ON INFERENTIAL TECHNIQUES USED IN STUDIES ON TEACHING STATISTICS

Wenqi Liu and Von Bing Yap
National University of Singapore, Singapore
stayapvb@nus.edu.sg

Amidst the current debate on the practice of statistical inference, more attention should be directed to the random-sampling assumption. We demonstrate the ubiquity of the assumption in sample surveys, controlled experiments and regression models, and argue that the assumption is crucial to inference. If it fails, as is often the case, not only the P value, but also the confidence interval may not be meaningful. An informal survey of papers that present include inferential data analysis found a low rate of a mention of the issue: 1 in 10 papers in two volumes of Journal of Statistics Education, and 2 in 8 papers in one volume of Statistical Education Research Journal. None of the three papers discuss implications of the failure on the conclusions. We make several recommendations to help the statistics education community improve the practice of inferential data analysis.

INTRODUCTION

Recent commentaries, like Halsey et al. (2015), Wasserstein & Lazar (2016) – the statement of the American Statistical Association (ASA), Amrhein, Greenland & McShane (2019), renew the cultural struggle against misuse of statistical inference. The problem is vexing: while the pros and cons of standard techniques are well-understood and have been accessible to undergraduates for decades (see Freedman, Pisani & Purves (2007) and earlier editions), flawed practices persist in the research literature. The aim of this article is to focus on the random-sampling assumption in inference and to report a preliminary survey of how it informs research practice in three volumes of research papers on statistics education.

THE RANDOM-SAMPLING ASSUMPTION

Every inference assumes a statistical model, that essentially, the data are like the result of drawing some numbers at random from a population. This is an obvious technical point, but its practical significance seems largely ignored. It is part of the "underlying assumptions" in the ASA statement, which is distinct from the null hypothesis. For a broad insightful exposition, we recommend *Statistical Assumptions as Empirical Commitments* (Berk & Freedman, 2003; Freedman, 2010). Here, we focus on the role of the random-sampling assumption on the evaluation of null hypothesis in several types of commonly used models.

Consider two assumptions:

(a1) Data $x_1,\ldots,x_n$ are realisations of independent and identically distributed normal random variables $X_1,\ldots,X_n$ with mean μ and variance σ^2

(a2) The value of μ is 0

It is a mathematical consequence of (a1) and (a2) that $T = \sqrt{n}\, M / S$ has a t_{n-1} distribution, where M and S^2 are respectively the random sample mean and variance.

Now suppose the statistic t computed from the data is far away from 0, such that the P value $P(|T| > |t|)$ is quite small, say 0.01. Since the chance is 1 in 100 that a more extreme statistic is observed, this raises questions about the assumptions, like a contradiction argument. Either (a1) is false, or (a2) is false, or both are false. If (a1) is true, then we confidently doubt (a2). However, if (a1) is false, then we cannot know anything about (a2). The logic is the same in all survey studies: If the random sampling assumption is not true, then a very small P value does not provide evidence of the null hypothesis being false. For instance, if an interviewer only approaches people working in a financial district, he is likely to get the impression that the average national salary has increased from the previous year, even if it has not. Since (a1) is often false, most judgements of null hypotheses via the P value are questionable. These issues are meticulously explained in the classic undergraduate textbook *Statistics* by Freedman, Pisani and Purves (2007). Besides devoting the last chapter to a critical assessment of hypothesis tests, the book is also unusual in offering exercises that turn a test

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

around to interrogate the random sampling assumption (a1), in the case where (a2) is known to be true, and an explicit frequency interpretation of probability (Von Mises, 1981).

In an experiment comparing two or more treatments, the investigator may attempt to enroll subjects that represent the target population, but the selection process is often non-random, due to practical constraints, such as the need for consent from subjects. Thus a formal generalisation to the population may not be reliable. However, a rigorous comparison of the treatments within the enrolled subjects is feasible, by randomly assigning subjects into the various groups, thus controlling for confounding. Suppose treatment group i contains n_i subjects, and that there are a total of I treatments, so that the total number of eligible subjects is $n = n_1+\ldots+n_I$. We may modify the assumptions as follows:

(a'1) The responses from treatment group i are like the result of n_i random draws without replacement from the population of n responses which will be obtained if every eligible subject were given treatment i

(a'2) All the I population means have the same value

This model, which goes back to agricultural experiments (Neyman, 1923), is a powerful way to think about the meaning of causation, and is an excellent framework for drawing causal inference from such randomised controlled experiments. Altogether there are nI parameters, namely the totality of n responses to treatment 1, …, n responses to treatment I, of which we observe only n. The same issue above applies: if (a'1) fails, then the data may suggest rejecting (a'2) even if it is true. For instance, if a new drug is given to healthier subjects, even if it is no better than a placebo, it will seem better. Freedman, Pisani and Purves (2007) is singular among introductory textbooks in devoting substantial pages to a careful treatment of the randomised controlled experiment.

On regression models, Berk and Freedman (2003) state "Generally, the error distribution is not empirically identifiable outside the model… The error distribution is an imaginary population…" Here, we explore a view of regression models as arising from a real population. In a population of size N, let the weight and height of individual i be w_i and h_i respectively. Suppose the association between weight and height is linear. Then by the least-square method, we can determine constants c and m such that $w_i = c + m\ h_i + d_i$, where the deviations d's sum to 0 over the population, and the sum of $h_i d_i$ over the population is also 0. All quantities are fixed numbers, including the deviations. Suppose we have the weight and height of n individuals. In order to estimate m or test a hypothesis about its value, we need to assume

(a''1) The data are like the result of n random draws without replacement from the population

The same caveat applies: If we do not have a random sample, then the inference on m may not be worth much. A similar issue applies to more sophisticated statistical models, such as logistic regression, structural equations models, etc.

In summary, we have the following description of statistical inference which covers the three classes of problems presented above, namely, sample surveys, controlled experiments, and regression models. In every case, there are commonly two assumptions:

(A1) [Random-sampling] The data are realisations of some random variables whose joint distribution is partially specified up to the actual values of some parameters.

(A2) The parameters of the joint distribution in (A1) take certain specified values

Under (A1) and (A2), the distribution of a test statistic is completely specified, which can be used to calculate a P value based on the data. If (A1) is true, then a small P value casts doubt on the mull hypothesis (A2). However, if (A1) is not true, then a small P value says nothing about (A2). In the current state of affairs, too much faith is put into making judgment about (A2) using the P value, when in most cases, (A1) is false.

How often is the random-sampling assumption given an appropriate consideration in research literature? We decided to answer this question by an informal survey of two statistics education journals.

METHOD & RESULTS

We examined research articles published in Volumes 25 (2017) and 26 (2018) of the *Journal of Statistics Education* (JSE), and Volume 16 (2017) of the *Statistical Education Research Journal* (SERJ). Of the 36 JSE articles, 21 perform data analysis, of which 10 apply inference techniques. Of the 46 SERJ articles, 33 perform data analysis, of which 8 apply inference techniques.

In all the 10 papers in JSE, the random-sampling assumption is not satisfied. Only one paper acknowledges this fact as potentially problematic: "One limitation of the study is that it relies on a sample of convenience." (Shinaberger, 2017, p. 127) In all the 8 papers in SERJ, the random-sampling assumption is not satisfied. Two papers state: "As with all observational studies, lurking variables may be present that could explain differences in course performance and anxiety between different types of students" (Hedges, 2017, p. 333); "Lastly, instructors were not randomly selected for participation in the study, nor were students randomly assigned to each curriculum. Even if researchers originally involved in the CATALST Project attempted to represent a meaningful diversity of instructors and institutions, without random selection there is risk that an important subset of the desired population has not been adequately represented. Furthermore, without randomly assigning students to each curriculum we cannot confidently rule out that any comparison of outcomes between curricula would be meaningfully influenced by confounding variables." (Beckman, DelMas & Garfield, 2017, p. 433) However, none of the three papers discuss how these observations bear on the interpretation of their statistical inferences.

DISCUSSION

Is it too much to expect statistics educators to be more forthcoming about the potential unreliability of inference when the random-sampling assumption is false? Perhaps not, because this is a subject close to our heart, more so than other empirical researchers. A more systematic literature review is needed to establish if statistics educators turn out to be more careful. To be fair, all fields of empirical enquiry suffer from this oversight, because drawing random samples from well-defined populations is very hard work. Many times, data have already been collected, and it is reasonable to try to learn something from it. Moreover, it is often unethical to conduct a randomised controlled experiment in education research, where a promising new pedagogy is deliberately denied from the control group. We also acknowledge the peer pressure to employ sophisticated modelling techniques to hunt for statistical significance, though some journals are adapting editorial policy accordingly. In any case, our findings should motivate our community towards a more thoughtful practice of statistical inference, and more importantly, to serve as better role models for colleagues and students. Some concrete steps that are immediately feasible include the following. First, do more exploratory data analysis (Tukey, 1977) and creative visualisations to bring out the prominent features of the data, without being concerned with the question of statistical significance. Given that many research articles fall short in this area, it is not difficult to do more. Secondly, the practical significance of the findings should be contextualised in the actual problem. Thirdly, if there is a need to do inference, describe the assumed statistical model explicitly, including the random-sampling assumption and the presumed target population where it applies, and critically assess the plausibility of the modelling assumptions to help the reader evaluate the conclusions of the inference. We predict the first two steps will be important for a long time, while the third step is like a tactical move that is necessary is most situations now, but it can transition to a future practice where inference is only presented when the modelling seems compelling.

We fully embrace the conclusion of the ASA statement:

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

On the common suggestion of using confidence intervals, however, we have reservations. Granted, a confidence interval is more informative than a *P* value, as it tells us the parameter estimate as well as an approximate uncertainty of the estimate. But for it to work as advertised, the random-sampling assumption (A1) is required. Given that in most studies, this assumption is questionable, a confidence interval may not be worth much more than a *P* value. For example, we found that out of the 33 papers that analysed data in SERJ Volume 16, 8 employed inference. To construct a confidence interval around 8/33 x 100% ≈ 24.2%, it is necessary to assume that we had a random sample of 33 papers from a population, say all SERJ papers that analysed data. Then 8 is a realisation of a binomial random variable, and the standard error in the estimate of 24.2% is approximately the square-root of 0.242 x 0.758 / 33, which is about 7.5%. But (A1) is false: we looked at a convenience sample, i.e., one volume of SERJ, not a random sample. So the interval 24.2% ± 15.0% should not be interpreted as a 95% confidence interval. More sophisticated techniques are unlikely to overcome the violation of (A1) (Freedman, 2010).

Finally, we strongly encourage every statistics educator to study the content and pedagogy of Freedman, Pisani and Purves (2007) and the companion book of Freedman and Lane (1981). Since the first edition of *Statistics* appeared in 1978, these authors have been unsparing in their effort to clarify the basic statistical issues to the undergraduate students. Much of the current debate could have been averted if researchers and educators had paid more attention to these resources.

REFERENCES

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists Rise up Against Statistical Significance. *Nature*, *567*, 305-307.

Beckman, M. D., delMas, R. C., & Garfield, J. (2017). Cognitive Transfer Outcomes for a Simulation-based Introductory Statistics Curriculum. *Statistics Education Research Journal*, *16*(2), 419-440.

Berk, R. A. & Freedman, D. A. (2003). *Statistical Assumptions as Empirical Commitments*. In Blomberg, T.G. & Cohen, S. (Eds.) *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger* (2nd edition) 235-254. New York, NY: Aldine.

Freedman, D. A. (2010). *Statistical Models and Causal Inference*. Collier, C., Sekhon, J.S. & Stark, P.B. (Eds.) Cambridge, UK: Cambridge University Press.

Freedman, D. A. & Lane, D. (1981). *Mathematical Methods in Statistics*. New York, NY: W.W. Norton & Company.

Freedman, D. A., Pisani, R. & Purves, R. (2007). *Statistics (4th edition)*. New York, NY: W. W. Norton & Company.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The Fickle *P* Value Generates Irreproducible Results. *Nature Methods*, *12*(3), 179-185.

Hedges, S. (2017). Statistics Student Performance and Anxiety: Comparisons in Course Delivery and Student Characteristics. *Statistics Education Research Journal*, *16*(1), 320-336.

Neyman, J. (1923). Sur les applications de la théorie des probabilitiés aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczki (On the Application of Probability Theory to Agricultural Experiments)*, *10*, 1-51. English translation by Dabrowka, D.M. & Speed, T.P. (1990) *Statistical Science*, *5*, 463-480.

Shinaberger, L. (2017). Components of a Flipped Classroom Influencing Student Success in an Undergraduate Business Statistics Course. *Journal of Statistics Education*, *25*(3), 122-130.

Tukey, J. W. (1977). *Exploratory Data Analysis*. New York, NY: Pearson.

Von Mises, R. (1981). *Probability, Statistics and Truth*. Mineola, New York: Dover Publications.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129-133.