $\hat{f}$ deviates greatly from the actual function f,

which links the predictors with the result variable. Nevertheless, the non-parametric approach has the disadvantage that it does not reduce the number of the estimated parameters to only the significant ones, and thus requires a much larger training set (James et al., 2013).

When selecting the best function $\hat{f}$ belonging to hypothetical space H, one should follow a specific quality criterion of fitting this function to the actual function f. The effectiveness of the statistical learning method with the specified $\hat{f}$ is measured by the mean square error of estimation (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2, \tag{4}$$

where
$\hat{f}(x_i)$ – the prediction of the actual f for the i-th observation.

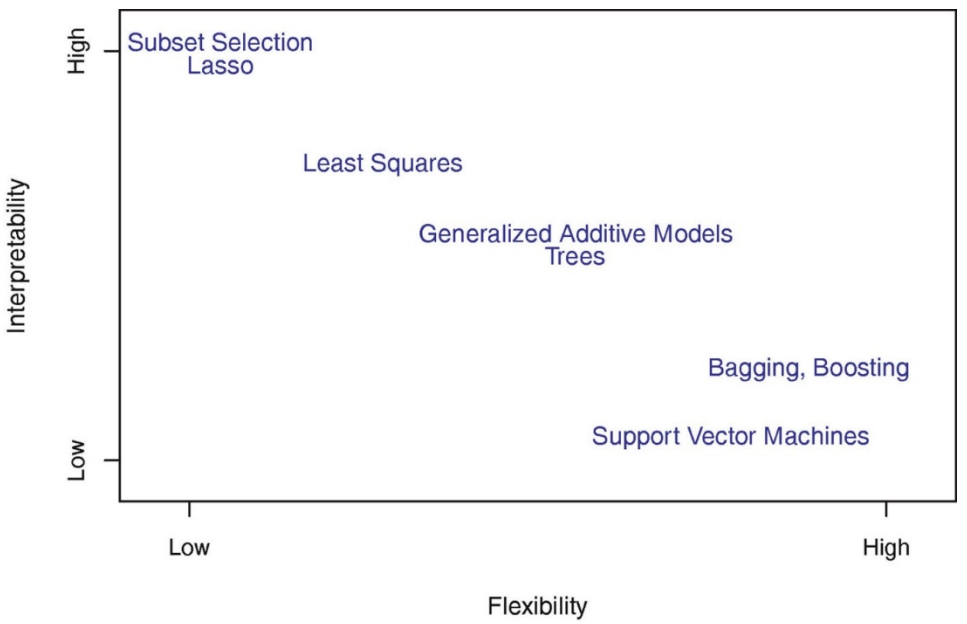
If the estimation error is calculated for the data from the training set, it is a measure of the goodness of fit of function $\hat{f}$ to empirical data y_i. However, we are actually interested in the accuracy of the predictions we obtain when applying a given statistical learning method to a previously unknown set of test data. Whether $y_i \approx \hat{f}(x_i)$ is of no particular interest in this context; what is important is that $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is an observation from the non-processed test set. We select the method which provides the lowest-test MSE as opposed to the lowest-training MSE. If a large number of test observations was available, the average squared prediction error for these test observations (x_0, y_0) could be computed. We select the model for which the average-test MSE is as small as possible. The MSE measure is treated as a generalisation error. In practice, the set of observations is usually divided into two subsets: the training set, on which we train the statistical learning method, and the test set, which is used to verify the effectiveness of the learning method. The proportion of the division of the data set into the training and the test part tends to be problematic, as on the one hand, it is assumed that the training set should be larger than the test set, but on the other hand, the test set cannot be too small. It is recommended that the proportion of the division into the training and test set is 8 to 2 (Géron, 2018). There are numerous studies comparing the prognostic abilities of forecasting models which use different proportions of the division of data sets into the training and testing part (cf. e.g. Pociecha et al., 2014).

Function $\hat{f}$ can fairly flexibly match the actual function f. However, when applying statistical learning procedures, the problem of the overfitting of the $\hat{f}$ function to the data may occur. This is related to the possible overtraining (overfitting) of the data learning model. This is the case when the MSE based on the

training set is clearly smaller than the MSE based on the test set. Along with the increase in the flexibility of the statistical learning method, a monotonic decrease in the MSE on the training set is observed. On the other hand, the distribution of the MSE on the test set is U-shaped and with the increasing flexibility of function $\hat{f}$, the MSE first decreases and then increases. These are the basic properties of statistical learning methods, independent of the specific data set and independent of the used learning method (James et al., 2013). This provides an opportunity to determine the optimal degree of flexibility of function $\hat{f}$ relative to the training set. The basic method of determining this optimum is cross-validation (Koronacki & Ćwik, 2005). It consists in separating the training set into mutually complementary subsets, and the models are trained in various combinations of these subsets and evaluated using the remaining, unused subsets; in result, the optimal model is determined.

The search for a compromise between the accuracy of prediction and the interpretation of the statistical learning model is an issue related to the one described above. Statistical learning methods involve functions which fairly flexibly adapt to the data from the training set. Linear regression is an example of an inflexible function, while e.g. spline functions are more flexible. The more restrictive functions, and therefore demonstrating less flexibility, allow for a deeper substantive interpretation of the obtained results. Flexible functions, on the other hand, can lead to such complicated estimates of the shape of the actual function f that it is difficult to interpret the relationship between the assumed predictors and the dependent variable. The relationship between the flexibility and interpretability of statistical learning methods is presented by James et al. (2013) as in Figure 1.

Figure 1. The relationship between interpretability and flexibility of statistical learning methods



Source: James et al. (2013, p. 25).

The authors indicate that the relation between the flexibility and interpretability of statistical learning methods is approximately inversely proportional. The Lasso regression is fully interpretable as it allows for the joint selection of explanatory variables and the assessment of their impact on the dependent variable (Kubus, 2014 or Tibshirani, 1996). Generalised additive models (GAM) are interpretable and at the same time more flexible, as they allow non-linear relationships between variables. Fully non-linear models, including bagging or boosting and the support vector method (SVM) are highly flexible, but difficult to interpret in terms of content. To sum up, if the aim of using statistical learning methods is to make the most precise prediction possible, then the most flexible learning method should be selected. If the interpretation of the relationship between the response variable and the explanatory variables is important, then the more classic methods should be applied (James et al., 2013).

Statistical learning methods are used to solve both regression and classification problems. If the result (response) variable is a quantitative (directly measurable) variable, then the explanatory variables' (predictors') influence on it is examined by regression. If the result variable is of a qualitative (directly non-measurable) nature, then the relationship between it and the predictors is examined by means of the classification method (Hastie et al., 2009).

A wide range of statistical learning methods are presented in the literature. There is no one best method in statistics and no method dominates all the others for all possible data sets. For a particular dataset, one method may work best, but another method may prove more efficient in relation to a similar and yet different data set. Therefore, statisticians face an important task of selecting the most effective method which – when applied for a given data set – gives the best results. In conclusion, the choice of the most appropriate statistical learning method is one of the most challenging decisions in statistical research practice (James et al., 2013).

5. Statistical learning paradigm

The previously characterised premises and principles of learning from data allow for the formulation of a statistical learning paradigm. The statistical learning paradigm will be presented against the classical paradigm of statistical inference. The starting point of the mathematical statistics paradigm is the probability theory and its basic concepts, including the random event, the axioms of probability theory, the random variable and its distribution. They are followed by the theorems of the probability of events, Bayes' theorem, the formalisation of particular types of distributions of a random variable and their characteristics in the form of distribution parameters for one- and multi-dimensional variables. The key elements of statistical inference include the concept of the distribution of statistics from a sample, the principles of estimation parameters and the principles of the verification of statistical hypotheses (Kot et al., 2011). The essence of the mathematical statistics paradigm is to start from the theory of probability and statistical inference and to check to what extent the empirical data can fit into the theoretical framework of mathematical statistics.

The statistical learning paradigm involves the opposite – the starting point is the available data set. The theory is based on the 'let the data speak for itself' and 'we learn from the data' concepts, which is consistent with neo-positivist beliefs, according to which all knowledge is based on empirical data, whereas anything that is not based on empirical facts is rejected. Neo-positivists assumed that experience is the source of all knowledge about the real world (Kołakowski, 2004).

The statistical inference paradigm is based on the concept of general population and sample. The condition for the correctness of inference about a population based on a sample is the random selection of the sample. The sampling method focuses strongly on sampling procedures so that the sample is representative of the entire population. The statistical learning paradigm, on the other hand, ignores the notions of population and sample. Instead, it assumes that we have a sufficiently large set of empirical data on the basis of which we can effectively make predictions and infer about the reality which these data come from. In the practice of applying statistical

learning procedures, it is often presumed that the training set has the properties of a random sample, but its actual randomness is not verified. Perceiving the training and test set automatically as random samples is in fact an unjustified transfer of the features of the statistical inference paradigm onto the statistical learning paradigm.

In the era of powerful computers, it is possible to collect, process and store large data sets, known as Big Data. However, such sets do not have the characteristics of random samples; they are said to be noisy, i.e. partially random and contain unreliable information which needs to undergo various data cleaning processes. It should be mentioned here that data from a random sample, selected in accordance with the rules of the sampling method, are not subject to 'cleaning' as, by definition, their appearance in the sample is determined by the probability of their occurrence in the population. However, the application of statistical learning procedures should not be limited to Big Data as they are known to be used in training sets with less than one hundred observations (James et al., 2013).

The essence of the statistical learning paradigm is the creation of self-learning systems, i.e. systems which improve automatically through experience. In the case of statistical learning in the supervised version, it involves providing the algorithm with a set of input-output pairs (x_i, y_i) in order to find unknown function f by mapping input data to output data, with the accuracy of the minimised mean square error of the estimate or mean prediction error. The actual function f in the statistical learning paradigm is understood as a black box – it can be a parameterised or non-parameterised function, it can even be a non-algorithmic procedure.

The purpose of statistical learning is to get closer to the real function by estimating it on the training set in which function $\hat{f}$ is taught how to recognise the actual function f as accurately as possible. The statistical learning effect is tested on a test set and its optimisation is performed in the process of cross-validation. The basic difference between the process of statistical estimation and the process of statistical learning is that in the former we estimate the parameters of a pre-determined function, and in the latter the form of this function and its parameters are selected by the learning method.

The statistical learning paradigm includes classical linear regression models, logistic regression, discriminant analysis, polynomial models, splined functions, generalised additive models, kernel classifiers, regression and classification trees, bagging and boosting methods, random forests, neural networks, support vectors machines, the *k*-means method, and other lesser-known learning procedures. As the list above suggests, the range of statistical learning tools is much wider than that of the classical mathematical statistics tools.

The concept and the theory of probability plays a key role in the paradigm of statistical inference. In the statistical learning paradigm, its role is secondary as there

are serious doubts whether the training set could be considered as a random data set. In this sense, the statistical learning paradigm is getting closer to the descriptive statistics paradigm.

The literature also emphasises the difference in terms of the research goals that can be achieved through both paradigms. Statistical research conducted within the mathematical statistics paradigm focuses primarily on explaining the relationships between the studied variables, i.e. on the implementation of analytical goals; thus, the forecasts built on their basis are often imprecise. Empirical research conducted within the statistical learning paradigm involves building on their basis forecasts which would be as accurate as possible; nevertheless, their analytical and interpretative role could be limited.

In conclusion, however, it should be emphasised that the statistical learning paradigm is a more universal research platform, as it has in fact absorbed the statistical inference paradigm at the expense of weakening its original assumptions. The statistical learning paradigm offers a great opportunity to use the computing power of modern computers and large data sets produced by contemporary socio--economic life.

References

Alexander, M., & Kusleika, R. (2019). *Access 2019 PL: Biblia* (R. Meryk & T. Walczak, Trans.). Gliwice: Helion.

Bracha, C. (1996). *Teoretyczne podstawy metody reprezentacyjnej*. Warszawa: Wydawnictwo Naukowe PWN.

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1977). *Fundations of Inference in Survey Sampling*. New York: John Wiley & Sons.

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

Géron, A. (2018). *Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow* (K. Sawka, Trans.). Gliwice: Helion.

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

Hołda, A., & Pociecha, J. (2009). *Probabilistyczne metody badania sprawozdań finansowych*. Kraków: Wydawnictwo Uniwersytetu Ekonomicznego.

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

Kim, W., Choi, B. J., Hong, E.-K., Kim, S. K., & Lee, D. (2003). A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*, *7*(1), 81–99. https://doi.org/10.1023/A:1021564703268.

Kołakowski, L. (2004). *Filozofia pozytywistyczna: Od Hume'a do Koła Wiedeńskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Koronacki, J., & Ćwik, J. (2005). *Statystyczne systemy uczące się*. Warszawa: Wydawnictwo Naukowo Techniczne.

Kot, S. M., Jakubowski, J., & Sokołowski, A. (2011). *Statystyka* (2nd edition). Warszawa: Difin.

Kubus, M. (2014). Propozycja modyfikacji metody złagodzonego LASSO. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu / Research Papers of Wrocław University of Economics*, (327), 77–84. https://www.dbc.wroc.pl/dlibra/publication/27745/edition/25111/content?&meta-lang=pl.

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

Kuhn, T. S. (2020). *Struktura rewolucji naukowych*. Warszawa: Wydawnictwo Aletheia.

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.

Onet. (n.d.). *Komputer Świat*. Retrieved January 21, 2021, from https://www.komputerswiat.pl/onet.

Pociecha, J. (2020). Philosophical foundations of statistical research. *Przegląd Statystyczny. Statistical Review*, *67*(3), 195–211. https://doi.org/10.5604/01.3001.0014.7109.

Pociecha, J., Pawełek, B., Baryła, M., & Augustyn, S. (2014). *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*. Kraków: Fundacja Uniwersytetu Ekonomicznego w Krakowie. https://im.uek.krakow.pl/wp-content/uploads/2020/03/Statystyczne-metody-prognozowania-bankructwa_online.pdf.

Pociecha, J., Podolec, B., Sokołowski, A., & Zając, K. (1988). *Metody taksonomiczne w badaniach społeczno-ekonomicznych*. Warszawa: Państwowe Wydawnictwo Naukowe.

Russel, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice Hall, Pearson Education, Upper Saddle River.

Steczkowski, J. (1995). *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych*. Warszawa–Kraków: Wydawnictwo Naukowe PWN.

Szymczak, W. (2018). *Praktyka wnioskowania statystycznego*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Tabakow, M., Korczak, J., & Franczyk, B. (2014). Big Data – definicje, wyzwania i technologie informatyczne. *Informatyka ekonomiczna / Business Informatics*, (1), 138–153. https://doi.org/10.15611/ie.2014.1.12.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer. https://doi.org/10.1007/0-387-34240-0.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory* (2nd edition). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4757-3264-1.