Review

Let’s Talk about Stats: Revising Our Approach to Teaching Statistics in Psychology

Psychological Reports
0(0) 1–29
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00332941211043447
journals.sagepub.com/home/prx
SAGE

Joshua J. Reynolds
Psychology Department, University of Scranton, USA

Abstract

Undergraduate statistics in psychology is an important, often challenging, course for students. The focus in psychology tends to be on hypothesis tests, such as *t* tests and analysis of variance. While adequate for some questions, there are many other topics we might include that could improve that data analytic abilities of students and improve psychological science in the long run. Topics such as generalized linear modeling, multilevel modeling, Bayesian statistics, model building and comparison, and causal analysis, could be introduced in an undergraduate psychological statistics course. For each topic, I discuss their importance and provide sources for instructor’s continuing education. These topics would give students greater flexibility in analyzing data, allow them to conduct more meaningful analyses, allow them to understand more modern data analytic approaches, and potentially help the field of psychology in the long run, by being one part of the strategy to address the reproducibility problem.

Keywords

Statistics, modeling, teaching, Bayesian, undergraduate education

Corresponding Author:
Joshua J. Reynolds, University of Scranton, 800 Linden St, Scranton, PA 18510, USA.
Email: Joshua.Reynolds@scranton.edu

Introduction

Statistics is one of the most important courses in psychology. Coupled with research methods, it forms the backbone of understanding scientific research. Therefore, how we teach statistics and the content we teach matters greatly. Here, I discuss why we should emphasize the importance of statistics in our major, some common topics covered in basic psychology statistics courses, and how we might revise the content, such as teaching generalized linear modeling. The purpose is ultimately not to argue that one specific set of content should be taught, but that we should keep up to date with advancements in statistics and have clear justification for including any content. Further, we should take an approach that is efficient; we should focus on teaching the smallest number of tools that can answer the greatest number of questions in the smallest amount of time and create a foundation for future student success. This paper is intended as an introduction to certain statistics topics for psychology statistics instructors to consider (as opposed to students with limited statistical knowledge or analysts who already use the approaches discussed) and provides additional resources for more in-depth explanations of each technique (e.g., Finch et al., 2019; McElreath, 2020; Pearl, 2009). Although this paper is not intended to teach instructors each topic discussed below, it is a useful starting point to discuss the benefits of teaching particular topics in a psychology statistics course and provide useful tools and resources for each topic presented.

Why statistics?

Statistics are essential in psychological research. However, the importance of a statistics course goes far beyond psychological science. The content and skills developed in a psychological statistics course, or any statistics course, are pertinent to everyday life and are a core part of the knowledge every citizen needs to function in today's world (Garfield & Ben-Zvi, 2007; Moore, 1998). Indeed, predictions from statistical models are everywhere, including the news, sports, and the weather (see Silver, 2012 for a discussion of predictions for a general audience). When you read a weather report, you are trying to understand the predictions from a model. At time of writing, we are in the midst of the COVID-19 pandemic. How we respond to this event is partly dependent on the predictions of models. The ability to understand such models, apply their predictions, and even be critical of them, comes from a foundation in statistics. Beyond predictions, we encounter statistics in other daily contexts. For example, if you get a medical test, such as for strep throat, and test positive, should you believe that you have strep throat given that you tested positive? When a jury hears testimony about evidence such as DNA, knowledge of statistics is essential (Koehler & Macchi, 2004). Lastly, the skills developed in a statistics course translate to marketable abilities that employers may find useful. If a statistics course emphasizes data analysis using computer software such as R, students gain valuable experience with programs that are used in many professions (R Core Team, 2020). Arguably, these features confer a particular importance to a statistics course. As most psychology students (approximately 76%) will not matriculate to graduate school, and statistics is not a requirement in high school, the mandatory statistics course psychology students take may be their only formal opportunity to learn this material (National Science Foundation, National Center for Science and Engineering Statistics, 2017).

The traditional content

There is variability in the statistical topics covered at both the graduate and undergraduate levels (Alder & Vollick, 2000; Friedrich et al., 2000). Content that is most likely to be covered in a mandatory statistics course includes descriptive statistics, variations of *t* tests (e.g., one sample *t* test, independent samples *t* test, and dependent samples *t* test), correlation, simple linear regression, and analysis of variance (ANOVA). Content that is sometimes covered includes factorial ANOVA, repeated measures ANOVA, multiple regression, and non-parametric statistics. There has been major progress in statistics, particularly in the last several decades. However, in a recent survey of topics taught in psychological statistics courses, Friedrich et al. (2018) found that aside from a greater discussion of effect sizes, little had changed from two decades ago. Similarly, in graduate psychology programs, despite major advancements in statistics, statistical training at the PhD level has largely stagnated (Aiken et al., 1990, 2008). This evidence indicates that while students are competent at basic hypothesis tests such as ANOVA, they are deficient in newer, more useful techniques.

Many, if not most, statistics textbooks in psychology and the social sciences (e.g., Adams & Lawrence, 2019; Gravetter & Wallnau, 2013; Kranzler, 2018; Warne, 2018; Welkowitz et al., 2012) and course syllabi (Project Syllabus – Society for the Teaching of Psychology, 2021) are based around teaching a catalog of statistical tests. That is, the common approach is "in X situation, apply Y test." One major problem with commonly taught statistical tests, such as ANOVA, is that they are rigid (McElreath, 2020). ANOVA is applicable when you have a very specific type of data with a very specific type of question. In ANOVA, if some assumptions are violated, the test may no longer be valid. While ANOVA is robust against some violations, like homogeneity of variance to a degree, if multiple violations are present, a new variant of the test must be used. This is not to say ANOVA is useless. In fact, it works remarkably well in many cases. However, considering the varied research questions that we might ask, it can only be applied to a very narrow range. If the outcome is a Likert scale, or dichotomous, or many other exceptions, ANOVA cannot be applied (Liddell & Kruschke, 2018). By teaching tools that are less flexible and can only be applied to a narrow class of data (e.g., metric data), we give students less resources and may even be implicitly training students to only ask certain questions (e.g., the questions for which techniques like ANOVA can used).

Another issue in the traditional content is null hypothesis testing. Students are taught to put their questions in the form of null and alternative hypotheses and then apply the test. However, what is being tested is the null hypothesis and this is never, or rarely, the goal of science, nor does it lead to intuitive ways of thinking about data. Furthermore, *p* values are misinterpreted so frequently, even by researchers, it is unclear if this material is being communicated effectively (Goodman, 2008). For example, reviewing 30 introduction to psychology textbooks, Cassidy et al. (2019) found that 89% incorrectly defined statistical significance. Given that *p* values have been a fixture of psychological statistics for over a century, this figure is alarming. Additionally, the focus nearly exclusively on hypothesis testing obscures the fact that the majority of applied statistics is about prediction and modeling.

The field of statistics has grown by leaps and bounds since Karl Pearson and Sir Ronald Fisher and some change is evident. For example, newer textbooks dedicate more time to effect sizes and confidence intervals (e.g., Warne, 2018). However, the majority of statistics books in psychology have arguably lagged behind the progress that has been made in the field. For example, there is an increasing reliance on bootstrapping methods in frequentist statistics, yet this is rarely (if ever) discussed in texts (Cobb, 2015). This lag in textbooks mirrors the data from Friedrich et al. (2018) in that instructors have changed the content of these courses very little. While all undergraduate textbooks may not represent the state of the art in that field, the lag in social science statistics texts spans decades. For example, while many areas are moving to Bayesian statistics, psychology texts rarely even mention Bayes theorem (Cobb, 2015; Page & Satake, 2017).

The justifications for the current content are not self-evident; however, I offer three potential justifications. One justification is historical. Some subjects in all disciplines are taught not to reflect the current state of thinking, but as historical interest. For example, many introductory psychology books discuss, at least to some extent, Freudian dream analysis (e.g., Schacter et al., 2019), which does not represent modern psychological science. However, introductory psychology books also discuss cognitive neuroscience and evolution, which do represent modern psychological science. Thus, students can get a sense of how the discipline has changed. On the other hand, in psychology statistics courses, modern data analytic approaches (e.g., bootstrapping, maximum likelihood estimation, Markov Chain Monte Carlo [MCMC] techniques, Bayesian statistics, and causal analysis techniques) are rarely mentioned. A second justification is that teaching a catalog of tests is manageable for students who may be intimidated by mathematics. There is a lack of data in this area, however, in a survey by collegestats.org, statistics was one of the most hated college courses. Researchers have even developed attitudes towards statistics and statistics anxiety measures (Cruise et al., 1985; Roberts & Reese, 1987; Yong & Rosli, 2020). In teaching psychological statistics, it has been my experience that students often dread the course, and students have reported that part of the reason they majored in psychology (or another social science) was that they thought no math was required. Again, there is a lack of data in this area. It is possible, and understandable, that the reason these courses have not been updated accordingly is that there is fear that students will not be able to handle the more complex analyses. Lastly, some instructors may genuinely feel that *t* tests and linear regression are the best places to begin. These and other justifications may indicate that the typical course content is adequate, and therefore does not need to be substantially revised. However, while the typical course content may be reasonable, it does not mean it is optimal.

The failure to use better data analytic approaches may mean not extracting all the necessary information from the research design and even incorrect conclusions. Within both the statistics community as well as psychology specifically, there have been calls to revise the content in statistics courses in major ways (Breiman, 2001). For example, Cobb (2015) argued that the entire undergraduate statistics curriculum needs to be redesigned from the ground up. Here, I focus on the content of undergraduate psychological statistics specifically.

Alternative content

Alternative content might include topics such as generalized linear modeling, multilevel modeling, Bayesian statistics, model building and comparison, and causal analysis. For each of these topics, I introduce the concept, argue why we might include it in our undergraduate statistics courses, and provide resources for faculty's continuing education on these topics.

Generalized linear modeling

Generalized linear modeling (GZLM) is an exceptionally flexibly tool that allows one to model many different distributions of an outcome variable. GZLM can model continuous, discrete, categorical, and ordinal data. Indeed, the general linear model (GLM) is essentially a special case of the GZLM with a normal distribution and identity link function (Hoffmann, 2004). GZLMs can also be useful in situations where certain assumptions of GLM are violated (e.g., normal distribution of errors). Thus, a major advantage of GZLMs is the ability to model many different kinds of data with a single approach without as many assumptions (as compared to ANOVA, for example).

In legal/forensic psychology, an analyst might be interested in using factors like race to predict verdict (guilty/not guilty). The analyst could use a GZLM with a binomial distribution and logit link function (i.e., logistic regression). In developmental psychology an analyst might want to test if different personality traits predict how many children someone has. Here a GZLM with a Poisson distribution and log link function might be used (i.e., Poisson regression). A clinical psychologist might be interested in examining the influence of different mindfulness practices on anxiety, as measured by a 5-point Likert scale. Here a GZLM with a cumulative distribution and probit link function can be used (i.e., ordinal regression). These examples illustrate that very different types of research questions with completely different types of data can be answered using one general approach. None of the traditional techniques taught in undergraduate psychology statistics would be appropriate for any of these questions. Further, applying regression or ANOVA in analyzing a 5-point Likert scale would be inappropriate and potentially lead to the wrong conclusions (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018). Therefore, a major advantage of GZLM is the high degree of flexibility in modeling many different kinds of outcome data and being able to apply the correct model given the data.

While a full discussion and presentation of GZLM may be beyond the scope of a basic undergraduate statistics course, students could receive an introduction to GZLM as part of a required course(s). An introduction to GZLM can be achieved by including some of the basic concepts. A useful place to start is through an understanding of modeling language. In the case of a continuous normally distributed variable (y or more often $\mathrm{y_i}$), we could write $\mathrm{y_i} \sim N(\mu_\mathrm{i}, \sigma^2)$, which translates as follows: our outcome variable is distributed as a normal distribution with mean μ_i and constant variance σ^2. This would be the random component of the model. We then replace the parameter μ_i, with a linear model, our systematic component (e.g., $\mu_\mathrm{i} = \beta_0 + \beta_1 \mathrm{X}$). We then need a function to link the linear model to the distribution, known as a link function. A link function thus maps the linear space of a model onto the non-linear space of some parameter (e.g., λ in the case of a Poisson distribution). In the GLM, this link function would be the identity link function, as no transform is required. Using this standard modeling language, students can first be shown that the GLM is essentially a special case of GZLM. Then students can be shown other distributions and link functions to model many different types of data.

While GZLM can be complex, by anchoring it in the standard modeling language and using concepts that students should already by familiar with, they can begin to understand the flexible and powerful GZLM strategy. For an excellent introduction to the concepts in GZLM, see Hoffmann (2004). Dunn and Smyth(2018) use R to work through many examples and provide an excellent introduction to working with many different types of distributions.

Multilevel modeling

Multilevel modeling (MLM) is another flexible and powerful approach to data analysis. When the data are nested structures, such as students (level 1) in classrooms (level 2), MLM can be used to account for meaningful variation (Finch et al., 2019; Gelman & Hill, 2007; Robson & Pevalin, 2016). MLM can also model longitudinal data. For example, a researcher might have multiple measures through time (level 1) of the same individuals (level 2). Whereas a typical approach taught in psychology might be to use repeated measures ANOVA, a MLM can be used instead, and can be used in many other cases of nested data where repeated measures could not. For example, MLM can account for many complex data structures including those that are cross-classified and interactions at any level.

MLMs also have numerous other advantages that likely contribute to their increased use over time. For example, MLMs have the advantage of more appropriately modeling the uncertainty in the parameter estimates of the model. The typical GLM makes some often unrealistic assumptions, which results in an overestimate in the precision of the parameters (Gelman & Hill, 2007). Consider a case where there are multiple measures from the same individuals, such as reaction times across trials or days. If the reaction times for each participant were averaged and a GLM used, it will have artificially low uncertainty compared to a MLM that can account for the dependencies in the data. By allowing the intercepts and slopes to vary, the MLM produces more realistic estimates. The same consequence occurs in GLM when a scale is pre-averaged and entered into a model; the GLM removes variation whereas MLM retains the variation (McElreath, 2020). In addition to modeling the uncertainty at the population level, MLMs also have the advantage of modeling group level effects, as separate regression lines can be calculated for each participant. This is very intuitive, as individuals might start out with different reaction times (e.g., some start out faster) and the change over the trials or days is also different for each individual. This variability is not captured in a typical GLM. There are numerous other advantages of MLM, including that it can avoid errors such as the ecological fallacy, when applicable. Given the likelihood of students encountering nested data structures and the advantages of MLM, students would benefit greatly from at least an introduction.

An introduction might include a demonstration of how we can modify the linear model. A linear model with one predictor might be written as: $\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{X}_i + \mathrm{e}$. It is clear that there is one intercept and slope for all groups. To simplify, we can write the equation without any predictors, which is called the null or intercept only model: $\mathrm{y_i} = \beta_0 + \mathrm{e_i}$. Although simplified, there is still only one intercept. Again, this is an unrealistic assumption in many cases. The equation can be rewritten, first, to allow for the intercepts to vary: $y_{ij} = \beta_0 + \mu_j + e_{ij}$. The *i* and *j* subscripts refer to levels in the model: in this case a 2-level model, where *i* refers to the level 1 units (e.g., students, time, etc.) and *j* refers to the level 2 units (e.g., classroom, same individuals, etc.). The μ_j term allows a different intercept for different groups/clusters. This small yet profound change does not require any advanced mathematics. Thus, while

MLM as a whole is a complex topic, the fundamentals should be grasped by undergraduate psychology students. Accommodating varying slopes is also simple. First, the varying intercept equation can be rewritten to include a predictor: $y_{ij} = \beta_0 + \mathrm{\mu_j} + \beta_1 \mathrm{X}_{1ij} + e_{ij}$. The new part of the equation $\beta_1 \mathrm{X}_{1ij}$ is the predictor variable, but while the intercepts are free to vary, the slopes are still fixed. The intercepts and the slopes can be varied in the following equation: $y_{ij} = \beta_0 + \mu_j + \beta_1 \mathrm{X}_{1ij} + \mu_{1j}\ \mathrm{X}_{ij} + \mathrm{e}_{ij}$. The $\beta_1 \mathrm{X}_{1ij} + \mu_{1j}\ \mathrm{X}_{ij}$ refers to the coefficient/slope and allows them to vary. Again, no additional mathematical knowledge is required to understand the fundamental logic of MLM.

MLM can also be modified to include any type of outcome. That is, we can use generalized multilevel modeling to understand nested data structures with an outcome that might be continuous, categorical, discrete, or ordinal. In contrast, the GLM is limited to continuous fixed effects. It can therefore be shown that the majority of models applied in psychology are in fact special cases of generalized multilevel modeling. While students will not master such complex approaches in a single course, being introduced to the concept and engaging in some simple data analysis exercises is valuable. For example, students can see the continuity in the different types of data analysis. Students can understand that if they learn a generalized multilevel modeling approach, they can answer many more types of questions. The ease of fitting these models is aided by excellent R packages such as lme4 (Bates et al., 2015). In fact, all the hypothesis testing that is taught in an undergraduate psychology statistics course (*t* tests, correlation, regression, and ANOVA) can be accomplished with the generalized multilevel model framework of lme4. For a brief conceptual text on multilevel modeling see Robson and Pevalin (2016). Finch et al. (2019) present the basic issues and work through many examples of multilevel models using R. Included is a chapter on Bayesian multilevel modeling using the package MCMCglmm, which will fit Bayesian generalized multilevel models (Hadfield, 2010). Gelman and Hill (2007) is a more comprehensive source for MLM which uses R and BUGS and includes a discussion of Bayesian inference.

Bayesian statistics

The most common approach in psychology is frequentist statistics, which often focuses on generating a point estimate and comparing the estimate to a null hypothesis, by way of a *p* value, that is, null hypothesis significance testing (NHST). The majority of undergraduate psychology statistics texts (e.g., Adams & Lawrence, 2019; Gravetter & Wallnau, 2013; Kranzler, 2018; Warne, 2018; Welkowitz et al., 2012) focus exclusively on NHST. Nevertheless, calls for the abolishment or limited use of NHST have been echoed for decades (see Cohen, 1994) with some journals banning or limiting the use of *p* values (e.g., Basic and Applied Social Psychology). The reasons include that NHST leads to false dichotomous thinking, misrepresentations of data, and bias in publication (Ioannidis, 2019; Kruschke & Liddell, 2018b). Szucs and Ioannidis (2017) argued that NHST is a contributing factor to the replication crisis in psychology and that given the major issues with NHST, it should no longer be the default in psychological research. Further, they argue that if NHST is used in research, it should have to be justified and use preregistered hypotheses. Despite such strong arguments, students continue to learn and equate statistics with NHST. It is worth reiterating, Szucs and Ioannidis (2017) and many others are not arguing that we should discard NHST. Szucs and Ioannidis (2017) demonstrate the ways in which NHST can be misused, the limitations, and that NHST should not be the default. It can still be worthwhile to teach NHST at both the graduate and undergraduate level.

While psychology currently uses predominantly frequentist approaches like NHST, Bayesian approaches are used exclusively in some fields and are becoming more popular in psychology (Page & Satake, 2017; Etz & Vandekerckhove, 2018). Bayesian statistics allows for the exploration of an entire distribution (i.e., the posterior) and there is no reliance on sampling distributions or *p* values to interpret coefficients (e.g., a slope). Kruschke (2015) describes Bayesian analysis as the reallocation of credibility/probability toward parameter values most consistent with the data, and allocation of credibility/probability away from parameter values that are inconsistent with the data (see also Dienes, 2011; Kruschke & Liddell, 2018a; Lindley, 1993). Interestingly, the frequentist focus tends to be on a likelihood (*p* value), which is essentially a probability of the data given the parameter, while in Bayesian statistics, the focus is on the probability of the parameter given the data (the posterior distribution). It is worth noting that the questions that scientific research often take, "what is the value of X, given the data," or "what is the probability this hypothesis is true, given the data" is precisely what Bayesian inference provides (Kruschke & Liddell, 2018b).

A more fundamental difference between Bayesian and frequentist approaches is their views of probability. The frequentist view of probability is objective, fixed, and based on long run frequency. Definitions of probability in the Bayesian approach include subjective and objective. However, one common definition is that probability is uncertainty expressed from 0 to 1, where 0 indicates certainty that something will not happen and 1 indicates certainty that it will happen (Lambert, 2018). Historically these views clashed and powerful proponents of the frequentist approach (e.g., R.A. Fisher) succeeded, at least for a time, in marginalizing the Bayesian views (McElreath, 2020).

There are many correct ways to analyze data. Indeed, in any given instance, a Bayesian model may come to the same conclusion as a frequentist model. However, in some cases the two approaches differ, and Bayesian analyses tend to provide richer information. For example, in frequentist statistics, confidence intervals can be used to understand the uncertainty in a parameter. Bayesian statistics uses a similar concept, credible intervals, as well as high density credible intervals. Both types of intervals can be used to understand the uncertainty in a parameter estimate such as an intercept, slope, mean, or effect size. However, the frequentist confidence intervals contain no distributional information (Kruschke & Liddell, 2018b). So, values closer to the point estimate are just as likely as those near the end of the interval. On the other hand, Bayesian credible intervals do contain distributional information, which is more informative. Furthermore, even in cases where the results are the same, a Bayesian solution may be preferred, as the parameter estimates are more intuitive. *P* values on the other hand are notoriously misinterpreted (Goodman, 2008). There are many more advantages of Bayesian statistics. For example, when pooling estimates from multiple imputation, the pooling in frequentist models is not particularly straightforward, but is simple in Bayesian statistics. Bayesian models are also generative. Wagenmakers et al. (2018) discuss a variety of the advantages of Bayesian statistics and notes that there is little justification in continuing to teach frequentist statistics in psychology as the default. Wagenmakers et al. (2018) suggests that one reason why *p* values continue to be more popular than Bayesian methods is that individuals will use what was taught to them, which becomes a self-perpetuating cycle. Wagenmakers et al. (2018) and others have argued that Bayesian statistics should be the default in psychology, and I agree. However, it is unlikely at this time that psychology as a field will make that large shift. If an instructor continues to teach frequentist statistics, students might still be exposed to the major concepts in Bayesian analysis and how they differ from frequentist approaches. This could be accomplished in one to two lectures.

If a required undergraduate statistics course in psychology used the frequentist rather than Bayesian approach, it would be highly beneficial to, at minimum, discuss the logic of, differences, advantages, and disadvantages of these two approaches. To understand the basic concept in Bayesian analysis, one could start with a historical overview of Bayes theorem. Satake and Murray (2014) have developed interesting methods for teaching Bayes theorem using a legal scenario. Next, the different approaches to calculating/estimating posterior distributions could be presented (e.g., grid estimation, quadratic estimation, and MCMC). Lastly, Bayes Factors could also be introduced for hypothesis testing and model comparison.

Bayes Factors are more informative than *p* values, and they indicate the strength of the evidence, which when conducting a hypothesis test, is often what the analysts wants to know (Page & Satake, 2017). *P* values do not have such information and are often incorrectly interpreted in ways to suggest that they do (Goodman, 2008). In analyzing 287,424 findings of 35,515 articles, and calculating Bayes Factors, Aczel et al. (2017) find that a large portion of psychological findings do not pass a level of $\mathrm{BF} = 3$, typically indicating only anecdotal evidence. This is attributed to the use of *p* values and setting a weak threshold for acceptance. Page and Satake (2017) argue that Minimum Bayes Factors should be part of undergraduate introductory statistics courses.

Thus, we should seriously consider at least some discussion and data analysis examples with Bayes Factor.

By presenting at least these basic aspects of Bayesian analysis, students can develop a richer understanding about statistics and critically think with data. Anecdotally, in teaching *p* values, the concept only becomes intuitive after learning Bayesian statistics, as students learn the difference between a likelihood and a posterior, and what a *p* value is not. By the same token, if the instructor used a completely Bayesian approach, it could be argued that they should include a section on frequentist statistics. By learning at least some of the basics of these different approaches, we can get a richer understanding of how to get information from data and be critical of our analyses. Moreover, there are some cases where only a Bayesian or frequentist solution is even possible. Therefore, it is not a matter of one versus the other, but when to use what approach. Lastly, by at least introducing Bayesian statistics, there are additional opportunities to interleave concepts.

Interleaving involves switching between topics and has shown to be an effective learning strategy (Rohrer, 2012; Rohrer & Taylor, 2007). For example, students can first be taught about the concept of credible intervals which do carry distributional information. Later, this might facilitate the learning of confidence intervals, which do not carry distributional information.

There are now many excellent resources for learning Bayesian statistics. Dienes (2011) provides a comparison of the two approaches and where they may result in different conclusions. Kruschke and Liddell (2018b) present an excellent overview of Bayesian data analysis for individuals starting out on the topic, as does Etz and Vandekerckhove (2018). van de Schoot and Depaoli (2014), as well as Depaoli and van de Schoot (2017) discuss what to report in a Bayesian analysis. Etz et al. (2018) provide a list of recommended readings for different topics in Bayesian data analysis.

More comprehensive sources and those that specifically discuss how to conduct Bayesian analyses can be found in several excellent textbooks. Lambert (2018) explains even some very basic concepts and uses many conceptual examples and figures to make the concepts more easily understood. This is one of the few textbooks on Bayesian statistics that might be used for an undergraduate psychology statistics course. However, in terms of practical data analysis there is not a significant source of material. Kruschke (2015) presents several excellent chapters on key concepts in Bayesian data analysis, such as MCMC. Kruschke (2015) uses JAGS and Stan, implemented in R. Gelman et al. (2014) is robust in its presentation of most of the key aspects of Bayesian data analysis and how to model many different types of data using GZLM and MLM, for example. However, it does not show readers how to conduct practical data analysis using particular software. McElreath (2020) is likely the single best source for attaining a working knowledge of Bayesian data analysis. Containing a wealth of information, the text discusses Bayesian analysis, how to compare models using information criterion, MLM, and graphical causal models. The examples are clear and well explained both conceptually and through R examples. Throughout the text there is R code with explanations on what each component means. Instead of pages of formulas and equations, there is more focus on how to process data in R and learn from it. McElreath (2020) uses the R package Rethinking (McElreath, 2016) to conduct most analyses. Part of the Rethinking package uses Stan to build models, however users do not need additional knowledge of Stan to build their own models. For instructors who are interested in learning what Bayesian analysis is and learning practical data analytic skills, there is no better text currently available. Lastly, instructors may also be interested in the R package brms. While there are several R packages that make conducting Bayesian data analysis easier, one of the most flexible and user friendly is brms (Bürkner, 2017). Brms can model most distributions and can model multilevel structures. Also built in are model comparisons and plots of conditional distributions. Moreover, it uses similar syntax as lme4, and therefore, is an easier transition for those who might conduct frequentist multilevel modeling. Given the user-friendly syntax, support for many models, and built-in functions such as plots and model comparisons, instructors might use brms for much of the course.

Model building and comparison

In undergraduate psychology statistics books (e.g., Adams & Lawrence, 2019; Kranzler, 2018) and even many psychology graduate statistics texts (e.g., Myers et al., 2010) the focus is almost exclusively on hypothesis testing. In syllabi provided online by the Society for the Teaching of Psychology, hypothesis testing is either explicitly mentioned on the course schedule or implied as the main type of analyses to be covered in every undergraduate syllabus. If the goal is to test a specific hypothesis, such as, “does this treatment reduce the symptoms of the disorder?”, then a hypothesis test is applicable. However, most applied statistics are not about testing a hypothesis, but rather estimating or building a model. For example, we might be interested in, “how much does this treatment reduce the symptoms of the disorder?”. An analyst might also want to explore several variables in order to build the best predictive model. For example, “How do we best predict recidivism?”. These goals/questions cannot be accomplished in hypothesis testing. It is therefore unfortunate that hypothesis testing is often the sole focus. As hypothesis testing is always frequentist, the approach we teach is even more narrow. When regression is taught, there may be some basic aspects of model building that are covered; however, most texts focus on using problematic model indices like R^2 and problematic approaches such as stepwise regression, which often result in overfitting.

Students would benefit greatly if they were taught explicitly about different parameter estimation techniques, how to build strong models, and how to compare models. In frequentist statistics, parameter estimation is often achieved via maximum likelihood estimation (MLE). Instead of interpreting if a parameter value is "significantly" different than zero, the analyst can simply estimate different parameters and use confidence intervals to understand the plausible range of values. MLE involves finding parameter values that maximize the likelihood of obtaining the data (Hoffmann, 2004). In Bayesian statistics, posterior distributions of parameters are typically estimated through an MCMC algorithm, and highest density intervals calculated on parameters. The parameters in Bayesian statistics represent the probability of the parameter given the data.

While there are many important aspects of model building, including predictive checking, an important aspect that I focus on here is model comparison. Model comparison involves fitting different models, for example, one with interactions compared to one without. The relative performance of the models can then be examined, and the analyst can better understand why some models perform better. Several different model comparison techniques also penalize (in some way) the inclusion of additional parameters. For example, in regression, the amount of variance in the outcome attributable to the predictors will almost certainly increase with the inclusion of additional parameters. As a consequence, the model becomes too tailored for that particular dataset. Needlessly including variables or adding irrelevant complexities makes the model fit the noise and not the signal. It is no surprise then when these models fail to replicate on another sample with its own quirks. Model comparison allows the analyst to critically think about advantages of some models and have more justification for increasing model complexity.

Model comparison is closely linked with one of the most important, but rarely mentioned concepts, in undergraduate psychology statistics: overfitting/underfitting. When a model is too tailored to the sample and thus learns too much from the data, the model can be overfit. On the other hand, when the model has not learned enough from the data, such as when there is an interaction and it is not in the model, the model is underfit. The desire, typically, is to have a model that is neither over nor underfit. In teaching this concept to undergraduate students, a dress analogy may be useful. A wedding dress, which fits very well on the wedding day but poorly other times, is a useful metaphor for overfitting (i.e., the model fits wells but only for that specific data). A baggy dress is tailored for no one and is a metaphor for underfitting (i.e., the model needs to be more tailored to the data). The "little black dress" fits well, but is not so tailored or specific to an occasion that it cannot be worn again, is used to illustrate the goal in model comparison.

While there are a variety of measures for model comparison that allow the analyst to measure overfitting, some of the most effective are information criteria. There are several key concepts, such as divergence, that are necessary to understand what information criteria are. Kullback-Leibler divergence

(KL divergence, or simply, divergence) is a way to quantify the additional uncertainty from using probabilities of one distribution to describe another (McElreath, 2020). The equation for divergence states that the divergence is the average difference in log probability between the two models: the target and the proposed model. Of course, divergence assumes the target distribution is already known. Thus, deviance is estimated instead. Deviance is the log predictive density of the data, given a point estimate from that particular model, multiplied by -2 (so now smaller values are better). Specifically, in-sample deviance provides an estimate of how well the model does at predicting the given data. Out-of-sample deviance provides an estimate of the "future" predictive accuracy of the model (i.e., the expected out of sample predictions). Deviance can be used to estimate the relative predictive abilities of candidate models. While it cannot tell us that some model is the "true" model, it can tell us which model is relatively better at predicting, given the number of parameters. Therefore, information criteria are ways to measure overfitting.

There are a variety of information criteria available on most statistical software including the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The AIC was the first information criteria developed and can be used for a variety of models such as regression and times series analysis (Akaike, 1973, 1974, 1981, 1987). There are now several information criteria available, each with their own assumptions and applications. The widely applicable information criterion, or Watanabe–Akaike information criterion (WAIC), is the most generalizable approach. WAIC is calculated by taking averages of log-likelihood over the posterior distribution and provides an estimate of the out-of-sample deviance. However, information criteria are only one strategy for assessing overfitting and model predictions. Another strategy is cross-validation, including the approximate leave-one-out cross-validation approach, for models fit using MCMC (LOO; Vehtari et al., 2017).

The mathematical knowledge required to fully understand information criteria or cross-validation approaches is beyond the scope of an undergraduate psychology statistics course. Just as with any undergraduate course, students at this level do not have to fully understand all the mathematics to be able to understand why these approaches are useful. Students can be taught the importance of assessing overfitting/underfitting and how to interpret metrics like AIC or WAIC. For understanding the basic concepts in using information criteria or cross validation see McElreath (2020) and Konishi and Kitagawa (2007).

It is important to understand that information criteria and cross-validation are used to measure overfitting, but they do not *address* overfitting. To address overfitting, in Bayesian statistics, regularizing priors can be used as a natural aspect of model building. In frequentists statistics, regularizing regression techniques can be used, such as ridge regression, lasso regression, or elastic net. Using regularizing regression techniques are simple and accomplished, for example, in R via functions such as lm.ridge in the MASS package

(Venables & Ripley, 2002). Regularizing priors or techniques such as ridge regression could easily be incorporated into a course that discusses overfitting and students would benefit from at least a simple data exercise using one of these approaches.

Causal analysis

Causality is one of the most important topics in data analysis. The issues in most texts and courses is that 1) there is often relatively little on the subject and 2) only one particular and problematic view of causality is all that is presented. Several statistics texts for psychology provide examples.

In Gravetter and Wallnau (2013), there is less than one page dedicated to causation, with the discussion centered on the adage that correlation is not causation. Warne (2018) has a brief, but more detailed discussion of causation, with some examples of third-variable problems. However, Warne (2018) focuses on the point that correlation is not causation. Lastly, Welkowitz and colleagues (2012) have a more detailed discussion of causality for a psychology statistics text, but there too, the focus is on correlation and causation. To the credit of Welkowitz et al. (2012), there is some discussion of the nuance of the statement, "correlation does not imply causation", and in a footnote, they comment that methods such as structural equation modeling can allow us to understand causal relations without random assignment. Nonetheless, these and other undergraduate texts give students very little information about our modern understanding of causality. These "traditional" views can likely be traced to three giants in the field of statistics: Sir Francis Galton, Karl Pearson, and Sir Ronald Fisher, with Pearson potentially having the greatest impact.

Karl Pearson, who was a disciple of Galton, believed that causation was essentially a special case of correlation (Pearson, 1892), meaning, correlation was the larger category. Pearson believed that data were all there is to science, and that everything can be reduced to a contingency table. A quote from Karl Pearson's Grammar of Science (1892) clearly indicates what he thought of causation: "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect." (p. iv).

In the early 1900s, Pearson's laboratory, Biometrics lab, was the center of the world for statistics (Pearl & Mackenzie, 2018; Porter, 2004). Pearson, along with Galton and Fisher, have had a major impact on statistics not only in the form of equations and tests, but philosophy. Their views of causation dominated the first half of the 20th century. Not surprisingly, when advancements were made in causal modeling, such as Sewall Wright's path analysis, they had little impact on most fields. Despite the advantages of path analysis, it was not used heavily, and some dismissed path analysis out of hand. Henry Niles (1922), who was a student of a student of Karl Pearson, famously published a brutal rebuttal of path analysis and criticized Wright's understanding of causation, while incorrectly interpreting Wright's results. Somewhat parallel to the adoption of NHST in statistics rather than Bayesian statistics, a relatively small but influential number of people had a major impact on the field. That most statistics texts in psychology have very little to say on causation and always in reference to the correlation, suggests that, as a field, the psychological view of causation is not dissimilar to what Karl Pearson espoused over 100 years ago. Causation is more discussed in methodology texts and courses (e.g., Cozby & Bates, 2018). However, the focus still tends to be on correlation is not causation, or experimental results vs. non-experimental, and there is rarely if ever mathematical justification for the arguments presented. The heart of the problem is that significant progress has been made in our understanding and investigation of causation and psychological statistics courses have not been updated.

In the area of causation, major contributors include Donald Campbell, Clive Granger, Donald Rubin, and Dawid Philip. Judea Pearl and his colleagues have had some of the biggest impact on the modern causal revolution, tracing back to the 1980s. Pearl and his colleagues have introduced a range of important concepts and algorithms to understand causation (Pearl, 2009). While some of these concepts would be beyond the scope of a psychology statistics course, some are indeed well suited and are critical for modern scientific research.

One concept that is likely to be beneficial to undergraduate psychology students is what Pearl (2009) calls the ladder of causation. Pearl (2009) distinguishes three levels of causation. Association, the lowest rung, is akin to observing. Intervention is the second rung on the ladder, akin to doing or intervening. Counterfactuals are the highest rung and involve imagining and retrospection. Counterfactuals allow us to understand "was it X that caused Y?" Studies and results from these different levels tell us fundamentally different information. Pearl (2009) shows mathematically, using do calculus/do operator, how each of these rungs or levels are different. While students may not understand all the mathematics, they can benefit greatly from understanding the types of questions posed at each level and therefore improve their critical and quantitative thinking skills which are not only important for conducting and digesting research, but in daily life as well.

As an example of the types of questions posed by each level, consider the COVID-19 pandemic. "Does difficulty breathing tell us about having COVID-19?", is a question of association. "If those with COVID-19 are administered hydroxychloroquine, will the symptoms be reduced?" is a question of intervention. "What if mask mandates had been instituted earlier in the pandemic, how many lives would have been saved?" is a counterfactual. Understanding this causal hierarchy does not require advanced mathematic abilities, the nuance far exceeds that of "correlation is not causation," and it is justified mathematically.

A second aspect of causal analysis that students could benefit greatly from is graphical causal models, such as Directed Acyclic Graphs (DAGs). DAGs involve vertices/nodes/variables that are connected together with edges/lines. Edges can be absent, indicating no causal relation, or single headed, indicating an asymmetrical relation such as A causes B. Edges can also double headed in graphs, as shorthand indicating that the vertices are not connected such that one causes the other, but that they share some common unknown cause. In Figure 1, there are three vertices with two edges, indicating that A causes B and subsequently B causes C. Sewall Wright's path analysis brought together directed graphs and data. Given a causal graph, certain relations must hold true. If no relationship implied by the causal graph is found, then we have evidence against a causal model. Readers will no doubt recognize the similarity to mediation analyses and structural equation modeling. While causal graphs are an inherent part of many statistical approaches, such as structural equation modeling, they are also inherently intuitive. The intuitive nature of causal graphs is arguably a major advantage. Thus, with a few easily understood principles, students can learn to think and communicate causally.

Arguably there is a great thirst for cause in psychology. The Baron and Kenny (1986) paper on mediation is one of the most cited papers in psychological science (cited over 97,000 times overall and over 17,000 in psychology). By introducing students to graphical causal models, not only are they prepared for understanding mediation and more modern techniques like latent variable modeling, but they are also prepared for understanding more complex aspects of modern causal analysis.

More importantly, using graphical causal models like DAGs, coupled with causal algorithms like d-separation, allow an analyst to get a causal estimate of an effect from observational data (Pearl, 2009). d-separation is a criterion for deciding, given a causal graph, whether some set of variables (X) are independent of another set (Y), given a third set (Z). Meaning, given the DAG is true, you can mathematically determine which variables you need to include or not

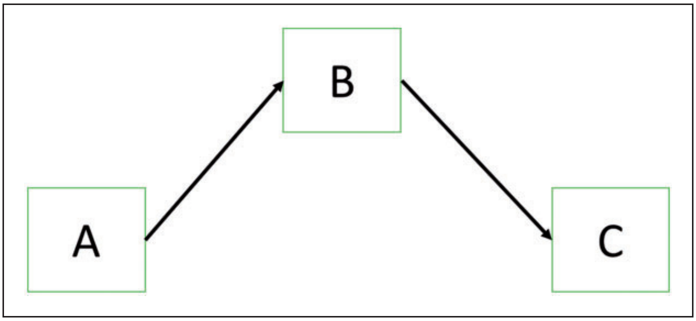

Figure 1. A basic directed acyclic graph.

include to get a causal estimate. Many undergraduate psychology statistics texts mistakenly write or imply that only experimental methodology can yield a causal estimate. As most scientific questions cannot be answered using experimental methodology, the ability to get a causal estimate from observational data is a major achievement and this should be a core part of the statistics education of psychologists.

Using graphical causal models like DAGs, students could also learn the four elemental confounds. In some statistics texts (e.g., Warne, 2018) as well as many research methods texts (e.g., Cozby & Bates, 2018), there is a discussion of what is posed as the "third-variable problem". For example, in Cozby and Bates (2018), the authors discuss how an association between exercise and anxiety could be caused by exercise causing anxiety, anxiety causing exercise, or a third-variable, such as income, causing both exercise and anxiety. When such an uncontrolled variable is operating, we might term it a confounder or confounding variable (more technically, confounding is anything that makes P (Y|do(X)) differ from P(Y|X); Pearl, 2009). Third variables are important considerations and allow us to think more critically about the cause-and-effect relationships in our model. However, the issue with traditional statistics texts is that this is the only type of confounder mentioned, when in fact, there are three more elemental confounds.

The familiar third-variable problem in most statistics texts of a common cause of two other variables is often referred to as a fork. In the language of DAGs, the relationship can be shown as $X \leftarrow Z \rightarrow Y$. In a fork, if we condition on Z, then learning X tells us nothing about Y; that is, X and Y are independent, conditional on Z. A second type is called a chain/pipe, expressed as $X \rightarrow Z \rightarrow Y$, meaning X causes Z, and Z causes Y. If we condition on Z in a chain, the path from X to Y is blocked. However, if the intent is to learn about the relationship between X and Y, conditioning on Z stops the flow of information from X to Z. For example, in a regression model, if there is a chain, and X and Z are predictors, the path from X to Y will be blocked, and it will appear that there is no relationship between X and Y. A third type is the collider, expressed as $X \rightarrow Z \leftarrow Y$. Here, there is no association between X and Y, *unless you condition on* Z. Conditioning on Z, the collider variable, opens the path. Once the path is open, information flows between X and Y. This can be highly problematic. For example, if a model is analyzed in a regression with X and Z as predictors and Y as the outcome, but Z is a collider, it can appear that there is a relationship between X and Y when no direct or indirect relationship exists (they are independent). Lastly, there is the descendant. Let's assume the following causal model, $X \rightarrow Z \rightarrow Y$, $Z \rightarrow K$. In the causal model, K can be said to be a descendant of Z. Controlling for K, will also to a lesser extent control for Z, which will partially block the path from X to Y. So, for example, conditioning on a descendent of a collider, will weakly condition on the collider.

Not understanding the four elemental confounds and their consequences has major ramifications. For example, research in epidemiology on the effects of smoking suggested that smoking might be *beneficial* to the infants of mothers who smoked, but only those born underweight (Pearl & Mackenzie, 2018; Yerushalmy, 1971). The reason for this paradoxical finding was not well understood for many years. Arguably, VanderWeele (2014) and others have now resolved this paradox by understanding that birthweight is a collider. By conditioning on birthweight, the effect of smoking on infant mortality is biased so much that it can appear that smoking is beneficial (see also Pearl & Mackenzie, 2018). Examples such as this should make it clear that without causal analysis, progress is substantially hindered. Knowing the difference between a collider and a fork, for example, is fundamental to conducting meaningful data analyses, yet this topic is rarely discussed in any psychology research methods or statistics classes or texts. Applying Bayesian statistics, MLM, or other approaches like machine learning, will not solve issues in causality. A major advantage of DAGs is that they make doing causal analysis intuitive and easy to communicate. Additionally, there are excellent R packages, such as dagitty, for making DAGs and using DAGs to engage in causal analysis with techniques like d-separation (Textor et al., 2016).

Many excellent texts exist on causality. Pearl and Mackenzie (2018) discuss the history of causation in science and what led to the modern causal revolution. They also discuss and have many examples of key concepts in causal analysis, such as the d-separation and the back-door criterion. Written for a general audience, it is an excellent introduction to modern causal analysis. McElreath (2020) discusses how to use DAGs and has several excellent real data exercises. Shipley (2016) has several introductory chapters that discuss basic aspects of causal analysis, but the focus is on using R for path analysis, structural equations, and causal inference. Shipley (2016) also discusses how to use R to build DAGs and test for probabilistic independence via d-separation. For a more detailed discussion of these issues with mathematical formulations see Pearl (2009). Lastly, for an overview of some these issues in article format, see Rohrer (2018).

Impediments and objections

Modifying our content in major ways is not without obstacles and one could raise many reasonable objections. First, statistics instructors may not be convinced that revision is necessary. Whether it is the limitations of GLM, the narrow scope of hypotheses tests, or problems in interpreting p values, I have outlined some issues with the current content. Space limitations prevents more in-depth discussion of these issues, but readers should refer to the citations for more developed critiques within each topic. While some may not be convinced that any one area of the alternative content deserves inclusion, I would encourage those instructors at a minimum, to clearly justify why the current content is adequate.

Second, limited time in a course may preclude some topics or may require some content to not be taught. If a department only offered one statistics course and this was mandatory, it is likely that some topics such as *t* tests, one-way ANOVA, factorial ANOVA, and repeated measures ANOVA could be removed. Given the greater flexibility and usefulness of generalized multilevel modeling, this is arguably a valuable tradeoff. To assist in planning of such a course, in the supplementary material there is an example course schedule. If a department offered multiple undergraduate statistics courses and all were mandatory, the traditional content would not necessarily have to be reduced. The second course could focus on fitting Bayesian generalized multilevel models, for example. Indeed, given the usefulness of statistics in science and in daily life, having two mandatory statistics courses would be highly beneficial.

Third, some instructors may feel that the suggested alternative content is too complex for undergraduate statistics or that it is beyond the capability of the students. It is true that some topics, such as GLZM, are necessarily more complex than GLM. However, the added complexity is arguably justified, given the benefits of GLZM. Statistics instructors might point out that students already struggle with topics like regression (i.e., GLM). However, rather than teach to the average student's ability, one could argue that we should teach the material that students need to know. This could mean a minor reduction in psychology majors, that students have to take the course multiple times, or that a university may have to increase the number of basic math courses required so students start with a stronger mathematical background. However, once again, this is arguably preferable than teaching older, albeit less complex, material.

On the other hand, most of the alternative content that has been suggested is not more complex. It is not clear that Bayesian statistics are inherently more complex than frequentist statistics, or that model building is inherently more complex than hypothesis testing. Any topic, for example ANOVA, requires much time and dedication to understand. Undergraduate students are not expected to master ANOVA, and simple rather than complex ANOVA examples are often used (e.g., 2 x 2 between-subjects ANOVA). Similarly, in teaching alternative content, mastery of multilevel modeling is not expected. What is expected is an introduction to the material and teaching students how to fit simple multilevel models, for example. For those students who progress to graduate school or who gain employment in data science related fields, they will have the fundamentals to then further their education.

Fourth, there is a lack of textbooks that discuss all of the suggested alternative content and that is directed towards undergraduate psychology students. This may be the biggest impediment currently, and a very reasonable objection. If only one mandatory statistics course was offered by the department, there is currently no one text that could be used. However, it is possible to supplement the required textbook with readings, including chapters from other books, such as those suggested here. Further, if a two-sequence statistics course was mandatory, McElreath (2020) could be used, as the text uses Bayesian analysis and shows how to compare models using information criterion, uses multilevel modeling, and DAGs. Further, the text uses the R package Rethinking, which was designed as a teaching tool. Lastly, if more instructors teach the suggested alternative content, publishers may be motivated to develop more comprehensive textbooks.

Fifth, the suggestions herein for data analysis have involved the statistical software R, rather than the more commonly used SPSS in psychology. Learning the alternative content while also learning less user-friendly software like R may seem like a monumental challenge to instructors. Furthermore, for those instructors who only know SPSS, there would be a substantial time investment in reworking data examples and assignments to R. While companies like DataCamp are an invaluable resource for instructors learning R and learning some analyses, there is no easy solution. Again, the added effort arguably pays dividends. Using statistical software like R, in particular packages such as Rethinking, allow the analyst (whether an undergraduate or instructor) to better understand the model, as more components and options have to be set explicitly (see McElreath, 2020). SPSS is advantageous in that it makes running analyses simple. The downside is that the model assumptions and predictions may be harder to understand. This is an empirical question, and to the author's knowledge, no data exists on this topic. An advantage of R is that it is free, while SPSS is a costly program. Lastly, if students do need to analyze data in their daily life or future career, they can do so in R. But it is unlikely that their employer would have SPSS. Further, job positions that involve data analysis specifically mention programs like R and Python more than programs like SPSS. There is other free software such as JASP and jamovi and both have point and click interfaces. JASP in particular is intuitive to use, produces APA style tables, and offers some Bayesian and frequentist options. Students are likely to find JASP easier to use than R, and thus from a teaching standpoint there are some major advantages. Given the cost, JASP and jamovi are likely superior alternatives to SPSS. R, however, is the recommended software. R is a more powerful and flexible software. There are packages in R for causal analysis, for example, whereas there are none for SPSS, JASP, or jamovi. Teaching R gives students skills for the job market. Lastly, the analyst has to specify more in R, and since many packages require writing the actual formula for the model, it can be used to reinforce the learning of basic information like writing regression equations. In particular packages such as Rethinking, force the analyst to explicitly identify all priors and parameters. Thus, the analyst must have a strong understanding of the model that is being built. In other software it is possible to easily build a model but not understand it.

Addressing reproducibility

While there are many justifications for modifying the content in our statistic courses, a particularly serious justification concerns the field itself. The undergraduate students we teach today will be the data analysts and professors of tomorrow. If the same material is taught, the same issues will keep developing. One such issue that has raised concern is reproducibility.

Most scientists recognize that there is a reproducibility problem, and there are many contributing factors such as selective reporting and pressure to publish (Baker, 2016). It may be a particular problem in some fields, such as psychology and medicine. No easy solution exists, and there are multiple fronts on which this needs to be addressed, including at the department level with tenure and on the publisher side. Part of the solution is changing our data analysis practices. There are many ways that we can do this, but it must start early. Therefore, the mandatory undergraduate statistics course(s) in psychology is potentially a cornerstone for improving our data analysis practices in the long term and potentially addressing the reproducibility issue within psychology.

To be sure, the main purpose of introducing the suggested content is not to address the reproducibility problem. Bayesian statistics will not eliminate false positives in the literature. If data collection and analysis is not conducted cautiously and thoughtfully, results can still be meaningless, using Bayesian models or otherwise. However, in addressing the reproducibility problem, updating our approaches and having sufficient justification for our content is reasonable and offers some advantages. Consider the advantages of understanding causality. Suppose a clinical researcher conducts a study in a mental health facility, investigating the relationship between two variables, X and Y, which might be disorders or symptoms. A relationship is found, and the study is published. However, upon replication in the general population, no relationship is found between the symptoms. What could be happening in this case is by sampling in a mental health facility, it could have the effect of conditioning on a collider. When the collider is conditioned upon, it introduces a spurious relationship. For example, this can happen when a study is sampled from a hospital and two diseases that are independent are believed to be associated; however, the association is spurious. Both diseases independently cause hospitalization. A researcher can condition on a collider through their sampling strategy. By using DAGs and thinking causally, we might reduce these types of issues.

Empirical examples also make this clear. One issue that has received worthwhile attention is police shootings of Black Americans. A substantial literature on this topic has been developed across the social sciences. Some studies find seemingly paradoxical findings (partly due to differences in benchmarking) and without a strong mathematically justified argument, it is difficult to sort out this literature (see for example, Cesario et al., 2019; Fryer, 2019; Scott et al., 2017). Doing just that, Ross et al. (2018) and Ross et al. (2021) use causal analysis with mathematically defensible benchmarking and Bayesian modeling to show that there is evidence of an anti-Black bias in police shootings. Understanding whether there is evidence of a bias in police shootings is not a trivial issue. Techniques such as those used in Ross et al. (2021) can help move the literature forward and students can start learning this material in their undergraduate statistics course(s).

Model comparison and information criteria might also help. Overfitting is a serious issue in building models. If a model is overfit, new data are not likely to generate similar results. In some statistics courses, measures such as R^2 are taught as indicators of model performance. However, these will result in overfitting, as it is only based on the sample data. By teaching students early on the importance of overfitting and how to use information criteria to compare models, the models that are published are likely to be less overfit and might have a higher probability of replicating.

An advantage of Bayesian statistics is that there is no dichotomous decision making, as in *p* values, thus *p* hacking would be avoided. While a researcher could change the priors to such a degree to make an effect appear reliable, priors are set explicitly and should be justified. Thus, this would be more transparent and likely easier to detect than in *p* hacking. Additionally, as the uncertainty in the parameter estimates are very clear and easier to interpret in Bayesian statistics, this might help researchers not develop overconfidence in the effects. This would also be an advantage of multilevel modeling. If there are multilevel structures and they are not accounted for, the model has artificially lower uncertainty. By using multilevel modeling, more realistic parameters estimates can be generated. Lastly, using an appropriate model given the type of data is an issue and using GZLM might help address that. For example, in psychology Likert data is often analyzed using metric models such as ANOVA and regression and doing so can lead to both type I and type II errors (Liddell & Kruschke, 2018). By teaching GZLM students can learn early on to apply more appropriate models given the type of data.

If these types of strategies can address the reproducibility problem to any degree is an empirical question. However, given the advantages of generalized multilevel modeling, Bayesian statistics, model building and comparison, and causal analysis, it is worth exploring. There are many other aspects of teachings statistics that could also be improved. For example, statistics texts rarely mention data organization and curation practices such as how to keep version control of the data, how to document how the data were analyzed, and how to communicate that (e.g., using R Markdown files). Improving these data practices in the scientific community and teaching these skills to students early on might also help to address the reproducibility problem.

Conclusions

Methodology and data analysis are central to understanding research, and thus is part of the core of psychological science. They are also central to daily life. As statistics are not required in high school and most psychology students do not progress to graduate school, it is likely that the only data analysis class they will receive is their mandatory undergraduate statistics course. In an introductory psychology class, students learn about the different sub-disciplines in psychology and the major issues in the field. This prepares them to take courses in other subdisciplines, such as cognitive psychology. The undergraduate statistic course in psychology has similar goals. For example, it gives students the basic knowledge in statistics to then apply to future courses. However, the content covered in most psychology statistics courses does not appear to be setting students up for future success. The focus on hyper-specific tests like *t* tests and little discussion of causation cannot prepare students well for the larger area of data analysis and scientific inquiry. Furthermore, as suggested by Wagenmakers et al. (2018), teaching the same topics continually may create a self-perpetuating cycle, where psychological scientists only use these basic tests.

Data analysis, specifically statistics, has progressed over many decades and has become a field in and of itself. As instructors, there is a duty to keep material up to date and to justify the material that is covered. I have suggested some topics including generalized linear modeling, multilevel modeling, Bayesian statistics, model building and comparison, and causal analysis. However, there are many more areas that arguably should be taught to students at the undergraduate level. Ideally, we might completely restructure our psychology statistics course(s) from the ground up, focusing more on Bayesian statistics and using generalized multilevel models, for example. However, at a minimum, undergraduate instructors might at least consider adding some of the topics outlined here. Nowhere in the arguments here have I suggested that these are the best content areas or that we should seek uniformity across the discipline. There is no uniformity even among statisticians. Indeed, I have discussed the advantages of including particular topics and encouraged undergraduate psychology statistics instructors to justify whatever material is taught. There are many other content areas we might consider teaching in the psychology undergraduate curriculum including bootstrapping techniques, structural equation modeling, factor analysis, ridge regression, network analysis, and missing data techniques like multiple imputation.

While mastery of any of the topics discussed here is difficult, resources have been provided to aid in instructors continuing education. We must think about the future of data analysis in the behavioral sciences, not just on the past and what is easiest. Having stronger justifications for our included content and continually updating much of the content may aid us in the long run to address the reproducibility problem, improve the image of psychological science, and produce citizens and scholars who can critically think about data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Joshua J. Reynolds https://orcid.org/0000-0001-9205-0395

References

Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PLoS One*, *12*(8), 1–8. https://doi.org/10.1371/journal.pone.0182651

Adams, K. A., & Lawrence, E. V. (2019). *Research methods, statistics, and applications* (2nd ed.). Sage Publications.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology. *American Psychologist*, *62*, 32–50.

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist*, *45*(6), 721–734.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Academiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*(1), 3–14. https://doi.org/10.1016/0304-4076(81)90071-3

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332. https://doi.org/10.1007/BF02294359

Alder, A. G., & Vollick, D. (2000). Undergraduate statistics in psychology: A survey of Canadian institutions. *Canadian Psychology/Psychologie Canadienne*, *41*(3), 149–151. https://doi.org/10.1037/h0086864

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*(3), 233–239. https://doi.org/10.1177/2515245919858072

Cesario, J., Johnson, D. J., & Terrill, W. (2019). Is there evidence of racial disparity in police use of deadly force? Analyses of officer-involved fatal shootings in 2015–2016. *Social Psychological and Personality Science*, *10*(5), 586–595. https://doi.org/10.1177/1948550618775108

Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, *69*(4), 266–282. https://doi.org/10.1080/00031305.2015.1093029

Cohen, J. (1994). The earth is round (p <.05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cozby, P. C., & Bates, S. (2018). *Methods in behavioral research* (13th ed.). McGraw-Hill.

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *American Statistical Association*, *4*, 92–97.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065.supp

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

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R*. Springer.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234. https://doi.org/10.3758/s13423-017-1317-5

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

Finch, W. H., Bolin, J. E., & Kelley, K. (2019). *Multilevel modeling using R*. CRC Press.

Friedrich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology*, *27*(4), 248–257. https://doi.org/10.1207/S15328023TOP2704_02

Friedrich, J., Childress, J., & Cheng, D. (2018). Replicating a national survey on statistical training in undergraduate psychology programs: Are there “new statistics” in the new millennium? *Teaching of Psychology*, *45*(4), 312–323. https://doi.org/10.1177/0098628318796414

Fryer, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, *127*(3), 1210–1261. https://doi.org/10.1086/701423

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*(3), 372–396. https://doi.org/10.1111/j.1751-5823.2007.00029.x

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org10.1053/j.seminhematol.2008.04.003

Gravetter, F. J., & Wallnau, L. B. (2013). *Essentials for statistics for the behavioral sciences* (8th ed.). Thomson Wadsworth.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22.

Hoffmann, J. P. (2004). *Generalized linear models: An applied approach*. Pearson Education.

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with p values? *The American Statistician*, *73*(sup1), 20–25. https://doi.org/10.1080/00031305.2018.1447512

Koehler, J. J., & Macchi, L. (2004). Thinking about low-probability events. *Psychological Science*, *15*(8), 540–546.

Konishi, S., & Kitagawa, G. (2007). *Information criteria and statistical modeling*. Springer.

Kranzler, J. H. (2018). *Statistics for the terrified* (6th ed.). Rowman and Littlefield.

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (2nd ed.). Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-016-1221-4

Kruschke, J., & Liddell, T. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. doi:10.3758/s13423-016-1221-4

Lambert, B. (2018). *A student's guide to Bayesian statistics*. Sage.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*(1), 22–25. http://doi.org/10.1111/j.1467-9639.1993.tb00252.x.

McElreath, R. (2016). *Rethinking: Statistical rethinking book package*. R package version 1.59.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press/Taylor & Francis Group.

Moore, D. (1998). Statistics among the liberal arts. *Journal of the American Statistical Association*, *93*(444), 1253–1259. https://doi.org/10.1080/01621459.1998.10473786

Myers, J. L., Well, A. D., & Lorch, R. F. (2010). *Research design and statistical analysis* (3rd ed.). Routledge.

National Science Foundation, National Center for Science and Engineering Statistics. (2017). National survey of college graduates public use microdata file and codebook. https://sestat.nsf.gov/datadownload/.

Niles, H. E. (1922). Correlation, causation and wright's theory of "path coefficients. *Genetics*, *7*(3), 258–273.

Page, R., & Satake, E. (2017). Beyond p values and hypothesis testing: Using the minimum Bayes factor to teach statistical inference in undergraduate introductory statistics courses. *Journal of Education and Learning*, *6*(4), 254–266. http://doi.org/10.5539/jel.v6n4p254

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pearson, K. (1892). *The grammar of science*. Adam & Charles Black.

Porter, T. M. (2004). *Karl Pearson: The scientific life in a statistical age*. Princeton University Press.

Project Syllabus – Society for the Teaching of Psychology. (2021). http://teachpsych.org/otrp/syllabi/index.php#stats

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roberts, D. M., & Reese, C. M. (1987). A comparison of two scales measuring attitudes towards statistics. *Educational and Psychological Measurement*, *47*(3), 759–764. https://doi.org/10.1177/001316448704700329

Robson, K., & Pevalin, D. (2016). *Multilevel modeling in plain language*. Sage.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*(3), 355–367. https://doi.org/10.1007/s10648-012-9201-3

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498. http://doi.org/10.1007/s11251-007-9015-8

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Ross, C. T., Winterhalder, B., & McElreath, R. (2018). Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police. *Palgrave Communications*, *4*(61), 1–9. https://doi.org/10.1057/s41599-018-0110-z

Ross, C. T., Winterhalder, B., & McElreath, R. (2021). Racial disparities in police use of deadly force against unarmed individuals persist after appropriately benchmarking shooting data on violent crime rates. *Social Psychological and Personality Science*, *12*(3), 323–332. https://doi.org/10.1177/1948550620916071

Satake, E., & Murray, A. V. (2014). Teaching an application of Bayes' rule for legal decision-making: Measuring the strength of evidence. *Journal of Statistics Education*, *22*(1), 1–29. https://doi.org/10.1080/10691898.2014.11889692

Schacter, D. L., Gilbert, D. T., & Wegner, D. M. (2019). *Psychology* (5th ed.). Worth Publishers.

Scott, K., Ma, D. S., Sadler, M. S., & Correll, J. (2017). A social scientific approach toward understanding racial disparities in police shooting: Data from the department of justice (1980–2000). *Journal of Social Issues*, *73*(4), 701–722. https://doi.org/10.1111/josi.12243

Shipley, B. (2016). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference with R* (2nd ed.). Cambridge University Press.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. The Penguin Press.

Szucs, D., & Ioannidis, D. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, *11*(390), 1–21. https://doi.org/10.3389/fnhum.2017.00390

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: The R package 'dagitty'. *International Journal of Epidemiology*, *45*(6), 1887–1894. https://doi.org/10.1093/ije/dyw341

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist*, *16*, 73–82.

VanderWeele, T. J. (2014). Commentary: Resolutions of the birthweight paradox: Competing explanations and analytical insights. *International Journal of Epidemiology*, *43*(5), 1368–1373. https://doi.org/10.1093/ije/dyu162

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. (2018). https://doi.org/10.3758/s13423-017-1343-3

Warne, R. T. (2018). *Statistics for the social sciences: A general linear model approach*. Cambridge University Press.

Welkowitz, J., Cohen, B. H., & Lea, R. B. (2012). Introductory statistics for the behavioral sciences (7th ed.). John Wiley & Sons.

Yerushalmy, J. (1971). The relationship of parents' cigarette smoking to outcome of pregnancy – Implications as to the problem of inferring causation from observed associations. *American Journal of Epidemiology*, *93*(6), 443–456. https://doi.org/10.1093/oxfordjournals.aje.a121278

Yong, C. S., & Rosli, R. (2020). Validity and reliability of the survey of attitudes toward statistics (SATS) instrument. *Malaysian Journal of Education*, *45*, 17–24. http://dx.doi.org/10.17576/JPEN-2020-45.01SI-03

Author Biographies

Joshua J. Reynolds is an assistant professor at the University of Scranton. His research is multidisciplinary in nature. Topics include homicide, rape, exploitative and deceptive strategies, self-control, jury decision making, police legitimacy, and fourth amendment interactions.