A Comparative Educational Study of Statistical Inference

Manfred Borovcnik
University of Klagenfurt, AUSTRIA
e-mail: manfred.borovcnik@uni-klu.ac.at

Abstract

In his “Comparative Statistical Inference”, Barnett (1982) investigates the various approaches towards statistical inference from a mathematical and philosophical perspective. There have been a few isolated endeavours to develop varied teaching approaches of statistical inference. ‘Comparative statistical inference from an educational perspective’ is long overdue. After discussing Barnett, we give an overview on various attempts to simplify the concepts for teaching. Informal inference is a major endeavour among such projects; resampling and Bootstrap is a newer development in statistical inference, which has also some appeal for teaching. In the light of Barnett’s comparative evaluation we develop some essential alternatives for teaching like Bayes or non Bayes. References to Barnett will illustrate that simple solutions might bias the concepts. Rather than optimizing isolated approaches towards teaching statistical inference, a comparative educational study is suggested. The aim of such an investigation is to highlight and compare relative merits and disadvantages of various approaches as a consolidation piece to guide further research on teaching inferential statistics.

Key words: Statistics education, statistical inference, schools of probability, errors and pitfalls, statistical thinking.

1. Introduction

Inferential statistics is the scientific method for evidence-based knowledge acquisition. However, the logic behind statistical inference is difficult to understand. The methods created for the purpose are based on advanced concepts of probability in combination with different epistemological positions developed in the history of scientific reasoning. Approaches to statistical inference have been developed over the years. The classical significance test of Fisher and the statistical tests by Neyman and Pearson (including sequential tests) were followed by decision theory introducing loss functions. There are at least two more approaches: the Bayesian approach and the re-randomization and bootstrap strand.

2. Comparative statistical inference

Barnett (1982) distinguishes the approaches to “statistical inference” and “decision making”. Inference is a statement about a parameter of a distribution; decision making includes a decision about such a parameter or a future action and involves the additional criterion of utility: while in inference one ignores the question of ‘how probable is it that the estimation or test result is correct (or wrong)?’ in decision making probability and utility (impact) become central issues. The second criterion for grouping the approaches is the way probability is perceived: whether probability is restricted to the frequentist interpretation or is open to qualitative information, which will differ from person to person.

The great controversy on the foundations of probability and a scientific justification of methods for the inverse inference, i.e., the conclusion from statistical data to parameters of the underlying probabilistic model, is signified by the characterization of probability. Stegmüller (1973) has written the most comprehensive compendium from the perspective of analytic science. He – in line with Kolmogorov’s intention – interprets the usual axiomatic theory as justification of a frequentist interpretation of probability.

From this assured basis, he analyzes various approaches to the inverse inference: amongst others, Fisher's significance test, Neyman and Pearson's test policy, Fisher's fiducial probability, and (comparative) Likelihood tests.

For all approaches, Stegmüller puts their flaws to the fore, including his (favoured) likelihood tests; he compares the Bayesian inverse inference (based on Bayes theorem) against the other approaches, which are usually named classical statistics and which refer to a conception of probability that is solely based on Kolmogorov's axioms (1933). The Bayesian inference has to go beyond a frequentist interpretation and is derived from axioms on preferences and the so-called coherence, which lead to a different *axiomatic* theory of probability. The de Finetti (1937) approach justifies a qualitative perception of probability, which all-too easily is interpreted as arbitrary as it has genuine personal ingredients (see more on the historic development in Borovcnik & Kapadia, 2013). The crux is that this conception of probability is linked to a complete different paradigm of science: as probability may also represent *qualitative* information it can entirely be disconnected from (repeated) experiments, which are seen as key to validation and improvement of models in the classical paradigm.

The gaps in rationality of all approaches towards the inverse inference thus lead to a dilemma: either solve them by switching to a different (non-experimental) conception of probability and open the way for subjective ingredients in scientific models, or remain within a conception of probability, which is closed within a theory that can be validated by experiments. The decision of Stegmüller (1973) was shared by the main part of the statistics community: accept that methods for the inverse inference are imperfect and reject subjective elements in science. From Barnett (1982) one may describe the diversity of comparative inference by the following table, which classifies the approaches by two categories, the type of probability involved and the type of statements, which describe the conclusion from the data.

		Framework	
Probability	Decision theory	←→	Inference
Theoretical	Wald (1950)		Fisher significance test; fiducial intervals
frequentist	Neyman-Pearson testing policy; repeated confidence intervals		
subjectivist	Bayes decision theory		Bayesian inference

According to Fisher, $P(E \mid H)$ is a discrepancy measure between the hypothesis H and the data E and lacks a frequentist interpretation. Alternatively, Neyman and Pearson (NP) developed a quasi-decision-theoretic framework and compared the null H_0 and alternative hypothesis H_1. Their rationale is to use a rejection rule V (which consists of a subset of the sample space) that guarantees $P(V \mid H_0) = \alpha$ (size or type I error of the test) and optimizes the type II error function $1 - P(V \mid H_1)$. In justifying the procedure, Neyman reduced the involved conditional probabilities to plain relative frequencies of a life-long testing policy. The more recent habit of using p values to describe the impact of data on hypotheses is a strange hybrid between Fisher' s non-frequentist discrepancy measure and Neyman and Pearson' s size of a test, which was reduced completely to a frequency interpretation.

A similar frequentist interpretation was attached to confidence intervals. Fisher developed his fiducial probability for confidence intervals, which should amount to a theoretical probability that can be applied to the single intervals while the NP method allows a probability statement for the long run rather than for single intervals. However, Fisher's fiducial probability refers either to an awkward argument or to the implicit use of prior distributions which would shift it to Bayesian inference (with an inherent 'logical' or objective prior which should supersede a subjective prior). The simplicity of the NP approach outweighed its evident flaws.

Stegmüller (1973) showed clearly the gaps of rationality of the approach; however, the so-called (favoured) likelihood tests share similar problems. Yet the final decision in the foundations was in favour of a theoretical probability with a strong preference for a frequentist interpretation and the classical procedures of statistical inference based on such a probability concept.

3. Informal inference

While Barnett's (1982) monograph might be perceived as a plea to use the methods in parallel, teaching of statistics followed the narrow path of a pure frequentist interpretation of probability and NP statistical inference. There was a fierce discussion about teaching inferential statistics to students – the Bayesian way or in the classical tradition (Berry, 1997; Albert, 1997). Moore (1997) focussed his critique on his judgement that Bayesian methods are too difficult for teaching. For applications, a more subtle perception of Bayesian models in the sense of Berger (1985) paved the way to apply them whenever they are useful, i.e., in case that there is not enough empirical data to base the inverse inference upon.

As it also desirable to introduce methods of statistical inference to students in high school, simplified approaches were advocated and developed. Via new technologies, simulation has become accessible and replaced the mathematics of the inferential methods, which lead to a strong connection of the involved (conditional) probabilities to a *primitive* frequentist interpretation. Carranza and Kuzniak (2008) have analyzed the consequences of such a biased attitude towards probability: It may confuse learners that the concepts introduced are frequentist (objective) while examples ask for solving strategies that are linked to Bayesian methods and probabilities derived are essentially subjective. Gigerenzer (2002) has searched for adequate representations to simplify the solution. He transfers all (conditional) probabilities to expected numbers which are presented either in tree diagrams or in two-way tables. By the embodiments used, probabilities involved have an even higher (and false) degree of objectivity, which is misleading in interpreting the relevance of the results. Borovcnik (2012) has analyzed conditional probability as a fundamental concept and refers to the importance of going beyond a frequentist interpretation to understand statistical inference.

The nonparametric approach (Noether, 1967) removed the requirement for families of parametric probability distributions. First, one could derive a statistical test only by calculating the test statistic on a finite set, which was established from equally likely cases. Second, the null (effect) hypothesis attracts a natural embedding in the context.

Example. If data under two conditions (treatment and control) are used to test whether treatment has a significant effect then the two data sets can be pooled under the hypothesis of no effect. From this pool, a selection of the treatment group (the rest constitutes the control group) establishes one possible result of the test statistic (difference of means of original values or of ranks). Such a selection is also named re-randomization. Under the hypothesis of no effect any such selection that is possible has the same chance whence the test reduces to inspect all possible recruitments of the treatment group and calculate the test statistic. For $n_1 = 20$ and $n_2 = 15$ this requires looking at (35, 20) potential samples. The tedious search for all possible samples can be substituted by random selection of a number of samples, which supply an empirical estimate of the null distribution. The obvious advantage is the natural embedding of H_0 (null effect) into the context and the way to describe it by equally likely cases instead of complicated distributions. The drawback lies in the lack of direct power considerations (complement of type II error) as there is no adequate representation of various 'distances' (in location) between the two groups without describing them by parametric distributions (a task that was intended to avoid by the method).

The relations between α *and* β errors (type I and II) are essential to comprehend statistical tests. As the basic ingredients are missing (about the type II error), the approach can be no more than a transient state in teaching. The idea of nonparametrics was taken up in what is called informal inference (e.g., Zieffler, et al, 2008). An informal (as opposed to a formal) use of probability models may be seen in Borovcnik (2011). Here, the parametric models are not circumvented; instead they are made accessible by simulation in order to get an idea of how big a discrepancy between the presupposed (hypothesized) model and the empirical data has to be in order to *decide to exclude* this model from further considerations (which establishes a risk as the model could still apply to the situation investigated) in order to simplify the range of models.

4. Resampling and Bootstrap

Key to resampling methods is that the original sample is used for an estimation of the 'true' *distribution* of the population. From this sample an estimate for one (nearly arbitrary) parameter can be calculated. To evaluate this estimate, a resampling interval is empirically derived from the Bootstrap distribution, which is gained by sampling from the estimate of the distribution instead of the (unknown) distribution of the population. The approach was developed by Efron and Tibshirani (1993). Its simplicity is attractive so that there is no wonder that statistics education took up the idea and tried to develop teaching approaches: Borovcnik (2007) used also the analogy of repeatedly 'measuring' the unknown parameter (like measuring a physical quantity). More recently, Engel (2010) and Pfannkuch and Wild (2012) have contributed suggestions. The trouble with the approach seems to be that Bootstrap intervals are slightly biased; to repair this defect, complicated mathematics is required (Lunneborg, 2000) so that the approach leads to a dead-end. However, future research may well change such a judgement on the merits of resampling for teaching.

5. Essential alternatives in teaching inferential methods

Following Barnett's characterisation of comparative statistical inference, we go on to discuss teaching issues formulated as alternatives.

Bayes or non Bayes. Since the controversy in the foundations has remained unresolved, the alternative is wrongly put. The statistics community usually reduces probability to a purely frequentist concept and derives the methods of statistical inference upon this basis. Moore (1997) considered classical statistical methods as much easier to understand – a very pragmatic view. Bayesians still promote the perception of probability exclusively in a *fundamental* subjectivist sense. The unresolved controversy indicates that neither approach has priority and probability genuinely has both objective and subjective features, and to ignore either side would bias the concept. It might well be appropriate to try a *hybrid* approach (as this author did for the Second Bayesian Meeting in Valencia in 1994). However, this would cause confusion about the character of the statements involved as the basic view on what is a scientific concept is so different between these schools. Vancsó (2009) has decided to teach both approaches *in parallel* and join them in subsequent applications: A learner can see what is missing with either approach and study the derived statements.

Decision theoretic or inferential perspective. From a teaching perspective, the question is 'Why do decision-theoretic viewpoints in the education of statistical inference help to understand the concepts involved?' More recently, endeavours may be traced to link even introductory probability with risk, i.e., to integrate considerations of impact of possibilities (or decisions) right from scratch (see Nikiforidoru & Page, 2011). Kahneman and Tversky (1979) have shown that judgements of probability are biased especially if gain or loss is involved. For the introduction of probability, this author prefers to calibrate the 'feeling' of probability without utility to establish a clear vision of what specific probabilities really mean and introduce concepts of utility at a second stage only. However, in order to demystify the essential concepts of type I and II errors and the consequences of test decisions, it is quite revealing to embed the inferential situation into a decision-making context. It would also help to recognize that questions about the probability of wrong decisions or about the probability of a hypothesis after data have led to its rejection have to remain *unanswered*. In the decision-theoretic framework such questions arise naturally and prompt the modeller to find adequate answers.

Tests or confidence intervals within classical statistics. Some statisticians claim that it is better to teach confidence intervals as the 'logic' of statistical tests is so complicated. Maybe that such a view also arises from a desire to avoid the rationality gaps for statistical tests. However, the same gaps apply to confidence intervals (see Berger, 1985). Furthermore, a purely frequentist interpretation of coverage probability of confidence intervals applies only to the *repeated* intervals; a transfer to a *single* interval lacks any justification. There is a variety of situations where a statistical test is essential. One case is for testing for probabilistic assumptions like a test for the type of the distribution or for independence. Another case is methods like the analysis of variance by which it becomes possible to test whether a specific factor is significant or not, i.e., whether its null effect can be rejected by a significance test or not. Confidence considerations do not apply in such cases. As they include essential fields of applications, a restriction to confidence intervals in teaching seems inappropriate.

6. Comparative inferential statistics from an educational perspective

We highlight essential issues in developing a comparative statistical inference for teaching in the light of Barnett (1982).

Utility and loss. Barnett (1982, p. 99) refers amongst others to the following specific consequences of integrating utility and loss functions into the framework "The rewards may contain many components. [...] Preferences will often be personal." A further complication arises out of infinite loss functions as are used normally.

The objectivity of models and the self-consistency of an objective approach. "The whole question of model validation is a major one. All we will say here is that, in any real-life study of such a problem, it is often not feasible to carry out a thorough validation. It is unlikely that adequate information would be available, and the model might at best be justified on a combination of subjective and quasi-objective grounds. Independence might be justified by arguments about the physical properties of the [phenomenon studied]." (Barnett, 1982, p. 31).

Chicken and egg – a plea for conceptual flexibility. Kendall (1949, referenced in Barnett, 1982, p. 94) refers to a basic dilemma: "The frequentist seeks for objectivity in defining his probabilities by reference to frequencies; but he has to use a primitive idea of randomness [...]. The non-frequentist begins by taking probability as a primitive idea but he has to assume that the values which his calculations give to a probability reflect, in some way, the behaviour of events. ... Neither party can avoid using the ideas of the other in order to set up and justify a comprehensive theory." As neither conception is self-contained, the question 'which is better?' is wrongly put; for enhancing probability a conceptually more flexible approach seems appropriate.

Decision theoretic perspective. Classical statistical procedures investigated from a decision theoretic perspective reveal basic drawbacks as Barnett (1982, p. 261) states introducing first a critique by Lindley and Smith (1972): "[...] many techniques of the sampling-theory [that is, classical] school are basically unsound [...]. In particular the least-squares estimates are typically unsatisfactory; or, [...] inadmissible in dimensions greater than two." Barnett continues: "Here is a strange juxtaposition!" Even if classical statistics is formulated without reference to losses, it reveals an unsatisfactory feature as it would unnecessarily shut the potential connection to decision theory – a connection that is basically used to establish Neyman and Pearson's test theory, which is one of the main schools of testing within the classical position towards statistics.

Barnett's conclusion for teaching. Barnett (1982, p. 307) summarizes his comparative studies by: "Various attempts [...] to describe the role of the statistician. [...] One solution to the ubiquitous demands on the statistician is to encourage multi-disciplinary team-work [...]." Newer approaches such as cross-validation shift to data analysis for a short time but will require even more knowledge about its probabilistic modelling. He concludes with "The teaching of statistics must continue to place major emphasis on basic principles and concepts, and on their implications in the form of practical statistical techniques. Exposure to the *range* of philosophical and conceptual attitudes to statistical theory and practice must be an essential ingredient." (p. 309).

Teaching statistical inference. Barnett remains unheeded. The statistics education community strived to simplify the probabilistic foundation of statistical inference towards a primitive frequentist interpretation; teaching at school level stopped dealing carefully with conditional probability as it requires complicated calculation and includes other connotations of probability that hinder a straightforward progress. More recently, the approach of informal inference has attracted some attention. While it makes some aspects palatable for teaching, it reduces the complexity of statistical inference to such a degree that it becomes difficult to discuss its drawbacks. Yet, a subtle knowledge of conditional probability is decisive to perceive the involved errors. Errors of type II (or, equivalently, power) can only be indirectly introduced. However, power considerations are fundamental to evaluate a 'decision' made upon the empirical evidence used. The question for 'a decision to be correct' cannot be placed and answered without integrating prior probabilities of hypotheses under test but these are 'excluded' as personal, subjective, and thus non-scientific. Too easily it is forgotten that models can never be 'objectively' validated. A prime goal for statistics education is to develop a *comparative* study of statistical inference from an educational point of view as Barnett (1982) did for the scientific community. One promising idea may be to link statistical inference to the more general process of scientific inference as was suggested by Wild and Pfannkuch (1997). By our 'late-breaking session' we have formulated a long-term project for the future: Rather than developing further isolated approaches towards teaching inferential statistics in a simplified manner, this project should elaborate on relative merits of various ways in the light of empirical research on the impact of teaching *and* in the light of philosophy and applications.

References

Albert, J. (1997). Teaching Bayes' rule: a data-oriented approach. *The American Statistician*, 51(3), 247-253.

Barnett, V. (1982). *Comparative statistical inference*. 2nd ed. New York: Wiley.

Berger, J.O. (1985). Statistical decision theory and Bayesian analysis. 2nd ed. New York: Springer.

Berry, D. A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician*, 51(3), 241-246.

Borovcnik, M. (2007). On outliers, statistical risks, and a resampling approach towards statistical inference. *Paper presented at* CERME 5). Larnaka.

Borovcnik, M. (2011). Key properties and central theorems in probability and statistics – Corroborated by simulations and animations. *Special issue in Statistics of Selçuk J. of Applied Mathematics*, 3-19. Online: www.sumam.selcuk.edu.tr/specialissue-s.html.

Borovcnik, M. (2012). Multiple perspectives on the concept of conditional probability. *Avances de Investigación en Didactica de la Matemática*, 2, 5-27. Online: www.aiem.es/index.php/aiem/issue/view/2.

Borovcnik, M., & Kapadia, R. (2013). A historical and philosophical perspective on probability. In E. J. Chernoff, & B. Sriraman (Eds.), *Probabilistic thinking: presenting plural perspectives.* New York: Springer.

Carranza, P. & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education.* Monterrey: ICMI and IASE. iase-web.org/Conference_Proceedings.php?p=Joint_ICMI-IASE_Study_2008.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York, Chapman & Hall.

Engel, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society.* Voorburg: International Statistical Institute. Online: iase-web.org/Conference_Proceedings.php?p=ICOTS_8_2010.

Finetti, B. de (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales Institut Henri Poincaré*, 7, 1-68. Foresight: Its logical laws, its subjective sources. In S. Kotz, & N.L. Johnson (1992), *Breakthroughs in statistics. Volume I. Foundations and Basic Theory* (pp. 134-174). New York, Berlin: Springer.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, XLVII, 263-291.

Kendall, M.G. (1949). On the reconciliation of theories of probability. *Biometrika*, 36, 101-116.

Lindley, D.V., & Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. of the Royal Statistical Society, Series B*, 34, 1-41.

Lunneborg, C.E. (2000). *Data analysis by resampling: concepts and applications*. Pacific Grove, CA: Duxbury Press.

Moore D. S. (1997). Bayes for beginners? Some reasons to hesitate. *The American Statistician*, 51(3), 254-261.

Nikiforidou, Z. & Page, J. (2011): Risk taking and probabilistic thinking in preschoolers. In D. Pratt (Ed.), *Working Group on Stochastic Thinking. CERME 7.*

Noether, G.E. (1967). *Elements of nonparametric statistics*. New York: Wiley.

Pfannkuch, M., & Wild, C. (2012). Laying foundations for statistical inference. *Proc. of* ICME 12. Seoul. Online: www.icme12.org/upload/submission/1883_F.pdf.

Stegmüller, W. (1973). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, vol.4, first part: Personelle Wahrscheinlichkeit und Rationale Entscheidung, second part: Personelle und statistische Wahrscheinlichkeit.* Berlin-New York: Springer.

Vanscó, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *International Electronic J. in Mathematics Education*, 4(3), 291-322. Online: www.iejme.com/032009/main.htm.

Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. With discussion. *International Statistical Review*, 67(3), 223-265.

Zieffler, A., et al. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research J.*, 7 (2), 40-48. Online: iase-web.org/documents/SERJ/SERJ7(2)_Zieffler.pdf.