Journal of University Teaching & Learning Practice

Volume 18
Issue 2 *Innovations and challenges in the teaching of statistics to non-specialists*

Article 08

2021

Ten tips for statistical educators in response to a constructive review of the book How to Analyze Data

Margaret MacDougall
University of Edinburgh, Scotland, margaret.macdougall@ed.ac.uk

Ten tips for statistical educators in response to a constructive review of the book How to Analyze Data

Abstract

This paper comprises a review of the Pocket Study Skills book *How to Analyze Data* which I was invited by MacMillan International Higher Education to provide as a contribution to the current special issue. This includes both a preliminary review and an extension of this review to provide ten tips for educators to enhance the statistical learning of non-specialists in statistics. The tips are intended to provide a constructive approach to the review process through safeguarding educators, including dissertation supervisors, and learners, from being misled by some of the shortcomings that I have identified in this book. It is in this sense that I refer to these tips as "remedial tips". As such, these tips should not be regarded as exhaustive in their own right in reflecting the statistical learning needs of non-specialists. Nevertheless, one of the important responsibilities professionally trained statisticians face in the teaching of statistics to non-specialists is that of empowering learners to stem the tide of misuse of statistics within their own disciplines, and one of the corresponding challenges is that of identifying published material that is potentially unhelpful and even counterproductive in achieving this goal. Thus, it is also my intention that these tips will be of broader pedagogical value as a handy reference, including for those uninitiated teachers of statistics to non-specialists who may be tempted to reach out for a pocket guide to support them in balancing the demands of teaching and research in an effort to make statistics less mystifying to their students.

Keywords

statistical learning, non-specialists, book review

1. Preliminary Review

1.1. What are the Bibliographic Details of the Book under Review?

The bibliographic details for the book are provided below.

How to Analyze Data, by Catrin Radcliffe. MacMillan International Higher Education ('MacMillan') and Red Globe Press. London, 2020. pp. xi, 160. ISBN 978-1-137-60846-8 (Paperback version: £7.49)

How to Analyze Data belongs to the Pocket Study Skills series of books edited by Kate Williams of Oxford Brookes University and as such is rather small in dimensions: approximately 14 x 11 cm. This series focuses mainly, although not exclusively, on generic skills development, including poster design and presentation, stress management and successful groupwork.

1.2. Who is the Intended Readership and What is the Intended Scope of this Book?

In the front matter of the book, the reader is advised that "For the time-pushed student, the *Pocket Study Skills* pack a lot of advice into a little book.", while on the reverse cover, *How to Analyze Data* is described as a "beginner's guide to understanding, analyzing and interpreting data, from getting started to completion of analysis" and the reader is advised that it "will equip you with the skills and confidence to succeed in statistics." It is also made clear in the introduction that interpretation of data in the above sense is intended to refer both to the reader's own data for the purpose of report writing and findings reported in published work. (p. ix) In the former case, the author refers specifically to writing the methods and results section of a report. In the introduction, they also claim that this book should complement "traditional statistics books", including "through making the content of these more accessible". (p. viii)

1.3. Where Does this Publication Sit in Relation to Statistics Books with Similar Goals?

Given the reputation of statistics as a daunting and, indeed, unpopular subject for non-specialist learners in further and higher education, (Slootmaeckers & Kerremans, 2013) it is unsurprising that many attempts have been made over the years to reach such leaners through the provision of accessible textbooks, including:

a) theoretical books which provide foundational learning which could motivate students to engage better with curricular learning in higher education; and

b) step-by-step guides for the use of statistical packages, which are designed to support competent usage of statistical models and interpretation of output from statistical packages and are of relevance to students engaged in projects involving data analysis and reporting.

Examples of type a) include *Medical Statistics at a Glance* (which is published both as a textbook (Petrie, 2020) and a companion workbook (Petrie, 2013)), *Statistics at Square One* (now in its 11th edition) (Campbell & Swinscow, 2009) and *Statistics Without Tears: An Introduction for Non-Mathematicians* (Rowntree, 2018). Examples of type b) include *SPSS Survival Manual* (Pallant, 2020) and *SPSS for Psychologists* (Harrison & Brace, 2020), both of which are currently in their 7th edition and are supported by complementary webpages providing practise data, and *Statistical Analysis with Excel for Dummies* (Schmuller, 2016). In addition, the book *Statistics with R: A Beginner's Guide* (Stinerock, 2018) provides a thorough explanation of basic statistical theory, while integrating theory with practice through inclusion of the relevant commands from the

programming language and free software environment R. The book is also supported by a wealth of complementary online resources for instructors and learners.

I should emphasize that none of these books appear in pocket-sized form and that in the case of the books of type b) which I have listed, the attention to detail in terms of assumptions testing for choice of statistical procedures, interpretation of statistical output and use of worked examples necessitates this difference. It should also be noted that a key goal of all of the above alternative publications has been that of providing an *accessible* statistics learning resource for non-specialist learners in statistics and that their popularity in achieving this goal, among others, is reflected in several cases through progression to multiple later editions.

Additional pocket statistics (or data analysis) books have an alternative, specialist, readership in mind, as is the case with The ASQ Pocket Guide to Statistics for Six Sigma Black Belts, (Barsalou, 2015) with a target readership of those involved in Six Sigma methodologies (for improving business processes).

Thus, it appears that it is in its status as a Pocket Study Skills book, which economises on time and effort expended on accessible statistical learning, specifically in data analysis, that *How to Analyze Data* could serve as a unique contribution to existing literature. This is evident from the fact that the scope of the book does not include replacing or adding to the statistical ideas in texts of the sort listed, above and that the authors of these texts have already sought to rise to the challenge of making statistics accessible without loss of attention to detail.

I therefore consider the work of critically reviewing *How to Analyze Data* as important in staying abreast with the categories of book of potential value in enhancing the statistical learning experiences of non-specialist learners in higher education. In expressing this viewpoint, I am simultaneously noting the clear messages from the book, as expressed above in section 1.2, that its content is intended both to contribute to and complement the learning of statistics proper, although not to contribute new ideas to statistics as a discipline per se.

1.4. In What Ways is this Book Helpful and What are the Competing Issues?

Presentation of Content
The book contains five parts (1. *Getting started*, 2. *Understanding and describing your data*, 3. *How do statistical tests work?*, 4. *What statistical test do you need?* and 5. *The statistical process*) across which the 16 chapters are divided. The chapter contents are presented with an attractive layout, inclusive of eye-catching boxes which help in signposting the reader to questions they should address, depending on the statistical task they have in mind. In Part 1, which relates to project design and data preparation, this includes guidance on which parts of the book to navigate to, depending on the reader's answers to such questions. The signposting and navigational support in this book is exemplary in terms of sensitivity to the learning styles of students with specific learner difficulties (SpLDs), including dyslexia. (MacDougall, 2009) There is also a comprehensive subject index (pp. 155-160) which includes good coverage of the subject matter of the book.

Research Design

With relevance to getting started with research (Part 1), it is promising to see that in the chapter on defining a research question (Chapter 3), the reader is warned against data collection before the identification of "a sufficiently focused set of research questions" (p. 13). This advice is fundamental to all research contexts, including the use of surveys, where research questions ought to drive, but not be confused with, survey questions for respondents.

Indeed, given that survey design and, where anticipated, future statistical analysis, are inextricably linked, it is pleasing to note that a chapter is devoted to coverage of survey ("questionnaire") design tips (Chapter 4). A few highlights are:

- recognition of the potential for paper-based questionnaires to attract a higher response rate; and
- pictorial illustrations for choice of response categories to avoid pitfalls, such as overlapping numerical endpoints, which could obstruct future analysis of response data.

It would have been helpful, however, if, in this chapter, the author had provided some words of caution by explaining what was intended in their recommendation to pilot a questionnaire. A legitimate concern here is that the learner will carry forward this lack of rigour into later research without due awareness of the role of pilot studies in rehearsing study conditions using a well-defined preliminary sample. For example, if the sample is not representative of the target population, this will in turn undermine the utility of their pilot studies, so-called, in informing sample size calculations for future studies. The importance of designing a pilot study involving a representative sample can be illustrated particularly well within the context of designing questionnaires to develop measurement scales for a range of skills, attributes, abilities or conditions, referred to as "constructs". For example, it has been previously noted that "Pilot testing involves testing the scale to a representative sample from the target population to obtain statistical information on the items, comments, and suggestions" (Kyriazos & Stalikas, 2018). Within the context of a survey, items may be thought of as questions in a questionnaire used to represent individual components of a construct. The statistical information obtained from pilot testing can include results from techniques known as item analysis and factor analysis. (Price, 2017) As these results are designed, among other things, to determine which questions to include in the final questionnaire to ensure that the construct is accurately represented, the issue of representativeness is non-trivial. The representativeness of the sample is key to the reliability of the results, which, for the above methods, demands large sample sizes, some authors would suggest 10 participants per item. (Moilanen et al., 2019)

Data Preparation

Given the theme of Chapter 4, a few tips on the management of multiple response data and design of corresponding survey questions would have been an appropriate addition to the chapter on data preparation in spreadsheets (Chapter 5). This is especially true, given that preparation of such data for construction of graphs and for analysis can present problems for students. Nevertheless, in Chapter 5, the author includes some helpful advice on the naming of variables, which could serve a valuable role in averting error messages on reading Excel data into statistical packages, such as IBM SPSS (Statistical Package for the Social Sciences). As a useful contribution to management of the data preparation process in this chapter, they also provide a few excellent basic tips for data entry in Excel using clear annotations, including advice on freezing the top row, which, though a simple step, can be a deterrent to human errors when seeking to cross-reference data across rows of a dataset. The gentle introduction to use of Excel functions includes a listing of common operations and corresponding Excel syntax and in turn, a few screenshots of the use of simple formulae within an Excel spreadsheet to perform calculations. On account of its accessibility, this content could entice the reader to explore the wide range of available Excel functions further, thus developing their

skills in efficient data management in lieu of circumventing an otherwise formidable world of syntax.

Data Types and Choice of Descriptive Statistics

In Chapter 6, the overview of the standard types of data for statistical analysis is supported by appropriate illustrations, such as the final positions of athletes in a 100m sprint as an illustration of ordinal data.

In their account of descriptive statistics (Chapter 7), the author's coverage of summation notation presents clear and useful foundational learning, as does their illustration of the use of Excel operations across cells of a spreadsheet to calculate the above measures. However, their recommendation of the usage of the range in their roadmap summary table as the complement exclusively to the mode for ordinal data (p. 59) invites a few words of caution, despite the inevitable simplifications that can arise from a basic summary table.

It is important to bear in mind that ordinal data are data which represent an ordering but for which the arithmetic difference and mean are not meaningful. Typically, ordinal data arise from or give rise to numerical scores, such as on a pain scale. In the latter case, these scores are derived from descriptive categories. For example, the categories (*no pain*, *mild pain*, *moderate pain*, *strong pain*, *worst pain imaginable*) on a 5-point pain scale may be assigned the integer values 1 to 5, respectively to support statistical analysis of the data. Whether we choose to focus on the descriptive labels or the numerical tags, the data are still ordinal. We have no justification for asserting that the extent of the difference between *strong pain* and *worst pain imaginable* is identical to that between *mild pain* and *no pain*, as the numerical values are only indicators of order. For the purpose of illustration, let's assume that the above pain score categories represent response categories in a questionnaire. It is desirable to gain a sense of spread for the response data and the absence of a meaningful arithmetic difference and mean precludes the use of the standard deviation for this purpose, as it relies heavily on both these statistics. Nevertheless, the problem of appropriateness does not disappear by resorting to the range, as the range is itself an arithmetic difference. Thus, given that the range is commonly used as a measure of spread for ordinal data, it is evident that the quest for an informative measure of spread for ordinal data has led to a degree of compromise.

In the above example, if all categories but *worst pain imaginable* were assumed by the sample of questionnaire respondents, we would have a range of 3, and what would this tell the reader, exactly? (The reader may wish to confirm that there are alternative ways of arriving at a range of 3 for the above example, depending on the response categories which are assumed by the sample of respondents.) This example highlights the importance of recommending additional information, namely the minimum and maximum, to support the use of the range, so as to indicate where the value for the range came from.

However, even if the categories *no pain* and *strong pain* were specified as those corresponding to the minimum and maximum, respectively for the study sample in this case, the reader would still be at a loss to know how well populated all of the assumed categories were across the sample or, indeed, if the categories *mild pain* and *moderate pain* were assumed at all. A simple barchart is a particularly useful tool in compensating for this lack of clarity, provided there is a modest amount of distinct ordinal values to display. This is true, not only on account of the opportunity to display the frequency of individuals falling under each ordinal value but also, on account of the scope for matching the order of the bars in the chart with the natural order (for example, according to severity of pain) of these data. At a glance, one can also observe the mode, which the author lists as a measure to use with ordinal data.

Nevertheless, the lingering sense of mismatch between data type (ordinal data, for which there is no meaningful arithmetic difference) and measure of spread (range, which is defined as an arithmetic difference) must point to the fact that, mathematically speaking, the range is more meaningfully used within the context of continuous data, of which interval and ratio data are types, as the arithmetic difference *is* meaningful for such data. Granted, there is some literature, of which Psychology Statistics for Dummies (Hanna & Dempster, 2012) is an example, which promotes the inter-quartile range as preferable to the range for continuous data, where there are extremes in the data or the data are skewed. However, the range remains *appropriate* for such data. The inter-quartile range is itself a range – the range of the middle portion of the data, when the data are ordered by value, and is as such itself a limited source of information.

The author of *How to Analyze Data* does not express a preference for the inter-quartile range for continuous data or for special cases of such data involving non-Normal data (data that are not Normally distributed). They, in fact, omit the range altogether for continuous data and recommend the inter-quartile range alone as a measure of spread for non-Normal continuous data. By contrast, I would recommend the range for more general use with non-Normal continuous data and the inter-quartile range as a particularly valuable complement to the range to gain a more precise perspective on the data where the range is likely to be influenced by extremes. However, in such cases, I would also advocate using a) the minimum and maximum with the range and b) the quartiles (which include the median) with the inter-quartile range, to improve clarity concerning the distribution of the data.

It is possible that it is with the intention not to mislead in such cases that the author omits the range altogether for continuous data. However, it is important not to over-generalise by restricting usage of the range to ordinal data and, simultaneously, pairing off the range as a measure of spread exclusively with the mode as a measure of central tendency. Consider, for example, a surgeon requiring summary data on post-operative change in haemoglobin levels for a small sample of patients, further to surgery. If these continuous data turn out to be skewed in the direction of large declines in haemoglobin levels and there is no evidence to suggest that the measurement data are unreliable, ignoring the range, and the corresponding minimum and maximum values, could be rather costly in identifying existing, and potential future, patients at risk of mortality. Given the above, a key take-home point in relation to the author's aforementioned summary table is the appropriateness of including the range for consideration alongside continuous (interval and ratio) data that are not Normally distributed, not just (and conceptually speaking, certainly not at best) ordinal data.

I shall also simply note in passing that in the associated table and a later extension to this table (p. 75) in Chapter 8 (on choice of plot), use of the term 'Central tendency' would seem to serve as a more appropriate header than those of 'Averages' and 'Average value', respectively, to represent the entries 'Mean', 'Median' and 'Mode'.

Simple Charts for Presentation of Findings

In the above extended table, which the author attempts to use to match methods for displaying data to data types, their effort to be concise clouds interpretation. For example, it seems misleading that in the row used for interval and ratio (collectively, continuous) data, the barchart is listed as a possible choice of plot. The author does not explain this choice, which compounds the issue. However, several of the key concerns are highlighted by Weissgerber et al. as follows:

> *[B]ar graphs are designed for categorical variables; yet they are commonly used to present continuous data in laboratory research, animal studies, and human studies with small sample sizes. Bar and line graphs of continuous data are "visual tables" that typically*

> *show the mean and standard error (SE) or standard deviation (SD). This is problematic for three reasons. First, many different data distributions can lead to the same bar or line graph. The full data may suggest different conclusions from the summary statistics.*
> *Second, additional problems arise when bar graphs are used to show paired or nonindependent data. Figures should ideally convey the design of the study. Bar graphs of paired data erroneously suggest that the groups being compared are independent and provide no information about whether changes are consistent across individuals. Third, summarizing the data as mean and SE or SD often causes readers to wrongly infer that the data are normally distributed with no outliers. These statistics can distort data for small sample size studies, in which outliers are common and there is not enough data to assess the sample distribution. In contrast, univariate scatterplots, box plots, and histograms allow readers to examine the data distribution. This approach enhances readers' understanding of published data, while allowing readers to detect gross violations of any statistical assumptions. The increased flexibility of univariate scatterplots also allows authors to convey study design information. In small sample size studies, scatterplots can easily be modified to differentiate between datasets that include independent groups and those that include paired or matched data.* (Weissgerber & Winham, 2015)

I would encourage my readers to consult the above reference for more details, including examples illustrating the above shortcomings of using barcharts with continuous data.

Appropriately so, in the same table the author also lists the histogram for continuous data. However, in presenting a histogram earlier in the same chapter, they have also advised that, "Histograms are continuous: no gaps between bars." (p. 69) While this advice is probably intended to distinguish the nature of a histogram from that of a barchart, it is not sufficiently clear.

In particular, it is important not to overlook the utility of the histogram more broadly in data exploration to inform choice of analyses through consideration of cases where the data may appear continuous according to the scale of measurement, but there are gaps in the scale in terms of which values are actually assumed by the individuals within the sample. Through identifying gaps in the coverage of particular value ranges for a sample of numerical data, histograms can be useful in informing categorisation of the data for subsequent data analysis. In the prior exploration of tumour density data for analysis, for example, such gaps may be extensive and warrant the categorisation of the tumours according to size through first banding the densities into value ranges to make the interpretation of the data more meaningful. Of course, gaps of the above sort can be the product of having a small or non-representative sample, and this possibility must be considered before categorising in the above way. Either way, the above possibilities should not be discounted.

More generally, in relation to basic Excel graphs, the author introduces a few illustrated tips on best choice through contextual examples in Chapter 8. However, with a view to the cultivation of sound interpretation of statistical findings, I would advise against displaying pie-charts with percentages (p. 64) but no frequencies, at least without recommending that the report writer include the individual frequencies for the slices, or even the total frequency, in a caption. Arguably, a chart should be a self-contained item, enabling the reader to make a sound judgement based on the information displayed and not to be swayed by appearance. As is well known, where sample data are meant to reflect patterns or relationships in a larger population, the accuracy of summary data, including percentages, as estimates of the population data, will increase with increasing sample size, provided that the corresponding samples remain representative of the parent population, and thus depending on the quality of the sampling procedure. Often, there will be flaws in the sampling procedure that compromise the degree of representativeness. However, in good scientific reporting,

the reader will have the full picture regarding strengths and limitations of the study and be well placed to make an overall judgement concerning the usefulness of the chart based on the quality of the study design and the individual group sizes pertaining to the slices of the pie. While it could be left to the discretion of the student where to include frequency information for a pie chart in a report, the difficulty lies in omitting to highlight the need for frequencies to be provided in a sufficiently accessible way when viewing a pie chart. The main concern here is that graphs can mislead and, while the book is designed to be concise, there is a danger that missing out advice on frequencies, particularly with reference to small groups sizes, provides the basis for prospective data analysts drawing misleading conclusions from observed patterns or, perhaps unwittingly, persuading their reader to do so.

1.5. How Might the Remaining Shortcomings of this Book Be Put to Good Use?

In addition to the issues discussed in section 1.4, I have identified some key shortcomings in *How to Analyze Data* that relate to Parts 3 and 4 of the book (Chapters 9 – 14, pp. 76 – 133) together with supporting content from earlier chapters. The key areas include the practice and interpretation of hypothesis testing, recognition of the importance of confidence intervals (CIs) and treatment of the closely allied topics of sampling procedures and management of outliers. Several of the shortcomings reflect or at least, have the potential to foster, common misuses of statistics within non-specialist disciplines in higher education. Occasionally, the author also points the reader to references for supplementary reading which I do not consider to be fit for purpose.

These observations provide the impetus for offering a coherent and principled approach to avoiding bad practice being perpetuated. I am therefore presenting the remainder of this book review mainly in the form of ten tips emerging purely from my observations on areas for improvement in the treatment of the above key areas in *How to Analyze Data*. It is in this sense that these ten tips serve as remedial tips. A more extensive publication will be required to offer teaching material which puts these tips effectively into practice. In providing these ten tips, I am not assuming that the intended readership of *How to Analyze Data* is necessarily students within the mathematical sciences, as the author does not suggest that this is the case. Also, I do not intend these tips to be considered in their own right as complete. I do hope, however, that they will serve as a handy gatekeeping tool for sound statistical practice to support educators, including uninitiated and time-constrained teachers of statistics to non-specialists who are themselves non-specialists in statistics.

2. Ten Tips for Statistical Educators to Support Student Learning in Statistics

2.1 Make Explicit the Place of Statistical Hypothesis Testing and Confidence Intervals in Inferential Statistics.

In Chapter 2 on “how to successfully analyze your data” (p. 5), the recommendation, “You might also need to perform a statistical test (inferential statistics)” (p.11), does not provide the reader with a clear sense of what statistical inference entails. Although the author does not say so explicitly, the model of hypothesis testing which they represent in their book is null hypothesis significance testing (NHST). (Goodman, 1999) This, the most pervasive model for hypothesis testing across the disciplines, is characterised by an approach involving the null hypothesis as the test hypothesis and the use of a significance level as an upper threshold value for deciding when the *p-value* (defined in this book) is low enough to support rejecting the null hypothesis in favour of the study (or, alternative) hypothesis. In preparing the way for the learner to engage with this model, it is vital to stress that while it is sample data that are tested, statistical hypotheses refer to the corresponding

parent populations from which these, ideally representative, sample data have emerged. An absence of awareness of this fundamental truth will inevitably leave the learner inept in the articulation of their findings. The seed for raising awareness of this truth is actually contained in the first few sentences of Chapter 9, "What is a statistical hypothesis?". Here, the author states,

> *A key purpose of statistics is to generalize results or make predictions based on the sample data available. This type of statistics is called inferential statistics, where sample data is used to make inferences about a larger population.* (p. 77)

However, regrettably they omit to carry forward this early mention of inferential statistics into something tangible, including in relation to the interpretation of the findings from statistical hypothesis testing they provide in illustrations.

Relatedly, the author's consistent neglect of CIs when referring to inferential statistics should not be regarded lightly. In the teaching of basic statistics, there is much to be gained from uniting the teaching of hypothesis testing with that of CIs, including through exploring the role of the null hypothesis value in determining whether a CI is representative of statistical significance, not to mention what the CI can add. I should emphasize that this is not a novel perspective on my part, but, rather, a reflection of the fundamental nature of CIs and the tendency to arrive at misleading conclusions through relying solely on *p-values* from hypothesis testing. For the interested reader, there is a huge literature on this topic, examples of which I reference herewith (Gardner & Altman, 1986; Ranstam, 2012; Sim, 1999; Wasserstein, 2016).

By way of illustration, Sim and Reid provide the following recommendations:

> *A CI should be included whenever a sample statistic such as a mean (or difference in means) is presented as an estimate of the corresponding population parameter ...*
> *Confidence intervals should be provided in addition to (or even instead of) the results of hypothesis tests, with the level of confidence for the CI matched to the level of statistical significance for the hypothesis test.* (Sim, 1999)

Their justifications rest on the fact that in NHST testing, whereas the sample estimate of the true (or, population) effect size, as an estimate on a numerical scale of the strength of the relationship being tested, is one taken from one of many possible samples, the need to be addressed is that of capturing the true effect size with a given level of accuracy and confidence. The CI does this by providing a margin of error within which the true value lies with a specified level of certainty. It also includes the effect size estimate taken from the study sample. Furthermore, with the exception of any possible anomalies arising from efforts to develop non-standard CIs under special conditions not covered in *How to Analyze Data* or mismatches between methods, interpretation of the CI is consistent with that of statistical significance. Specifically, for a statistical significance level α, the $(1-\alpha)$ % CI excludes (includes) the null hypothesis value for the estimated effect size if and only if the corresponding hypothesis test yields (does not yield, respectively) a statistically significant effect. By way of illustration, for a significance level of 0.05, consider the fictitious results 0.51 (95% CI: (-0.36, 1.38), n = 84) for difference in mean test scores across two classes of students (psychology and sociology students) on subtracting the mean score for psychology students from that of sociology students. The results reveal a difference in means of 0.51 at the sample level. As the 95% CI includes the null hypothesis value of 0 for the difference in means, we also already have the result from the corresponding hypothesis test (an independent samples t-test) that for $\alpha = 0.05$, there is no significant difference in mean scores between the two groups of students. Furthermore,

we can with 95% certainty conclude that the true difference in mean scores lies somewhere between -0.36 and 1.38. Thus, we can see that the CI delivers the information required from a *p-value* based on hypothesis testing, namely what we need to decide on the presence or absence of statistical significance.

Additionally, we can see that, relative to the magnitude (0.51) of the sample mean difference, the magnitude of the width of the 95% CI (1.74) is high. This suggests that a larger study would be useful to improve the accuracy of the sample difference in means as an estimate of the true difference in means. However, if we consider a difference in mean scores of 10 to be of practical importance (scientifically significant), we can also see at a glance that, provided all forms of bias during the sampling process were minimised previously, one would not expect an increase in sample size of any order to capture a difference of practical importance. Thus, our CI can also be a useful means, independently of statistical significance, of informing us that the true effect is unlikely to be of practical importance. This can, of course, inform future practice within a given field. For example, in the above educational study, based on the similarity of the test scores on average, there might not be a case for changing the approach to student learning across the two disciplines.

Relatedly, as far back as 1986, Gardner and Altman had already offered the following advice:

> *Presenting P values alone can lead to them being given more merit than they deserve. In particular, there is a tendency to equate statistical significance with medical importance or biological relevance. But small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant only because the number of subjects studied was small.* (Gardner & Altman, 1986)

The use of CIs in the above senses is incredibly important for reporting purposes. The need for CIs arises from the fact that a single application of NHST cannot capture sampling variation for an effect size estimate. By virtue of the latter impediment, a single application of NHST cannot provide a representation of the accuracy of a given sample effect size as an estimate of the true (or, population) effect. By contrast, a CI can provide the limits within which the true effect size lies with a high level of certainty. It is also possible to see whether this corresponding range of values reflects, not only statistical significance but also, scientific significance in the sense that the population effect size is likely to assume a value of practical importance.

It is also important to note that the generation of CIs for the most basic effect size estimates, such as the difference in means can be performed with great ease using appropriately chosen statistical packages, such as through the simple selection of an additional option when performing the corresponding hypothesis test, assuming users are taking a menu-based approach to using a statistical package.

While the debate on the role of *p-values* in scientific reporting may simmer for many years to come, those who continue to advocate their use should guard against giving pride of place to the *p-value*, even when using NHST. (Wasserstein, 2016)

For completeness, it is also worth mentioning that the use of *p-values* and corresponding CIs are each grounded on theory which assumes that the study samples are random samples of the parent populations from which they originate. For a variety of reasons, some of which relate to research conduct, this scientific paradigm is rarely achieved. However, the awareness of such a condition

may help learners in deciding whether the use of the above methods is justified based on their own research conduct, including where the study sample is a convenience sample.

2.2. Motivate the Student to Ask, Are the Sample and Target Population Distinct Entities?

The question *Are the sample and target population distinct entities?* is simple to frame. In my regular experience, the need for research students to be prompted to ask it originates in a lack of awareness of the importance of remedial tip 2.1, above and in particular, the following corollary:

> *If the sample and the population are the same entity, the sample descriptive statistic, for example, mean, odds ratio or difference in proportions, equates with that of the population, and the application of statistical hypothesis testing is superfluous. Indeed, in such cases, such testing could prove misleading based on statistical power limitations in relation to a hypothesised larger population which is superfluous to the needs of the study*.

In contexts where students are assessing, for example, responses of specific patients or clinicians to a telemedicine intervention across a fixed time period, the concept of a (separate) target population may not be appropriate as the individuals involved in the experiment and indeed, the experimental conditions, are difficult to generalize to everyday practice settings. While this problem can rear its head even within the context of randomised controlled trials, including on account of participant characteristics and experimental settings, (Mulder et al., 2017) there are less sophisticated experimental design contexts where the results of data analysis are intended to reflect performance of an intervention within a restricted timeframe and where there is no apparent rationale for extrapolation to other settings. These include clinical audits, which are considered important for UK undergraduate medical students to participate in as part of their training in preparation for becoming junior doctors. (Hexter, 2013)

Given these observations, readers of *How to Analyze Data* could have benefited from pointers for determining whether the nature and quality of their study design guarantee that a generalisation from sample to population level statistics is meaningful, and thus that the use of hypothesis testing is appropriate. Addressing the hiatus, even briefly, in instructing the learner on how to take ownership of the above decision is important. Noting the diversity of learning situations which student data analysts may find themselves in, it should not be assumed that "the tutor" has the answer (cf. p. 11).

2.3. Raise Awareness of the Roles and Limitations of Different Sampling Procedures.

The current tip, as stated, is inextricably linked with the idea of assessing the suitability of study design for engaging in inferential statistics, and as such, is closely tied to the needs already expressed under the previous tip.

Respecting the role of *How to Analyze Data* as a Pocket Study Skills book, pointers to key references for future reading on the roles and limitations of sampling procedures would have been adequate. However, the pointer to *Statistics in a Nutshell* for guidance on "how to select samples" (p. 39) does not seem fit for purpose. As the full title of this reference, *Statistics in a Nutshell: A Desktop Quick Reference* (Boslaugh, 2012), suggests, on more careful examination of the content, the learner will find themselves in need of further references to obtain the guidance they require to put theory into practice. For this reason, I recommend Cochran's *Sampling Techniques* (Cochran, 1977) as a more suitable reference. Noting that the author of *How to Analyze Data* states that their own book "is designed to complement traditional statistics books", making the content of these more accessible, a useful add-on to complement the work of Cochran would have been advice on constructing simple and stratified random samples in Excel.

2.4. Motivate Students to Explore Whether There is a Theoretical Warrant for the Removal of Outliers.

While it is appropriate for the author to raise the question, "How will you deal with outliers?" (p. 11), it is important to note that there is considerable variation in practice across the disciplines in choice of warrants for removal of outliers. Correspondingly, it seems hopeful that the author provides a pointer to supporting literature, given the absence of explicit guidance on management of outliers in *How to Analyze Data*. The reader is directed to Upton and Cook's Oxford dictionary of statistics. (Upton & Cook, 2014) Interestingly, this reference provides comprehensive details on the detection of outliers. However, support, including cautionary notes, for decision-making on the removal of outliers remains lacking. It is important to bear in mind that this reference is a dictionary, not a guide on good practice, and that the shortcoming here lies more specifically in the exclusive choice of this reference rather than the quality of the dictionary per se. In *How to Analyze Data*, therefore, a warning for the uninitiated regarding the dangers of outlier hacking would have been apposite.

Non-specialists ought at least to be warned about the susceptibility of small samples to outlier detection, with the possibility of new outliers being generated iteratively on removal of old ones, leading to an outlier hacking exercise with no guiding rule for initiating the process.

A guiding rule in the latter sense, may, for example, require removal of outliers exclusively when biochemical measurements are off-scale, pointing to contamination of equipment. The importance of including early warnings on this topic in *How to Analyze Data* should not be undermined by appeal to the intended scope and size of the book. For example, as has been examined elsewhere, outlier hacking can promote *p*-hacking. (Pollet & van der Meij, 2016) The term *p-hacking* refers to a range of inappropriate behaviours involving the manipulation of data and the analysis of such data to arrive at statistically significant results and is one of a family of behaviours aimed at cherry-picking statistically significant results. In the case of outliers, manipulation of data can include making post-hoc decisions as to whether to include outliers based on the results of statistical significance testing, so as to favour statistically significant outcomes and in turn, to increase the likelihood of the final 'results' being published. On account of its widespread prevalence, (Simonsohn et al., 2014) *p*-hacking appears to be regarded in some communities as a respectable form of scientific fraud. However, it is already recognised that, "*p*-hacking can allow researchers to get most studies to reveal significant relationships between truly unrelated variables." Therefore, the potential cost to key stakeholders of research findings should not be underestimated. (Simonsohn et al., 2014)

2.5. Endeavour to Guard Against Cherry-Picking Through Post-Hoc Changes to Research Questions.

With reference to a single "investigation", the author presents research by means of a diagram as a "cyclical process" (p. 16) involving refinement of the original research question based on the results of research. In response to this diagram, which lacks any preliminary words of caution, it needs to be said that *Statistics for non-specialists should be made as simple as possible, but not at the potential cost of scientific integrity.* On account of the arguably non-intuitive nature of statistics, from the perspective of student learning, tension between simplicity and integrity can arise. Nevertheless, for reasons which I shall now explain, if research findings based on data analysis are to be trusted, greater clarity is needed than that afforded by the message forthcoming from the author's cyclical diagram.

Many student research projects can be opportunistic in the sense that a research team has a pre-existing dataset available from which a new student project can be developed. Developing new

questions from observed baseline data can be fraught with bias, as a scan of the data can give a sense of which study hypotheses could yield significant correlations and differences. By contrast, scientific research ought to be informed by what is scientifically meaningful, and this can include a wide range of possibilities, based on both what is already known about the field, but needs to be confirmed in the current study, and what is novel but scientifically meaningful. Nevertheless, much can be learnt from retrospective studies, including case-control studies, provided that researcher transparency is practised and cherry-picking of variables and corresponding analyses is discouraged. Sadly, formal procedures for the widespread rigorous assessment of full protocols in advance of research and for cross-checking dissertation content with protocol content do not typically feature in undergraduate student research modules, to the detriment of graduate understanding of scientific integrity.

For this reason, in preparing students for research involving data analysis, the author could have set the standard by explicitly advising against the malpractice of redefining a research question based on results unless the redefined research question is to form the basis for a new study, *involving transparent reporting from the current study*. Where this standard is compromised, publication bias in studies can include the failure to report 'negative' findings, where, in fact, such findings may be needed for the moderation of positive findings from other studies and indeed, in identifying inherent disparities across different study populations. This has a carry-over effect to the validity of systematic reviews, which rely on a synthesis of available studies, both for informing policy making and for evaluating the evidence for the effectiveness of prospective interventions. Setting the standard for research in the above sense is therefore of ethical relevance within the wide range of fields in which statistical findings inform decision making for the public good, including business, education and Medicine.

For a more detailed discussion regarding the problem of defining questions within a given study further to obtaining results, the interested reader may wish to refer to Kerr's 1998 article on a relevant by-product of this behaviour knowing as HARKing. (Kerr, 1998) As a final note, I wish to emphasize that the above recommendations rest on the assumption that the researcher is relying on NHST as their approach to hypothesis testing. With specific reference to the application of Bayesian inference, which the author does not consider, it has been argued that, "the evidence for a theory is just as strong regardless of its timing relative to the data." (Dienes, 2011)

2.6. Consider the Conditions for Use of and the Scope of Simple Tests of Association Between Categorical Variables.

The role of Fisher's Exact test as a more conservative alternative to the chi-square test of association when testing for an association between two categorical variables, and the conditions under which use of the chi-square test of association is considered to be legitimate, are helpful contributions to basic learning on hypothesis testing. Therefore, it is commendable that Fisher's Exact test is highlighted in *How to Analyze Data*. I say this bearing in mind that this book is *not* designed to take on board the philosophical and mathematical debates concerning the use of this test. It is also not an academic piece with a new thesis for consideration, but focuses on core traditional methods in basic statistics. Readers of *How to Analyze Data* should therefore expect the specified conditions for use of the chi-square test of association to follow traditional practice unless a clear justification and suitable references are provided. Furthermore, in a book on data analysis, such content, inclusive of explaining the scope of Fisher's Exact test, should be aligned with available options for students in popular statistical packages, such as SPSS.

Following traditional practice, the rule of thumb for use of the chi-square test of association instead of Fisher's Exact test can be stated as follows:

> *if in a two-way contingency table, at least 80% of the expected counts (not including the totals) are of size 5 or more (that is, of size ≥ 5), use the chi-square test of association; otherwise, use Fisher's Exact test*.

For ease of reference, I will refer to the above version of the rule of thumb as the traditional rule of thumb. This is consistent with the following advice from Clark-Carter :

> *The usual rule of thumb is that all of the expected frequencies in a 2 x 2 table should be at least five. In the case of tables which are larger than 2 x 2, at least 80% of expected frequencies should be at least five.* (Clark-Carter, 1997) (pp. 232 – 233)

Interestingly, specification of an upper limit of 10, which has its roots in debates dating back to the 1940s (Cochran, 1952), appears in the Concise Oxford Dictionary of Mathematics (Clapham, 2014), but it is less common.

As Clark-Carter notes, an application of the traditional rule of thumb to the 2 x 2 case *requires 100% of the expected counts to be of size at least 5* for the chi-square test of association to be used. Also, the requirement for the 2 x 2 case that 100% of cells (4 out of 4 cells in a 2 x 2 contingency table, excluding the totals) should have expected count of size at least 5 is the standard one assumed in statistics textbooks (Petrie, 2020). Yet, for this case, (the only case which they consider) the author of *How to Analyze Data* states the following:

> *… in order to use a chi-squared test, we need the expected values (E) to be greater than 5 in at least 75% of cases. Otherwise, Fisher's Exact test can be used*. (p. 115)

The author's intention to use the rule as they have presented it is evident from their inclusion of the content, "Chi-squared test of association (E ≥ 5 in 75% of cases)" in their flowchart for "Categorical data". (p. 103)

I have not come across this variant of the above traditional rule of thumb elsewhere. For this reason, I am surprised to find it in such an elementary book on statistics without any reference to the literature. Its use accommodates the outcome of having only 3 out of 4 cells in a contingency table, excluding the totals, with a count of at least 5 as a satisfactory condition for use of the chi-square test of association. Being exposed to it could leave the learner in a dilemma as to whether or not to use the chi-square test of association for a given 2 x 2 case, where the requirement for use of this test in the traditional rule of thumb is not satisfied. It is possible that the choice to follow the author in this context could have a tendency to increase the likelihood of incorrectly rejecting the null hypothesis (a Type I error), although this would need to be investigated through simulation studies.

The generalisability of Fisher's Exact test beyond the 2 x 2 case is recognised in SPSS by the available functionality for hypothesis testing, although, strictly speaking, the test used beyond the 2 x 2 case is called the *Fisher-Freeman-Halton Test* (https://www.ibm.com/support/pages/fisher-exact-test-rxc-table-fisher-freeman-halton-test). As is implicit from the above quotation from Clark-Carter, and the use of the above traditional rule of thumb beyond the 2 x 2 case is already recognised in the literature. (Clark-Carter, 1997) It would have been appropriate for the author to have included these generalisations in *How to Analyze Data* in preparation for use of a reputable statistical package. For completeness, I would like to stress here, however, that I am omitting consideration of higher level discussions regarding the shortcomings of Fisher's Exact test (Lin, 2008) and the traditional

rule of thumb for use of the chi-square test of association. These invite a level of statistical and philosophical understanding that is beyond the scope of *How to Analyze Data* and the associated recommendations have not to my knowledge found their way into other introductory statistics books. I am simply responding to the author of the above book in their own terms.

2.7. Exercise Care in the Classification of Hypothesis Tests.

More generally, the author's presentation of "Categorical tests" in a flowchart (p. 103) as each involving a null hypothesis of the type "Equal proportions" is potentially limiting. It gives the impression that such tests are limited to the 2 x 2 case, whereas this is not the case. Indeed, as we have seen in the case of Fisher's Exact test, statistical software, such as SPSS, commonly used by higher education students across the disciplines, can accommodate more generalised basic-level statistical hypothesis tests. Of those tests which the author lists, this applies to the chi-square test of association, the McNemar test (which, in its generalised form, is referred to as the McNemar-Bowker test) and Fisher's Exact test.

The author's choice to use the classifications 'Categorical, Parametric and Nonparametric' to distinguish between hypothesis tests according to the type of data is consistent with their misclassification of chi-square tests, Fisher's Exact test and the McNemar test through excluding them from the class of non-parametric tests. My concern here is not merely a matter of semantics. It is a reasonable expectation that students should appreciate what is meant by the term 'non-parametric data', not least where, as in this case, this concept is assumed in a book on data-analysis. By way of further motivation for correct classification of hypothesis tests, following menu pathways within statistical packages can include opting for the choice *non-parametric* before choosing any one of a range of non-parametric tests. This is the case, for example, with SPSS when opting for the chi-square test of association or Fisher's Exact test, although an alternative menu pathway also exists. In simple terms, non-parametric data are data that do not follow, or approximate to following, a known parameterised distribution. Correspondingly, within the context of hypothesis testing, non-parametric tests are hypothesis tests which are suited to such data. In keeping with Chin and Lee's interpretation of non-parametric data (Chin, 2008), it is advisable that learners should be pointed to the following conditions:

- the data are not Normally distributed;
- the data are not continuous (e.g., nominal or ordinal data);
- Normality testing proves inconclusive, but it is known from other studies that the data are likely to originate from a population of non-parametric data.

At an introductory level, it is usually appropriate to assume that the only parametric distribution which the learners are likely to encounter is the Normal distribution and that therefore any one of the above conditions is representative of parametric data.

The above recommendations support sticking to standard nomenclature in statistics that has stood the test of time, so as to avoid confusing the learner, including, as explained, above in the use of statistical packages.

With reference to the McNemar test in particular, the availability of the McNemar-Bowker test as an extension to this test beyond the 2 x 2 case would have been a welcome addition to the aforementioned flowchart for categorical tests, not least, since, as with the Fisher-Freeman-Halton, test. this test has been accommodated by SPSS for over a decade.

On applying the chi-square test of association later within the context of a case study, it would have been appropriate for the author to have at least pointed learners to recent literature raising concerns on the conservative nature of Yates's correction rather than applying it as normal practice for the two-by-two case (pp. 119 - 121). In balance, the author points the reader to Field's book, *An Adventure in Statistics: The Reality Enigma* (Field, 2016), which covers the issue of conservativeness in a clear manner. However, a pointer in *How to Analyze Data* regarding this specific issue would have been helpful so as to encourage learners to stay connected with current practice in use of corrections to the chi-square test of association rather than adopting Yates's correction without questioning it, as does the author.

2.8. Exercise Caution in the Interpretation of Non-Parametric Tests.

Under the previous tip, I highlighted the importance of including tests of association between two categorical variables among the class of non-parametric tests and recognising their generalisability beyond the two-by-two case. The interpretation of non-parametric tests lends itself to a wide range of misconceptions of a more profound nature. A common one is to assume that generally speaking, the Mann-Whitney U-test represents a comparison of medians. The author demonstrates this misunderstanding in a broader sense by unifying the above test, the Wilcoxon signed rank test, the Kruskal-Wallis test and the Friedman test under a single flowchart header involving the common null hypothesis type '$\text{median}_a=\text{median}_b$' (p. 106). It is inappropriate to use this type of paper to fully explore the problems with this over-simplification or to address them. However, by way of illustration, it should serve as food for thought that it is possible to generate instances where the Mann-Whitney U-test yields a statistically significant result across two groups, where the medians for the corresponding numerical data across these groups are identical. Such examples serve as counterexamples to the above statement of the null hypothesis as presented by the author. More generally, oversimplifications of the above sort contribute to the prevailing poor usage in non-statistical journals of effect size estimates.

Rightly so, the author states, "If you don't ensure that your dataset meets the assumptions of the statistical test, you may end up with meaningless output." (p. 107) With reference to their suggested type of null hypothesis, they should also note, however, that *if you misinterpret the purpose of the statistical hypothesis test, you are unlikely to arrive at a meaningful interpretation of your findings*. This oversight may in turn explain the neglect of a table of effect sizes as a complement to the material on non-parametric tests in this Pocket Study Skills book.

2.9. Ensure That Study Hypotheses are Well-Aligned with the Original Research Questions and that the Research Questions are Well-Posed.

Within Chapter 14 (entitled 'Case studies'), which the author describes as illustrating "the *statistical process* in action" (p. 108), they present a research question in the following manner:

> *The students think that more staff bring their own cups because they have office space to keep them. Is this the case?* (p. 114)

Before progressing to any related content on hypothesis testing, I am immediately struck by the need to seek greater clarity. In particular, what is the intended reference for "this" in "Is this the case?" – what the students are alleged to think, the alleged reason for staff bringing more cups, or something else? Aside from this, what is the comparator for "more staff". Turning to the null hypothesis which the author presents, in contrast to the research question, there is no mention of a reason for staff bringing more cups. Instead, they state the null hypothesis as follows:

> *There is no association between university role (staff, student) and BYO (yes, no).*

Unfortunately, on account of a lack of precision and clarity, it is difficult to see why this hypothesis should have been a natural derivative of the research question provided. Also, those unfamiliar with colloquial English may find the undefined term 'BYO' rather mystifying. Where hypotheses are meaningful in research, the framing of research questions and the transition to framing hypotheses for statistical testing has in my experience frequently proven to be challenging for students embarking on research projects. Sadly, the author's approach to defining a research question and corresponding null hypothesis reinforces this problem.

2.10. Explain the Role of Exploratory Tests for Choice of Statistical Procedures.
I am curious about the author's intention behind their insertion of blank flowcharts (pp. 115, 123 and 128) for choice of hypothesis tests. In fact, the lack of prompting begs the questions as to whether they are the product of a typesetting error and whether they represent best use of space for what is intended to serve as a compact study guide. In my view, it would have been a much more valuable use of space to expand the content on parametric testing by introducing readers to the Central Limit theorem and its compensatory role in relation to testing for Normality, together with the importance of Levene's test as a test of equality of variances, in deciding whether there is a need to resort to Welch's version of the independent samples t-test. The Central Limit Theorem tells us that for sufficiently large samples (of size at least 30, say) (Weiss, 1994), the distribution of the sample mean approximates to Normality regardless of whether or not the parent population is Normally distributed. This is a handy finding, as for sufficiently large groups, a variety of statistical procedures can be used while waiving the requirement for the sample data to be Normally distributed and thus the need to test for Normality.

For sufficiently large samples, having permission to perform techniques for Normally distributed data where the sample data is not itself Normally distributed can make reporting simpler, including in terms of effect sizes. For example, a difference in means arising from a t-test is simpler to compute and to explain in report writing than some of the alternative effect size estimates available for use under the assumption that the data are not Normally distributed.

Application of the Central Limit should, of course, follow assumptions checking in relation to conditions other than Normality for application of individual statistical procedures.

With respect to assumptions testing, the author appears to contradict themselves in relation to the requirements for use of the Pearson Correlation Coefficient (PCC). They specify these assumptions with reference to the sample data being: "independent random observations from normally distributed data" and "linearly related". (p. 130) They then proceed to provide an example of testing for a correlation between "ice-cream sales" and "shark attacks". As is indicated in a table (p. 131) these variables refer to monetary values and frequencies, respectively. The idea that number of shark attacks is Normally distributed, or even on a continuous scale, cannot be justified. The data pertain to a particular day and the only available counts are 1 to 4, with a sample size of nine. Moreover, the fact that the author fits a regression equation to nine data points, conveniently supported by a good straight-line fit, (p. 127) begs the question as to whether this is the best choice of example for introducing the reader to simple linear regression. In particular, the question of Normality of residuals (differences between observed and predicted values) which is fundamental to data exploration prior to embarking on a linear regression analysis receives no mention, although this is unsurprising, given the invalid choice of data type and the fact that there are only nine points. Admittedly, such procedures can be used informally with non-continuous data to give a crude sense of patterns in the data, but why publish such a case as exemplary? Such issues could perhaps have been avoided through greater involvement of disciplinary experts capable of providing a successful

union between statistical rigour and well-researched real-world learning contexts of value to students within different higher education disciplines.

For completeness, I note here that there is some confusion in academic circles concerning the need for the data for both variables to be Normally distributed in order to simply calculate the PCC. It may indeed only be necessary for the data to be Normally distributed for at least one of the two variables (both of the variables) when testing for statistical significance (calculating a CI for the PCC, respectively). (Freeman, 2009) It would be inappropriate to be overly decisive here regarding the correctness of the assumption for simply calculating the PCC that the data are Normally distributed for both variables. The important point is that the author has stipulated this assumption but omitted to provide a suitable exemplar for its use.

3. Final Recommendations.

Of the Pocket Study Skills books published by the series editor to date, *How to Analyze Data* is the only one which is marketed with a focus on meeting the needs of a specific academic discipline (even at a basic level). This, together, with the hiatus more generally speaking in pocket skills books in elementary statistics could serve as a driver to meet the basic statistical learning needs of higher education students across a range of disciplines in an accessible manner. However, it is critical to bear in mind that such learning should involve professional accountability in terms of the ideas that students are likely to carry forward into the workplace and utilise in research publication or the review thereof on graduation. This includes proactively safeguarding the discipline in relation to neglect of assumptions testing and underestimating the conditions for use of existing basic statistical procedures.

While it is clear that a study skills book cannot serve the purpose of a traditional recommended textbook for students, readers ought to be at liberty to evaluate a book based on its own claim, including, as in this case, that of equipping readers "with the skills and confidence to succeed in statistics." (reverse cover) While this is a student learning outcome which all statistical educators should be aspiring to, it cannot be achieved with integrity in the absence of statistical rigour. As is evident from the initial review and ten remedial tips above, there is a considerable amount of key learning content which has been omitted from this book at the expense of accuracy and appropriate engagement with, or referencing to, core statistics in the literature.

Hopefully, the above tips provide useful pointers, suggestions and reminders to teachers of statistics to non-specialists for use in their own teaching with a view to encouraging students to think soundly in the application and reporting of statistics. While such readers may find several of the positive features of *How to Analyze Data* which I have highlighted in section 1.2 of interest, I have yet to be convinced that use of a pocket-sized book, either as a standalone or a complementary resource, is a safe approach to learning introductory statistics. Correspondingly, I would advise against relegating the learning of statistics to the domain of study skills development, as this explicitly dismisses the place of statistics among the sciences, at the potential expense of expert input to teaching in the discipline. The uninitiated learner in statistics may enthusiastically embrace the teaching in this book on grounds of its accessibility, conciseness and attractive presentation, not to mention the convenience of popping the book in their pocket! However, there is a danger that this can lead to a false confidence in the non-expert educator and non-specialist learner in statistics alike, thus perpetuating student misunderstandings, including within the context of statistical consultancy, while stunting the educator's own growth in statistical understanding. While the author lists some references and other resources at the back of the book (pp. 151 - 154), there is no evidence that they have taken up the challenge of remedying the above issues through cross-referencing to these or

other types of published work, even in the form of additional reading (as might be expected in the case of emphasizing the importance of CIs).

While the author does not recommend or engage with any statistics packages, as noted earlier they do encourage readers to engage with Excel for the practice of a range of data skills. I have commended the author on some of this content. However, statistical educators who are interested in taking their learners further with Excel may appreciate some exposure to the literature on the limitations of Excel for use in data analysis. It is beyond the scope of this review to discuss this topic. However, I provide the following references here by way of example as complementary guides to use with books designed to support statistical analysis with MS Excel: *Using Excel for Statistical Data Analysis – Caveats* (Goldwater, 2007) and *The Pros and Cons of Using Excel for Statistical Calculations*. ("The Pros and Cons of Using Excel for Statistical Calculations," 2020) For accuracy, the interested reader may wish to cross-check the most recent features and relevant add-ons within MS Excel with the advice provided in these references.

References

Barsalou, M. A. (2015). *The ASQ guide to statistics for Six Sigma Black Belts*. ASQ Quality Press.

Boslaugh, S. (2012). *Statistics in a nutshell: A desktop quick reference guide* (2nd ed.). O'Reilly.

Campbell, M., & Swinscow, T. (2009). *Statistics at square one* (11 ed.). Wiley-Blackwell.

Chin, R. L., Bruce Y. (2008). *Principles and practice of clinical trial medicine*. Academic Press.

Clapham, C. N., James. (2014). *Concise Oxford dictionary of mathematics* (5 ed.). Oxford University Press.

Clark-Carter, D. (1997). *Doing quantitative psychological research: From design to report*. Psychology Press Ltd.

Cochran, W. G. (1952). The c2 goodness-of-fit test. *The Annals of Mathematical Statistics*, 23(3), 315 - 345.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.

Dienes, Z. (2011). Bayesian versus orthdox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3), 274–290. https://doi.org/10.1177/1745691611406920

Field, A. (2016). *An adventure in statistics: The reality enigma*. SAGE Publications Ltd.

Freeman, J., & Young, T. (2009). Correlation coefficient: Association between two continuous variables. *Scope*, 33, 31-33. https://www.sheffield.ac.uk/polopoly_fs/1.43991!/file/Tutorial-14-correlation.pdf

Gardner, M. J., & Altman, D. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *Statistics in Medicine*, 292, 746-750.

Goldwater, E. (2007). *Using Excel for statistical data analysis - Caveats*. https://people.umass.edu/evagold/excel.html

Goodman, S. N. (1999). The p-value fallacy. *Annals of Internal Medicine*, 130(12), 995-1021.

Hanna, D., & Dempster, M. (2012). *Pyschology statistics for dummies*. John Wiley & Sons Ltd.

Harrison, V. K., Richard, & Brace, N. S., Rosemary. (2020). *SPSS for Psychologists* (7 ed., Vol. Hampshire). MacMillan International Higher Education.

Hexter, A. T. (2013). *How to conduct a clinical audit: A guide for medical students*. Retrieved 28 January 2021, from http://sites.cardiff.ac.uk/curesmed/files/2014/10/NSAMR-Audit.pdf

Kerr, N. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kyriazos, T., A, & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology*, 9, 2531 - 2560. https://doi.org/10.4236/psych.2018.911145

Lin, C. Y. Y., Ming-Chung. (2008). Improved p-value tests for comparing two independent binomial proporitons. *Communications in Statistics - Simulation and Computation*, 38(1), 78-91. https://doi.org/10.1080/03610910802417812

MacDougall, M. (2009). Dyscalculia, dyslexia, and medical students' needs for learning and using statistics. *Medical Education Online*, 14(1). https://doi.org/10.3402/meo.v14i.4512

Moilanen, T., Pietilä, A.-M., & Coffey, M. K., M. (2019). Developing a scale: Adolescents' health choices related rights, duties amd responsibilities. *Nursing Ethics*, 26(7-8), 2511-2522. https://doi.org/10.1177/0969733019832952

Mulder, R., Singh, A. B., Hamilton, A., Das, P., Outhred, T., Morris, G., Bassett, D., Baune, B. T., Berk, M., Boyce, P. L., Bill, & Parker, G. M., Gin S. (2017). The limitations of using randomised controlled trials as a basis for developing treatment guidelines. *Evidence-Based Mental Health*, 21(1), 4-6. https://doi.org/10.1136/eb-2017-102701

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7 ed.). Open University Press.

Petrie, A. S., Caroline. (2013). *Medical statistics at a glance workbook paperback*. Wiley-Blackwell.

Petrie, A. S., Caroline. (2020). *Medical statistics at a glance* (4 ed.). Wiley Blackwell.

Pollet, T. V., & van der Meij, L. (2016). To remove or not to remove: The impact of outlier handling on significance testing in testosterone data. *Adaptive human behaviour and physiology*, 3, 43-60. https://doi.org/10.1007/s40750-016-0050-z

Price, R. (2017). *Psychometric methods: Theory into practice*. The Guildford Press.

The Pros and Cons of Using Excel for Statistical Calculations. (2020). *GraphPad Knowledgebase, 1406*. https://www.graphpad.com/support/faq/the-pros-and-cons-of-using-excel-for-statistical-calculations/

Ranstam, J. (2012). Why the p-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and Cartilage*, 20, 805 - 808. https://doi.org/10.1016/j.joca.2012.04.001

Rowntree, D. (2018). *Statistics without tears*. Penquin.

Schmuller, J. (2016). *Statistical analysis with Excel for dummies* (4 ed.). John Wiley & Sons.

Sim, J. R., N. (1999). Statistical inference by confidence intervals: issues of interpretation and utilization. *Physical Therapy and Rehabilitation Journal*, 79(2), 186-195. https://doi.org/10.1093/ptj/79.2.186

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143, 534-547. https://doi.org/10.1037/a0033242

Slootmaeckers, K., & Kerremans, B. A., Johan. (2013). Too afraid to learn: Attitudes towards statistics as a barrier to learning statistics and to acquiring quantitative skills. *Politics*, 34(2), 191-200. https://doi.org/10.1111/1467-9256.12042

Stinerock, R. (2018). *Statistics with R: A beginner's guide*. London.

Upton, G., & Cook, I. (2014). *Oxford dictionary of statistics*.

Wasserstein, R. L. L., Nicole A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129 - 133. https://doi.org/10.1080/00031305.2016.1154108

Weiss, N. A. H., Matthew J. (1994). *Introductory statistics* (4th ed.). Addison-Wesley Publishing Company.

Weissgerber, T. L. M., Natasa M, & Winham, S. J. G., Vesna D. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLOS Biology*, 13(4). https://doi.org/10.1371/journal.pbio.1002128