EKONOMISTA | **ARTYKUŁY**

2024, 2, 138–154
DOI: 10.52335/ekon/188076

Krzysztof Jajuga
Katedra Inwestycji Finansowych i Zarządzania Ryzykiem, Uniwersytet Ekonomiczny we Wrocławiu, https://orcid.org/0000-0002-5624-6929,
krzysztof.Jajuga@ue.wroc.pl

Józef Pociecha
Katedra Statystyki, Uniwersytet Ekonomiczny w Krakowie, https://orcid.org/0000-0003-3140-481X,
pociecha@uek.krakow.pl

Mirosław Szreder
Katedra Statystyki, Uniwersytet Gdański, https://orcid.org/0000-0002-7597-0816, miroslaw.szreder@ug.edu.pl

Statistical inference and statistical learning in economic research – selected challenges

Wnioskowanie statystyczne i uczenie statystyczne w badaniach ekonomicznych – wybrane wyzwania

Abstract

This paper presents the main trends and the most important challenges related to the use of classical and modern statistical methods in economic research. The first part describes the key changes in quantitative economic and social research, including those related to technological progress. The main features of this evolution are the development of research infrastructure and changes in relations between researchers. The second discusses the need to find a compromise between easily accessible statistical data sets, advanced statistical software to analyse them, and the formal requirements of statistical inference. The third part details the essence and principles of statistical learning and presents a panorama of statistical learning methods. This results in the formulation of a paradigm of statistical learning for conducting modern statistical research.

Keywords: quantitative methods, statistical inference, technological progress, hypothesis testing, statistical learning.

JEL: A12, C10, C12, C18, C38

Streszczenie

W pracy przedstawiono główne tendencje i najważniejsze wyzwania, związane ze stosowaniem klasycznych i współczesnych metod statystycznych w badaniach ekonomicznych. W pierwszej części został przedstawiony opis kluczowych przemian w ilościowych badaniach ekonomicznych i społecznych, między innymi związanych z postępem technologicznym. Głównymi cechami tej ewolucji są: rozwój infrastruktury badawczej oraz zmiany w relacjach między badaczami. Część druga poświęcona została dyskusji o konieczności znalezienia kompromisu pomiędzy łatwo dostępnymi zbiorami danych statystycznych i zaawansowanym oprogramowaniem statystycznym do ich analizy a formalnymi wymogami wnioskowania statystycznego. W trzeciej części przedstawiono istotę i zasady uczenia statystycznego oraz panoramę statystycznych metod uczenia. W efekcie prowadzi to do sformułowania paradygmatu uczenia statystycznego dla prowadzenia współczesnych badań statystycznych. Sprostanie współczesnym wymogom stojącym przed badaniami ekonomicznymi stanowi wyzwanie dla zastosowań metod ilościowych.

Słowa kluczowe: wnioskowanie statystyczne, metody ilościowe, postęp technologiczny, testowanie hipotez, uczenie statystyczne.

JEL: A12, C10, C12, C18, C38

1. Introduction

There have been significant changes in the application of statistical methods in economic research over the last two decades. To a large extent, this applies to quantitative methods, including statistical methods. These changes pose challenges for researchers using previously commonly used statistical methods (as well as other quantitative methods). This article reviews and discusses the main challenges of using statistical methods. The work focuses on three main challenges. The first is adapting to the changes resulting from the development of economic theories and the resulting philosophy and methodology of economic research. These are closely related to technological changes in conducting empirical research intended to verify hypotheses. The second challenge is to deepen the understanding of the principles and rules of statistical inference used in economic research. The third challenge concerns the need to implement current achievements in the development of IT tools in statistics.

The article is accordingly divided into three main parts. The first part discusses current trends in the development of quantitative methods. These are mainly the result of technological progress. One of them is the growing role of research conducted in the framework of the positive economy. This, however, is in no way meant to depreciate the normative economy. The second trend is the growing role of interdisciplinary research in the economic sciences. The third clear trend is the increasingly rigid formalization of these sciences. The second part discusses the main challenges associated with classical statistical inference. These have to do with guaranteeing the reliability of statistical inference, and the downside of the increasing popularity of statistics in scientific research. Attention was also paid to the implications for statistical research of the use of online panel surveys and studies which are not based on random sampling and incorporate administrative records. The third part is devoted to statistical learning as a response to the challenges associated with the use of artificial intelligence (AI) methods in economic research. The essence of statistical learning and its general principles are presented. These methods significantly expand the scope and depth of statistical analyses. This leads to a new paradigm of statistical research, viz., the paradigm of statistical learning.

2. Some current trends in the development of quantitative methods

Two fundamental tendencies have influenced research in the economic sciences and some other disciplines formally classified as social sciences.

The first tendency is derived from the fact that mainstream research in the economic sciences, including economics and finance, has not been able to generate accurate forecasts or formulate convincing explanations of certain economic processes, particularly during periods of disruptive changes in the economy and the financial markets. Nor has it come up with a satisfactory explanation for the behaviour of economic agents. Criticism of classical economic and financial theories, predicated on rationality (*homo oeconomicus*), intensified after the global financial crisis of 2007–2008 (see e.g., Colander *et al*, 2009; Marglin 2010). The second tendency, discussed below, is derived from rapid and transformative advances in technology.

The present authors contend (see Jajuga, 2019) that the three main features of the application of quantitative methods in contemporary economic research will continue to influence future research.

1) The increasing importance of positive (descriptive) research in the economic sciences (see e.g., Banerjee, Duflo, 2009; 2014; Nordhaus, 2018; Card, 2009; Finkelstein, Gentzkow, Williams, 2016).

Normative economic research may not reflect real-world problems because:

- it may be based on counterfactual assumptions;
- it is often inflexible and unadaptable to dynamic changes in the economy and the financial markets;
- it pays insufficient attention to solving real, as opposed to abstract, economic problems, which are more amenable to a purely quantitative approach.

This statement can be supported by the fact that over the last few years, the Nobel Memorial Prize in Economic Sciences has been awarded for scientific achievements that have impacted real problems (which is a feature of positive research). Take the following years:

2023 – Claudia Goldin – 'for having advanced our understanding of women's labor market outcomes';

2022 – Ben Bernanke, Douglas Diamond, Philip Dybvig – 'for research on banks and financial crises';

2021 – David Card – 'for his empirical contributions to labor economics';

2019 – Abhijit Banerjee, Esther Duflo, Michael Kremer – 'for their experimental approach to alleviating global poverty';

2015 – Angus Deaton – 'for his analysis of consumption, poverty and welfare';

2013 – Eugene Fama, Lars Peter Hansen, Robert Shiller – 'for their empirical analysis of asset prices'.

2) The interdisciplinarity of research in the economic sciences (see e.g.: Angrist *et al.,* 2020; Fourcade *et al*., 2015; Lanzano *et al*., 2021; Truc *et al*., 2021; Burnewicz, 2021).

Research on economic processes is increasingly drawing on the achievements of other scientific disciplines. This is due to the following reasons:

- some scientific disciplines allow for a more in-depth analysis of the behaviour of economic agents, e.g., psychology, neurophysiology (the impact of brain functioning on human behaviour), sociology (the impact of interacting in social networks on human behaviour), and cultural anthropology (the impact of culture on human behaviour);
- changes in the global economy impose methodological requirements, forcing the use of other areas of knowledge.

3) Formalization of the economic sciences.

This feature is essential as far as quantitative methods are concerned. The history of scientific achievements, notably those recognized by awarding the Nobel Memorial Prize in Economic Sciences, shows an extensive formalization of economic and financial theories using a quantitative approach. Obviously, an appropriate level of mathematical formalization is necessary in order to maintain an acceptable level of precision and to avoid the chaotic narratives that all too often pass for 'qualitative' research. Clearly, both qualitative and quantitative research are required. However, quantitative research is sometimes omitted, even when crucial to the research topic, with the result that unsubstantiated generalities are produced and labelled 'qualitative' research.

However, excessive formalization can lead to model risk. A model is a theoretical construct that represents an analysed case with only its essential elements included. Therefore, a model can be considered a simplification of a part of the real world. It is an essential element of the scientific method. The term 'model risk', coined in finance by Derman (1996), denotes the risk of a formal model being inadequate to solve real-world problems. However, the problem of model risk was known before the term was coined. It is worth mentioning the statement of Robert Merton (Merton, 1994) on finance:

'At times the mathematics of the models become too interesting and we lose sight of the models' ultimate purpose. The mathematics of the models are precise, but the models are not, only approximations to the complex, real world. Their accuracy as a useful approximation to that world varies considerably across time and place. The practitioner should therefore apply the models only tentatively, assessing their limitations carefully in each application.

There are many reasons why models proposed in economics and finance research may not work in the real world. First, theoretical insights and models are not universal ('There is nothing like one size fits all'); what is suitable for one case may not work in another. Secondly, human behaviour is unpredictable; people (not models) run the economy and the financial markets. Real-world results will therefore only correspond with model predictions in some cases.

The models should comply with two essential requirements: robustness to market changes and transparency for the model's end-user.

As mentioned, the rapid and transformative technological progress is the second fundamental tendency (Bogdanienko, 2018; Jajuga, 2023). This has had a major impact on research in the economic sciences and other social science disciplines. Technological progress has made the use of quantitative methods more accessible and more effective. The main characteristics of this progress are:

- faster computer operations;
- more data available for processing (Big Data);
- increased transmission speeds and bandwidths;
- more interpersonal relationships through social networks.

The technological development of the last two decades has changed scientific research. The main characteristics of this evolution can be divided into two groups: changes in research infrastructure; and changes in the relationships between researchers. The evolution of research infrastructures includes the following elements.

1) Increasing the availability of scientific texts.

This has been made possible by:

- an increase in the number of journals publishing in the Open Access system;

- unlimited possibilities for sharing research results (including draft versions) on various portals, e.g., ResearchGate, Academia.edu, SSRN;
- increased access to publications through subscriptions to scientific institutions;
- systems providing information on the latest publications tailored to individual preferences.

2) Increasing the availability of data.

Access to data is crucial for conducting empirical research in the economic sciences. Increases in the scale of data acquisition, characterized by volume, variety, velocity, and veracity (the 4Vs of Big Data), have made it more available. However, reliability remains the main problem.

3) Increasing the availability of data analysis tools.

The availability of IT tools containing statistical and econometric software is enormous today. Some are free (e.g., the R program, which is crucial for statistical calculations), and most are accessed through cloud computing.

4) Increasing the availability of tools facilitating the preparation of a scientific text.

The tools in this group include automatic literature reviews and linguistic verifications of texts.

5) Increasing the ability to conduct research in the cloud.

This is a result of developing cloud computing solutions. These are becoming a standard tool for classic calculations and data processing and storage operations. The second group of evolutions in how scientific research is conducted are changes in the relationship between researchers. This process began a long time ago with the development of the internet and increased contacts between researchers. Initially, the primary source of contact was e-mail; now, there are various communicators and conference systems, such as Zoom and MsTeams. These systems were known before the pandemic, but the pandemic accelerated and facilitated scientific discussions (e.g., through online conferences).

Technological progress has impacted the development of the quantitative methods used in economic research in recent decades. This has resulted in:

- facilitating the application of statistical methods;
- developing increasingly complicated exploratory methods. However, these are still mathematical methods based on classical statistical concepts, e.g., correlation (within the broad meaning of the term).

Data analysis methods (called statistical learning methods by statisticians and machine learning methods by computer scientists) have likewise evolved. The methods used in the economic sciences (research) and the other social sciences are a conglomeration of two research trends, viz.:

- a trend developed over several decades in statistics and econometrics, both in statistical inference methods and exploratory methods (the latter in the descriptive approach);
- a trend based on combining the efforts of scientists in applied mathematics and statistics with those of IT scientists. This has resulted in the development of artificial neural networks (and similar methods).

For scientific research purposes, exploratory methods have a significant advantage in predictive power. However, their severe limitation (or complete lack) of cognitive values is a major disadvantage. It is these values that enable phenomena, and not merely their practical effectiveness, to be understood. This is especially true of artificial neural networks operating on a 'black box' basis.

The need to weigh the ready availability to quantitative methods against the difficulty in explaining the results they generate constitutes a major challenge to the use of quantitative methods in economic research.

3. Statistical inference – assumptions and practices

3.1. Does statistics pay a price for its increasing popularity?

Over the last few decades, statistics has become increasingly popular among researchers representing various sciences. Additionally, it attracts the curiosity and interest of those who wish to analyse many data sources independently. The development of data processing methods and technologies, the online availability of ever greater quantities of data, and the increasing tendency to replace qualitative descriptions and analyses with quantitative ones have all significantly contributed to the popularity of statistics. With slight exaggeration, this phenomenon is a rebirth of statistics. However, this is about more than just rediscovering what has been known since the works of Ronald Fisher, Jerzy Neyman, and Egon Pearson at the beginning of the 20th century. There are new opportunities for using statistics but also new challenges. Large data sets require new methods and analysis techniques, including big data. Many researchers point out that big data should be primarily understood as a probabilistic method for identifying patterns in large data sets.

Devising new data analysis methods and techniques necessitates a return to the very essence of statistical description and inference. Regardless of the size of the data sets and the methodology used to analyse them, detecting regularities of the kind that create new knowledge remains the principal objective of statistics. This refers to both statistical description and statistical inference. The latter is beset with more challenges, as sample studies are becoming more common in more areas than ever before.

Any tool that becomes increasingly popular in science is liable to misuse. This is because it may promise more than it can deliver, and because researchers may be tempted to use it universally, regardless of its limitations. This is the case with modern statistics. Easy access to internet data files of varying quality, combined with user-friendly statistical software, encourage the application of statistical methods without any consideration of the theory of statistics and the strict mathematical assumption of inference. Users sometimes ignore those foundational assumptions or may simply be unaware of them. This mindset can lead to unreliable or unjustified conclusions about the populations being analysed unless the requirements connected with the sample are met.

Medical researcher John P. A. Ioannidis was one of the first scientists who noticed and discussed the cost of the increasing popularity of statistics. Ioannidis presumably intentionally gave his 2005 paper the ominous title: 'Why Most Published Research Findings Are False' (Ioannidis, 2005). In the following years, some of his followers, after studying the issue carefully, claimed that errors arising from practical applications of statistical methods accounted for the lack of credibility of a significant portion of scientific research (see Gelman and Loken, 2014). In 2016, the American Statistical Association issued a statement on p-values and statistical significance, indicating that the statistical inference problem was recognized as severe (see Wasserstein and Lazar, 2016). In addition, *The American Statistician*, a widely respected journal, published more than 40 papers on statistical significance and the role of p-value in hypothesis testing in March 2019. There seems to be increasing awareness among scientists about the consequences of improper utilization of statistical methods and inadequate interpretation of research findings based on statistical inference. This problem is undeniably one of the main challenges facing empirical statistics. Nevertheless, many of those who use statistical techniques are either ignorant of, or unconcerned with, the mathematical background and theoretical assumptions underpinning estimation and hypotheses testing. One of the most widely discussed consequences of this mindset is the replication crisis affecting the social sciences, policy analysis, and other branches of science (see e.g. Gelman, 2018).

3.2. Underlying problems with the credibility of statistical inference

One of the main reasons for growing concerns about the credibility of research findings based on statistical inference is – perhaps paradoxically – the rapidly increasing popularity of statistical methodology in practice. This is a result of the sustained progress in IT technologies designed to collect, process, and analyse data sets of various kinds and sizes. The enormous advances in both statistical software and computer hardware offer unprecedented ways for quick and efficient statistical data analysis. Perhaps even more critical is that the software is relatively simple to use and does not require an advanced knowledge of statistics. As a result, an increasing proportion of both scientists and non-scientists are using statistical software without exhibiting any particular interest in the theoretical background of the methods this software employs. This confidence in statistical software is bolstered by effective promotion and other marketing activities aimed to present it as being useful in all circumstances, regardless of the assumptions and requirements of statistical theory. All this creates an environment in which statistical software is strongly advertised and promoted, and is increasingly user-friendly. However, many potential users have access to various statistical data sets and are looking for methods to facilitate understanding of the data or discover new knowledge. If the users are not interested in the background theory of statistics, the inevitable results will be misunderstandings and even scientific fraud.

Another reason for concern about the reputation of statistical surveys is the insistence of some researchers on using non-probability samples for statistical inference accompanied by measures of accuracy and diversity which refer to the probability model of repeated random sampling. The most popular non-random samples used nowadays are purposive samples, convenience samples, and quota samples. Although they may be helpful and effective in some circumstances, none of these techniques allows for using probability concepts to interpret survey findings. This means that sampling distributions of sample statistics cannot be obtained, and the inference results lose their probability context. This is a consequence of generating sample observations by a nonprobability mechanism which does not allow for frequency interpretation of probability. '*In such cases, we should accept the hard truth that statistical inference is not possible*', conclude Hirschauer *et al.* (2021, p. 24). It is imperative that researchers be made aware of this limitation, and for their part, statisticians have to accept that there is a great need for other statistical methods and techniques which would not be controversial if applied to non-random samples. This seems to be one of the most critical challenges at present.

A slightly subtler problem that arises in applications of statistical inference relates to statistical significance – one of the fundamental notions present in the inference model. Statistical communities worldwide have expressed growing concern about the impact of statistical significance in research and about the role of p-values in deciding whether or not to reject tested hypotheses. The impact of p-values (and their dichotomization) in testing hypotheses is criticised for being overstated. It should be stressed that deciding whether to reject the null hypothesis on the sole criterion of whether the p-value is less than an arbitrarily assigned significance level (usually 5%) is a simplification that is difficult to rationalise. There is reason for the p-value, which is only a measure of discrepancy between the observed data and the null hypothesis, to be fixed, as is often the case in statistical software and practical applications. It needs to be borne in mind that the p-value calculated for a given sample only applies to that particular sample. No inference can be made about any other sample. The inherent uncertainty of any statistical test automatically precludes any certainty in the correctness of the decision of whether or to accept the hypothesis. Relying on dichotomous decisions based on the inequality $p<0.05$ without analysing the actual p-values has almost eliminated our ability to distinguish between statistical results and scientific conclusions, as Goodman (1999, p. 996) correctly points out. The p-value obtained in testing a hypothesis is usually the result of a random variation and factors that reflect deviations from the assumptions of the mathematical inference model. Therefore, a small p-value may indicate that the null hypothesis should be rejected, but sometimes such a value will result from other (nonprobability) errors affecting sample observations.

On the other hand, a p-value which exceeds 0.05 cannot be considered incontrovertible proof that the null hypothesis is true. This value only measures the degree of compatibility between the sample results and the value of the tested population parameter in the null hypothesis. No p-value is capable of deciding whether the null hypothesis is true. Confidence intervals are worth mentioning in this context as they may provide valuable information about the inference process without recourse to the p-value. For one thing, they allow for a careful examination of the size effect. A p-value, on its own, can result in a situation where a small size effect in a large sample generates the same p-value as a large size effect in a small sample[1].

Neglecting the above challenges of statistical significance and relying entirely on statistical software that does not verify any theoretical assumptions may cause serious problems, including a loss of trust in statistics or science in general.

3.3. Inference based on internet panels of respondents

Social media, electronic mailing lists, and other electronic devices offer enormous opportunities for contacting potential respondents representing given populations, and are consequently used for sampling. There is a tendency to use less costly and less time-consuming non-probability sampling techniques rather than stricter and more demanding probability ones. Opt-in online surveys have recently become the most popular sampling method. These are online surveys completed by volunteers recruited to fill in online questionnaires, sometimes in exchange for specific gifts. They constitute a convenient way of recruiting respondents and forming a statistical sample but they are not without their pitfalls and shortcomings[2].

First, there is the risk of committing coverage errors when non-probability sampling is frequently applied, as this method requires a high-quality sampling frame. The survey sample may consist solely of that portion of the population with internet access. This is likely to lead to biased estimates. Moreover, it would be extremely difficult to evaluate the size of the bias and its sign (positive or negative). Increasing the sample size will not reduce the error (or bias) in these cases. Hirschauer *et al.* (2020) emphasize that even if it were reasonable to assume a complete sampling frame, it would still be difficult to show that those who agreed to cooperate and those who refused were not systematically different. It only takes a few characteristics associated with the key study variables to differ between the two segments of the population to make a sample biased. A large sample size cannot remedy this situation. 'A large n means nothing if the sampling is biased' (Cochrane, 2015, p. 17). A sampling error consists of two components: random error, which is a diminishing function of sample size; and non-random error, sometimes systematic (bias), which is not a function of sample size.

Secondly, using a non-probability sample excludes any interpretation of the properties of estimators or results of inference in terms of probability or long-run frequencies. In particular, this holds true for the confidence level in interval estimation, and the significance level and p-value in hypothesis testing. This consequence should be given serious consideration because statistical inference

[1] See Goodman (1999). For more on the impact of large sample sizes in statistical hypothesis testing, see Szreder (2022).

[2] A remarkable document on the properties of opt-in online surveys was issued by the American Association for Public Opinion Research (AAPOR) in 2010 (see: AAPOR Report on Online Panels, 2010). It includes some warnings related to inference based on such panels.

relates directly to the model and not the real world. As Kass (2011, p. 2) explains, 'Conclusions are drawn by applying a statistical inference technique, which is a theoretical construct, to some real data'. Amrhein, Trafimow and Greenland (2019, p. 262) provide a similar understanding of statistical inference and claim that 'statistical inference is a thought experiment, describing the predictive performance of models about reality'. There is therefore no justification for claiming that statistical inference is not concerned with the assumptions underlying the model and its requirements. Internet panels may violate some of its crucial assumptions.

3.4. Non-sample data and their impact

Statistics was never meant to be a purely theoretical science. Therefore, statisticians should be ready to respond to the changing needs of researchers who use statistics to understand better and improve the real world. Consequently, if non-probability sampling techniques are used to obtain, e.g., opt-in online samples, statisticians cannot confine themselves to criticism of these sorts of samples. Fortunately, they do not. There are various recommendations given by statisticians in this respect. One of the most obvious is to look at both the sample and other useful secondary data sources in order to determine how representative the sample is and/or to make it more representative.

There are other approaches that aim to reduce selection bias (e.g., propensity score matching). However, it is most commonly recommended to use external (non-sample) information in order to improve or strengthen the sample data. External data sources may include administrative records and official registers. Although the theory of statistical inference does not give substantive support to making inferences based on such combinations of data sets, it is a developing branch of statistics (see, e.g., Wallgren A., Wallgren B., 2007). Weights and calibration techniques are also used for adjusting for nonresponse or correcting the sample structure to improve inference precision.

Access to an increasing number of internet data files and data sets will continue to stimulate the use of combined sample and non-sample information to improve the quality of inference. This also applies to big data. As is the case with some other internet-derived statistical data, big data may be unstructured, messy, and of poor quality. However, if processed carefully, it may constitute a valuable external data source in statistical studies. Big data, mainly administrative records, may complement sample data and improve inference quality. The main challenge in this respect is the lack of theoretical models for making inferences from large data sets. This is why a new data science has evolved in parallel to classical statistical analysis and inference methods over the last few decades. This data science concentrates on mining large data sets to identify regularities and patterns. This methodology not only employs statistics but also machine learning algorithms, neural networks, and increasingly advanced AI techniques. Statistics and data science have so far been complementary rather than competitive.

4. Statistical learning as a response to the challenges facing economic investigations

4.1. The essence of statistical learning

The formal definition of statistical learning was introduced by Vladimir Vapnik (1998). It describes the general statistical learning model by giving three of its components:

- a generator of random vectors drawn independently from a fixed but unknown probability;
- distribution, defined by a probability distribution function *F*(*x*);
- a supervisor who returns an output value *y* to every input vector *x*, according to a conditional;
- distribution *F(y/x)*, also fixed but unknown;
- a learning machine capable of implementing a set of functions $f(x,\alpha)$, $\alpha \in A$, where *A* is the set of parameters.

The problem with learning is having to choose the function that best approximates the supervisor's response from the set $f(\mathrm{x},\alpha)$, $\alpha \in A$. The selection of the desired function is based on a training set of independent and identically distributed observations. The learning machine plays a critical role in the given definition. As the data generator is usually random, the function $f(x,\alpha)$, combining input data with output, usually has an indeterministic, stochastic character, machine learning is most often understood as statistical learning, based on probabilistic rules (Hastie, Tibshirani, Friedman, 2009).

The choice of the best available approximation of the actual function to the supervisor's response is based on solving the problem of risk minimization. This requires some measure of the loss or discrepancy $L(y, f(x,\alpha))$ between the response *y* of the supervisor to a given input *x* and the response $f(x,\alpha)$ provided by the learning machine. The expected value of the loss is given by a risk functional. The mathematical task is to find the function $f(x,\alpha_0)$ that minimizes the given functional over the classes of functions $f(x,\alpha)$, $\alpha \in A$ when the joint probability distribution function $F(x,y)$ is unknown and the only available information is contained in the learning set. In order to minimize the risk functional (which has an unknown distribution function), its theoretical form is replaced by an empirical risk functional constructed based on the training set. This is known as the empirical risk minimization inductive principle. An inductive principle defines a learning process in which a learning machine chooses an approximation using this principle for any set of observations. The empirical risk minimization principle plays a crucial role in learning theory. This principle is quite general. The classical methods for solving a specific learning problem, e.g., the least-square method in the problem of regression estimation or the maximum likelihood method in the problem of density estimation, are realizations of this general inductive principle for specific loss functions (Vapnik, 1998). This formulation of the learning problem is rather broad, and encompasses many specific problems, including pattern recognition, regression, density estimation, and classification problems.

The above brief outline of the essence of statistical learning shows that the probabilistic nature of socio-economic processes is the most general platform for combining data analysis methods based on AI models with the classical statistical methods of modelling and forecasting these processes.

4.2. General principles of statistical learning

The classical approach to statistical learning uses a dependent variable Y, understood as the response variable, and k explanatory variables (predictors) X_1, X_2, X_k. A relationship between Y and $X = (X_1, X_2, X_k)$ is assumed and generally write as $Y = f(X) + \xi$, where f is an unknown function associating Y with X and ξ is a random component. The essence of statistical learning is to estimate the function f using the function h, which is one of the hypotheses belonging to the hypothesis space H, concerning the unknown function f (Hastie, Tibshirani, Friedman, 2009).

Approximating the actual f function is a critical statistical learning problem. It is estimated using a data set (called the training data or training set) that contains input and output information (x_{ij}, y_i). In other words, a function, for which $Y \approx$, is detected for any pair of observations from the set (X,Y). Either a parametric or non-parametric approach can be used.

The parametric statistical learning method involves specifying the analytical form of the function. This is assumed to be a linear multivariate model in the simplest and most common case. The partial regression coefficients are then estimated using the data from the training set, most often by using the least squares method. Obviously, parametric statistical learning provides 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, a form of the function f that fits as closely as possible to the training set data has to be found. The non-parametric approach can have a significant advantage over the parametric approach in that it can fit the empirical data more accurately by avoiding the assumption of a specific analytical form of the function f. The parametric approach carries the risk that the analytical form of function deviates significantly from the actual function f, which links the predictors with the result variable. Nevertheless, the non-parametric approach has the disadvantage of not reducing the number of estimated parameters by eliminating the insignificant ones. It therefore requires a much larger training set (James *et al.*, 2013).

The advantage of statistical learning methods is that they include both parametric and non-parametric approaches. Therefore, they go far beyond the standard applications of classical analysis. Different statistical learning methods can be used for a single training set and the best method, based on the empirical criteria, can be selected. In statistical learning, no one method dominates in every possible data set. One method may work best for a particular dataset, but another method may work better on a slightly different dataset. Therefore, deciding which method will yield the best results for a specific dataset is an important task. However, it is also one of the most challenging decisions when applying statistical learning methods.

4.3. Learning methods as an extension of the scope of statistical analysis

The overview of statistical learning methods presented below is incomplete. However, it is intended to show that the palette of statistical learning methods is much broader than the collection of classical mathematical statistics methods. The parameters of the linear regression model, logistic regression model, and linear discriminant function can be estimated from the random sample data and learned from the training set by a machine. An alternative to a parametric, linear approach is non-parametric non-linear models, e.g., the k-nearest neighbour method. Hence, there are alternatives between parametric and non-parametric approaches and between classical and machine learning methods in empirical investigations.

Extensions and generalizations of the parametric methods are tending towards combining the parametric and non-parametric approaches. This is expressed by the generalized linear models and generalized additive models. These are all semiparametric. These models can also be estimated by the sample or learned by the training set.

Another means of replacing the parametric approach with non-parametric ones is through kernel estimators and classifiers. Kernel estimation enables the distribution density of a random variable to be determined from the a sample or learning set. Estimating kernel density by unsupervised learning leads to a family of non-parametric classification procedures called kernel classifiers. The kernel classifier enables non-linear relationships in the data sets being analysed to be reconciled with the linear nature of the classifiers. An alternative is to use mixture models.

An important group of statistical learning methods creates spanning tree methods based on the principle of recursive division. These include the classification and regression methods based on decision trees, which are implemented using various algorithms. The multi-model approach is an extension of individual models, and has a relatively simple architecture. It aggregates several individual models into a model that is usually more accurate than any of the models it contains. The result is improved predictive power. Among the many known algorithms of the multi-model approach, the most important are the bootstrap aggregation method (bagging), the boosting method, and random forests.

The support vector machine (SVM) method is an original and widely used method for classification and regression. It applies the idea of classification using discriminatory hyperplanes. This is a very flexible method. It also makes it possible to find a solution when classes in the training set cannot be separated linearly, i.e., it can be considered in a linear and non-linear variant.

Artificial neural networks are the oldest and most commonly used AI statistical learning methods. A properly constructed artificial neural network can solve many regression, classification, and forecasting problems.

Choosing an adequate network architecture is essential in building a proper neural network. There are three main types of architectures: one-way, recursive, and cellular networks. Each type is built using different procedures. However, network learning has always been indeterministic, i.e., the result of the learning process is never completely unambiguous.

It should be noted that learned spanning tree methods, multi-model procedures in the form of bagging, boosting, and random forests algorithms, as well as support vector machines and artificial neural networks, have no equivalents in statistical inference. Thus, these methods enrich the commonly used set of statistical methods in socio-economic research, making an essential contribution to the development of economic research.

4.4. Statistical learning paradigm

The premises and principles of learning from data, presented above, allow for the formulation of a statistical learning paradigm (Pociecha, 2021). In the mid-1990s, Vladimir Vapnik (1998) drew attention to the revolution taking place in statistical research methodology by formulating a new paradigm of this research, viz., the paradigm of statistical learning. Vapnik noticed that the classical (Fisherian) paradigm of mathematical statistics, formulated in the 1920s and 1930s, is being replaced by a new paradigm. According to Vapnik, the problem of learning from data is so general that almost any statistical problem can be formulated in machine learning theory. Because data are random, machine learning becomes statistical learning.

The principles of the statistical learning paradigm will be presented in the context of the differences between it and the classical paradigm of statistical inference. The starting point of the classical paradigm is probability theory and its basic concepts, including random events, the axioms of probability theory, the random variable, and its distribution. The essence of mathematical statistics 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 starting point in the statistical learning paradigm is the opposite: ‘Let the data speak for itself’, ‘we learn from the data’, and the starting point is the dataset. This paradigm is based on neo-positivist beliefs, according to which all knowledge comes from empirical data. Neo-positivists assume that experience is the source of all knowledge about the real world (Pociecha, 2020).

A key concept in the paradigm of statistical inference is the notion of population and random sample. The statistical learning paradigm ignores the notion of population and sample. Instead, it assumes that there is a set of empirical data which is sufficiently large for making accurate predictions and drawing valid inferences about the reality from which these data come. This set can be a sample, even a small one, drawn according to the rules of the sampling method, it can be a set of cleaned data, according to the rules of data cleaning, and it can also be a Big Data set, including streaming data.

The essence of the statistical learning paradigm is the creation of self-learning systems, i.e. systems that improve automatically through experience. Statistical learning in the supervised version involves providing the algorithm with a set of input-output pairs to find an unknown function by mapping input data to output data, with the accuracy of the minimized mean square error of the estimate or the mean prediction error. The actual function in the statistical learning paradigm is a black box as a parametrized or non-parametrized function; it can even be a non–algorithmic procedure. The primary difference between statistical estimation and statistical learning is that the former involves estimating the parameters of a predetermined function, whereas the latter selects the form of this function and its parameters.

The concept and probability theory play a crucial role in the statistical inference paradigm. The statistical learning paradigm's role is secondary as there are severe doubts as to whether the training set can be considered a random data set. In this sense, the statistical learning paradigm is approaching the descriptive statistics paradigm.

Bradley Efron and Trevor Hastie, in their groundbreaking *Computer Age Statistical Inference,* emphasize that 'Statistical inference is an unusually wide ranking discipline, located as it is at the triple-point of mathematics, empirical science and philosophy' (Efron and Hastie, 2016, p. 15). They also noticed that the centre of statistical methods has shifted from a traditional mathematical and logical approach to a more computational focus over the past sixty years. There is room for both the classical paradigm of mathematical statistics and the statistical learning paradigm based on AI models in conducting statistical research.

5. Concluding remarks

Increasing the quantity of data available to researchers, together with efficient statistical software for collecting, processing and analysing data, has brought new challenges to anyone who makes use of statistical inference. These challenges relate mainly to those theoretical (mathematical) assumptions of statistical inference that tend to be neglected or ignored by some researchers when they have access to variety of data files. As a consequence, not all scientific findings based on non-random samples or a combination of administrative data and sample data can be regarded as reliable.

The theory of statistical inference has not yet created a sufficiently robust mathematical background for incorporating every kind of data set, including big data. For this reason, it is important to stress the formal requirements that have to be met in applications of estimation techniques or hypothesis testing. Moreover, inferences should be treated with caution and the underlying assumptions of the methods used need to be taken into account.

There are differences in the research goals that can be achieved through statistical inference and statistical learning paradigms. Statistical research using statistical inference focuses primarily on explaining the relationships between the variables being analysed. Empirical investigations conducted within statistical learning involve building on their basic forecasts which would be as accurate as possible; nevertheless, their analytical and interpretative power could be limited.

The statistical learning paradigm is a universal research platform as it has absorbed the statistical inference paradigm at the expense of weakening its original assumptions. This paradigm offers an excellent opportunity to use the computing capabilities of modern computers and large data sets produced by contemporary socio-economic life.

References:

American Association for Public Opinion Research. (2010). *AAPOR Report on Online Panels*. https://www.aapor.org/Education-Resources/Reports/Report-on-Online-Panels.aspx.

Amrhein, V., Trafimow, D., Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, *73*(Sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Angrist J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2020). Inside Job or Deep Impact? Extramural Citations and the Influence of Economic Scholarship. *Journal of Economic Literature*, *58*(1), 3–52.

Banerjee, A., Duflo, E. (2009). The Experimental Approach to Development Economics. *Annual Review of Economics*, *1*, 151–178.

Banerjee, A., Duflo, E. (2014). Do Firms Want to Borrow More? Testing Credit Constraints Using a Directed Lending Program. *Review of Economic Studies*, *81*(2), 572–607.

Bogdanienko, J. (2018). *Istota i problemy poznania naukowego*. CeDeWu.

Burnewicz, J. (2021). *Filozofia i metodologia nauk ekonomicznych*. PWN.

Card, D. (2009). Immigration and Inequality. *American Economic Review*, *99*(2), 1–21.

Cochran, J. (2015). *ASA Leaders Reminisce. Peter (Tony) Lachenbruch*.

Colander, D., Follmer, H., Haas, A., Goldberg, M.D., Juselius, K., Kirman, A., Lux, T., Sloth, B. (2009). The financial crisis and the systemic failure of academic economics. *Kiel Working Paper*, 1489.

Derman, E. (1996), *Model risk*. Quantitative Strategies Research Notes. Goldman Sachs.

Efron, B., Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press.

Finkelstein, A., Gentzkow, M., Williams, H. (2016). Sources of Geographic Variation in Health Care: Evidence from Patient Migration. *Quarterly Journal of Economics*, *131*(4), 1681–1726.

Fourcade, M., Ollion, E., Yann, A. (2015). The Superiority of Economists. *Journal of Economic Perspectives*, *29*(1), 89–114.

Gelman, A. (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*, *44*(1), 16–23.

Gelman, A., Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, *102*, 460–465.

Goodman, S.N. (1999). Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine*, *12*, 995–1021.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Hirschauer, N., Grüner, S., Mushoff, O., Becker, C., Jantsch, A. (2020). Can p-values be meaningfully interpreted without random sampling? *Statistics Surveys*, *14*, 71–91.

Hirschauer, N., Grüner, S., Mushoff, O., Becker, C., Jantsch, A. (2021). Inference using non-random samples? Stop right there! *Significance*, *18*(5), 20–24.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine, 2.*

Jajuga, K., (2019). Nauki ekonomiczne – dylematy klasyfikacji dyscyplin. Tendencje zmian. *Ewolucja nauk ekonomicznych,* 140–150. PAN.

Jajuga, K., (2023). Badania naukowe w ekonomii i finansach a rozwój technologiczny. W: M.J. Nowak, R. Rakoczy (red.), *Nauka polska. Szanse, bariery i wyzwania* (s. 153–170). Wydawnictwo Naukowe Scholar.

James, G., Witten D., Hastie T., Robert Tibshirani R. (2013). *An Introduction to Statistical Learning.* Springer.

Kass, R.E. (2011). Statistical Inference: The Big Picture. *Statistical Science*, *26*(1), 1–9.

Lanzano, C., Navarra C., Vallino E. (2021). Interdisciplinarity and the future of development studies after the 2019 Nobel Prize in economics. *Anthropologie & développement*, 315–329.

Marglin, S. (2010). *The Dismal Science: How Thinking Like an Economist Undermines Community*, Harvard University Press.

Merton, R.C. (1994). Influence of Mathematical Models in Finance on Practice: Past, Present and Future. *Philosophical Transactions of The Royal Society*, *347*(1684), 451–463.

Nordhaus, W D. (2018). Evolution of Modeling of the Economics of Global Warming: Changes in the DICE model, 1992–2017. *Climatic Change*, *148*(4), 623–640.

Pociecha, J. (2020). Philosophical foundations of statistical research. *Przegląd Statystyczny*, *67*(3), 195–211.

Pociecha, J. (2021). The paradigm of statistical inference and the paradigm of statistical. *Przegląd Statystyczny*, *68*(1), 1–16.

Szreder, M. (2022). Szanse i iluzje dotyczące korzystania z dużych prób we wnioskowaniu statystyczny. *Wiadomości Statystyczne. Statistical Review, 8*, 1–16.

Truc, A., Santerre O., Gingras Y., Claveau F. (2021). *The Interdisciplinarity of Economics*, www.ssrn.org

Vapnik, V. (1998). *Statistical Learning Theory,* Wiley.

Wallgren, A., Wallgren B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*, Wiley.

Wasserstein, R, Lazar N. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician*, 70, 129–33.