The Eight Steps of Data Analysis: A Graphical Framework to Supplement (or Replace) Null Hypothesis Significance Testing

Dustin Fife[1]

[1] Rowan University

Abstract

Data analysis is a risky endeavor, particularly among those who are unaware of its dangers. In the words of Cook and Campbell (1976; see also Shadish, Cook, & Campbell, 2002), "Statistical Conclusions Validity" threatens all experiments that subject themselves to the dark arts of statistical magic. Although traditional statistics classes may advise against certain practices (e.g., multiple comparisons, small sample sizes, violating normality), they may fail to cover others (e.g., outlier detection and violating linearity). More common, perhaps, is that researchers may fail to remember them. In this paper, rather than rehashing old warnings and diatribes against this practice or that, I instead advocate a general statistical analysis strategy. This graphical eight step strategy promises to resolve the majority of statistical traps researchers may fall into without researchers having to remember large lists of problematic statistical practices. These steps will assist in preventing both Type I and Type II errors, and yield critical insights about the data that would have otherwise been missed. I conclude with an applied example that shows how the eight steps highlight data problems that would not be detected with standard statistical practices.

Keywords: statistical assumptions, NHST, confirmatory data analysis, graphical data analysis, fishing, p-hacking

Word count: X

The field of psychology has of late been forced to participate in methodological introspection, of sorts. This introspection began late in the 20th century as methodologists vehemently protested the knee-jerk focus on p-values that Null Hypothesis Significance Testing (NHST) encourages (Cohen, 1994; Harlow, Mulaik, & Steiger, 2016; Jones, 1952; Rozeboom, 1960). The American Psychological Association (APA) suggested several alternatives, including a much stronger focus on estimation (i.e., identifying the strength of the

Correspondence concerning this article should be addressed to Dustin Fife, 201 Mullica Hill Road Glassboro, NJ 08028. E-mail: fife.dustin@gmail.com

effect, as well as the size/direction of the parameters of interest; Wilkinson, 1999), rather than the probability of no effect in the face of the data (a strange hypothesis indeed!).

This forced introspection not only continues today, but the recent replication crisis has heightened its necessity. A recent attempt to replicate 100 different studies (from three top journals in psychology) resulted in poor replication rates (Open Science Collaboration, 2015); estimates of effect sizes were half as strong in the replications than in the original studies and 61% of the attempted replications failed to produce statistical significance. (For an alternative perspective on the replicaiton crisis, see Shrout & Rodgers, 2018 and Maxwell, Lau, & Howard, 2015). Some have suggested this replication crisis was caused (at least partially) by researchers' overreliance on NHST (and other statistical practices; Cumming, 2014; Pashler & Wagenmakers, 2012). In this paper, however, I will not rant and rave about the absurdity of NHST (others have already highlighted the problems with NHST; Cohen, 1994; Cumming, Fidler, & Thomason, 2001; Schmidt, 1996; Trafimow, 2017). Rather, I introduce a framework for data analysis that will not only protect against false conclusions, but will also free researcher's minds from rigid NHST thinking that is endemic in psychology. But first, I review several reasons why NHST practices are pervasive in psychology, then discuss potential causes for the replication crisis. Next, I review how the eight steps of data analysis encourages a greater focus on estimation and "listening" to one's data. Finally, I conclude with an example where I show how the eight steps prevented false conclusions.

Should Psychology Abandon p-values?

For decades, some methodologists have suggested significance tests ought to have no place in psychological journals (Cohen, 1994; Harlow et al., 2016; Schmidt, 1996; Valentine, Aloe, & Lau, 2015). Yet there seems to be no evidence of p-value extinction in psychological journals, nor does there seem to be much of a visible shift in statistical practices. Despite passionate and cogent arguments against NHST, several obstacles remain and *will* remain, no matter how red-faced methodologists get. These include:

- **Social pressures**. The entire field of psychology understands p-value speak and a researcher may decide not to venture outside NHST practices for fear of having an otherwise publishable paper relegated to the file drawer.
- **Habit**. For many researchers, they have been doing NHST for decades. For these people, shifting away from such rote practices is counterintuitive (and difficult).
- **Learning**. Abandoning p-values in favor of some other statistical practice may require considerable time and effort that most researchers do not have.
- **P-values reduce ambiguity**. Without p-values, it would open the field to disagreement about what constitutes a relationship. A rigid cutoff of 0.05 acts as an operational definition for a relationship that ought to be noticed (and published).
- **P-values provide a "translational mechanism" from theory to data**. As argued by Cortina and Landis (2011), NHST bridges theoretical language into data analysis, and back again into theoretical language. This translation, they argue, is ill-defined, at best, and non-existent, at worst, in other statistical frameworks.

The framework I introduce side-steps (and sometimes addresses head-on) all of these concerns. No researcher needs to learn new software, additional statistical techniques, or

remove p-values from their method section. Rather, I advocate a simple shift in focus that will add richness to one's statistical practices. Doing so will, over time, shift the culture away from p-values and toward a greater focus on estimation and attending to messages our data have long been trying to tell us.

Potential Causes of Replication Crisis

To diagnose the cause of the replication crisis, it is advantageous to think of the current predicament as a systematic, discipline-wide habit of committing Type I errors. In other words:

$$p(\text{rejecting } H_0|H_0) > 0.05$$

Once framed as a Type I error, the diagnosis of the problem is quite simple. Shadish, Cook, and Campbell (2002), noted two characteristics of data in particular that may inflate Type I error rates: (1) Violated assumptions of statistical tests (e.g., normality, homogeneity, linearity, independence), and (2) Fishing/multiple testing. Various authors have commented on both of these issues and I will briefly review each in turn.

Violated Assumptions of Statistical Tests

Linear models (e.g., regression, ANOVA, t-tests, structural equation models) are the most common models in psychology. For these models to behave appropriately (i.e., for p-values to actually reflect the probability of a Type I error under the null), the data must meet four key assumptions: independence[1], normality, homoskedasticity, and linearity.[2]. Some of these assumptions are more critical than others (e.g., if normality is violated, the probability of committing a Type I error, given the null, tend to stay close to 0.05; while violations of independence will lead to significant departures from 0.05). When violated, Type I error rates may remain fairly close 0.05, or may deviate significantly. Likewise, when violated Type II errors may also be inflated.

The previous paragraph is how most textbooks write of statistical assumptions. I tend to think of them differently. If these assumptions are violated, it means the researcher has simply chosen the wrong statistical model, a condition easily fixed by choosing another. Yes, p-values will not remain at 0.05 if assumptions are violated, but it also means the researcher has simply chosen a model that is not appropriate. It would be like choosing to compute the mean on highly skewed data; one can do it but the information gleaned may be misleading. If the wrong model is chosen, one might inflate Type I or Type II error rates.

The sensitivity of linear models to these assumptions have been well documented (e.g., Maxwell & Delaney, 2004; Osborne, 2013) as well as the consequences of violating these assumptions (Micceri, 1989). Yet rarely do researchers mention whether they checked for the appropriateness of linear models (Hoekstra, Kiers, & Johnson, 2012). And because most

[1]Independence is a serious assumption that, when violated, will result in substantial bias. However, independence is more of a design issue than a characteristic of the data. Consequently, I will not address how to visualize independence.

[2]Linearity is actually not an assumption of ANOVAs/t-tests.

researchers do not provide the datasets used for analysis[3], it is unknown the degree to which psychological research has been corrupted by violated assumptions.

The solution to the problem, as I mention in detail later, is a greater focus on estimation and graphical data analysis. Graphics allow the researcher to determine at a glance whether assumptions have been violated.

Fishing/Multiple Testing/P-Hacking

Most researchers are likely familiar with the problem of multiple testing: when there are four groups, for example, it would be unwise to perform a t-test comparing each and every group (group 1 vs. group 2, group 2 vs group 3, etc.). Though the probability of one test being significant under the null is 0.05, the probability of rejecting one among several is much higher (much like the probability of rolling at least one six over the course of 10 rolls is much higher than the probability of getting it on the first roll). Indeed, this problem is sufficiently well-known that if any researcher were to submit a paper and report they performed 107 t-tests, the paper would likely be summarily rejected.

Yet multiple testing likely happens all the time in psychology, but in more nuanced ways (Simmons, Nelson, & Simonsohn, 2011). Suppose, for example, a researcher collected 10 covariates that could potentially muck up the relationship between the IV and the DV. Said researcher might include a covariate, then run the analysis. If the treatment effect is non-significant, the researcher may decide another covariate is more appropriate to control for. This practice may continue until one of the covariates yields statistical significance on the treatment effect. This is another form of multiple testing. Likewise, if the researcher measured several dependent variables, then performed statistical analysis on each DV until significance was achieved, this too constitutes multiple testing, yet it is not so explicitly and universally condemned as running multiple t-tests.

Simmons, Nelson, and Simonsohn (2011) coined a term that encapsulates the new ways in which some researchers participate in multiple testing. They called it "p-hacking," and p-hacking includes not only the practices I have mentioned (measuring multiple DVs and covariates and running several models until significance is obtained), but also several others, including adding more observations until significance is reached and dropping an experimental condition. With a discipline so focused on p-values, it is simple to see why so many researchers exercise "researcher degrees of freedom" (knowingly or unknowingly) to achieve the "gold standard" of $p < 0.05$.

Simmons, Nelson, and Simonsohn (2011) suggest researchers voluntarily report their planned statistical analysis *a priori* (e.g., through the "as predicted" platform: https://aspredicted.org/), then be explicit in the paper about any modifications to the original plan. This is a promising idea, but will not address the statistical issues mentioned previously (i.e, violating of statistical assumptions), nor will it invite researchers to conider the uncertainty associated with statistical analysis. The eight steps of data analysis I propose, on the other hand, incorporate Simmons, Nelson, and Simonsohn's (2011) suggestions into a new framework for data analysis that targets the other sources of Type I errors as well, all while inviting greater caution in interpreting results.

[3]This statement is based on extensive (and frustrating) personal experience.

Type II Errors and the File-Drawer Problem

The replication crisis highlights the fact that psychology may be inundated with Type I errors. Unfortunately, it is more difficult to estimate the frequency of Type II errors. Researchers may spend months collecting data only to have a p-value not reach statistical significance. Some may abandon the project, while others might participate in p-hacking until significance is acheived.

The framework I propose will alleviate the problem of both Type I and Type II errors. For example, a pattern may not reach statistical significance for several reasons that would be detected under this framework, including outliers that pull means to be more similar, strong non-linear patterns that are poorly represented by a straight line, and violated statistical assumptions that render traditional tests overly conservative. In either case, whether we publish spurrious fndings or abandon non-significant results, spending more time with our data via the eight steps of data analysis will aid in shifting our focus to what the data are actually trying to tell us.

Guiding Principles of Data Analysis

Before I explain the eight steps, let me first introduce three guiding principles of data analysis:

1. *Visuals of raw data are almost always preferred over summaries.* One can easily be deceived when a graphic displays only summaries of the data (e.g., means as dots and standard errors as lines). For example, consider Figure 1. Suppose these data came from an experiment where subjects were randomly assigned to watch a neutral video or a violent video. Further, suppose the subjects were subsequently measured on aggression. If one were to simply interpret the left plot, they might believe that the type of video had a large effect on aggression. Yet when we overlay the "jittered"[4] raw datapoints (right panel), we see that the aggression scores are bimodal in the violent group. Perhaps that bimodality is caused by gender (e.g., males report higher aggression after watching the video, while females do not). This would be an important discovery that would be masked if one simply graphed the summaries rather than the raw data. Furthermore, the right panel shows the mean for the violent video group is quite misleading; the mean falls at a place where data are quite sparse.

```
## Warning: package 'tidyverse' was built under R version 3.4.2

## Warning: package 'tidyr' was built under R version 3.4.2

## Warning: package 'purrr' was built under R version 3.4.2

## Warning: package 'dplyr' was built under R version 3.4.2

## Warning: package 'cowplot' was built under R version 3.4.2
```

[4] Jittering means to add random noise to a categorical variable (in this case, Neutral and Violent, which may be coded as 1 and 2). Jittering categorical values prevents overlap of datapoints and makes it easier to see the distribution of the datapoints. See xxx for how to jitter variables in both R and SPSS.

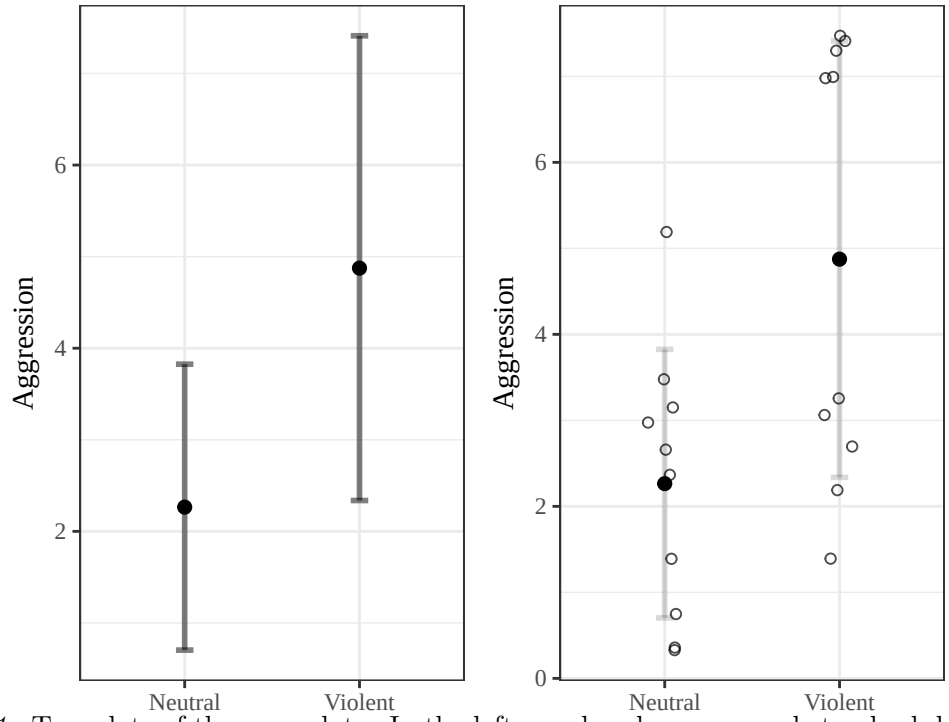


Figure 1. Two plots of the same data. In the left panel, only means and standard deviations are reported. In the right panel, the raw data points overlay the standard deviation bars. When possible, graphical displays of raw data are always preferred over graphical summaries.

2. *Sensitivity analyses are advisable whenever a non-standard decision is made.* Often times our data throw us curve balls that require making a non-standard decision. For example, our data may require a transformation to render the residuals more normal, an outlier may require deleting, or a missing value may require imputing. One would hope the conclusions gleaned from data remain largely unaffected by whatever decision we make. The only way to determine whether our results are sensitive to our decisions is to run the analysis both ways. For example, suppose a researcher decides an outlier ought to be deleted. The researcher should then run the analysis both with and without the outlier deleted and determine the degree to which the results change. The researcher might also investigate other strategies for dealing with the outlier (e.g., treating it as missing then imputing that value). Under this situation, I strongly recommend the researcher report the results of all three analyses (at least in a supplemental section) and comment on whether the results are sensitive to the decision made.

   Note that this sensitivity analysis is comparing two models that test the *same* hypothesis, rather than two models that test *different* hypotheses (i.e., a model comparison; Rodgers, 2010). Model comparisons are excellent tools for teasing out competing explanations of the data, but these sorts of model comparisons are outside the scope of this paper.

3. *Explicit admission of exploratory vs. confirmatory data analysis.* Exploratory data analysis is the process whereby researchers analyze data without preconceived notions of what patterns they might find. Such practices may lead to behaviors characteristically maligned as "fishing", such as exploring the relationship between a multitude of variables on the same dataset, or even performing multiple comparisons. There is actually nothing wrong with this practice, provided researchers are explicit about having no *a priori* hypothesis in their report (Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012). What *is* troubling, on the other hand, is when a researcher reports the results of an exploratory analysis as if it were the hypothesis all along. Such practices are problematic.

   These eight steps are designed to assist researchers in avoiding mistakes related to *confirmatory* data analysis. However, often times interesting patterns present themselves which a researcher may want to include. I advocate researchers be explicit about whether the results were discovered through confirmatory or exploratory practices. One might choose to blend the two into one paper, provided the researcher is again explicit about which results were confirmatory and which were exploratory. Returning to Figure 1, for example, the researcher may have decided before collecting data that the video would make participants more aggressive, yet did not anticipate the bimodality of scores in the treatment group. In the report, the researcher might say, "Upon graphing the results, a bimodal distribution was discovered among the participants in the treatment group. These results were unanticipated *a priori* so we decided to explore this relationship further. . . "

The Eight Steps to Data Analysis

In this section I will describe each of the eight steps. For simplicity, I have included a table (Table 1) that shows the eight steps and the function each step serves. For each of the steps that follow, I will address why I recommend performing said operation and what weaknesses it aims to overcome.

1. State the theoretical hypothesis and state a decision criteria

The first step to data analysis ought to be to state the theoretical hypothesis. Ideally, this would take place long before the researcher actually collects data (i.e., the hypothesis is "pre-registered"). Doing so in advance will assist in preventing researchers from "bending" their original hypothesis to fit the actual analysis performed (e.g., "Oh yeah. I, uh, meant to include that as a covariate originally. I just forgot"). This step is entirely voluntary and is a token of the researcher's inherent interest in advancing science (rather than pushing a publication through the pipeline). If the theoretical hypothesis is stated *a priori*, the probability of that one hypothesis being statistically significant will remain at nominal alpha levels (if the null is true).

To maximize the utility of these hypotheses, however, I recommend three additional strategies: (1) mapping hypotheses to specific statistical parameters, (2) stating strong hypotheses, and (3) developing decision criteria for clinical "significance."

Table 1
A Summary of the eight Steps of Data Analysis

Step	Purpose
1. State the theoretical hypothesis	Helps to minimize "fishing" for statistical significance
	Provides a translational map from theory to data
	Allows users to specify their own decision criteria
2. Assess psychometric properties of variables	Invites researchers to think about measurement
3. Plot univariate distributions	Helps identify outliers
	Helps identify issues with non-normality
	Assists in identifying coding errors
4. Plot a graphic to match the theoretical hypothesis	Directs focus toward the size of effects
	Helps identify potential problems with non-linearity/heteroskedasticity
	Improves cognitive encoding of results
5. Study residuals	Helps identify problems with normality (e.g., through histogram of residuals)
	Helps identify problems with non-linearity/homoskedasticity (e.g., through a residual dependence plot)
6. Interpret parameter estimates/effect sizes	Encourages the researcher to focus on estimation before significance
7. Determine statistical significance (if appropriate)	Assists in identifying whether visual patterns are "real"
8. Replicate on a new dataset	Encourages cumulative and reproducible science

Mapping hypotheses to parameters. Thee theoretically-derived hypotheses ought to be tied to specific parameters in a model. Too often researchers write well-crafted introduction sections, providing strong theoretical rationale for their chosen verbal hypothesis. Unfortunately, there seems to be a disconnect between the well-crafted introduction and the results section; the hypothesis points to a particular parameter (e.g., the interaction term in a model or the main effect of a predictor after controlling for another), yet the results report gobs of results and corresponding tests of significance. Not only does this dilute the message (because the parameter of interest is buried between other tests), but this constitutes fishing. Each reported p-value is the result of a tested hypothesis and, as such, each reported p-value ought to be clearly supported by strong theoretical rationale. If the introduction section only develops arguments for testing one parameter, then only one parameter ought to be reported.

Granted, some analyses require entering other parameters in the model. For example, if the researcher's hypothesis concerns an interaction effect between two independent variables, the main effects must be included in the model as well. However, these main effects need not be tested (or rather, reported since software packages tend to report significance for

all parameters) for significance because, again, the researcher's hypothesis is not concerned with these parameters.

Stating strong hypotheses. Years ago, Meehl (1967) criticized the use of zero as a tested hypothesis. That's a rather low bar to pass. Instead, he advocated for "strong" hypotheses, where researchers specify numeric values for the parameter. For example, rather than testing whether a correlation is different from zero, a researcher can test whether the correlation is different from +0.4. This amounts to reversing the role of the null and the alternative and can lead to some logistic problems (e.g., researchers might be inclined to collect small samples so they don't have power to reject their cherished hypotheses). With some modification, we might instead hypothesize the parameter of interest falls within a particular range (e.g., from r = 0.2 to 0.4). Better yet, researchers might abandon p-values and instead use the values of the parameters themselves to set their own decision criteria, which I will discuss next.

Developing decision criteria. As mentioned previously, one of the purported advantages of NHST is that it provides a bridge from theory to conclusion via a p-value computation (Cortina & Landis, 2011). Clearly, certain situations call for such judgments (e.g., does this finding have scientific merit? Should this treatment be used? Are side-effects of medication small enough to merit implementation?) and alternative frameworks have no clear route from theory to judgment.

Unfortunately, using a universal criteria ($p < 0.05$) has, in a way, "hijacked" decision making from the scientific community. A p-value is a function of both the sample size and the effect size. In certain domains (e.g., neuroscience) a large N is simply not feasible, and yet the culture of NHST does not permit flexibility in considering other decision criteria that allow for more lenient p-values. On the other hand, alternatives to NHST (e.g., effect sizes) have been criticized because they do not provide simple rules for deciding whether a finding has (or has not) scientific merit (Cortina & Landis, 2011).

I advocate, instead, for the decision criteria to be left in the hands of the researcher. Researchers may choose, if they wish, that a significant finding is one that reaches $p < 0.05$. Other researchers, on the other hand, may decide a clinically significant finding is one where $d > 0.83$, or a mean difference between treatment and control group is greater than 10 points, or expenditures are reduced by $10,000. In short, any metric may serve as the basis for making decisions; one is not limited to p-values.

One might criticize this approach by asking what is to stop researchers from setting even lower bars than $p < 0.05$. Researchers have a vested interest in a paper reaching "significance" (however it is defined) and if they can lower the threshold for reaching significance to even less than what is currently acceptable, they will abuse that. They might, for example, state in advance that any correlation greater than 0.0000001 is clinically significant.

Fortunately, the researcher will eventually have to defend their decision criteria when their paper is submitted for review. If they set a low bar for their decision criteria at pre-registration, they will have to answer to a skeptical community of reviewers at a later date. I suspect this knowledge will severely limit the degree to which researchers seek to abuse the practice of setting their own decision criteria. In addition, by placing these sorts of decisions back in the hands of the researcher and the scientific community, the significance of results will not have to be qualified as "statistically but not clinically significant."

2. Assess psychometric properties.

Many have suggested the "replication crisis" is a result of attempting to make conclusive answers on noisy data. The obvious solution to the noise is simply to increase the sample size. In some situations, however, this is not feasible, nor is it always the most practical approach. Gelman (2018) noted that doubling the reliability of a test will yield equivalent gains in precision as quadrupling one's sample size. In other words, we might get more "bang for your buck" by spending a bit more time with measurement.

By assessing the psychometric properties of our measures, it will invite deeper thinking on measurement issues and how they might affect data analysis. If our measure fail psychometrically, no amount of sophisticated modeling will yield any insights that have scientific merit.

3. Plot the univariate distributions

The second step of data analysis is to plot the univariate distributions of the variables of interest. For quantitative variables, boxplots and histograms are good candidates. (Quantile-quantile plots may also be beneficial, but they are less common and thus less interpretable). For categorical variables, bar charts or association plots are appropriate. A visual of the distribution will inform the researcher of several potential pitfalls in the coming analysis, such as:

- Incorrectly coded values (e.g., with a second wave of data collection, the researcher accidentally changes the treatment group designation from "Treatment" to "TRT", then later aggregates the two data waves and accidentally treats those labeled as "Treatment" and "TRT" as separate groups)
- Improperly coded missing values (e.g., a -999 is treated as a value, rather than a missing variable)
- Non-normality (e.g., if there are excessive zeroes in the sample)
- Outliers

Beginning with such plots may prevent embarrassing retractions later. For example, Hofmann, Fang, and Brager (2015) wrote an article that suggested oxytocin reduced psychiatric symptoms, but later had to retract the article. When entering the effect sizes for a meta analysis program, they assumed all effect sizes were positive. Because the program they used required specifying which effects were negative (and because they improperly assumed they were all positive), the aggregated effect size was inflated. This could be avoided by simply plotting the univariate distribution of effect sizes. (See more at Chawla, 2016).

At this point it may not be necessary to address the outliers and/or non-normality of the data. Remember the assumption of linear models (e.g., regression, ANOVA, t-tests) are that the residuals of the model are normally distributed. The outcome variable itself need not be normal (though it usually helps). Often times including one's predictors in a model will render the residuals normal, even if the variable itself was not normal. Likewise, an outlier in univariate space may not be an outlier in multivariate space. However, plotting the univariate distributions in advance will inform the researcher of potential problems that may occur later in the analysis.

4. Plot a graphic to match the theoretical hypothesis

Once the univariate distributions are plotted (and any coding errors are handled), the next step is to plot a graphic to match the theoretical hypothesis. If one were to perform simple linear regression, for example, a scatterplot would help the reader visualize the results. For ANOVAs/t-test, boxplots or mean plots would be most appropriate. Table 2 maps appropriate graphics for different types of analysis, along with an example in the last column. Although some of the strategies used (and types of plots) may be new, all of them are easy to produce in either R or SPSS.[5] For an article on determining which graphic is most appropriate and for instructions on creating each of these types of plots, see xxxxx.

[5]SPSS can perform most of these plots, but unfortunately (as far as I know) does not allow the user to overlay the raw datapoints over boxplots, mean plots, multiway dot plots, etc.

Table 2

Appropriate Graphics to use for Various Types of Statistical Analyses

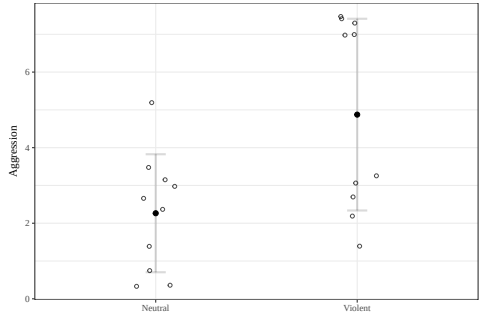
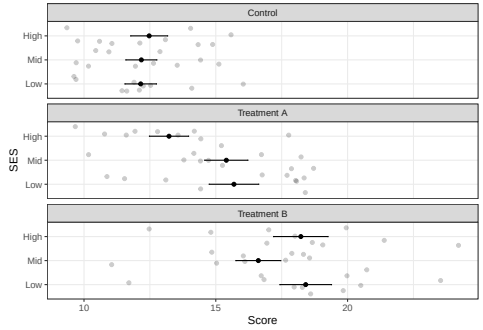
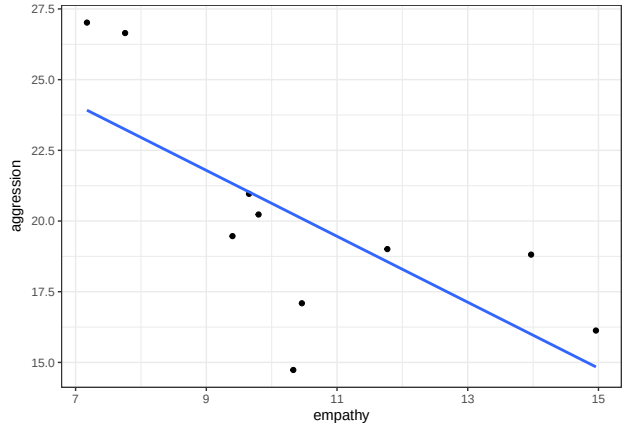
Analysis	Plot	Example
T-Tests/ANOVA	Boxplots/Mean Plots	
Factorial ANOVA	Multiway Dot Plot	

Table 2

Appropriate Graphics to use for Various Types of Statistical Analyses

Correlation/Regression	Scatterplot
ANCOVA	Coded Scatterplot



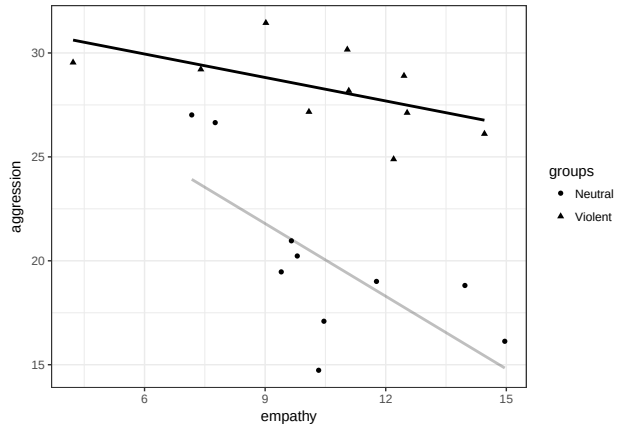
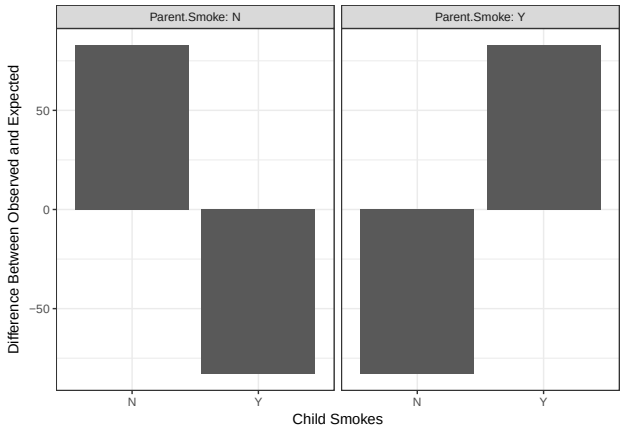


Table 2

Appropriate Graphics to use for Various Types of Statistical Analyses

χ^2	Association Plot
Multiple Regression	Coplots or Added Variable Plots



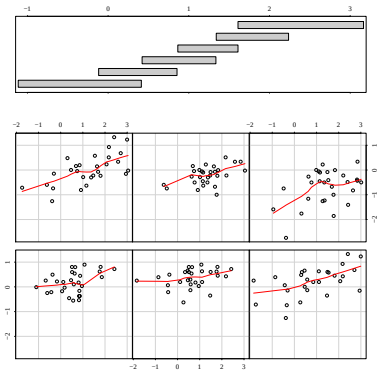
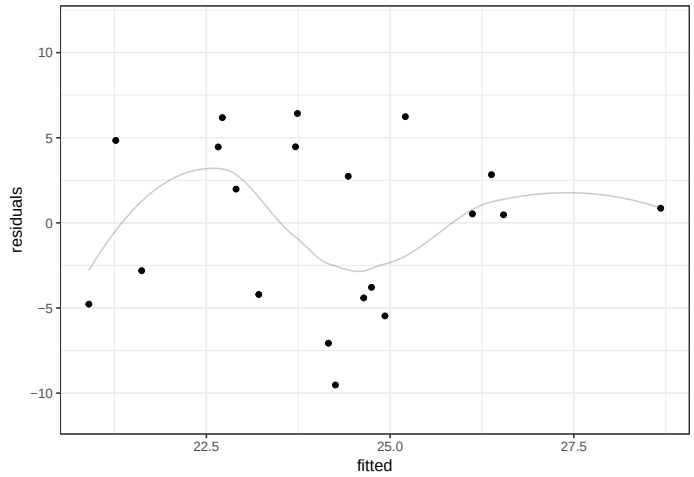


Table 2

Appropriate Graphics to use for Various Types of Statistical Analyses

Assessing Linearity/Homoskedasticity	Residual Dependence Plots



The advantage of graphing one's analysis is that it makes it nearly impossible to deceive one's readers (or one's self, for that matter, especially if raw data are displayed). Graphics are essential for identifying two problems in particular: (1) outliers, and (2) non-linearity.

If one or more outliers are present, it is possible they are driving statistical significance. If so, this will be easy to determine from a graphic. Likewise, if the researcher attempts to fit a straight line to data that are clearly non-linear, usually a graphic will show the error of one's ways. If this is the case, one need not hang their head in defeat. It simply means the researcher has chosen the wrong model and must select another (such as performing polynomial regression, analyzing a transformed dependent variable, using non-parametric procedures, or estimating generalized linear models).

To better detect departures from linearity, I recommend adding loess lines to a graphic (visit xxxx to read a tutorial on how to do this). Loess lines are non-parametric curves that are allowed to "bend" with the data. They can assist in detecting non-linearities that a standard model (which forces a straight line) would not detect.

Other options for graphics include coplots for multiple regression (added variable plots are also a nice alternative), coded scatterplots for ANCOVAs (where the categorical variable is displayed as a different symbol/color/plot/regression line), multiway dotplots for factorial ANOVAs, and barplots for χ^2 analyses.

Recall our original goal: we are trying not to commit a Type I (or Type II) error. At this point, it is impossible to make a false conclusion. We have not yet even computed an estimate, let alone made a decision about statistical significance. In addition, the researcher may decide that computing the p-value on a dataset is not necessary. If the visual analysis shows a whopping effect, who really cares about whether it is statistically significant? (Though the converse is also true. I once had a graduate student produce scatterplots and find the predicted relationship was non-existent. She immediately quit the analysis and concluded, "If there is an effect, it's so small I don't even care about it." I promptly gave her an A on her assignment).

In short, plotting the data before computing the analysis will prevent researchers from deceiving themselves, render tests of statistical significance less necessary, and force the reader to think in terms of the size of the effect, rather than its existence.

Unfortunately, not all statistical analyses lend themselves nicely to graphical display. Structural equation models, factor analysis models, and Hierarchical Linear Models, for example, are more difficult to map onto a single plot. These may require multiple plots and, unfortunately, the fragmented nature of these graphics may divorce the visuals slightly from the actual analysis. Future research ought to attempt to bridge that gap and find intuitive, graphical representations of these more complex models.

5. Study the residuals

After plotting the data, the researcher may not know if the chosen analysis (e.g., linear regression) is appropriate. The data may violate the assumption of linearity, normality of residuals, or homoscedasticity. Yet in order to properly diagnose the problem, we ought to compute the residuals (and thus have to actually perform the analysis). Unfortunately, before extracting the residuals, the model has to actually be fit to the data. I would advise the reader to close one's eyes (metaphorically or otherwise) until after the residuals have been studied before studying the results of a statistical test.

With residuals in hand, the researcher is now ready to properly assess the assumptions of the model. It is common (at least in the bio-medical literature) to perform statistical tests of normality or homoscedasticity, or independence. I would advise against it. Like other tests of significance, these tests of assumptions are sensitive to sample size. With a small enough N, even large departures from statistical assumptions are not detected, while with large N, even trivial differences are flagged. These statistical tests tell us whether our distributions depart from what is expected. They do not tell us whether they are different enough to muck up our analysis. The latter is better done through visual interpretation of results (as well as through an assessment of robustness).

I recommend a visual inspection of the residuals. Histograms will inform the researcher whether the normality assumption has been approximately met. Residual dependence plots (see Table 2) assist in determining whether linearity and homoscedasticity have been met. Either of these plots will assist in flagging outliers. For instructions in how to diagnose problems using these plots, see Kutner, Nachtsheim, Neter, and Li (2004).

If the visual inspection of the residuals signals problems, one may have to iterate through steps 2-4 until the assumptions have been met, each time making a modification to the model (such as transforming the DV, removing outliers, utilizing weighted least squares, or using generalized linear models). Furthermore, problems at this early stage may suggest the researcher is not yet ready for confirmatory data analysis. Again, there is nothing wrong with exploratory data analysis and if a researcher finds the parametric version of a model is not appropriate, it may be best for the researcher to explicitly state their analysis has turned from confirmatory to exploratory. Regardless, the researcher may proceed to step 5 once the assumptions have been met.

6. Interpret Effect Sizes/Parameter Estimates

At this point the researcher has far more information about the data than what is typically reported in psychological journals; the researcher knows that outliers are not driving the analysis, the model chosen is appropriate, and has a visual that illustrates the strength of the relationship between the variables of interest. After step five, the researcher ought to be confident the model chosen is appropriate (i.e., the assumptions of the model have been met). Once again, I emphasize that it is impossible to commit a Type I (or Type II) error because statistical significance has not yet been evaluated. Rather, I recommend the researcher study and interpret effect sizes and parameter estimates. We all have been cautioned against making mountains out of molehills, or emphasizing statistical significance at the expense of practical significance. This is why the APA recommended researchers report effect sizes in addition to statistical significance. I would argue practical significance is far more important than statistical significance. Studying the effect size (and parameter estimates) before statistical significance is a conscious choice aimed at reminding the researcher of this preference for estimation rather than significance.

Most statistical packages offer readily available estimates of effect sizes, including f^2, part and partial correlations, r^2, and Cohen's d. To determine which measure of effect is appropriate, I recommend the concise and effective article by Cohen (1992).

In addition, the researcher ought to study and interpret the parameter estimates (which, themselves are non-standard effect sizes). In a regression, the parameters of interest are the slopes (and occasionally the intercept). For ANOVAs/t-test, the parameters of

interest are the mean differences between groups. For structural equation modeling, the parameters of interest are the path coefficients. For logistic regression and other generalized linear models, the researchers may have to perform mental gymnastics as they attempt to interpret things in terms of log odds (or in terms of odds ratios).

Studying these parameters adds another layer of depth at which the researcher can make sense of one's data. Not only will it inform the researcher about the direction of the effect (e.g., males scored higher in aggression than females, anxiety is positively predictive of depression, performance is inversely related to mood), but also offers a mathematical equation that maps predictors onto outcomes. For example, suppose a researcher performs a regression that predicts weight loss from experimental condition, controlling for motivation. Further suppose the regression equation is as follows:

$$weight\ change = 1.2 - 0.8 \times motivation - 4.5 \times treatment - 1.2 \times motivation \times treatment$$

This regression equation would be an interesting result indeed. This suggests the following:

- Those in the control group who have no motivation will actually gain an average of 1.2 pounds (because control group is the reference group)
- Every time an individual increases their motivation by a point, they can expect to lose 0.8 pounds
- The treatment group averages 4.5 pounds more weight loss than the control group
- The relationship between motivation and weight loss is stronger for the treatment group than for the control group, such that for the treatment group, for every point increase in motivation, they lose 2 pounds (i.e., 1.2+0.8)

Granted, much of this information could be gleaned from a graphic, but the estimates put the visual interpretation into concrete mathematical terms that are interesting of their own right. Furthermore, the effect sizes and parameter estimates reduce ambiguity inherent in visual interpretation. To further reduce ambiguity (and marry the ideas of significance testing with estimation), a researcher may decide to pair confidence intervals with these estimates. Doing so will further reduce ambiguity (especially if the confidence interval does not contain zero), while explicitly recognizing the degree of uncertainty.

7. Make a decision (if applicable)

Recall we have plotted univariate distributions to flag potential data recording errors, assess normality, and identify potential outliers. We have also created a visual representation of our analysis that shows both the size of the effects and the direction. We have also thoroughly assessed whether our model is appropriate through residual analysis, estimated effect sizes, and interpreted parameter estimates. In short, we have much more thoroughly familiarized ourselves with our own data (Tukey, 1969).

If, at this stage, the reader feels it rather pointless to assess statistical significance, I have successfully made my point. This second to last step is entirely optional and ideally makes it clear that our data have long been trying to tell us much more than we have allowed them. Simply computing statistical significance without doing the previous steps is akin to

eating a single sprinkle off a large birthday cake. With so much richness remaining, it is a shame that we limit ourselves to a single test that is largely uninformative.

Earlier, I advocated that researchers instead set their own decision criteria. At this point, making a decision of significance is easy; one has already pre-specified what clinically significant is and now they simply compute the numbers and identify whether significance was reached.

8. Repeat with new data

The decision made in the previous step is always provisional. Few single studies have the power (statistically or otherwise) to make conclusive statements about the truthfulness of a hypothesis. Rather, these findings are tentative and ought to invite closer scrutiny and replication. Better yet, the estimates obtained for the parameter of interest ought to serve as a prior in a bayesian analysis, or ought to be aggregated into the next study (and others like it) via a meta-analysis. Such cumulative methods will invite a greater sense of humility about one's own role on the scientific process and consequently invite deeper attention to development of theory.

Reporting results

I recommend every researcher perform the eight steps when doing confirmatory data analysis. However, it may not be necessary to report every step in a journal article. Not only would this increase the length of most articles (a trivial problem as journals become more digitized), but it may detract from the purpose of the article (to test the original hypothesis). However, at minimum, I strongly recommend a researcher's final report contain:

1. One or more graphical depictions of the analysis of interest (Step 3, see Table 2).
2. A comment on how the researcher determined the appropriateness of statistical assumptions.
3. Parameter and effect size estimates, with confidence intervals.
4. A supplemental section containing all graphics and sensitivity analyses *or* a link to a website where these can be viewed.

In other words, I am not advocating for a complete ravamping and replacing of how statistics are reported in journal articles. Rather, I suggest we add these few pieces of information so that the richness of our data is more visible.

Example

In the section that follows, I decided *not* to re-analyze existing datasets of previously published papers for two reasons. First, it is difficult to find studies where researchers have actually uploaded their data for public scrutiny. Second, I would hate to pick on researchers conscientious enough to actually offer their dataset by highlighting their data analysis mistakes. I want to encourage openness, and becoming a public data vigilante would be counterproductive. Consequently, I will analyze a publicly available dataset, the 2014 National Survey of Drug Use and Health (Health & Services, 2014) and offer my own hypothesis that, when analyzed using traditional NHST methods, yields misleading results.

Table 3
ANOVA Summary Table for the Example Hypothesis, Tested Using Standard NHST Practices \label{tab:anova}

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
health.rating	4	1,141	285	10	0
MI	1	1,531	1,531	52	0
Residuals	211.0	6,245.1	29.6	NA	NA

1. State the theoretical hypothesis of interest and set a decision criteria.

Suppose I am a drug counselor that has had only marginal success in assisting heroin addicts overcome their addictions; those who use heroin experience more psychological distress, which in turn motivates them to escape their distress via heroin. Now let us suppose I believe promoting healthy behaviors (e.g., exercise, nutritious eating) will reduce psychological distress and help break that negative feedback loop. Ideally, one would perform an experiment, but perhaps as a preliminary study, I decide to use an existing dataset to perform an observational analysis to assess the potential efficacy of promoting healthy behaviors in a full experiment.

However, I may consider controlling for mental illness. One may be experiencing psychological distress because of their mental illness, which may make them more likely to escape such distress through drug use. Stated differently:

Among heroin users, those who report having more healthy behaviors will report less psychological distress, after controlling for mental illness

I might also predict, based on my review of the literature, an effect size between $d = 0.21$ and d=0.32 (i.e., the standardized difference between excellent health and poor health) in psychological distress. Furthermore, I might also decide that if d is less than 0.15, it is not clinically significant.

For illustrative purposes, I will test this hypothesis using standard NHST methodology using an ANCOVA model. Based on these data, I could conclude:

> Self-reported health rating was significantly associated with psychological distress $F(4, 211) = 9.64$, $MSE = 29.60$, $p < .001$, $\hat{\eta}^2_G = .155$.

(See also Table **??**). As I will show, the above statement is both misleading as well as incomplete.

Also, I make explicit (in the spirit of the third guiding principal of data analysis) that my analysis is confirmatory.[6]

2. Psychometrics

[to be completed. . .]

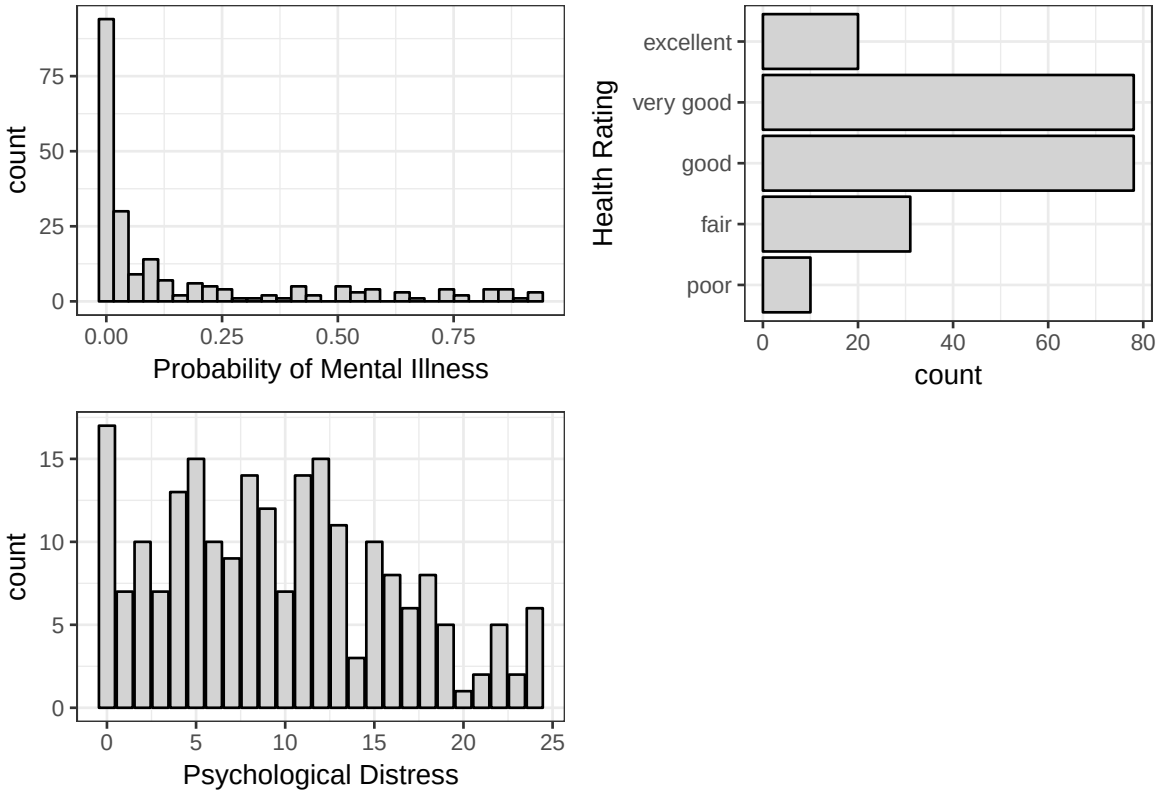


Figure 2. Univariate distributions of probability of mental illness, health rating, and psychological distress.

3. Plot Univariate Distributions

I began by plotting the univariate distributions of the three variables of interest: Probability of Mental Illness (MI), Health Rating, and Psychological Distress. These distributions are shown in Figure 2.

The plots reveal several potential issues with the data:

1. The dependent variable (psychological distress) is far from normally distributed. The mode of the distribution is at zero (i.e., the data are zero-inflated). This could certainly be problematic for linear models (such as an ANCOVA)
2. The mental illness (MI) variable is severely skewed. Note that linear models make no assumptions of normality for the independent variables (or the dependent variables for that matter, rather the assumption is about the *residuals* of the dependent variable). However, in my experience, if both the IV and the DV are skewed, the assumption of linearity will almost certainly be violated.

At this point I am primed to look for serious issues with normality, linearity, and likely heteroskedasticity. As I mentioned earlier, if these assumptions are violated, it simply means we have chosen the wrong model to fit the data.

[6] In full disclosure, I did a little "fishing" of my own in order to find a hypothesis that was particularly illuminating.

4. Plot a Graphic to Match the Analysis of Interest

Recall that the fictitious researcher has decided to perform an ANCOVA. An ANCOVA essentially performs a standard linear regression between the covariate and the DV, extracts the residuals, then performs an ANOVA on the residuals (though this is all done simultaneously with an ANCOVA). In Table 2, I mentioned a "coded scatterplot" would be an appropriate graphic to match the ANCOVA, where each category of a factor is displayed using a different symbol, plot, line, and/or color. For these data, I plotted each level in a separate panel (from lowest value in the top-left to highest value in the bottom-right) in Figure 3, with loess lines overlaying the data. Once again, there are a few things worth noticing:

1. These loess lines are not "parallel," meaning that the assumption of homogeneity of regression has been violated (indicating that ANCOVA is not appropriate).
2. These curves are not linear, indicating that linear models will not be appropriate.
3. It is not easy to determine whether one's health reduces stress; those who report "very good" health seem to have less distress (on average) than those who report "fair" health. However, the relationships for the other health categories are less clear.
4. Relatively few heroin users report either poor or excellent health.

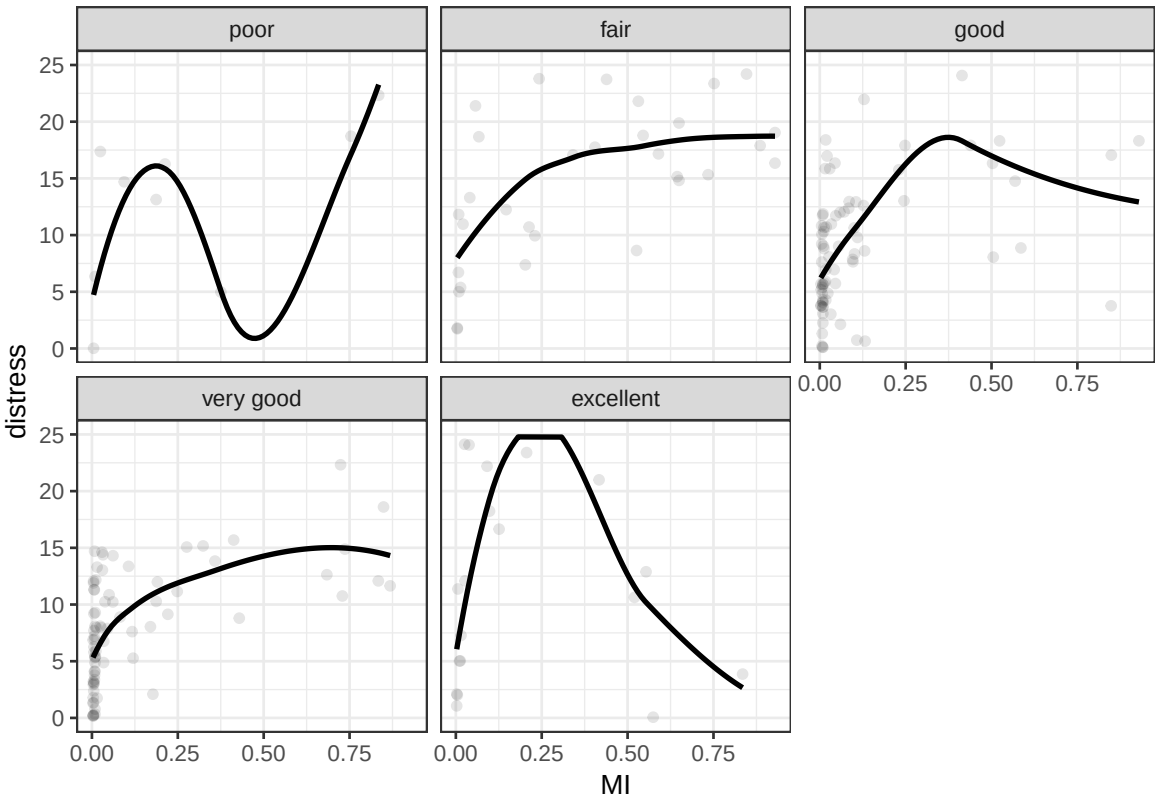


Figure 3. An ANCOVA-based graphic of the NSDUH dataset. The lines are loess lines mapping the relationship between mental illness and distress, conditional on self-reported health rating.

5. Study the residuals

Figure 3 already revealed a problem with fitting linear models to the data. Yet, for the sake of completeness, I plotted a histogram of the residuals, as well as a residual dependence plot. These are shown in Figure 4. Not surprisingly, the residual dependence plot shows some curvilinear effects the linear model missed.

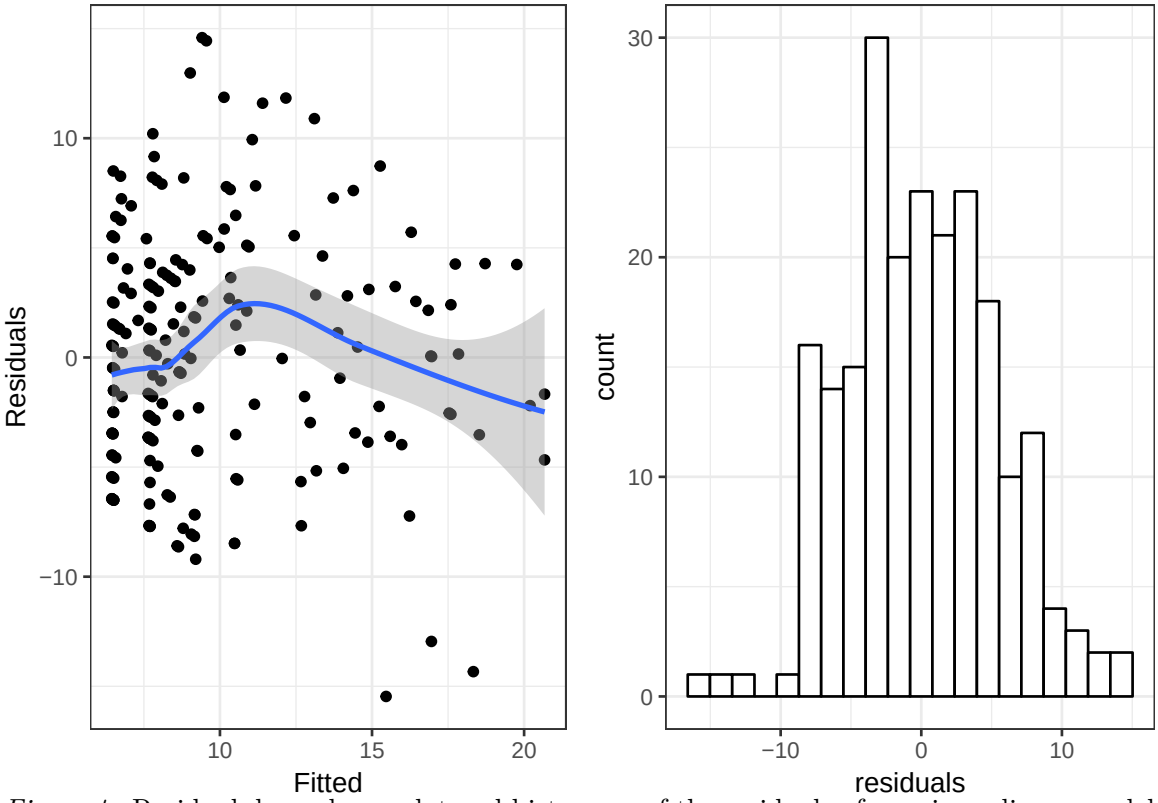


Figure 4. Residual dependence plot and histogram of the residuals of running a linear model on the NSDUH dataset.

At this point, it is clear a linear model is not appropriate for the data. I have a few options:

1. I can transform the dependent variable and attempt to "linearize" the relationships. This generally fails when the mode of the DV is the minimum (as it is in this case). I confirmed "off camera" that this is indeed the case.
2. I can attempt to use non-parametric procedures, such as rank transformations of the dependent variable. Unfortunately, rank transformations fail to preserve interaction effects (which are clearly happening, as seen in Figure 3). As an alternative, I could decide to limit my analysis to only those with a low probability of a mental illness (e.g., less than 10%), then do *bivariate* non-parametric procedures (such as a Kruskal Wallace Test). Since the original hypothesis is concerned only with *controlling* mental illness and not the variable itself, I think this strategy is promising.
3. I can perform more "modern" robust methods (Erceg-Hurn & Mirosevich, 2008). These methods essentially replace mean-based estimates (including conditional means) with

trimmed means, standard variances with winsorized variances, and standard confidence intervals (CIs) with bootstrapped CIs that are computed from the trimmed/winsorized estimates. Unfortunately, these methods would not work since more than 10% (the "traditional" degree of trimming from either tail) of the tail of the distribution is contained at the approximate mode of the distribution. (In general, modern robust methods do not work well for zero-inflated data).

4. I can use generalized linear models. These are parametric procedures that relax the assumptions of normal-based models by fitting non-normal distributions (e.g., Poisson, Gamma, negative-binomial). Unfortunately, they tend to be more difficult to interpret (because coefficients are often expressed in logs, log odds, inverses, etc.).

I favor the second and fourth strategies. Once again, I prefer a visual approach to modeling these data, so for both options two and three, I will present graphics that show how the models compare in terms of their predictions.

Before I do so, however, I will aggregate the poor and fair health categories, as well as the very good and excellent. Otherwise, the poor and excellent categories may be unduly influenced by outlying datapoints.[7]

The results of these two models are shown in Figure 5. Recall that for the non-parametric model, I only studied those with a low probability (<10%) of MI. The top model shows the results for a generalized linear model (in this case, an ordinal logistic regression model), while the bottom model shows the results for the non-parametric model (including interquartile ranges). In both cases, the predictions tend to occur in the densest parts of the data (though in the ordered logistics, the prediction for those in very good health seems to underestimate distress for those low in MI). The models also yield two slightly different pictures of how health relates to mental illness: the polytomous model suggests good health staves psychological distress for those with mental illnesses, but only up to a point. For those who have a 20% or more probability of mental illness, good health seems to have little mitigating effect. On the other hand, the non-parametric model suggests there is very little difference in psychological distress between the different health categories (at least for those with a low probability of having a mental illness).

I will forego choosing a model until after we see how they differ in terms of estimates, effect sizes, and p-values.

6. Study Parameter Estimates/Effect sizes

Studying the estimates for the non-parametric model is simple. The medians of the three health conditions are 7, 6.5, and 6 for fair, good, and very good health, respectively. In other words, moving from a poorer health category to a better one will decrease one's distress by half a point. The bootstrapped 95% confidence interval for the difference between very good and fair was -7.5 to 2.

For the nonparametric model, on the other hand, the parameters are much more difficult to interpret because we must interpret coefficients on a logit scale (or convert them

[7] I did perform a sensitivity analysis on this decision as well (i.e., performing the analysis both before and after aggregating those categories). Combining the categories makes the picture more clear. Whether that clarity is spurious, I leave it to the reader to decide.

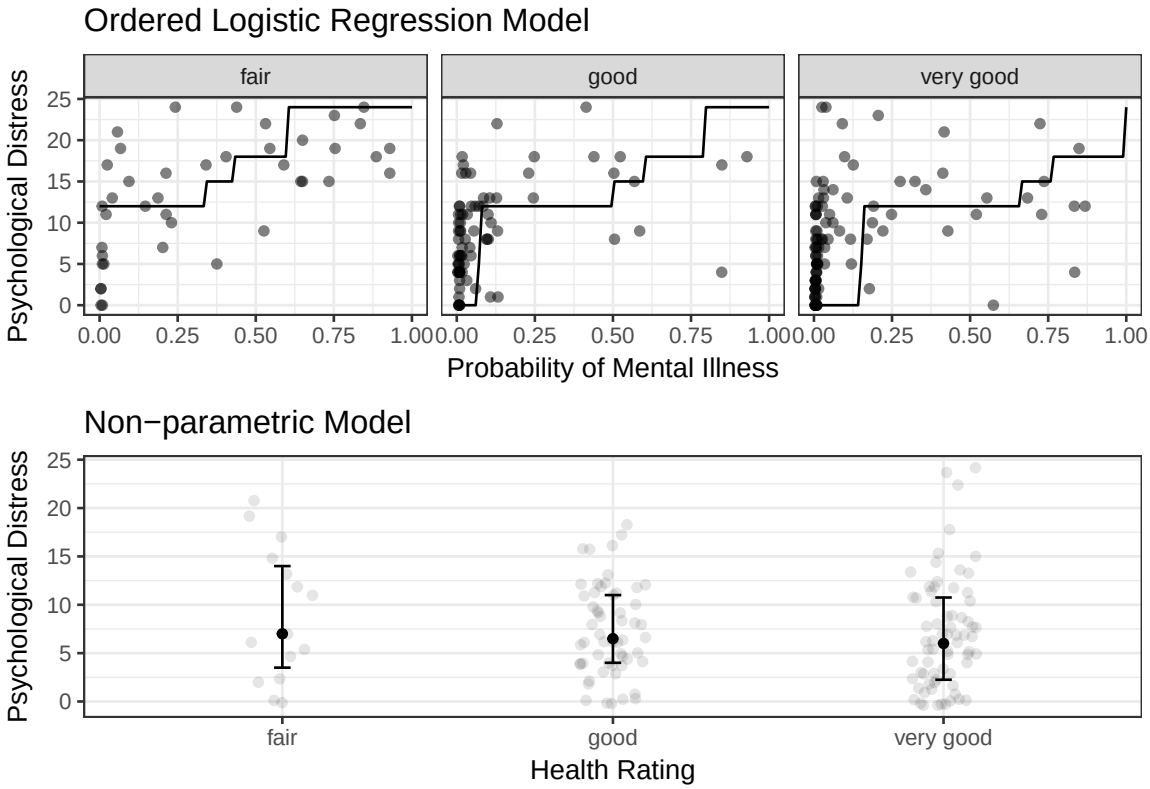


Figure 5. Predictions for the non-linear models of the NSDUH data. The top model shows predictions (as lines) and observed (as dots) for the polytomous logistic regression model, while the bottom model shows the predicted/observed values for the non-parametric model (as well as interquartile ranges).

to odds ratios). I find it much more beneficial to actually visualize the predicted values rather than interpret odds ratios (especially if interactions are involved).

As an alternative, it may be beneficial to generate predictions for a few critical points of interest. For example, if an individual has a 50% chance of MI, how much does the model suggest they would differ in psychological distress? For those of good health, their predicted distress is 18, while the predicted distress is 12 for those of fair health. For those with a 25% chance of MI, the values are both 12.

```
## Warning: package 'coin' was built under R version 3.4.3
```

To compute the effect size for the non-parametric model, we can convert the test statistic from a Wilcoxon Test into a more familiar metric of the correlation coefficient, but must limit the comparison to two groups. In this case, I wanted to determine the *maximal* effect size by comparing the difference between the "fair" and the "very good" group. That effect size, in the familiar correlation coefficient metric is -0.07, with a bootstrapped 95% confidence interval between -0.29 and 0.15, which is quite small (and consistent with what the visuals present).

For the polytomous model, it is customary to report odds ratios as estimates of effect sizes. As I mentioned previously, odds ratios are difficult to interpret, especially when

interactions are present in the model and the outcome is polytomous. I tend to favor effect sizes with understandable and familiar metrics. One metric we might use is classification accuracy. In other words, how accurate is the polytomous regression model at classifying people's distress levels, based on the two covariates? Or, more importantly, how much does health rating improve the accuracy of classification, above and beyond mental illness?

Unfortunately, any estimate of effect size based on classification accuracy in this case is fighting an uphill battle. Given that there are 25 categories one could be classified into, there is a lot more room for error than if there were, for example, only two categories one could be classified into. Because of this, I decided a classification was "correct" if it predicted a distress level within 5 points of their actual distress level. Yes, I may be inflating the accuracy of the model and my choice of within five points is admittedly arbitrary, but the purpose of the effect size is not to determine how well my model fits the data (I have already done that visually). Rather, my purpose is to determine the merits of *health* in prediction accuracy (relative to mental illness).

The within five point accuracy of the model with both health and MI is 0.46. When we remove health from the model, accuracy drops to 0.41, a drop of 0.05. When we drop MI from the model, on the other hand, prediction accuracy drops by 0.14.

In short, comparing the highest vs. the lowest in health, our correlation between health and distress is -0.07 and the improvement in classification that is attributable to health is 0.05. Regardless of the metric we use to quantify the size of the effect, the size of that effect is quite small.

7. Determine clinical significance based on a decision criteria

At the first step, I decided that health was a clinically significant predictor of psychological distress if the standardized difference between the highest and lowest health rating exceeded $d = 0.15$. Unfortunately, that metric is no longer relevant since we have moved to non-parametric/generalized linear models. Furthermore, our results unexpectedly changed from confirmatory to exploratory data analysis. Given these problems, it does not make sense to apply our decision criteria at this stage of the analysis. We could, however, use these results to inform future confirmatory analyses and corresponding decision criteria.

8. Repeat with a new dataset

Fortunately, the NSDUH routinely reports results of their survey of drug and alcohol use every few years. Unfortunately, space constraints do not allow me to revisit this hypothesis (and, frankly, few researchers would likely be interested in pursuing this further, given the rather small effect sizes).

Results Section

Earlier I stated that, at a minimum, a researcher should report (1) A graphical depiction of the analysis, (2) A comment on the appropriateness of statistical assumptions, (3) parameter/effect size estimates with confidence intervals, and (4) a supplemental section or link where the reader can view all graphics/sensitivity analyses. For this example, the results section may read as follows:

Upon visual inspection of the residuals, it was determined that linear models were not appropriate. This discovery forced a change from confirmatory to exploratory analysis. Consequently, we performed a Wilcoxon Rank Sum analysis, but limited our analysis to only those with less than a 10% probability of Mental Illness*. We also aggregated the poor/fair, as well as the good/very good groups because data were quite sparse at the extremes. A visual depiction of the statistical test is shown in Figure 6. The largest difference between groups was between those who reported good/very good vs. poor/fair groups, which had a one distress point (95% bootstrapped confidence interval: -7.5, 2), with an effect size of -0.07 ($-0.29 \leq r \leq 0.15, p = 0.33$). More details (such as graphics of distributions and sensitivity analyses) can be viewed at http://www.quantpsych.net/8-steps

* We also analyzed the data using a polytomous logistic regression model where Mental Illness was explicitly modeled. The substantive results were very similar between the two models. To see the results of this analysis, visit http://www.quantpsych.net/8-steps.

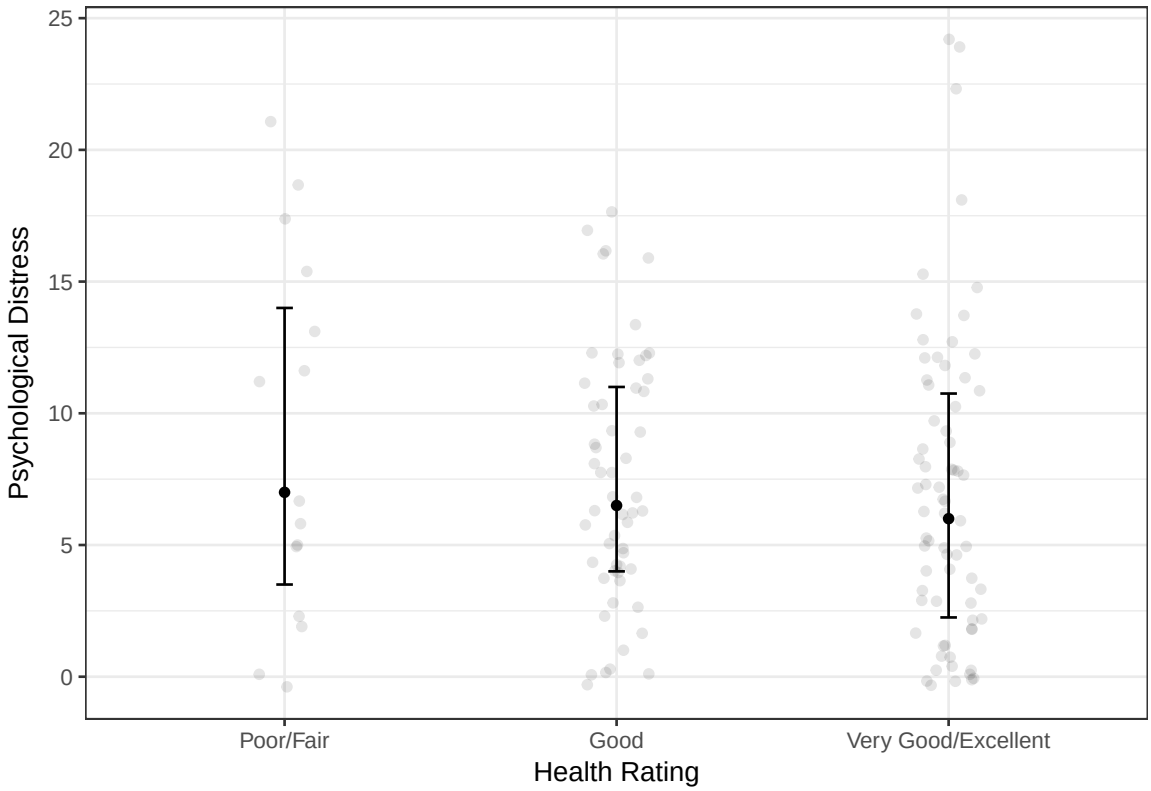


Figure 6. Graphical depiction of the final model that investigates the relationship between health and distress, for those with mental illness probability lower than 10%. Black dots represent medians and bars represent interquartile ranges, while the light gray dots represent the raw datapoints.

Summary

My first naive NHST analysis revealed a significant effect of health on psychological distress among heroin users, with a respectable effect size. However, by following the eight steps of data analysis we have learned many things we otherwise would have missed, such as:

- Distress is highly skewed
- The relationship between distress and mental illness is highly curvilinear; as the probability of mental illness increases, one's distress increases rapidly, then levels off with higher levels of MI.
- The relationship between distress and mental illness may change slightly as a function of one's health rating
- Once we appropriately handle the non-linearity, there *may* be a very small effect of health, but there is too much noise to determine whether the observed effect is any different from chance.

It is important to note that the issues noted above led to *substantially different conclusions* about the data and a more thorough understanding of what the data are telling us.

Throughout the analysis, I compared two different models and mentioned I would forego choosing between the two. Likewise, it does not seem to matter which model we choose; the substantive conclusions are essentially the same. However, given that the non-parametric model is simpler, I think many people would prefer that model.

Discussion

The recent "replication crisis" suggests there are statistical practices within the field of psychology that inflate the probability of Type I (or Type II) errors. These practices include "p-hacking," failing to meet statistical assumptions, and a narrow focus on statistical significance rather than interpreting what the data are actually telling us. In this paper, I have suggested a framework under which researchers might perform data analysis that easily fits within current data analysis practices, while inviting a greater focus on estimation and data visualization. In addition, this framework provides step-by-step guidance for researchers that aims to empower analysts to focus on what the data are actually saying.

I have also highlighted these principals and practices with the use of an actual dataset. My analysis revealed that performing the NHST ritual, even when effect sizes were reported, yielded a Type I error. My re-investigation of the same hypothesis revealed patterns more nuanced than a single NHST p-value (or effect size) captured. In the end, regardless of which of the two models I investigated, the conclusions were essentially the same: healthy behaviors seemed to have little (if any) mitigating effects on psychological distress among heroin users.

It is my hope the example I provided was illustrative. I suspect most researchers do not have the background (or interest) to utilize generalized linear models. Allow me to offer some hope. First, I intentionally chose a dataset I knew would severely violate the assumption of linearity. I would hope most researchers would not encounter such "zero-inflated" models where non-linear relationships are almost inevitable. On other hand, some constructs psychologists investigate (e.g., frequency of rape occurrences, number of times

one has attempted suicide) should not be investigated with standard linear models. Those researchers who study these types of data perhaps ought to dust off (or purchase) their generalized linear model books to best interpret what their data are saying. However, it may be optimistic to think researchers will learn these complex modeling techniques themselves. In these situations, it might be best to collaborate with those who are familiar with these techniques.

Secondly, the median-based, non-parametric model was far less complicated and offered quite comparable results. Though it would be preferable if researchers performed sensitivity analyses by comparing the predictions of two models, a non-parametric model will almost always be better than a standard linear model when data behave as they did in this example. (Of course, there is no rule against using the standard linear model for comparison in a sensitivity analysis. Such a comparison will inform the researcher how violating the assumptions affect the substantive conclusions reached).

Finally, despite my best efforts to emphasize this framework easily fits within current statistical practices, I suspect there may be some resistence. For example, editors everywhere may lament that performing these eight steps will double the length of the average article. I agree. However, I do not think it necessary every plot and sensitivity analysis make it to the final version of the paper. Researchers already frequently omit details about data cleanup and how models were decided. However, I do strongly suggest this information be publicly available, either through supplemental material or through an author's website. Doing so will allow future consumers of the research to understand what decisions were made, why they were made, and how these decisions may (or may not) have affected the analysis. In addition, it provides additional tools to consumers that allows them to judge the verisimilitude of the research themselves.

Another obstacle to incorporating these suggestions may be a lack of knowledge. Many researchers may not know how to graph loess lines, jitter categorical variables, or create multiway dotplots. Because of this, I have created a step-by-step tutorial that shows researchers how to visually represent the most common analyses, including regression, multiple regression, factorial ANOVAs, and t-tests. This tutorial demonstrates how to perform these in both SPSS as well as R and can be found at xxxxxx.

In conclusion, the discipline of psychology is at a crossroads. We can continue to participate in NHST-based psychology and the problems we have recently encountered will persist. Or we can revolutionize the way we think about analysis, listen to the messages the data are trying to tell us, and uncover truths previously buried behind ANOVA summary tables and p-values.

References

Chawla, D. S. (2016, October). Oh, well - "love hormone" doesn't reduce psychiatric symptoms, say researchers in request to retract. *Retraction Watch*. Retrieved from http://retractionwatch.com/2016/10/04/oh-well-love-hormone-doesnt-reduce-psychiatric-symptoms-says-researchers-in-request-to-retract/

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, J. (1994). The earth is round (p-less-than. 05). *American Psychologist*, *49*(12), 997–1003.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. *Handbook of Industrial and Organizational Psychology*, *223*, 336.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966

Cumming, G., Fidler, F., & Thomason, N. (2001). The Statistical Re-Education of Psychology ®. In.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *The American Psychologist*, *63*(7), 591–601. doi:10.1037/0003-066X.63.7.591

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2016). *What If There Were No Significance Tests?: Classic Edition*. Routledge.

Health, U. S. D. of, & Services, H. (2014). *Substance abuse and mental health services administration. center for behavioral health statistics and quality. national survey on drug use and health, 2014*. Ann Arbor, MI: Inter-University Consortium for Political; Social Research [distributor].

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are Assumptions of Well-Known Statistical Techniques Checked, and Why (Not)? *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00137

Hofmann, S. G., Fang, A., & Brager, D. N. (2015). Effect of intranasal oxytocin administration on psychiatric symptoms: A meta-analysis of placebo-controlled studies. *Psychiatry Research*, *228*(3), 708.

Jones, L. V. (1952). Test of hypotheses: One-sided vs. two-sided alternatives. *Psychological Bulletin*, *49*(1), 43.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied linear statistical models. In *Applied linear statistical models*.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (Vol. 1). Psychology Press.

Micceri, T. (1989). The Unicorn, The Normal Curve, And Other Improbable Creatures. *Psychological Bulletin*, *105*(1), 156–166.

Osborne, J. W. (2013). Is data cleaning and the testing of assumptions relevant in the 21st century? *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00370

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on*

Psychological Science: A Journal of the Association for Psychological Science, *7*(6), 528–530. doi:10.1177/1745691612465253

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *The American Psychologist*, *65*(1), 1–12. doi:10.1037/a0018326

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Schmidt, F. L. (1996). *Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers.* American Psychological Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Trafimow, D. (2017). Using the Coefficient of Confidence to Make the Philosophical Switch From A Posteriori to A Priori Inferential Statistics. *Educational and Psychological Measurement*, *77*(5), 831–854. doi:10.1177/0013164416667977

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*(2), 83.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life After NHST: How to Describe Your Data Without "p-ing" Everywhere. *Basic and Applied Social Psychology*, *37*(5), 260–273. doi:10.1080/01973533.2015.1060240

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Maas, H. L. J. van der, & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. doi:10.1177/1745691612463078

Wilkinson, L. (1999). Statistical Methods in Psychology Journals. *American Psychologist*, *54*(8), 594–604. doi:http://dx.doi.org/10.1037/0003-066X.54.8.594