REVISITING FUNDAMENTAL IDEAS FOR STATISTICS EDUCATION FROM THE PERSPECTIVE OF MACHINE LEARNING AND ITS APPLICATIONS

Rolf Biehler
Paderborn University, Germany
biehler@math.upb.de

Fundamental ideas (Burrill & Biehler, 2011) have been specified to structure the curriculum across different age levels and to focus on and revisit the most important ideas. Such ideas must be based on an epistemological analysis of the scientific domain of statistics and its applications. Because these domains are changing, fundamental ideas have to be revisited from time to time. In particular, data science and machine learning have led to new methods and applications in society that must be considered for updating fundamental ideas.

INTRODUCTION

In mathematics education, specifying fundamental ideas of a discipline dates back to Jerome Bruner's educational philosophy and his vision of a spiral curriculum to be structured along with fundamental ideas that can be revisited and deepened at various grades (Bruner, 1960). In curriculum guidelines such as the GAISE (Guidelines for Assessment and Instruction in Statistics Education) reports (Bargagliotti et al., 2020; Franklin et al., 2005/2007), we find a similar philosophy. The same framework of statistics is used for all levels, and its aspects are revisited and deepened progressively at higher age levels. However, these guidelines do not use the notion of "fundamental ideas." In the mathematics education of the 1960s and 1970s and the new math reform movement (Phillips, 2015), fundamental ideas were narrowly taken from a view of mathematics as a discipline that was developed by the Bourbaki group (Bourbaki, 1950), emphasizing sets and mathematical structures. A newer interpretation assumes that fundamental ideas must be based on an epistemological analysis of the discipline, including its applications (Heymann, 2003). From this perspective, Burrill and Biehler (2011) specified fundamental ideas for statistics. This paper aims to update some of these ideas based on the new developments around data science, machine learning, and artificial intelligence. Because the field is rapidly developing, the contribution of this work is meant as a first draft to stimulate further discussion in the community of data science educators (statistics and computer science educators). We cannot see a consensus on the essence of data science. The characterization depends, among others, on whether one's background is more in computer science or statistics.

The presented analysis is based on recent curriculum work in *Project Data Science and Big Data at School* (www.prodabi.de/en) (Biehler & Schulte, 2018). Burrill and Biehler (2011) highlight, on the one hand, a process view of statistics such as specified in the PPDAC (Problem, Plan, Data, Analysis, and Conclusion) cycle (Wild & Pfannkuch, 1999) and determine, on the other hand, a set of fundamental ideas: data, variation, distribution, representation, association and modeling relations between two variables, probability models for data generation processes, sampling, and inference. We will briefly reconsider the process view and only three of the fundamental ideas because of limited space.

STATISTICAL PROCESSES

The process description of data science is often specified by the CRISP-DM (Cross Industry Standard Process for Data Mining; see Figure 1) cycle (Chapman et al., 2000). A similar scheme is presented by Berthold et al. (2020, p. 9). CRISP-DM has several differences from the PPDAC cycle. The following are essential:

- Data may be "already there" and may not be collected according to a plan; available data are considered in the light of a study problem.
- Business and data understanding are emphasized as a first step. The step may involve dialogue with clients, not only individual and autonomous statistical researchers.
- Data preparation and cleaning are emphasized as separate steps and need much more time and attention than in traditional applications (Wilkerson et al., 2021).
- Modeling is added as a separate step in the process cycle. Data science uses new types of algorithmic models. Prediction as a goal for modeling gets a new emphasis.

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

- Validation of the model is emphasized as a step. This includes distinguishing training and test data on which the initial model is validated, taking up the statistical idea of cross-validation (Refaeilzadeh et al., 2016).
- "Conclusions" as a final process step is highlighted. Whereas statistics aims at "knowledge," data science and computer science "deploy" models in a business and social context. This includes further monitoring an application with social responsibility and being aware of ethical questions, for instance, when implementing an automatic decision system to decide who gets a particular medical treatment or who gets a loan from a bank.

Statistics education must rethink which scheme is used to structure students' activities.

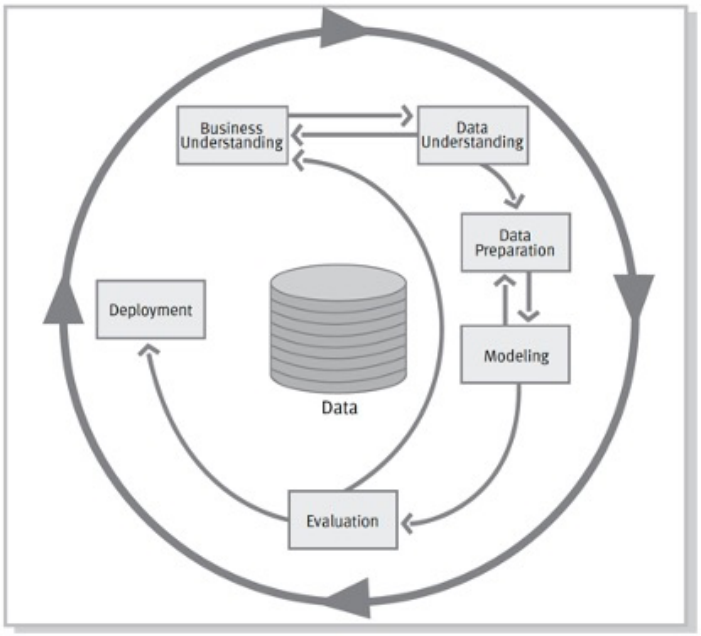

Figure. 1 CRISP-DM cycle

FUNDAMENTAL IDEAS: DATA

The role of data has fundamentally changed in business, society, and culture. Books such as Weigend (2017) or journals such as *Big Data & Society* (https://journals.sagepub.com/home/bds) reflect these developments. The multivariate nature of data is much more prominent. Tidy data as rectangular data tables with different variable types are an important standard, but data may have to be transformed into this scheme. Non-traditional data include data collected by sensors or by personal devices, transactional data (traffic, supermarket buys), images and texts, data scraped from webpages, data with geographic information, big data, and open data (data that is openly accessible, exploitable, editable, published with CC license).

Traditional statistics education often focuses on data from randomized controlled experiments or random samples of a well-defined population. Exploratory data analysis in the tradition of John Tukey (1977) has always been open to "dirty data" while remaining aware that the kind of conclusions one can draw depends on the quality of the data and deep knowledge of meta-data. This problem is exacerbated by the many available open and big data sets on the internet, whose origin and quality are often unclear.

Metadata and problems of measurement, including bias, require more emphasis. These problems are already well emphasized in statistical literacy education (e.g., see the Volume 16, Issue 1 special issue of the *Statistics Education Research Journal* in 2017). Instead of a direct measurement, often a proxy is measured. These problems can also lead to enormous distortions in interpretations or misuses in machine learning (ML) applications. A relevant example is provided by Obermeyer et al. (2019): the need for actual medical treatment of a person was approximately measured by the costs of past medical treatments for that person, which led to racial bias and discrimination against less affluent persons. This kind of bias is related to the measurement of a single variable. But distortion and bias can

also come from the set of variables that are selected for representing an object if this set is not carefully considered as a "model" that is different from reality. Models are simplified representations of an object for specific purposes. Although one could talk about the value of height as a person's model, nobody would take this as a replacement for a person. However, this may be different if a person is represented by a set of values of a dozen variables. In ML applications, persons are treated as if they only consist of this set of values. An important question is whether this model-like representation by a set of variables is adequate for a particular purpose. This has to be checked in the validation phase of the CRISP-DM cycle.

FUNDAMENTAL IDEAS: MODELING

In Burrill and Biehler (2011), modeling was split into two fundamental ideas of statistics education: (a) probability models for data-generating processes and (b) association and modeling relations between two variables. An essential notion in machine learning is predictive modeling. How can this be related to these fundamental ideas? Predictive modeling includes—as a rule—more than two variables, emphasizes model validation on new data as a crucial component, and emphasizes prediction more than other types of goals for models such as explanation. The role of probability models for data-generating purposes for assessing uncertainties in the model is often neglected, and statisticians criticize that dealing with probabilistic uncertainty is usually not given enough attention in ML practice (Friedrich et al., 2021). School students typically encounter probability models for data-generating processes in univariate situations only: the model is a (theoretical) probability model for a single random variable or a random experiment. Model validation with empirical data must already be emphasized even in this relatively simple situation. It is challenging, however, to make probability model validation accessible to school students when they cannot use statistical tests such as the χ^2-test or the Kolmogorov-Smirnov test. Gelman and Nolan (2002) discuss this problem concerning dice and coins, and Biehler (2005) and Biehler et al. (2017) refer to situations where a binomial probability model has to be validated against empirical data.

Predictive modeling comes—if we simplify—in two versions. From a set of predictor variables, either (a) a numerical variable is predicted (regression problem) or (b) a categorical variable is predicted (classification problem). Classification problems have not been part of statistics education but have recently gotten new attention. For instance, classification is not mentioned in the original GAISE framework (Franklin et al., 2005/2007), but in GAISE II (Franklin & Bargagliotti, 2020), an example of classification (which may lead to decision trees) is included at level C (high school).

Association and modeling related to two variables also depend on the type of the variables. Let us start with two numerical variables. At the school level, a basic approach is to fit a function to a scatterplot showing two numerical variables, either by eye-fitting or by some methods such as the least-squares criterion or the median-median line. If the least-squares line is regarded as the solution to an optimization problem, as "the best line," further validation may not be considered necessary. Visual residual analysis can do elementary validation of the model (see chapter 14 in Gould et al., 2020). However, validation of the model on new data is rarely practiced but would be essential as a preparation for the predictive modeling approach. Overfitting is a central problem in most ML approaches. This means that an ML algorithm provides good predictions on the training data but performs worse on test data. The fitting procedures take structures of the training data into account, which may be specific to the training data set or are just random noise and do not generalize to the test data (Tong, 2019). See also the discussion on overfitting and its remedies ("pruning") in the context of decision trees as a particular ML method for classification (Fleischer et al., 2022).

This facet can be experienced in the two-variable regression context if polynomials of increasing degrees are fitted to data in a scatterplot, where fit quality (measured, for instance, by the sum of squared residuals) increases by degree. However, as a rule, the fit of a higher degree polynomial is much worse on a new data set, whereas a linear function may better generalize to new data (https://towardsdatascience.com/too-many-terms-ruins-the-regression-7cf533a0c612). The sum of squared residuals or the variance of residuals could be reinterpreted and serve as a quality measure of a model fit even in ML contexts.

From a traditional point of view in statistics, the next step after a descriptive fit of least-squares lines would be to make a data generating model such as $Y_i = a \cdot X_i + \varepsilon_i$, where the ε_i are stochastically independent random variables with an expected value of 0, the same variance, and normally distributed.

This model (if valid) would allow inferences from the sample to the population, for instance, calculating confidence intervals. Students may conclude that this is not necessary to test the model on other data and to identify additional sources of uncertainty. This problem is related to a more general inference problem in ML than just sample-to-population inference (see the next section).

The next step in teaching modeling from the perspective of machine learning would be to predict the value of a "target variable" Y from several predictor variables X_1, …, X_n. Formulated abstractly, a ML algorithm is a multivariable function that computes the value of the target variable from a set of values of predictor variables. This can be visualized as a machine doing this computation. We may better speak of the *device representing an algorithm* than using the notion of a mathematical function. In mathematics education, the connotation of a function is assigning an input number by a simple formula to an output number.

Moreover, functions of several variables are usually not taught in any secondary mathematics curriculum. This usually is different in computer science education, where the algorithm-as-a-data-processing-machine metaphor is often used, even if algorithms are designed in off-line or "unplugged" activities (Battal et al., 2021). The notion of an *algorithmic model is essential here* (Breiman, 2001). In sum, the fundamental idea of modeling has to consider the *validation of models on new (test) data, criteria for quality of fit, overfitting, and algorithmic models*.

The case of two categorical, especially binary, variables is different from two numerical variables. In traditional approaches at the school level, on the one hand, measures of association for two categorical variables are taught, including the concept that two categorical variables may be statistically independent. On the other hand, problems related to Bayes' rule are often introduced in secondary curricula. In the context of tests in medicine (e.g., HIV or Corona), one binary variable "health status" has the values infected/not infected and the other binary variable "test result" has positive/negative results. We can reinterpret such a testing situation as a classification problem: if the test is positive, we predict "infected," and if the test is negative, we predict "not infected." Sensitivity (probability of a positive test if someone is infected) and specificity (probability of a negative test if someone is not infected) of such a test can be used to determine the specific misclassification rates and the overall misclassification rate, which is a weighted mean of the specific rates, with the weights coming from the infection/non-infection rate (the base rate). These types of errors can be directly related to what is called the *confusion matrix* in classification problems in ML contexts. The (inverse) probability of being infected when the test is positive is called "precision" in machine learning, where it is also a quality measure. Thus, a reinterpretation of "traditional" notions and their applications to ML problems would be a natural extension of current fundamental ideas and curricular approaches.

An extension of prediction by just one dichotomous variable to using a stepwise set of dichotomous, categorical variables or numerical variables is the idea of using decision trees with several predictor variables (binary, categorical, or numerical). This is exemplified in the GAISE report (Bargagliotti et al., 2020, p. 97). ML as such, however, starts if a computer program automatizes the process of finding a tree. We developed teaching material for decision trees for grades 6, 8/9, and 12, with different levels and facets (Biehler & Fleischer, 2021; Fleischer et al., 2022; Podworny et al., 2021).

FUNDAMENTAL IDEAS: SAMPLING AND INFERENCE

Another source of bias lies in the aim of ML to make inferences or predictions beyond the data set used for training. In statistics, the sample-to-population inference is an essential feature. With a random sample from a well-defined population, an inference can be made about the population. The uncertainty of this inference can be calculated as significance levels or confidence interval levels. For the validity of the inference, the sampling must not be biased. The calculation of uncertainties relies on the random character of the sampling process. Dealing with this problem in statistical education is essential and is included in many modern courses. However, an inference may cause further issues even when sampling is not biased. In practical applications, inferences are often made, at least tentatively, beyond the population. This is not justified by applying the statistical method alone (see Hacking, 1965, for a profound philosophical analysis of these limitations). It has to rely on evidence and theoretical considerations made in empirical sciences, whether they use the sample-to-population inference or not. Replication of an experiment and theoretical explanations are essential elements (Diaconis, 1985).

The recent replication crisis in the psychological and educational sciences shows the problem (Wiggins & Christopherson, 2019). Another remedy in statistics suggested above is "cross-validation,"

i.e., partitioning the data set into training and test data (to use modern terms), where the models and findings of the analysis of the training data set are checked and modified on new data. This repartitioning can be repeated several times. Replication and cross-validation are seldomly discussed in elementary statistics courses. Still, they should become an element of an updated fundamental idea of inference, nearer to "scientific inference" than to "sample-to-population" inference.

CONCLUSIONS AND PERSPECTIVES

The paper has tried to show how some of the fundamental ideas considered for structuring the secondary or tertiary curriculum and the conceptualizations of the steps in statistical analysis can be reinterpreted and updated to integrate new ideas, methods, and process conceptualizations of the emerging field of data science. For practical implementations at all educational levels, the question of adequate digital tools that support students' learning and applying data science is essential. This was not discussed in this paper for reasons of space (but see Biehler, 2019; Biehler et al., 2013). There are digital tools specifically developed for students, such as iNZight, TinkerPlots, Fathom, and CODAP, and professional tools such as R and Python, embedded in Jupyter notebooks that can be adapted for educational purposes (Biehler & Fleischer, 2021; Fleischer et al., 2022). It is a challenge for future developments to experiment with these different tools and explore how far they can support the fundamental ideas and processes of statistics and data science.

REFERENCES

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II)—A framework for statistics and data science education*. American Statistical Association; National Council of Teachers of Mathematics.

Battal, A., Afacan Adanır, G., & Gülbahar, Y. (2021). Computer science unplugged: A systematic literature review. *Journal of Educational Technology Systems*, *50*(1), 24–47. https://doi.org/10.1177/00472395211018801

Berthold, M. R., Borgelt, C., Höppner, F., Klawonn, F., & Silipo, R. (2020). *Guide to intelligent data analysis—How to intelligently make sense of real data* (2nd ed.). Springer. https://doi.org/10.1007/978-1-84882-260-3

Biehler, R. (2005). Authentic modelling in stochastics education: The case of the binomial distribution. In G. Kaiser & H.-W. Henn (Eds.), *Festschrift für Werner Blum* (pp. 19–30). Franzbecker.

Biehler, R. (2019). Software for learning and for doing statistics and probability–Looking back and looking forward from a personal perspective. In J. M. Contreras, M. M. Gea, M. M. López-Martín, & E. Molina-Portillo (Eds.), *Proceedings of the Third International Virtual Congress of Statistical Education*. University of Granada. www.ugr.es/local/fqm126/civeest.html

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 643–689). Springer. https://doi.org/10.1007/978-1-4614-4684-2_21

Biehler, R., & Fleischer, Y. (2021). Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. *Teaching Statistics*, *43*(S1), S133–S142. https://doi.org/10.1111/test.12279

Biehler, R., Frischemeier, D., & Podworny, S. (2017). Elementary preservice teachers' reasoning about modeling a "family factory" with TinkerPlots—a pilot study. *Statistics Education Research Journal*, *16*(2), 244–286. https://doi.org/10.52041/serj.v16i2.192

Biehler, R., & Schulte, C. (n.d.). *ProDaBi-Project: Project Data Science and Big Data at School.* https://www.prodabi.de/en/

Bourbaki, N. (1950). The architecture of mathematics. *The American Mathematical Monthly*, *57*(4), 221–232. https://doi.org/10.1080/00029890.1950.11999523

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Bruner, J. S. (1960). *The process of education*. Harvard University Press.

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics–*

Challenges for teaching and teacher education. A Joint ICMI/IASE Study: The 18th ICMI Study (pp. 57–69). Springer. https://doi.org/10.1007/978-94-007-1131-0

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

Diaconis, P. (1985). Theories of data analysis: From magical thinking through classical statistics. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends and shapes* (pp. 1–36). John Wiley & Sons.

Fleischer, Y., Biehler, R., & Schulte, C. (2022). Teaching and learning data-driven machine learning with educationally designed Jupyter notebooks. *Statistics Education Research Journal*, *21*(2). Article 7. https://doi.org/10.52041/serj.v21i2.61

Franklin, C., & Bargagliotti, A. (2020). Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.246107bb

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005/2007). *Guidelines for assessment and instruction in statistics education (GAISE) report—A pre-K–12 curriculum framework*. American Statistical Association. http://www.amstat.org/education/gaise/

Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Ickstadt, K., Kestler, H. A., Lederer, J., Leitgöb, H., Pauly, M., Steland, A., Wilhelm, A., & Friede, T. (2021). Is there a role for statistics in artificial intelligence? *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-021-00455-6

Gelman, A., & Nolan, D. (2002). You can load a die, but you can't bias a coin. *The American Statistician*, *56*(4), 308–311. https://doi.org/10.1198/000313002605

Gould, R., Wong, R., & Ryan, C. N. (2020). *Introductory statistics: Exploring the world through data* (3rd ed.). Pearson.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge University Press.

Heymann, H. W. (2003). *Why teach mathematics? A focus on general education*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-017-3682-4

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447–453. https://doi.org/10.1126/science.aax2342

Phillips, C. J. (2015). *The new math—A political history*. The University of Chicago Press. https://doi.org/10.7208/chicago/9780226185019.001.0001

Podworny, S., Fleischer, Y., Hüsing, S., Biehler, R., Frischemeier, D., Höper, L., & Schulte, C. (2021). Using data cards for teaching data based decision trees in middle school. In O. Seppälä & A. Peterson (Eds.), *Proceedings of 21st Koli Calling International Conference on Computing Education Research* (Article 39). Association for Computing Machinery. https://doi.org/10.1145/3488042.3489966

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 1–7). Springer. https://doi.org/10.1007/978-1-4899-7993-3_565-2

Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician*, *73*(sup1), 246–261. https://doi.org/10.1080/00031305.2018.1518264

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Weigend, A. S. (2017). *Data for the people: How to make our post-privacy economy work for you*. Basic Books.

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–265. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Wilkerson, M. H., Lanouette, K., & Shareff, R. L. (2021). Exploring variability during data preparation: A way to connect data, chance, and context when working with complex public datasets. *Mathematical Thinking and Learning*. Advance online publication. https://doi.org/10.1080/10986065.2021.1922838