WHAT CAN STATISTICS EDUCATION OFFER TO DATA SCIENCE?

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

Data science relies heavily on statistical ideas, though it seems more concerned with prediction than statistics, which is more focused on modeling the data production process. This article will argue that the data scientist will do well to pay more attention to the likely disconnect between the chosen statistical model and the process it tries to emulate. Three learning goals are proposed and illustrated with elementary examples to help students grasp the idea. The disconnect is relevant to the replication crisis, yet is inadequately discussed in statistical communities. The lessons here are applicable to the education of statisticians.

INTRODUCTION

In its essence, data science seems a new term for data analysis. Novel data types from digital innovations drive technical innovations in data analysis, but have little impact on its fundamental character. Decades ago, digital representations of images and videos encouraged statisticians to learn new manipulation tools, yet "data analysis" did not change its name. However, it is true that statisticians have focused much energy on fitting a good model, if not the best model, to a data set, putting prediction on the backburner, while the machine learning community made much progress in making good predictions. Data science may be a synthesis of the two cultures (Breiman, 2001). Frequently, the data set has been collected without having a fixed question to answer. Statisticians ought to take a more fluid view of data analysis. It need not proceed in the sequence of formulating question, collecting data, analysing data, interpreting results, such as described in GAISE (2016). To be fair, a list is a sequence that forces certain choices. Perhaps it is better to present the four components in two dimensions.

Clearly, exploratory data analysis (EDA) and model-fitting are indispensable statistical ideas for the data scientist to make sense of the data and to make predictions. These are rightly embraced by the typical data science course. The thoughtful discussion of the pedagogical issues by Cobb and Moore (1997) is still very compelling. If any update is needed, it might be that the wide availability of software packages behooves us to reinforce students' acquisition of the subtler concepts in data analysis. This article will focus on the connection of data production and statistical models. Its importance to the data scientist will be discussed, and some relevant learning goals will be proposed.

THE IMPORTANCE OF DATA PRODUCTION

If the data set is already on the desk, it is too late to rue any flaw in the collection process. Still, the data scientist ought to find out about how the data were produced. Prediction of future observations typically means estimating the means of some subpopulations or strata defined by the predictor variables. The caveat is elementary and obvious: garbage in, garbage out. Introductory statistics textbooks talk about the importance of random samples for learning about a population. Media reports on surveys based on convenience sampling can occasionally be critical, though more should be done. However, if the data is multivariate, the caveat seems to be ignored, even by seasoned analysts. This could be due to a widespread belief that the problem goes away by modelling, for instance by regression. The statistics community seems rather reticent on this issue, a notable exception being David Freedman, whose brilliant theoretical and practical contributions can be found in Freedman (2011) and other works. At the bottom, the issue is not hard to grasp. If one has a sample of convenience, it is quite unlikely that any subsample is representative of the corresponding stratum defined by the predictor variables. As a result, biases of unknowable magnitude creep into parameter estimates, casting serious doubt on the validity of inference conclusions. I think this is a major cause of the replication crisis, besides the technical issues concerning the p-value that came under intense scrutiny in the last decade (Wasserstein and Lazar, 2016). More attention should be paid to it, starting in the classroom.

A secondary consideration is about doing inference using software. Given data sets, practically all software packages produce estimates, standard errors, confidence intervals, and P-

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

values based on some standard models. But the output rarely carries a warning that the procedures may not be valid if the underlying model assumptions are inadequate "empirical commitments" (Freedman, 2011). Thus the numbers appear definitive, as if the algorithms apply universally, like taking the logarithm of a positive real number. It is necessary for the data scientist to know the contingency of such outputs on assumptions about how the data were produced, which is a potential impetus to look for an alternative analysis if necessary.

LEARNING GOALS

Here are three learning goals pertaining to data production that should be useful for the future data scientist. Each goal will be explained in some detail and illustrated with examples.

[A] Standard inference procedures do not always work

From a mathematical point of view, this point is not surprising. The properties of inference procedures, such as the claim that the standard 95% confidence interval (CI) for a probability of success indeed has a coverage probability of 95% with a sufficiently large sample, depend on the fact that the data were generated from independent and identically distributed (IID) Bernoulli random variables. If it were not so, the properties may not hold. Students need to know this fact as early as possible, ideally when they first encounter an example on inference, which is likely before they see (if ever) a logical justification of the inference procedures. In order for students to know, it is better for them to experience concrete examples, even for theoretically prepared students.

Example A1. Suppose a simple random sample of size 100 was taken from a population of 5 million adults, to estimate the proportion who violated some law for controlling an epidemic. Since 100 is much smaller than 5 million, the draws can be regarded as independent, so that the responses (0: no violation, 1: some violation) are like realisations of IID Bernoulli(p) random variables X_1, X_2, …, X_{100}, p being the population proportion of violators. Unknown to the interviewers, a proportion $d < 1$ of those who violated the law will admit to it. It is intuitive that the standard procedure will underestimate p. Some algebra will convince the student that the estimate will fluctuate around dp, and that as the sample size increases, the coverage probability of the standard 95% CI will go to zero.

Example A2. A large population consists of two subpopulations whose sizes are proportional to w_1 and w_2, where $w_1 + w_2 = 1$. Simple random samples of sizes n_1 and n_2 are taken from the subpopulations, where n_i is not proportional to w_i. The $n_1 + n_2$ responses, all 0's and 1's, are displayed as an unannotated column. Then the standard analysis will yield a biased estimate of the population proportion of 1's.

Both examples must be demonstrated by computer simulations, which are excellent tools for students to learn coding, and also crucially to appreciate frequency-based probability, random variables, and to fully grasp the learning goal. For more complicated data sets, it is usually possible to propose multiple models, which might be compared in some automated fashion. However, in both our examples, the only sensible model is the standard one postulating IID Bernoulli random variables. The success probability is not the parameter of interest, and the problem cannot be fixed in an automated way. Further investigation into the process of data production is necessary to obtain an appropriate procedure. Although extremely simple, binary data occur so frequently in practice and their analyses are reported so widely, that it is a social duty to help students get the ideas clearly.

In many interesting surveys reported in the media, the selection process is not random. For example, soliciting responses on the internet is unlikely to yield a representative sample; it is a sample of convenience. We are justified to believe that the standard inference will perform even worse than our examples. Since it is more challenging to simulate a non-random process on a computer, it is hard to use computer simulation to support this belief.

A more general lesson is that, just because some inference procedure has been applied to a data set does not mean its conclusion is trustworthy.

[B] Given a data set with a fitted statistical model, state the model in terms of random variables

In an abstract sense, a statistical model fitted is a statement about data production. Stating the model means saying which numbers are assumed to be realisations of certain random variables, i.e., generated from their joint distribution; these belong to the response or dependent variables. The joint distribution is to be specified up to the parameters. Ideally, given certain values for the parameters,

pseudo-data sets can be simulated in a computer to illustrate the randomness in the model and also to verify the inference results or to check the goodness-of-fit of the model. Insightful references for teaching this idea at introductory and advance levels are Freedman, Pisani and Purves (2007) and Freedman (2009). Example A1 contains a statement of the standard model producing a list of 0's and 1's, so will not be repeated here.

Example B1. Numerical data y_i, x_{i1}, x_{i2}, …, x_{ip} are obtained from individual i, where i ranges from 1 to n. Let y be the nx1 vector $(y_1,\ldots,y_n)$ and X be the nxp matrix with entry x_{ij} at row i and column j. The standard regression or linear model with normal errors can be stated in several ways.
By entry: For each i, y_i is a realization of $Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$. $Y_1,\ldots,Y_n$ are independent random variables, and $\beta_1,\ldots, \beta_p, \sigma^2$ are parameters.
As vector: y is a realization of $Y \sim N(X\beta, \sigma^2 I_n)$, where $\beta = (\beta_1,\ldots, \beta_p)$ and σ^2 are parameters. I_n is the nxn identity matrix.

Given the popularity of the regression model, asking students to make such statements might seem overboard. But without explicit articulation, the model assumptions will not take hold firmly in the mind. Powerful software tends to exacerbate the difficulty, since so many models can be fitted easily, leaving less time to scrutinise the internal workings of any model. Another important benefit is that model articulation emphasises the connection between probability theory and inference, which is hard to grasp, and not often taught. In order to see the equivalence of the two statements, the students need to know a special property, that uncorrelated normal variables are independent. Hence basic knowledge in probability theory also gets reinforced.

Example B2. In B1, if y's are either 0 or 1, the standard model is the logistic regression. For each i, y_i is a realization of $Y_i \sim \text{Bernoulli}(p_i)$, where

$$\log(p_i/(1-p_i)) = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$Y_1,\ldots,Y_n$ are independent random variables, and $\beta_1,\ldots, \beta_p$ are parameters.

Unlike the linear model, there is no easy way to state the distribution of the random vector Y of a generalised linear model (GLM), of which logistic and Poisson regressions are the most commonly used special cases. The underlying distribution of a GLM is more elusive than a linear model.

Example B3. In B1, if y's are non-negative integers, the standard model is the Poisson regression. For each i, y_i is a realization of $Y_i \sim \text{Poisson}(\mu_i)$, where

$$\log(\mu_i) = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$Y_1,\ldots,Y_n$ are independent random variables, and $\beta_1,\ldots, \beta_p$ are parameters.

Example B4. Suppose each of n adults is classified as either old (>60 years) or young, and either diabetic or not diabetic. There are then three commonly used models for the four counts: n_{11} (old diabetic), n_{12} (old non-diabetic), n_{21} (young diabetic), and n_{22} (young non-diabetic), which add up to n.

[i] $(n_{11}, n_{12}, n_{21}, n_{22})$ is a realisation of $(N_{11}, N_{12}, N_{21}, N_{22}) \sim \text{Multinomial}(n, (p_{11}, p_{12}, p_{21}, p_{22}))$, where the p's are proportions summing to 1.

[ii] (n_{i1}, n_{i2}) is a realisation of $(N_{i1}, N_{i2}) \sim \text{Binomial}(q_i, 1-q_i)$, where q_1 and q_2 are proportions. (N_{11}, N_{12}) and (N_{21}, N_{22}) are independent random vectors.

[iii] (n_{1j}, n_{2j}) is a realisation of $(N_{1j}, N_{2j}) \sim \text{Binomial}(r_j, 1-r_j)$, where r_1 and r_2 are proportions. (N_{11}, N_{21}) and (N_{12}, N_{22}) are independent random vectors.

[C] Does the model describe the data production process accurately?

This question usually cannot be answered definitively. Nevertheless, it is useful to try to answer it, so as to temper the faith placed on the inference conclusions. An effect supported by a very small p-value, even if its size is judged to be practically significant, is not worth much if the underlying model is suspect. The discovery might be largely driven by biased data production, and by using an unsuitable model, bias in the effect estimate is misinterpreted as genuine signal.

Generally speaking, model-based analysis of data from observational studies has to be approached very cautiously. Even if the data were obtained by a simple random sample from a well-defined population, the standard model can be wrong because of non-response or response bias, such as Example A1. Rigorous surveys typically employ other probability sampling methods than simple

random sampling. If such data are analysed by the standard model, say for the purpose of classroom teaching, the students must be told clearly that there is a divide between the model and the data production process, that there are technical fixes to bridge the gap, which they can learn later. The mass media regularly report sensational findings from surveys that rely on samples of convenience. Such articles almost never acknowledge that it is extremely unlikely for any model to mimic the rather complex mechanism of data production, and therefore, any inference is suspect.

The simplest association study involves two binary variables. Borrowing from epidemiology, the study can be cross-sectional, cohort or case-control. If the data were obtained via simple random sampling from either the whole population, the two exposure strata, or the two response strata, then [i], [ii] or [iii] in Example B4 is respectively the correct model, provided the strata are large. If the parameter of interest is the odds ratio, then remarkably for all three models, the formulae for the maximum likelihood estimate and the approximate standard error are the same. However, many studies do not employ any probability sampling; it is challenging enough to get any data. This does not mean the data are not useful. Many important medical discoveries were made with such data, including the harm of smoking. But it ought to put a seed of doubt into inference results. We should also watch out for sentences such as "Since the study is cross-sectional, we may assume that the four counts are realisations of a multinomial distribution." A cross-sectional study can be based on a sample of convenience.

The logistic regression model (B2) amounts to independent simple random samples from strata defined by distinct combinations of predictor values present in the data set. This view nicely separates the prediction challenge into two components: the random part from the simple random samples, and the systematic part from the fact that most likely not all strata defined by the predictors in the population have been sampled. Using a linear combination to make predictions for unseen strata can be quite wrong. For a simple example, let there be $p = 2$ binary predictors, and suppose individual i has $(x_{i1},x_{i2}) = (0,0)$, $(0,1)$ or $(1,0)$ only, i.e., the $(1,1)$-stratum is not sampled. Then the logistic regression model has to be additive in order to predict the proportion of 1's in the $(1,1)$-stratum. Obviously, it will be wrong if there is interaction. These points apply to the linear and Poisson regression models, with obvious adaptations.

DISCUSSION

A data set has been produced by some unknown process. Ideally, there is a method, operating on the data set alone, or some auxiliary information, to design a statistical model that mimics the process closely. Example A1 shows there is no such thing in the simplest case of binary data. The prospect is likely no better for more complicated data types. Hence it is so important for the data scientist to develop a good understanding of the distinction between data production and statistical models, and the distance that often separates them.

Pedagogically, B must be come before C: explicit statement of the model precedes and facilitates an attempt to inquire the extent to which the model assumptions are rooted in reality. A can be anywhere, though it is put in the beginning as I feel the inference of a proportion is sufficiently elementary to establish a beachhead for the more demanding tasks in B and C.

If the three suggested learning goals seem unfamiliar in statistics education, I hope this article has made sufficient arguments in favour of their inclusion. Perhaps another consequence of the emergence of data science is a message to statistics educators to relook at our beliefs and practice. On this note, it is also time for statisticians to adopt a more relaxed attitude towards the concept of a true model. Perhaps the main conclusion of a study should be specific predictions about future observations, rather than a best guess of the true model, which often degenerates into a set of competing models which are too close to distinguish.

The random variable is a constant appearance in the presentation here, which is suitable for students who have some acquaintance with elementary manipulation of random variables acquired through, for example, a course in probability theory; a measure-theoretic understanding is not necessary. When teaching less-prepared students inference, formal or informal, these lessons ought to be included. The box model (Freedman, Pisani and Purves 2007) is a powerful and effective pedagogical tool to convey the essential messages in a more elementary way.

REFERENCES

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*, 199-231.

Cobb, G.W. & Moore, D.S. (1997). Mathematics, Statistics, and Teaching. *The American Mathematical Monthly*, *104*, 801-823.

Freedman, D.A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Freedman, D.A. (2011). *Statistical Models and Causal Inference: A Dialogue with the Social Science*. Ed. Collier, D., Sekhon, J.S. & Stark, P.B. Cambridge University Press.

Freedman, D.A., Pisani, R. & Purves, R. (2007). *Statistics*. Norton.

GAISE College Reort ASA Revision Committee, *"Guidelines for Assessment and Instruction in Statistics Education College Report 2016,"* http://www.amstat.org/education/gaise.

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

Yap, V.B. & Liu, W. (2019). On Inferential Techniques Used in Studies on Teaching Statistics. *"Decision Making Based on Data," IASE Satellite Conference.*