Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data

One: Why Data Cleaning is Important: Debunking the Myth of Robustness

Contributors: Jason W. Osborne
Book Title: Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data
Chapter Title: "One: Why Data Cleaning is Important: Debunking the Myth of Robustness"
Pub. Date: 2013

Access Date: November 11, 2015
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412988018
Online ISBN: 9781452269948
DOI: http://dx.doi.org/10.4135/9781452269948.n1
Print pages: 1-17

http://dx.doi.org/10.4135/9781452269948.n1*Old Dominion University*
[p. 1 ↓]

One: Why Data Cleaning is Important: Debunking the Myth of Robustness

> *You must understand fully what your assumptions say and what they imply. You must not claim that the "usual assumptions" are acceptable due to the robustness of your technique unless you really understand the implications and limits of this assertion in the context of your application. And you must absolutely never use any statistical method without realizing that you are implicitly making assumptions, and that the validity of your results can never be greater than that of the most questionable of these*.
>
> (Vardeman & Morris, 2003, p. 26)
>
> *The applied researcher who routinely adopts a traditional procedure without giving thought to its associated assumptions may unwittingly be filling the literature with nonreplicable results*.
>
> (Keselman et al., 1998, p. 351)
>
> *Scientifically unsound studies are unethical*.
>
> (Rutstein, 1969, p. 524)

Many modern scientific studies use sophisticated statistical analyses that rely upon numerous important assumptions to ensure the validity of the results and protection from undesirable outcomes (such as Type I or **[p. 2 ↓]** Type II errors or substantial misestimation of effects). Yet casual inspection of respected journals in various fields shows a marked absence of discussion of the mundane, basic staples of quantitative methodology such as data cleaning or testing of assumptions. As the

quotes above state, this may leave us in a troubling position: not knowing the validity of the quantitative results presented in a large portion of the knowledge base of our field.

My goal in writing this book is to collect, in one place, a systematic overview of what I consider to be *best practices* in data cleaning—things I can demonstrate as making a difference in your data analyses. I seek to change the status quo, the current state of affairs in quantitative research in the social sciences (and beyond).

I think one reason why researchers might not use best practices is a lack of clarity in exactly *how* to implement them. Textbooks seem to skim over important details, leaving many of us either to avoid doing those things or having to spend substantial time figuring out how to implement them effectively. Through clear guidance and real-world examples, I hope to provide researchers with the technical information necessary to successfully and easily perform these tasks.

I think another reason why researchers might not use best practices is the difficulty of changing ingrained habits. It is not easy for us to change the way we do things, especially when we feel we might already be doing a pretty good job. I hope to motivate practice change through demonstrating the benefits of particular practices (or the potential risks of failing to do so) in an accessible, practitioner-oriented format, I hope to reengage students and researchers in the importance of becoming familiar with data *prior* to performing the important analyses that serve to test our most cherished ideas and theories. Attending to these issues will help ensure the validity, generalizability, and replicability of published results, as well as ensure that researchers get the power and effect sizes that are appropriate and reflective of the population they seek to study. In short, I hope to help make our science more valid and useful.

Origins of Data Cleaning

Researchers have discussed the importance of assumptions from the introduction of our early modern statistical tests (e.g., Pearson, 1931; Pearson, 1901; Student, 1908). Even the most recently developed statistical tests are developed in a context of certain important assumptions about the data.

[p. 3 ↓]

Mathematicians and statisticians developing the tests we take for granted today had to make certain explicit assumptions about the data in order to formulate the operations that occur “under the hood” when we perform statistical analyses. A common example is that the data are normally distributed, or that all groups have roughly equal variance. Without these assumptions the formulae and conclusions are not valid.

Early in the 20th century, these assumptions were the focus of much debate and discussion; for example, since data rarely are perfectly normally distributed, how much of a deviation from normality is acceptable? Similarly, it is rare that two groups would have exactly identical variances, so how close to equal is good enough to maintain the goodness of the results?

By the middle of the 20th century, researchers had assembled some evidence that some minimal violations of some assumptions had minimal effects on error rates under certain circumstances—in other words, if your variances are not identical across all groups, but are relatively close, it is probably acceptable to interpret the results of that test despite this technical violation of assumptions. Box (1953) is credited with coining the term *robust* (Boneau, 1960), which usually indicates that violation of an assumption does not substantially influence the Type I error rate of the test. Thus, many authors published studies showing that analyses such as simple one-factor analysis of variance (ANOVA) analyses are “robust” to nonnormality of the populations (Pearson, 1931) and to variance inequality (Box, 1953) when group sizes are equal. This means that they concluded that modest (practical) violations of these assumptions would not increase the probability of Type I errors (although even Pearson, 1931, notes that strong nonnormality can bias results toward increased Type II errors).

Remember, much of this research arose from a debate as to whether even minor (but practically insignificant) deviations from absolute normality or exactly equal variance would bias the results. Today, it seems almost silly to think of researchers worrying if a skew of 0.01 or 0.05 would make results unreliable, but our field, as a science, needed to explore these basic, important questions to understand how our new tools, these analyses, worked.

Despite being relatively narrow in scope (e.g., primarily concerned with Type I error rates) and focused on what then was then the norm (equal sample sizes and relatively simple one-factor ANOVA analyses), these early studies appear to have given social scientists the impression that these basic assumptions are unimportant. Remember, these early studies were exploring, and they were concluding that under certain circumstances minor (again, practically insignificant) deviations from meeting the exact letter of the assumption **[p. 4 ↓]** (such as exact equality of variances) did not appreciably increase Type I error rates. These early studies do not mean, however, that all analyses are robust to dramatic violations of these assumptions, or to violations of these assumptions without meeting the other conditions (e.g., exactly equal cell sizes).

Despite all our progress, almost all our analyses are founded on important, basic assumptions. Without attending to these foundations, researchers may be unwittingly reporting erroneous or inaccurate results.

Note also that the original conclusion (that Type I error rates were probably not increased dramatically through modest violation of these assumptions under certain specific conditions) is a very specific finding and does not necessarily generalize to broad violations of any assumption under any condition. It is only focused on Type I error rates and does not deal with Type II error rates, as well as misestimation of effect sizes and confidence intervals.

Unfortunately, the latter points seem to have been lost on many modern researchers. Recall that these early researchers on “robustness” were often applied statisticians working in places such as chemical and agricultural companies as well as research labs such as Bell Telephone Labs, not in the social sciences where data may be more likely to be messy. Thus, these authors are viewing “modest deviations” as exactly that—minor deviations from mathematical models of perfect normality and perfect equality of variance that are practically unimportant. It is likely that social scientists rarely see data that are as clean as those produced in those environments.

Further, important caveats came with conclusions around robustness, such as adequate sample sizes, equal group sizes, and relatively simple analyses such as one-factor ANOVA.

Some Relevant Vocabulary

Type I Error Rate: the probability of rejecting the null hypothesis when in fact the null hypothesis is true in the population.

Type II Error Rate: the probability of failing to reject the null hypothesis when in fact the null hypothesis is false in the population.

Misestimation of Effect Size: failure to accurately estimate the true population parameters and effects.

Robust: generally refers to a test that maintains the correct Type I error rate when one or more assumptions is violated. In this chapter, I argue that robustness is largely a myth in modern statistical analysis.

This mythology of robustness, however, appears to have taken root in the social sciences and may have been accepted as broad fact rather than narrowly, as intended. Through the latter half of the 20th century, the term came to be used more often as researchers **[p. 5 ↓]** published narrowly focused studies that appeared to reinforce the mythology of robustness, perhaps inadvertently indicating that robustness was the rule rather than the exception.

In one example of this type of research, studies reported that simple statistical procedures such as the Pearson product-moment correlation and the one-way ANOVA (e.g., Feir-Walsh & Toothaker, 1974; Havlicek & Peterson, 1977) were robust to even "substantial violations" of assumptions.[1] It is perhaps not surprising that robustness appears to have become unquestioned canon among quantitative social scientists, despite the caveats to these latter assertions, and the important point that these assertions of robustness usually relate only to Type I error rates, yet other aspects of analyses (such as Type II error rates or the accuracy of the estimates of effects) might still be strongly influenced by violation of assumptions.

However, the finding that simple correlations might be robust to certain violations is not to say that similar but more complex procedures (e.g., multiple regression) are

equally robust to these same violations. Similarly, should one-way ANOVA be robust to violations of assumptions,[2] it is not clear that similar but more complex procedures (e.g., factorial ANOVA or analysis of covariance—ANCOVA) would be equally robust to these violations. Yet as social scientists adopted increasingly complex procedures, there is no indication that the issue of data cleaning and testing of assumptions was revisited by the broad scientific community. Recent surveys of quantitative research in the social sciences affirms that a relatively low percentage of authors in recent years report basic information such as having checked for extreme scores or normality of the data, or having tested assumptions of the statistical procedures being used (Keselman, et al., 1998; Osborne, 2008b; Osborne, Kocher, & Tillman, 2011). It seems, then, that this mythology of robustness has led a substantial percentage of social science researchers to believe it unnecessary to check the goodness of their data and the assumptions that their tests are based on (or to report having done so).

With this book, I aim to change that. I will show how to perform these basic procedures effectively, and perhaps more importantly, show you why it is important to engage in these mundane activities.

Are Things Really that Bad?

Recent surveys of top research journals in the social sciences confirm that authors (as well as reviewers and editors) are disconcertingly casual about data **[p. 6 ↓]** cleaning and reporting of tests of assumptions. One prominent review of education and psychology research by Keselman and colleagues (1998) provided a thorough review of empirical social science during the 1990s. The authors reviewed studies from 17 prominent journals spanning different areas of education and psychology, focusing on empirical articles with ANOVA-type designs.

In looking at 61 studies utilizing univariate ANOVA between-subjects designs, the authors found that only 11.48% of authors reported anything related to assessing normality, almost uniformly assessing normality through descriptive rather than inferential methods.[3] Further, only 8.20% reported assessing homogeneity of variance, and only 4.92% assessed both distributional assumptions and homogeneity of

variance. While some earlier studies asserted ANOVA to be robust to violations of these assumptions (Feir-Walsh & Toothaker, 1974), more recent work contradicts this long-held belief, particularly where designs extend beyond simple one-way ANOVA and where cell sizes are unbalanced, which seems fairly common in modern ANOVA analyses within the social sciences (Lix, Keselman, & Keselman, 1996; Wilcox, 1987).

In examining articles reporting multivariate analyses, Keselman and colleagues (1998) describe a more dire situation. None of the 79 studies utilizing multivariate ANOVA procedures reported examining relevant assumptions of variance homogeneity, and in only 6.33% of the articles was there any evidence of examining of distributional assumptions (such as normality).

Similarly, in their examination of 226 articles that used some type of repeated-measures analysis, only 15.50% made reference to some aspect of assumptions, but none appeared to report assessing sphericity, an important assumption in these designs that when violated can lead to substantial inflation of error rates and misestimation of effects (Maxwell & Delaney, 1990, p. 474).

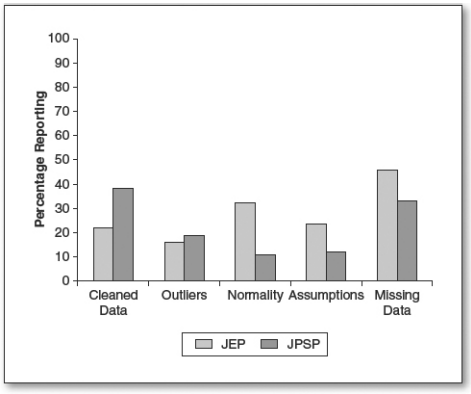
Finally, their assessment of articles utilizing covariance designs (N = 45) was equally disappointing—75.56% of the studies reviewed made no mention of any assumptions or sample distributions, and most (82.22%) failed to report any information about the assumption of homogeneity of regression slope, an assumption critical to the validity of ANCOVA designs.

Another survey of articles published in 1998 and 1999 volumes of well-respected educational psychology journals (Osborne, 2008b) showed that indicators of high-quality data cleaning in those articles were sorely lacking. Specifically, authors in these top educational psychology journals almost never reported testing any assumptions of the analyses used (only 8.30% **[p. 7 ↓]** reported having tested any assumption). Only 26.0% reported reliability of data being analyzed, and none reported any significant data cleaning (e.g., examination of data for outliers, normality, analysis of missing data, random responding).

Finally, a recent survey of recent articles published in prominent American Psychological Association (APA) journals' 2009 volumes (Osborne, et al., 2011)

found improved, but uninspiring results (see Figure 1.1). For example, the percentage of authors reporting data cleaning ranged from 22% to 38% across journals. This represents a marked improvement from previous surveys, but still leaves a majority of authors failing to report any type of data cleaning or testing of assumptions, a troubling state of affairs.

Figure 1.1 Percentage of Papers Reporting Having Checked for Each Data Cleaning Aspect



Similarly, between 16% and 18% reported examining data for extreme scores (outliers), 10% and 32% reported checking for distributional assumptions (i.e., normality), and 32% and 45% reported dealing with missing data in **[p. 8 ↓]** some way. Clearly, even in the 21st century, the majority of authors in highly respected scholarly journals fail to report information about these basic issues of quantitative methods.

Why Care about Testing Assumptions and Cleaning Data?

Contrary to earlier studies, it is not clear that most statistical tests are robust to most violations of assumptions, at least not in the way many researchers seem to think. For example, research such as that by Havlicek and Peterson (1977) shows one-factor ANOVA to be more robust to violations of distributional assumptions than violations of

the assumption of homogeneity of variance, but primarily when cell sizes are equal. One-way ANOVA appears to be less robust to violations of distributional assumptions when cell sizes are unequal, or to violations of variance homogeneity under equal or unequal cell sizes (e.g., Lix, et al., 1996; Wilcox, 1987). Yet this information about the robustness of simple one-way ANOVA, a relatively rare procedure in modern times, does little to inform us as to the relative robustness of more complex ANOVA-type analyses. In fact, recent arguments by research ethicists such as Vardeman and Morris (2003) state that statistical assumptions must be routinely assessed in order to ensure the validity of the results, and researchers such as Rand Wilcox (e.g., 2003, 2008) have made contributions by providing strong alternatives to traditional procedures for use when typical parametric assumptions fail the researcher.

One of the primary goals of this book is to convince researchers that, despite a seemingly ingrained mythology of robustness, it is in the best interests of everyone concerned to screen and clean data and test assumptions. While robustness research often focuses on Type I error rates (which are important), cleaning data and attending to assumptions also can have important beneficial effects on power, effect size, and accuracy of population estimates (and hence, replicability of results), as well as minimizing the probability of Type II error rates.

How can this State of Affairs be True?

So how is it that we have come to this place in the social sciences? In the beginning of the 20th century, researchers explicitly discussed the importance **[p. 9 ↓]** of testing assumptions. Yet contemporary researchers publishing in prominent empirical journals seem not to pay attention to these issues. Is it possible that authors, editors, reviewers, and readers are unaware of the importance of data screening and cleaning? Perhaps. It is true that most modern statistical textbooks seem to provide little concrete guidance in data cleaning and testing of assumptions, and it is also true that many modern statistical packages do not always provide these tests automatically (or provide guidance on how to interpret them). I have taught graduate statistics classes for many years now, and having surveyed many textbooks, I am troubled at how few seem to motivate students (and researchers) to focus on these issues. Even when texts do discuss these issues, it is often abstractly and briefly, giving the reader little concrete guidance on how to

perform these tests and how to think about the results of the tests of assumptions. It is possible that many students complete their doctoral training in the social sciences without focusing on these seemingly mundane issues.

It also is possible that some portion of researchers are faithfully testing assumptions and not reporting having done so. I would encourage all researchers to both *perform and report the results of* data cleaning and testing assumptions, even if no action is necessary. It gives the reader confidence in the results.

Data cleaning and testing of assumptions remain as relevant and important today as a century ago, and perhaps even more so. Data cleaning is critical to the validity of quantitative methods. Not only can problematic data points lead to violation of other assumptions (e.g., normality, variance homogeneity) but can lead to misestimation of parameters and effects without causing severe violation of assumptions. For example, in Chapter 7 I demonstrate that effectively dealing with extreme scores can improve the accuracy of population parameter estimates, decrease Type I and Type II errors, and enhance effect sizes and power.

There is good evidence that two of the most basic assumptions in many statistical procedures (that data come from populations that conform to the normal density function with homogenous variances) appear rarely met in practice (Micceri, 1989). This raises important concerns about the validity of conclusions based on these assumptions in the absence of overt information about whether they are met. Further, I will demonstrate how paying attention to basic issues such as distributional assumptions may protect researchers from errors of inference, as well as lead to strengthened effect sizes (and hence, power and significance levels). These are not only relevant to parametric statistical procedures, coincidentally. Meeting these distributional assumptions also can **[p. 10 ↓]** positively influence the results of nonparametric analyses (e.g., Zimmerman, 1994, 1995, 1998).

Additionally, I will review issues such as the importance of dealing with missing data effectively, response sets and how they can bias your results, the basic mechanics of identifying and dealing with extreme or influential scores, performing data transformations, issues around data cleaning when the data consist of repeated measures, and using data sets that involve complex sampling. In each chapter, my goal

is to use empirical evidence and theory to guide the quantitative researcher toward best practices in applied quantitative methods.

The Best Practices Orientation of this Book

It is my belief that quantitative researchers should be able to defend their practices as being the best available, much like medical doctors are encouraged to use the best practices available. In this spirit, I attempt to empirically demonstrate each major point in this book. For example, many authors have argued that removal of outliers (or influential scores) does harm to the data and the results, while others have argued that failure to do so damages the replicability of the results.[4] In my mind, it is less interesting to debate the philosophical aspects than to examine the evidence supporting each side. We, as quantitative researchers, should be able to definitively test which perspective is right and find evidence supporting a course of action. In the chapter on extreme scores (Chapter 7), I attempt to assemble a compelling empirical argument showing that it is a best practice to examine your data for influential data points, and to thoughtfully consider the benefits and costs of different courses of action. Similarly, there has been debate about whether it is appropriate to transform data to improve normality and homogeneity of variance. Again, I think that is something we can test empirically, and thus in Chapter 8 I attempt to persuade the reader through evidence that there are good reasons for considering data transformations. Further, in that chapter I present evidence that there are ways to perform transformations that will improve the outcomes.

Thus, the spirit of the book is evidence based. If I cannot demonstrate the benefit or importance of doing something a particular way, I will not recommend it as a best practice. Further, if I cannot clearly show you how to incorporate a practice into your statistical routine, I will not recommend it as a best **[p. 11 ↓]** practice. In other words, I propose that we as a field move toward a “survival of the fittest” mentality in our statistical practices. If we can show that, under certain circumstances, one practice is better than another, we should adopt it as a best practice, and shun others as less effective, at least in those situations where we have demonstrated a clear advantage of one technique over another.

As we move toward increasing specialization in the sciences, I believe it is unrealistic for scholars to remain current and expert in all areas. Thus, we need a cadre of statistical scholars who push the envelopes of innovation, who blaze the trail practitioners use, but we can no longer expect all researchers to be scholars of statistical methods. We must create clear, practitioner-oriented guidelines that help researchers get the best outcomes possible without assuming they are masters of matrix algebra and statistical theory. In this vein, my goal in each chapter is to make procedures explicit so that practitioners can successfully apply them. I encourage my colleagues to do the same. Just as practicing nurses and doctors need explicit, research-based guidelines on implementing best practices, practicing researchers need clear guidance in order to do the greatest good.

Data Cleaning is a Simple Process; However…

In conceptualizing this book, I intended to produce a simple series of procedures that researchers could follow. Yet the more deeply I delved into this world, the more I realized that this is often not a simple, linear process. There is an art to data cleaning and statistical analysis that involves application of years of wisdom and experience. Not all readers at this time have extensive wisdom and experience with quantitative data analysis. Thus, the best you can do is to use your best professional judgment at all times. Every data set presents unique opportunities and challenges, and statistical analysis cannot be reduced to a simple formulaic approach. To do so ignores the complexities of the processes we deal with in the research enterprise and opens the researcher to miscarriages of scientific justice. This book is a beginning, not an end, to your exploration of these concepts. I cannot anticipate every eventuality, so all researchers must take the advice contained within as a set of guidelines that (I hope) generally work in most cases, but may not be appropriate in your particular case. This is where the art of data analysis meets the science of statistics. Intimate familiarity with your own data, experience, and solid training in **[p. 12 ↓]** best practices will prepare you to be optimally successful in most cases, but only you can determine when it is appropriate to deviate from recommended best practices. The only thing I would

suggest is that whatever decisions you make in a particular analysis, you should be able to justify your course of action to a disinterested party (e.g., a qualified peer reviewer or dissertation committee member).

One Path to Solving the Problem

As my students (Brady Kocher and David Tillman) and I explored the mysteries surrounding statistical practice this past year, it has become increasingly clear that the peer review and publishing process itself can be part of the solution to the issue of data cleaning.

It may be the case that some portion of researchers publishing in the journals we examined did faithfully screen and clean their data and faithfully ensure that important assumptions were met prior to submitting the research for peer review. Perhaps these aspects of data analysis are viewed as too mundane or unimportant to report. Alternatively, some portion of researchers may be aware of the tradition of screening and cleaning data but for some reason may be under the impression that when using modern statistical methods and modern statistical software it is unnecessary to screen and clean data. In a perfect world, editors and peer reviewers would serve as a methodological safety net, ensuring that these important issues are paid attention to.[5]

Regrettably, the usual peer-review process implemented by most scholarly journals seems ill-prepared to remedy this situation. Elazar Pedhazur, in Chapter 1 of *Multiple Regression in Behavioral Research* (Pedhazur, 1997), is even stronger in indicting current research quality in the social sciences, and the failure of the peer review process:

> Many errors I draw attention to are so elementary as to require little or no expertise to detect…. Failure by editors and referees to detect such errors makes one wonder whether they even read the manuscripts. (p. 10).

Unfortunately, Pedhazur is not the only prominent scholar to question the quality of the traditional peer-review process (see also Kassirer & Campion, 1994; Mahoney, 1977;

Peters & Ceci, 1982; Weller, 2001). Reviews of the **[p. 13 ↓]** literature (e.g., Hall, Ward, & Comer, 1988) going back decades find that a disturbingly large portion of published educational research appears to contain serious methodological flaws. Many of these errors are unnecessary and largely the result of poor methodological training (e.g., Thompson, 1999).

Yet as problematic as peer review might be, in at least one specific instance it appears that the system may have worked as a powerful agent of positive change in statistical practice. In 1999 the APA released guidelines for statistical methods in psychology journals (Wilkinson & Task Force on Statistical Inference, 1999) that specified that effect sizes should be routinely reported. In response, many journals now include effect size reporting in their author guidelines and review criteria, and as a result, we have seen a substantial increase in the reporting of effect size, at least partly because journal gatekeepers were mandating it. In the same spirit, it would be simple for professional organizations such as the APA to mandate authors report on data screening, cleaning, and testing of assumptions.

Until that day, I hope this book encourages you, the reader, to change your practice to incorporate these easily-to-use techniques that can have unexpected payoffs. This book continues the spirit of best practices begun in my first edited volume (Osborne, 2008a) by presenting researchers with clear, easily implemented suggestions that are research based and will motivate change in practice by empirically demonstrating, for each topic, the benefits of following best practices and the potential consequences of *not* following these guidelines.

For Further Enrichment

- Review the author instructions for journals generally considered to be top tier or most respected in your field. See if any of them explicitly instruct authors to report testing assumptions, data cleaning, or any of the other issues we raise.
- On our book's website (http://best-practices-online.com/), I provide links to author instructions from journals in various fields. Which journals or fields have the most explicit author instructions? Which have the least explicit

instructions? Can you see any differences in the articles contained in journals that have more explicit directions for authors?

- Review a recent study of yours (or your advisor) where statistical assumptions were not tested and where the data are still available (we all have them, and I am as guilty as everyone else). As you work through this book, apply the various data cleaning techniques and test all assumptions for all statistical tests used in the study. Perhaps all the assumptions are met and your results now have even more validity than you imagined. Congratulations! Perhaps after cleaning the data and testing assumptions, your results are changed. Sometimes that can be a positive outcome, or sometimes that can be disappointing.
- If you have an interesting example of results and conclusions that changed after revisiting a data set and testing assumptions, I would love to hear from you at mailto:jasonwosborne@gmail.com. Send me a summary of what you found, and how things changed.

Notes

1. Yet again, it is important to point out that these studies are often focused narrowly on probability of Type I error rather than accuracy of parameter estimates or effect sizes. These latter aspects of analyses are often as important in modern research as the probability of making a Type I error.

2. To be clear, it is debatable as to whether these relatively simple procedures are as robust as previously asserted.

3. For more information on best practices in assessing normality, see Chapter 5.

4. These arguments are covered in greater depth in Chapter 7, and therefore are not reproduced here.

5. I must thank one of my doctoral committee members from years ago, Scott Meier, who gently reminded me to make sure I had done due diligence in cleaning my data and paying attention to extreme scores. Dr. Meier's gentle reminder salvaged what was turning out to be rather dismal results, allowing me to identify a very small number of

inappropriately influential scores that were substantially biasing my results. Removal of these few scores led to strong support for my original hypotheses, as well as a two-decade-long appreciation of the power of “sweating the small stuff.”

References

Boneau, C. A. The effects of violations of assumptions underlying the t test. Psychological Bulletin, vol. 57 (1960). (1), pp. 49–64. doi: 10.1037/h0041412

Box, G. Non-normality and tests on variances. Biometrika, vol. 40 (1953). (3–4), pp. 318–335.

Feir-Walsh, B. and Toothaker, L. An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. Educational and Psychological Measurement, vol. 34 (1974). (4), pp. 789–799.

Hall, B. W., Ward, A. W., and Comer, C. B. Published educational research: An empirical study of its quality. The Journal of Educational Research, vol. 81 (1988). (3), pp. 182–189.

Havlicek, L. L. and Peterson, N. L. Effect of the violation of assumptions upon significance levels of the Pearson r. Psychological Bulletin, vol. 84 (1977). (2), pp. 373–377. doi: 10.1037/0033-2909.84.2.373

Kassirer, J. and Campion, E. Peer review: Crude and understudied, but indispensable. JAMA, vol. 272 (1994). (2), pp. 96–97.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., and Levin, J. R. Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, vol. 68 (1998). (3), pp. 350–386.

Lix, L., Keselman, J., and Keselman, H. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Review of Educational Research, vol. 66 (1996). (4), pp. 579–619.

Mahoney, M. Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research, vol. 1 (1977). (2), pp. 161–175.

Maxwell, S. Delaney, H. (1990). Designing experiments and analyzing data: A model comparison perspective . Pacific Grove, CA: Brooks/Cole.

Micceri, T. The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, vol. 105 (1989). (1), pp. 156–166. doi: 10.1037/0033-2909.105.1.156

Osborne, J. W. (2008a). Best practices in quantitative methods . Thousand Oaks, CA: Sage.

Osborne, J. W. Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. Educational Psychology, vol. 28 (2008b). (2), pp. 1–10.

Osborne, J. W. Kocher, B. Tillman, D. (2011). Sweating the small stuff: Do authors in A PA journals clean data or test assumptions (and should anyone care if they do)? Unpublished Manuscript, North Carolina State University.

Pearson, E. The analysis of variance in cases of non-normal variation. Biometrika, vol. 23 (1931). (1–2), pp. 114–133.

Pearson, K. Mathematical contribution to the theory of evolution. VII: On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London, vol. A 195, (1901). pp. 1–47.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd. ed.). Fort Worth, TX: Harcourt Brace College.

Peters, D. and Ceci, S. Peer-review practices of psychological journals: The fate of published articles, submitted again. Behavioral and Brain Sciences, vol. 5 (1982). (2), pp. 187–195.

Rutstein, D. D. The ethical design of human experiments. Daedalus, vol. 98 (1969). (2), pp. 523–541.

Student. The probable error of a mean. Biometrika, vol. 6 (1908). (1), pp. 1–25.

Thompson, B. (1999). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas . In B. Thompson (Ed.), Advances in social science methodology (pp. 23–86). Stamford, CT: JAI Press.

Vardeman, S. and Morris, M. Statistics and ethics. The American Statistician, vol. 57 (2003). (1), pp. 21–26.

Weller, A. (2001). Editorial peer review: Its strengths and weaknesses . Medford, N.J.: Information Today.

Wilcox, R. New designs in analysis of variance. Annual Review of Psychology, vol. 38 (1987). (1), pp. 29–60.

Wilcox, R. (2003). Applying contemporary statistical techniques . San Diego: Academic Press.

Wilcox, R. (2008). Robust methods for detecting and describing associations . In J. W. Osborne (Ed.), Best practices in quantitative methods (pp. 266–280). Thousand Oaks, CA: Sage.

Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, vol. 54 (1999). (8), pp. 594–604.

Zimmerman, D. W. A note on the influence of outliers on parametric and non-parametric tests. Journal of General Psychology, vol. 121 (1994). (4), pp. 391–401.

Zimmerman, D. W. Increasing the power of nonparametric tests by detecting and downweighting outliers. Journal of Experimental Education, vol. 64 (1995). (1), pp. 71–78.

Zimmerman, D. W. Invalidation of parametric and nonparamteric statistical tests by concurrent violation of two assumptions. Journal of Experimental Education, vol. 67 (1998). (1), pp. 55–68.

http://dx.doi.org/10.4135/9781452269948.n1