A comparative study of statistical inference from an educational point of view

Manfred Borovcnik*
Dep. of Statistics, University of Klagenfurt, Austria – manfred.borovcnik@aau.at
Ramesh Kapadia
Dep. of Statistics, University of Klagenfurt and London, England – ramesh.kapadia0@gmail.com

Abstract

Inferential statistics is the scientific method for evidence-based knowledge acquisition. The underlying logic is difficult and the mathematical methods created for this purpose are based on advanced concepts of probability, combined with different epistemological positions. Many different approaches have been developed over the years. Following the classical significance tests of Fisher and the statistical tests by Neyman and Pearson, and decision theory, two more approaches are considered here using qualitative scientific argument: the Bayesian approach, which is linked to a contested conception of probability, and the rerandomization and bootstrap strand, which is bound to simulation. While Barnett (1982) analysed statistical inference from a mathematical/philosophical perspective to shed light on the various approaches, we analyse from the grand scenario of statistics education and investigate the relative merits of each approach. Some thoughts are developed to reconsider informal inference, which bases teaching on rerandomization and bootstrap and reduces probability to a frequentist concept. The ideas are designed to initiate a deeper discussion about learning paths towards inference.

Keywords: Schools of inference; statistical thinking; learning paths for inference; informal inference.

1. Introduction

A late-breaking session at the World Statistics Congress in Hong Kong in 2013 was devoted to the topic "*Statistical Inference – an Unresolved Issue in Statistics Education*". It promoted a suggestion to teach classical and Bayesian inference in parallel (and not to merge the diverging approaches) so that the learner would comprehend both approaches and their relative limitations better. Another lecture was devoted to decision-theoretic aspects of probability and statistics, which is also faced with using probabilistic information of widely diverging quality and sources so that it often cannot be interpreted as frequentist probabilities (because of its qualitative character). A further presentation designed a perspective for future research in statistics education by a *comparative* educational study of statistical inference. Following Barnett's (1982) monograph discussing the various attempts to develop statistical inference from a mathematical and philosophical point of view, the statistics education community must work on a comparison of methods related to different schools of inference to enhance the underlying ideas. One presentation enriched the perspective of independent trials of the classical approach (that are usually done under the same conditions) by informal investigations of a broader probabilistic process that might change the parameter of the distribution of the investigated phenomenon (it focused on recognizing when such a change point occurs).

The prime reaction of the audience was that everything beyond a frequentist approach to probability and – especially a comparative programme for statistical inference – is suitable only for a minority, as most students will not understand the inherent concepts. As the main effort of the statistics education community on statistical inference is devoted to promote an approach called "informal inference" (referred to as II-approach here) there seems to be a need to initiate further and deeper discussion on how far the various teaching approaches can reach and to find out where the specific difficulties of each approach are manifested. Otherwise we might end up with two conflicting classes of statistics, one for those who proceed to higher levels (and are required to relearn a method of statistical inference markedly different from the previous one) and the other for those who remain at basic levels, with few connections between the two types of "statistics". The simplified II-statistics could become the minimal programme of reading statistical results in exactly that way that experts explain to people at a

lower level. Following this approach, how can laypersons develop a critical stance on statistical results, which Gal (2005) refers to as a prime goal of probabilistic and statistical literacy?
This discussion was scheduled for the WSC in Rio as a late-breaking session again (an STS this time), but it was too late find a suitable representative of the informal approach. The goal of this paper is to argue for a more refined teaching approach either with carefully-designed measurements to continue the simplified II towards statistical inference or to support a different attitude towards formal inference by finding suitable informal pathways to formal methods of inference as learning goals.

2. Alternatives for developing statistical methods and for teaching

There are four key methods for statistical inference, which differ by the type of probability (frequentist or subjectivist) and the general perception of science and scientific truth (hypotheses, unknown statements, data without hypotheses as central focus of evidence): i. the Bayesian approach to inference (BI); ii. an approach to inference oriented towards decision theory (DT); iii. classical methods of statistical inference (Neymann-Pearson and Fisher), iv. using resampling for informal inference (II). Following Barnett (1982) and ideas from Borovcnik (2013), a comparative statistical inference raises key alternatives for statistics education.
i. Bayesian or non-Bayesian inference. BI involves the interpretation of probability as a frequentist (FQT) or as a subjectivist (SJT) concept (SJT is related to the preference system of a person). It also involves the question whether it is feasible to attribute probabilities to hypotheses or not. The classical decision is to link probability to a theoretical concept, which is tightly connected to frequencies. Consistent with that, any hypothesis does not have a probability. Unfortunately, the key concepts to describe the properties of classical statistics are commonly misunderstood (not only by students) as probabilities for hypotheses: i. The coverage of confidence intervals is erroneously directed towards the resulting single confidence interval even though it is a property of two random variables covering the unknown parameter. ii. The type I error of a test, or Fisher's p value are probabilities for observations (strangely enough, the latter probability for observations is calculated ex post, i.e., after the data is known) yet they are erroneously interpreted as probability that the null hypothesis is "true" given the data. This marks inconsistent thinking as the view on (experimentally based) science on the outset led to a restrictive basic assumption that probabilities have to be (at least ideally) open to a statistical test using relative frequencies from experiments. As a consequence this view of science denies that parameters of a model can have probabilities as there are no experiments with data on this parameter. In the end, the figures that describe essential properties of statistical methods are misunderstood in exactly that way that probabilities about these parameters are extracted yet no one opposes to that. Maybe the methods draw their popularity exactly from that basic misunderstanding and would not be accepted if it were made clear that the interpretation does not apply at all.
The common decision favours the non-Bayes alternative and consistently binds the concept of probability to frequencies with the classical methods of Neyman-Pearson test policy and confidence intervals including the well-known misinterpretations referred to earlier; even Fisher's significance test is re-interpreted to fit to this frame. The decision could well be in favour of Bayes with the known problems to justify the prior distribution on parameters that is required in order to derive the posterior distribution of the unknown parameter. This prior is not open to an experiment and there can be no empirical evidence for it. The programme of finding objective arguments (like invariance arguments) for the choice of priors has failed. The modelling group of Bayesians works also on investigating the influence of a model prior with the aim to find its impact on the final decision. For teaching there is also an interesting suggestion to develop both classical and Bayesian methods of statistical inference (Vancsó 2009) and to learn from both sides what is missing and what constitutes either side of statistical inference. Moore (1997) has argued that Bayesian methods are too complicated for teaching. However, with regard to the misconceptions that were mentioned above, classical methods are no easier. Neither of the alternatives of Bayes or non-Bayes is self-contained. Both have to refer to the other part (probability is not reducible to frequencies; subjectivist probability cannot ignore frequencies) so that the question 'which is better?' is wrongly put. Barnett (1982, p. 94) identifies it as fundamental dilemma and argues for a conceptually flexible approach to probability and inference.

ii Decision theoretic or inferential perspective. A decision-theoretic point of view does not need to begin with all the distributional problems (methods to find statements about parameters of a statistical model underlying the process of data generation). It could start with decision situations in every-day life. In fact, life is a sequence of decisions based upon data-driven beliefs. The ideas of decision-making are not particularly deep, but the application of ideas proves to be difficult in practice.
Medical diagnosis may serve as a context to enhance the concepts involved. What to do after a diagnosing test has yielded a positive result (indicating the disease under scrutiny) or a negative test (indicating that this disease is not present)? The "errors" of decisions occur as conditional probabilities and the probability $P(D|T^+)$ does not equal the reverse $P(T^+|D)$ and can only be calculated if a prior probability of the disease is known in addition to the quality parameters of the diagnosing method (as there are the sensitivity $P(T^+|D)$ and the specifity $P(T^-|\text{no D})$).
Many decisions are one-off decisions that do not comply with the frequentist concept of probability. A proper interpretation of the various probabilities (frequentist or subjectivist) can be discussed as decisions depend on subjective probabilities (a popular frequentist interpretation of the medical diagnosis from above is illustrated by calculating what happens if there are 1000 persons; however, there are not 1000 persons exactly like you); decisions also depend heavily on the criteria used. The perception of probabilities is biased, especially if probabilities are low and the potential impact is high. How to counter such biases by teaching? Another advantage of the decision-theoretic view is the ease of integrating the impact of decisions into the analysis, which leads to a popular strand of probability (not only in teaching), the analysis of risk (Gigerenzer, 2002; Nikiforidou, & Page, 2011).
The framework of decisions can be extended (or restricted) to the situation of a classical statistical test. Neyman and Pearson developed their concepts of a policy of testing statistical hypotheses within a decision-oriented situation, which implies *comparing* several probability distributions for their consequences on the decision. This complicates the situation for teaching as several probability models are used and compared (why not use the best model?) as it results in various values for the probability under scrutiny so that the link to reality via relative frequencies is lost (one does not have real datasets for each of the models used). A decision-theoretic framework may also clarify the differences between several schools of classical statistics that are permanently confused: the Fisher significance test and its modification to p values on the one side and the NP tests. Fisher defined p values as a discrepancy measure to the null hypothesis and would have denied any frequentist interpretation of it. The frequentist interpretation has been brought onto the scene by Neyman to give a more concrete meaning to type I and type II errors.
iii. Tests or confidence intervals within classical statistics – Fisher's significance test or NP test policy. Some statisticians favour confidence intervals over tests. Distinct from tests that discriminate between null and alternative hypothesis, confidence intervals avoid some of the difficulties above as they treat all parameters (and the linked models) symmetrically. While this is true, confidence intervals cannot answer the specific questions that are modelled by statistical tests and they are prone to the same basic misinterpretation: the success frequency of the method cannot be linked to single intervals and the latter is the only issue of interest. Also, in many applied problems (ANOVA, test for independence, test for a specific family of distributions, test for equality of variances, test for equality of covariances) there is simply only a null hypothesis which has to be put under test. Thus, not even the power (complement of type II error) can be calculated as it is neither possible to assign specific alternatives nor to order these alternatives by distance to the null hypothesis. This latter problem has induced a practice of statistical tests that may be called a hybrid test between Fisher and Neyman (with a null hypothesis only, but with a frequency interpretation of the p value). Hubbard and Bayarri (2003, p. 182) warn about the resulting confusion: "Measures of evidence (as p values or likelihood ratios) have radically different interpretations to frequentist measures of performance (as type I errors and power functions), and mixing both of them in the same analysis is always delicate. It certainly requires a fair amount of statistical training and sophistication, and we believe that it is way too dangerous to encourage its use by the casual, statistically untrained, user."

iv Resampling and bootstrap or statistical inference. Here we leave the alternatives of Barnett (1992). Given all the subtleties of statistical inference, there has been an endeavour for simplification starting from applications (to reduce required assumptions) and later taken up from statistics education (to reduce the complex situation). In applications, the new methods were developed to meet the need to simplify the *assumptions* (e.g., normal distribution) as it is hard to check the assumptions (and what to do if they are not fulfilled?). In education, the driving force was to simplify the *complexity* of a test situation: the formulation of a null hypothesis is represented by various probability distributions. First, it is difficult to decide whether to formulate a one-sided or a two-sided problem. Second, it is not obvious why these cannot be interchanged (e.g., for a binomial parameter testing $p \leq p_0$ against $p > p_0$ is not interchangeable with testing $p \geq p_0$ against $p < p_0$). The formulation of the null (and of the alternative) hypothesis is tightly bound to context *and* methodology. If the situation can be simplified so that alternatives play no direct role (for the test in question) and the null hypothesis seems *natural* then this is seemingly an advantage from an educational point of view. If furthermore probability distributions (as abstract entities) are substantiated by simulated data, a further step to enhance learning is accomplished.

We explain the methods in the case of comparing two distributions. The null effect hypothesis of no difference between the two distributions naturally leads to pooling the data sets as $x_1, x_2, ..., x_m$, $y_1, y_2, ..., y_n$, from which the *x*-data might be sampled repeatedly. If the pooled sample is downgraded in scale by ranks instead of original values and the distribution of the rank difference between the two populations is established by simulation (re-randomization), we get the nonparametric Wilcoxon-rank-sum test. If we sample with replacement from the original values of each of the samples separately, we can also calculate a mean difference and establish the (empirical) bootstrap distribution. The 2.5 % and 97.5 % quantiles then yield the bootstrap interval for the mean difference that imitates the classical confidence interval (but has features distinctly different from it).

3. Need for simplification in approaches towards statistical inference

To counter the well-known misinterpretations of classical methods, there have been ongoing endeavours to design learning paths to statistical inference at secondary level. Parallel to other approaches, checklists have been compiled to sequence the methods in steps in order to increase proper use and interpretation. Descriptive statistics has gained a role with the trend towards modelling and application from the 1980's onward but there has always been a desideratum to find palatable ways to teach statistical inference. In the 1980's EDA (Tukey, 1977) and nonparametrics (Noether 1967) have been suggested to improve the educational situation. EDA was an approach back to basics (exploring data alleviating assumptions) to find structures that can be interpreted in the context but EDA does not lead to statistical inference. Nonparametrics also allows for less stringent assumptions (using ranks instead of values) and yields robust methods but does not allow for an investigation of alternative hypotheses so that it is viable for specific cases only.

The first approaches (from the 1990's) towards teaching inference ended with limited success. Meanwhile simulation gained a key role to illustrate the essential properties of statistical inference methods. However, simulation could not solve the teaching problem as the inherent complexity marks the proper hurdle for understanding, which is reinforced by the embodiment of the models by probability distributions (especially of sample statistics like the mean).

In the 1990's resampling and bootstrap were suggested (Efron, & Tibshirani, 1993) as an approach that is totally data-centred. Since then there have been several attempts to use the potential of nonparametrics and resampling for teaching (Borovcnik, 2006; Cobb, 2007; Engel, 2010). Other more informal approaches have been advocated; some attempts simply use simulation to build up an intuitive understanding of the required decisions in test situations (Zieffler, Garfield, & DelMas, 2008). Wild, Pfannkuch, and Regan (2010) promote a graphically oriented approach towards more accessible conceptions of statistical inference. Wild and Pfannkuch (1999) mark an endeavour to root statistical inference within the more general process of scientific inference.

4. Informal inference – some thoughts to reconsider

Informal inference (II) is a term used more recently in statistics education to unify various approaches starting with simulation of probabilistic models (to illustrate their impact and to solve the problems), or using re-randomization methods to reduce the required assumptions (as is done in non-parametric statistics, e.g., Noether, 1967), and going on to use bootstrap methods to replace statistical inference by a data-oriented approach, as was the intention by Efron and Tibshirani (1993). These approaches have been developed from special needs in applications (to reduce assumptions, to find tractable solutions where parametric models do not fit or allow for tractable solutions, etc.). It is pertinent to analyse these approaches from an educational perspective.

There are two different approaches subsumed under the term resampling. The resampling approach is meant to be easy and intuitive. For *re-randomization* one samples (*without* replacement) from the *null* hypothesis and gets an approximation for the distribution of nonparametric test statistics in specific cases. However, it is not possible to introduce an alternative hypothesis in the same embodiment. For *bootstrap*, the initial data set is sampled (*with* replacement). Rather than from the cumulative distribution function *F*, sampling is done from the estimate of *F*. In the following, rerandomization and bootstrap are compared to key statistical ideas.

	Rerandomization	*Bootstrap*
Hypotheses	Only the *null* effect hypothesis	Not possible to conceptualize
Type I (or alpha) error	Yes	No
Type II (or beta) error	No	No
Alternative hypotheses	Not possible to conceptualize	Not possible to conceptualize
Methods	Only significance test of *null*	Only intervals

i. *Adaptation to classical intervals*. The corrections for bootstrap intervals (to let them asymptotically converge to confidence intervals) are quite complicated and far from being intuitive.

ii. *Coverage property.* Coverage coincides only for symmetric distributions and also only via coincidence (Lunneborg, 2000). Bootstrap intervals have a different conceptualization of „coverage". Neither can the interpretation of bootstrap intervals be transferred to confidence intervals nor are the values the same; and the intervals have other differing properties.

iii. *Significance test.* A link between confidence intervals and hypotheses tests is not straight-forward as spread and skewness of the distributions may differ by the value of the investigated parameter.

iv. *Alternative hypotheses*. There is no way to introduce alternative hypotheses except by using probabilistic *assumptions* (although simulation might replace probability calculations). None of the alternatives can be resampled as there is no initial sample to work with. Thus, resampling is not suitable for investigating alternative hypotheses.

v. *Hypothetical comparison of models*. Modelling involves *comparing* various scenarios / probabilities and not judging only a single one. In hypothesis tests different models (of the same family of distributions) are compared against each other. To perceive how this thought experiment is done is the key to understand statistical inference.

Bootstrap fails with low (tail) probabilities. If it is about the tails of a distribution, one would not get data about them so that the tails will not be resampled as they are not contained in the initial sample (a serious problem for applied statistics.) If the first sample is too small, more regions of the distribution cannot be sampled well enough. If the initial sample is large, then classical methods deliver better results via the central limit theorem. If one resamples then the resampling error is large unless one takes more than 10,000 re-samples. That makes it intractable for application (and for teaching).

Vital issues are: i. How well does the II-approach build on probability and does the approach reduce the character of probability towards a mere frequentist concept? ii. What are the consequences of II to the perception of modelling as all information is contained in the given data and how well is the idea of modelling a real situation by hypothetical models promoted? iii. How is the logic of statistical inference embedded in II since re-randomization does not allow type II error considerations and for bootstrap it is impossible to conceptualize both type I and II errors. iv. Bootstrap, while convincingly intuitive, leads in practice to biased methods and fails with tail probabilities (which are low and not well represented by data). v. How smooth is the educational path from II towards formal inference?

Rather than informally exploring probabilistic models by simulation, II is intended to replace traditional statistical inference. As the full complexity of the inference situation is not developed (resampling is intended to avoid that), it is hard to think in terms of *hypotheses* within the resampling approach. Furthermore, it is not possible to conceptualize errors of type I and II (as they depend on hypotheses). In addition to that, resampling (like simulation) marks a transition from probability models to simulated data and causes a shift in connotation from hypotheses to facts (data as facts) as models are absorbed in (resampled) data.

5. Conclusions – informal ways to formal statistical inference

A prime goal for statistics education is to develop a comparative study of statistical inference from an educational point of view as Barnett did for the scientific community. We continue exploring various pathways to make sound inferential reasoning accessible to students at any level. This long-term project can be fuelled by a series of articles with proponents of each educational approach with invited discussants. The benchmark for "success" of such approaches cannot solely be the short-term evidence of students who solve specific tasks better. Such an analysis should also consider the educational "cost" for the continuation of learning paths later on. The target of teaching should be to improve people's understanding the way how statistical inference can provide empirical evidence. We support Barnett's (1982, p. 309) final statement for teaching: "Exposure to the *range* of philosophical and conceptual attitudes to statistical theory and practice must be an essential ingredient."

References

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

Borovcnik, M. (2006). On outliers, statistical risks, and a resampling approach towards statistical inference. *Paper presented at CERME 5*. Larnaka.

Borovcnik, M. (2013). A comparative educational study of statistical inference. *Proceedings of the 59th World Statistics Congress of the ISI* (pp. 1114-119). The Hague: ISI.

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, 1(1).

Efron, B. Tibshirani, R.G. (1993). *An introduction to the bootstrap*. New York: Chapman and 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: ISI.

Gal, I. (2005). Towards "probability literacy" for all citizens: Building blocks and instructional dilemmas. In G. A. Jones (Ed.), *Exploring probability in school* (pp. 39-63). Dordrecht: Kluwer.

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

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing (with discussion). *The American Statistician*, 57 (3), 171-182.

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.

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

Vancsó, 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.

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

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research J.*, 7 (2), 40-48.