Reviewing of Applied Research with an Industry 4.0 Perspective

Ron S. Kenett

Contents

I. Background

This note provides a perspective on a questions checklist to be used in reviewing applied research using analytics and statistical analysis. The paper will also emphasize applied research related to Industry 4.0. It expands on a paper providing similar guidelines with a general focus on information quality (Kenett and Shmueli, 2016a). The note is methodological. It covers aspects of study design, algorithmic and inferential methods in frequentism analysis, Bayesian methods in Bayesian analysis, selective inference aspects, severe testing properties and presentation of findings. Information quality is based on good responses to questions about a specific report such as what is the goal of the analysis, is the data resolution adequate for the stated or implicit goal, how is data from different sources integrated, has a generalization claim been made, on what basis? etc etc…

A specific context for reviewing applied research is systems engineering and industrial applications and the growing interest in Industry 4.0. Specific analytic challenges in this area include: 1) Engineering design, 2) Manufacturing systems, 3) Decision support systems, 4) Shop floor control and layout, 5) Fault detection and quality improvement, 6) Condition-based maintenance, 7) Customer and supplier relationship management, 8) Energy and infrastructure management and 9) Cybersecurity and security. For more details see Kenett et al, (2016), chapter 13 in Kenett et al (2020) and Zonnenshain and Kenett, 2020).

Industry 4.0 offers wide opportunities for industry post COVID-19. Encouraging applied research in this area is therefore a forward-looking outlook on domains that could restructure for enhanced sustainability, competitiveness and productivity.

In reviewing studies done in Industry 4.0 areas, one finds data collected actively or passively, models developed with empirical methods, first principles or as hybrid models. Industry, as opposed to science, is less concerned with reproducibility of results, but it should. The industrial cycle provides short term opportunities to try out new products of new process set ups and, based on the results, determine follow up actions. Deriving misleading conclusions can be however very costly and time consuming.

With this background, consider the role of classical statistical methods, versus new age statistical methods, in order to devise guidelines for a statistical evaluation of applied research. The emphasis here is on statistical evaluation as opposed to impact assessment, economic returns or business risks. To address this, let me first summarize some of the recent developments I am interested in accounting for.

A much-debated document has been the ASA statement on p-values (Wasserstein and Lazar, 2016). This statement formulates six principles for statistical analysis focused on the interpretation of p-values. Other approaches mentioned in the ASA statement, without critical appraisal, include: i) Confidence intervals, ii) Prediction intervals, iii) Estimation, iii) Likelihood ratios, iv) Bayesian methods, v) Bayes factor and vi) Credibility intervals.

Are these guidelines useful to follow? A series of papers and blogs present contrarian and supporting views (e.g. Gelman and Loken, 2014, Amrhein and Greenland, 2018, Mayo, 2019, Ioannidis, 2019a, Anjum et al, 2020). In particular, Mayo (2018) has characterized these debates as "the statistics wars". In several scientific domains such as psychology, much discussion has focused on misuses of the null hypothesis testing process (NHTP) and low powered studies.

Does this imply that NHTP should not be used in industrial applications? Should studies that did not deliver on promised improvements be flagged as invalid?

II. Statistical Analysis

Efron and Hastie (2016) present a comprehensive review of statistical analysis over time. Classical statistics consists of an algorithmic and an inferential part. Frequentism (or "objectivism") is based on the probabilistic properties of a procedure of interest as derived and applied to the output of a procedure of interest for observed data. This provides us with an assessment of bias and variance. The frequentists interpretation is based on a scenario where the same situation is repeated, endlessly. To achieve this, within the frequentism framework, several methods can be applied: 1) the plug-in substitution principle, 2) the delta methods Taylor series approximation, 3) the application of parametric families and maximum likelihood theory, 4) the use of simulation and bootstrapping computer intensive numerical methods and 5) pivotal statistics. These distinctions are important for reviewers to make. The Neyman-Pearson lemma provides an optimum hypothesis testing algorithm where a black and white decision is made. You either reject the null hypothesis in testing for an alternative hypothesis, or not. This offers an apparently simple and effective way to conduct statistical inference. However, confidence intervals are considered by many as more informative, although they are as well the object of criticism in the ASA statement (Barnett and Wren, 2019). Alternatively, the statistical analysis can be conducted within a Bayesian framework by transforming a prior distribution on the parameters of interest, to a posterior, using the observed data. In computer age analytics one distinguishes between algorithms aiming at estimation, prediction or explanations of structure in the data. Estimation is assessed by accuracy of estimators, prediction by prediction error and explanations are based on variable selection using variance bias tradeoffs, penalized regression and regularization criteria.

Mayo (2018) presents a new perspective on statistical inference based on the concept of severe testing, she labels it "error statistics philosophy". For error statisticians, a claim, or research finding, is severely tested if it has been subjected to and passes a test that probably would have found flaws, were they present (Mayo, 2018, p. xii). If little or nothing has been done to rule out flaws in inferring a claim, then it has not passed a severe test. Mayo identifies three types of models: Primary models, experimental models, and data models. Primary models break down a research question into a set of local hypotheses that can be investigated using reliable methods. Experimental models structure the particular models at hand and serve to link primary models to data models. Data models generate and model raw data, as well as checking whether the data satisfy the assumptions of the experimental models. Error statistical assessments pick up on the effects of data dredging, multiple testing, optional stopping and a host of biasing selection effects. Biasing selection effects are blocked in error statistical accounts because they preclude control of error probabilities. Error statistical accounts require–with justification– preregistration. Long-run performance requirements are only necessary and not sufficient for severity. Long-run behavior could be satisfied while the error probabilities do not reflect well-testedness in the case at hand. Tools that are typically justified because they control the probability of erroneous inferences in the long-run are given an inferential justification. It's only when long-run relative frequencies represent the method's capability to discern mistaken interpretations of data that the performance and severe testing goals line up. Mayo (2018) proceeds to presents a range of conceptual methods such as Bad Evidence, No Test (BENT), Probabilism, Performance, and Probativeness. Insevere tests yield BENT. Performance is about controlling the relative frequency of erroneous inferences in the long run of applications. *Probabilism*, views probability as a way to assign degrees of belief, support, or plausibility to hypotheses. *Probativeness* is scrutinizing BENT science by the severity criterion. In interpreting confidence intervals (CI), one needs to connect actual experiments with the idealized concepts. Specifically, 'The set of all confidence intervals at different levels of probability. . . [yields a] confidence distribution" (Cox 1958, p. 363). The severity logic is the counterfactual reasoning: Were μ less than the 0.995 lower limit, then it is very probable (> 0.995) that our procedure would yield a smaller sample mean than 0.6. This probability gives, SEV, the severity (Mayo, 2018, p. 195). In general, the reported analysis should be able to pinpoint the sources of failed predictions and indicate what is/is not learned from negative results (Haig, 2020). Every reported inference should include what can't be reliably inferred, what potential mistakes were not probed or ruled out and what gaps would need checking in order to avoid various misinterpretations of results, Mayo (2018, p. 437).

Another aspect, to be evaluated in reviewing an applied research paper, is study design. Some studies are based on observational data and some on interventions, or experiments, designed by the researchers. There are many publications on statistical methods to design experimental interventions. The following illustration is adapted from Kenett and Zacks (2014). Interventions are determined by factor level combinations, the effects measures through responses. One particular aspect in this methodology is the use of blocking and randomization which aims at increasing the precision of the outcome and ensure the validity of the inference. Blocking is used to reduce errors. A block is a portion of the experimental material that is expected to be more homogeneous than the whole aggregate. For example, if the experiment is designed to test the effect of polyester coating of electronic circuits on their current output, the variability between circuits could be considerably bigger than the effect of the coating on the current output. In order to reduce this component of variance, one can block by circuit. Each circuit will be tested under two treatments: no-coating and coating. We first test the current output of a circuit without coating. Later we coat the circuit, and test again. Such a comparison of before and after a treatment, of the same units, is called paired comparison. Another example of blocking is the boy's shoes example (pp. 97 in Box et al, 1978). Two kinds of shoe soles' materials are to be tested by fixing the soles on n pairs of boys' shoes and measuring the amount of wear of the soles after a period of actively wearing the shoes. Since there is high variability between activity of boys, if m pairs will be with soles of one type and the rest of the other, it will not be clear whether any difference that might be observed in the degree of wear out is due to differences between the characteristics of the sole material or to the differences between the boys. By blocking by pair of shoes, we can reduce much of the variability. Each pair of shoes is assigned the two types of soles. The comparison within each block is free of the variability between boys. Furthermore, since boys use their right or left foot differently, one should assign the type of soles to the left or right shoes at random. Thus, the treatments (two types of soles) are assigned within each block at random. Other examples of blocks could be machines, shifts of production, days of the week, operators, etc. Generally, if there are t treatments to compare, and b blocks, and if all t treatments can be performed within a single block, we assign all the t treatments to each block. The order of applying the treatments within each block should be randomized. Such a design is called a randomized complete block design. If not, all treatments can be applied within each block, it is desirable to assign treatments to blocks in some balanced fashion. Such designs are called balanced incomplete block designs (BIBD). Randomization within each block is important also to validate the assumption that the error components in the statistical model are independent. This assumption may not be valid if treatments are not assigned at random to the experimental units within each block.

Yet another aspect of statistical analysis, with a potential strong impact on the results, is selective inference. Selective inference is inference on a selected subset of the parameters that turned out to be of interest, after viewing the data. This selection leads to difficulties in reproducibility of results and needs to be accounted for and controlled in the statistical analysis. We can distinguish between out-of-study and in-study selection. The former is not evident in the published work and is due to publication bias, p-hacking or other forms of significance chasing. The in-study selection is however evident in the published work. This is due to selection by abstract, table, figure or highlighting results passing a threshold (Ioannidis, 2019b, Benjamini, 2019).

Finally, findings have to be presented and generalized. Generalization can be achieved by a range of methods, some intuitive, some conceptual and some more formal, invoking, for example, causal arguments (Pearl, 2015, Kenett and Rubinstein, 2017]. Findings can be presented in different ways, including an approach based on alternative representations, some with meaning equivalence and some with surface similarity (Kenett and Rubinstein, 2017).

III. Information Quality

In Kenett and Shmueli (2016a), a set of guidelines to help reviewers assess information quality of an applied research paper are proposed. The information quality concept (InfoQ) presented in Kenett and Shmueli (2014, 2016b), is a general framework for planning, tracking and assessing information quality using four components and eight dimensions. InfoQ is defined as "the utility of a particular data set for achieving a given analysis goal by employing statistical analysis or data mining", Kenett and Shmueli (2014, 2016b). InfoQ is affected by the data (X), the data analysis (f) and the analysis goal (g), as well as by the relationships between them. Utility is measured using specific metric(s) (U). By examining each of the components and their relationships, we can learn about the contribution of a given project. The eight InfoQ dimensions are: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Generalizability, Chronology of Data and Goal, Operationalization, and Communication.

A first step in assessing the information quality of a particular applied research study is to identify the study goal (g), utility (U), data (X) and analysis method (f). The report on the study should present this explicitly. Following that, one can move to assessing the eight information quality dimensions. Table 1, derived from Kenett and Shmueli (2016a). presents questions offered as a checklist to reviewers interested in assessing the information quality dimensions of a specific research study.

Table 1: Questions for Reviewing Information Quality Dimensions

Dimension	Questions
1. Data Resolution	1.1 Is the data scale used aligned with the stated goal? 1.2 How reliable and precise are the measuring devices or data sources? 1.3 Is the data analysis suitable for the data aggregation level?
2. Data Structure	2.1 Is the type of the data used aligned with the stated goal? 2.2 Are data integrity details (corrupted/missing values) described and handled appropriately? 2.3 Are the analysis methods suitable for the data structure?
3. Data Integration	3.1Are the data integrated from multiple sources? If so, what is the credibility of each source? 3.2 How is the integration done? Are there linkage issues that lead to dropping crucial information? 3.3 Does the data integration add value in terms of the stated goal? 3.4 Does the data integration cause any privacy or confidentiality concerns?
4. Temporal Relevance	4.1 Considering the data collection, data analysis and deployment stages, is any of them time-sensitive? 4.2 Does the time gap between data collection and analysis cause any concern? 4.3 Is the time gap between the data collection and analysis and the intended use of the model (e.g., in terms of policy recommendations) of any concern?
5. Chronology of Data & Goal	5.1 If the stated goal is predictive, are all the predictor variables expected to be available at the time of prediction? 5.2 If the stated goal is causal, do the causal variables precede the effects? 5.3 In a causal study, are there issues of endogeneity (reverse-causation)?
6. Generalizability	6.1 Is the stated goal statistical or scientific generalizability? 6.2 For statistical generalizability in the case of inference, does the paper answer the question “What population does the sample represent?” 6.3 For generalizability in the case of a stated predictive goal (predicting the values of new observations; forecasting future values), are the results generalizable to the to-be-predicted data? 6.4 Does the paper provide sufficient detail for the type of needed reproducibility and/or repeatability, and/or replicability?
7. Operationalization	Construct operationalization: 7.1 Are the measured variables themselves of interest to the study goal, or is their underlying construct? 7.2 What are the justifications for the choice of variables? Strength of operationalizing results: 7.3 Who can be affected (positively or negatively) by the research findings? 7.4 What can the affected parties do about it?

8. Communication	8.1 Is the exposition of the goal, data and analysis clear? 8.2 Is the exposition level appropriate for the readership of this journal? 8.3 Are there any confusing details or statements that might lead to confusion or misunderstanding?

IV. Statistical Evaluation Checklist

Our goal here is to set a framework for a reviewer considering aspects related to the statistical analysis of an applied research paper. These are structured in six parts:

1. Study design
2. Algorithmic and inferential methods in frequentism analysis
3. Bayesian methods in Bayesian analysis
4. Selective inference aspects
5. Severe testing properties
6. Presentation of findings

Specific questions addressing these sections are listed in Table 2.

Table 2: Questions for Reviewing Statistical Analysis in Applied Research

Part	Questions
1. Study Design	1.1 Is the experimental set up clearly presented? 1.2 Have aliasing and power consideration been taken into account? 1.3 Is there reference to blocking, split plots and randomization?
2. Algorithmic and Inferential methods	2.1 Are the algorithmic and inferential methods uses clearly stated? 2.2 Is the analysis aiming at estimation, predictive or explanatory goals? 2.3 Is data and code available to replicate the analysis? 2.4 Are outcomes of inferential analysis properly interpreted?
3. Bayesian Analysis	3.1Are prior distributions justified using prior experience or data? 3.2 What are the Bayesian methods used in the analysis? 3.3 How are Bayes factors interpreted?
4. Selective inference	4.1 Has the study neem pre-registered? 4.2 Have any false discovery rate corrections been made? 4.3 Is the presentation of findings affected by selective inference?
5. Severe Testing	5.1 Have the findings been tested with an option of failing the test? 5.2 Is the study a first or is it replicating previous studies? 5.3 Have Probabilism, Performance, and Probativeness criteria been considered? 5.4 What type of model is used in the analysis: Primary models, experimental models or and data models? 5.5 If used, how are CI interpreted?
6. Presentation of Findings	6.1 How are the research findings presented? 6.2 Have the research findings been generalized? 6.3 Are there any causality arguments presented? 6.4 In a causal study, are there issues of endogeneity (reverse-causation)?

These questions provide guidelines to reviewers assigned the task of assessing the statistical analysis of an applied research paper. They are not meant to be prescriptive and are only deigned as a sort of review checklist.

V. Some final comments

This final note is about evaluating the statistical merit of a data driven applies research study. A key group of stakeholders who should be able to perform such assessments are organizational data scientists. As a minimum, data scientists should understand the questions listed in Table 1 and 2 so that they can plan and evaluate their work. For more on this aspect see Kenett and Redman (2019).

Typically reports summarize findings without being specific as to how the data analysis was performed. However, data analysis pipelines affect the outcomes of statistical analysis, Botvinik-Nezer et al (2019). Part of this is the handling of missing data and outliers. These are usually not documented. For an exception see openml.org Vanschoren et al (2013). Reviewers of analysis uploaded to this platform would be able to fully replicate the study. We anticipate that the future will require a documentation of the data analysis pipeline, beyond current practice to put data and code under configuration control.

The review and guidelines in this series aim at providing professional support to industrial statisticians and applied statisticians in general. Specifically, the questions in Tables 1 and 2 are designed to structure a discussion between the people involved in a study. Properly managed discussion usually contributes to enhancing the quality of the related work. Our goal is to percolate some of the knowledge and insights derived in the context of science in general to the industrial statistics context.

References

Amrhein, V and Greenland, S, Remove, rather than redefine, statistical significance. Nat Hum Behav 2, 4, 2018.

Anjum RL, Copeland S and Rocca E, BMJ Evidence-Based Medicine ;25:6–8, 2020

Barnett AG and Wren JD, Examination of CIs in health and medical journals from 1976 to 2019: an observational study. BMJ Open 2019;9:e032506. doi:10.1136/bmjopen-2019-032506

Benjamini Y, Selective Inference: The Silent Killer of Replicability, Rietz Lecture, Joint Statistical Meetings, Denver, Colorado, 2019.

Botvinik-Nezer R, Holzmeister F et al, Variability in the analysis of a single neuroimaging dataset by many teams, bioRxiv, 2019, https://www.biorxiv.org/content/10.1101/843193v1

Box GEP, Hunter, W amd Hunter S, Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley and Sons, 1978.

Cox DR, Some Problems Connected with Statistical Inference, Annals of Mathematical Statistics 29(2), 357–72, 1958.

Efron B and Hastie, T, Computer Age Statistical Inference: Algorithms, Evidence and Data Science, Cambridge University Press, 2016.

Gelman, A and Loken, E, The Statistical Crisis in Science. American Scientist 2: 460-5, 2014.

Haig BD, What can psychology's statistics reformers learn from the error-statistical perspective?, Methods in Psychology, https://doi.org/10.1016/j.metip.2020.100020, 2020.

Ioannidis J, The importance of predefined rules and prespecified statistical analyses: do not abandon significance. JAMA 321:2067-2068, 2019a.

Ioannidis, J, What Have We (Not) Learnt from Millions of Scientific Papers with p-values? The American Statistician, 73:sup1, 20-25, 2019b.

Kenett RS and Rubinstein A, Generalizing Research Findings for Enhanced Reproducibility: A Translational Medicine Case Study, 2017. https://ssrn.com/abstract=3035070

Kenett RS and Shmueli G, Helping Authors and Reviewers Ask the Right Questions: The InfoQ Framework for Reviewing Applied Research, Journal of the International Association for Official Statistics (with discussion), Vol. 32, pp. 11-35, 2016a.

Kenett RS and Shmueli G, *Information Quality: The Potential of Data and Analytics to Generate Knowledge*. John Wiley and Sons, 2016b.

Kenett RS and Shmueli G, On Information Quality, Journal of the Royal Statistical Society, Series A (with discussion), Vol. 177, No. 1, pp. 3-38, 2014

Kenett RS, Zacks S and Amberti D, Modern Industrial Statistics: with applications in R, MINITAB and JMP 2nd Ed., John Wiley and Sons, 2014.

Kenett RS and Redman,T,The Real Work of Data Science: Turning data into information, better decisions, and stronger organizations, John Wiley and Sons, 2019.

Kenett RS, Swarz R and Zonnenshain, Systems Engineering in the Fourth Industrial Revolution: Big data, Novel Technologies, and Modern Systems Engineering, John Wiley and Sons, 2020.

Kenett RS, Zonnenshain A and Fortuna G, A road map for applied data sciences supporting sustainability in advanced manufacturing: the information Quality dimensions, Procedia Manufacturing, 21, pp 141-148, 2016.

Mayo D, P-value thresholds: Forfeit at your peril. Eur J Clin Invest, 49: e13170, 2019.

Mayo D, Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars, Cambridge University Press, 2018.

Pearl J, Generalizing Experimental Findings, Journal of Causal Inference 3(2): 259–266, 2015.

Reis M. and Kenett RS, Assessing the Value of Information of Data-Centric Activities in the Chemical Processing Industry 4.0, AIcHe, *Process Systems Engineering*, 64, 11, pp 3868-3881, 2018.

Vanschoren J, van Rijn JN, Bischl B, and Torgo , OpenML: networked science in machine learning. SIGKDD Explorations 15(2), pp 49-60, 2013.

Wasserstein R and Lazar N,The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70, 129–133, 2016.

Zonnenshain A and Kenett RS, Quality 4.0—the challenging future of quality engineering, *Quality Engineering*, 2020. https://doi.org/10.1080/08982112.2019.1706744