International Journal of Nursing Studies 98 (2019) 87–93

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

Pervasive errors in hypothesis testing: Toward better statistical practice in nursing research

Vincent S. Staggs[a,b,*]

[a] Biostatistics & Epidemiology Core, Health Services & Outcomes Research, Children's Mercy Kansas City, 2401 Gillham Rd., Kansas City, MO, USA
[b] School of Medicine, University of Missouri-Kansas City, 2411 Holmes St., Kansas City, MO, USA

ARTICLE INFO

Article history:
Received 8 January 2019
Received in revised form 26 June 2019
Accepted 28 June 2019

Keywords:
Statistics
Statistical methods
Research methods
Nursing research

ABSTRACT

Background: In recent years several authors have documented common problems in the use of statistics in nursing research, including failure to consider the effects of multiple testing, inattention to clinical significance, and under-reporting of effect sizes and confidence intervals. More subtle forms of multiple testing are not as widely recognized, and abuse of researcher degrees of freedom has received little attention in the nursing research literature. These and other unsound practices in applying and interpreting statistics are problematic in themselves, and they arguably reflect an insufficiently clear understanding of statistical inference as a method for dealing with randomness among many researchers.

Objectives: The goal of this educational paper is to improve the understanding and practice of inferential statistics among nursing researchers. An accessible explanation of hypothesis testing is provided, including discussion of the crucial concept of repeated sampling. Several pervasive mistakes and misconceptions in statistical inference are examined in detail, including misinterpretation of "non-significant" *p*-values as evidence for the null hypothesis, failure to account for forms of multiple testing that arise in model selection, abuse of researcher degrees of freedom, and hypothesis testing for baseline differences between arms in randomized trials. Recommendations for better statistical practice are offered.

Conclusion: For the foreseeable future classical methods of statistical inference based on the idea of repeated sampling will be the primary tools for quantifying randomness in nursing research. The hypothesis testing framework, despite its limitations, can be helpful in ruling out chance as an explanation for observed effects. Nursing researchers who use quantitative methods, as well as journal reviewers and editors, should understand this framework well. Those involved in educating nursing researchers and those who teach statistics would do well to ask what changes need to be made to raise the level of statistical practice in nursing research.

What is already known about the topic?

- Errors in use and reporting of statistics are widespread in scientific journals
- Substandard statistical practices in nursing research are well-documented

What this paper adds

- Pervasive errors in hypothesis testing suggest that statistical inference is not universally well-understood as a means of quantifying randomness
- When subtle forms of multiple testing go unrecognized, observed *p*-values can be misleading
- Abuse of researcher degrees of freedom is a serious, under-appreciated problem
- Accessible explanations of these issues are provided

1. Background

In recent years several authors have raised concerns about the use of statistics in nursing research, including the problems of

* Correspondence to: Biostatistics & Epidemiology Core, Health Services & Outcomes Research, Children's Mercy Kansas City, 2401 Gillham Rd., Kansas City, MO, USA.
E-mail address: vstaggs@cmh.edu (V.S. Staggs).

https://doi.org/10.1016/j.ijnurstu.2019.06.012
0020-7489/

multiple testing, inattention to clinical significance, and under-reporting of effect sizes and confidence intervals (Floyd, 2017; Gaskin and Happell, 2013,2014; Polit, 2017). Statistical errors are by no means unique to nursing research. Citing the findings of dozens of studies, many from papers in leading biomedical journals, Lang and Altman (2016) write, "The truth is that the problem of poor statistical reporting is long-standing, widespread, potentially serious, concerns mostly basic statistics, and yet is largely unsuspected by most readers of the biomedical literature."

Not only are unsound practices in applying and interpreting statistics problematic in themselves, they often reflect an insufficiently clear understanding of statistical inference as a method for dealing with randomness. Journal reviewers and editors may catch obvious abuses of multiple testing, offer corrective guidance to authors who have relied on "statistical significance" without considering effect sizes and practical significance to identify meaningful findings, and request confidence intervals when they are not initially reported, though apparently some need to do so with more consistency. But more subtle forms of multiple testing and abuses of researcher degrees of freedom are not as widely recognized, and a firm conceptual grasp of statistical inference seems to elude researchers in a variety of scientific disciplines.

The goal of this educational paper is to improve the understanding and practice of inferential statistics among nursing researchers. The paper includes an accessible explanation of hypothesis testing, including discussion of the crucial concept of repeated sampling. The author also examines in detail several pervasive mistakes and misconceptions in statistical inference, including often ignored forms of multiple testing, and explains the concept of researcher degrees of freedom, which has received little if any attention in the nursing research literature.

2. The hypothesis testing framework

There is a large literature on hypothesis testing, and only a basic treatment is provided here. For further study, interested readers are directed to the American Statistical Association's Statement on Statistical Significance and P-Values (2016) and to the *The American Statistician's* special issue on statistical inference, beginning with the editorial by Wasserstein et al. (2019). These are freely available online.

The purpose of all statistical inference is to learn about some population of interest by studying a sample from that population, much like we draw conclusions about the quality of a piece of fruit based on the first few bites we take. Simply defined, statistical inference is drawing conclusions about a population based on data from a sample drawn from it. Hypothesis testing is a method of statistical inference that requires a null hypothesis about the population; contrary to what researchers may see in practice, the null and alternative hypotheses are not statements about the sample. Our goal is to draw conclusions about the population, and therefore this is what our hypotheses must be about.

In hypothesis testing we examine a sample for evidence against the null hypothesis, which is typically a statement that the difference or association of interest is zero. Traditionally, and unfortunately (as discussed below), this evidence has been used to make a yes/no decision as to whether to reject the null. In this decision-making framework, if the null is true, and we mistakenly reject it, we commit a *Type I error*, also known as a *false positive*. And if the null is false, and we mistakenly fail to reject it, we commit a *Type II error*. Of course, unless we know that the null is true or false, we cannot know for certain if we have committed one of these errors. We can, however, try to control the *probability* of making each type of error.

2.1. Motivating example

As a motivating example, consider a study designed to compare a standard treatment to a new treatment for some condition. Suppose we recruit a sample of 100 patients and randomly assign 50 patients to each of the two study arms (treatment groups). Ultimately what we want to know is not whether one treatment is more efficacious than the other in this particular *sample*, which is a question we could answer using only descriptive statistics, but whether we can expect one treatment to work better in the *population* of patients with the condition. In other words, we want the findings from our sample to hold more generally in the population. This is where methods of statistical inference come in.

Our null hypothesis might be that, *in the population*, the average improvement in symptoms for patients receiving the new treatment is equal to the average improvement for patients receiving the standard treatment. Of course we cannot administer treatment to the entire population, but we can hypothesize about what would happen if we did. Suppose our goal is to show that the new treatment leads to more improvement in the population, on average, than the standard treatment. Thus we want to find enough evidence against the null in our study to conclude that the new treatment is better, not just in the sense that the patients receiving it in our study did better than those receiving the standard treatment, but in the sense that we can expect patients *in the larger population* to do better on this new treatment than on the standard treatment.

We could test our null hypothesis by simply comparing the mean improvement for the two arms in our study (i.e., the sample means), but we have a problem: in a different sample of patients, the sample means would almost certainly be different and could even lead us to the opposite conclusion about which treatment is better. Perhaps, by chance, the new treatment appears more efficacious than the standard treatment in our study sample, but if we recruited another 99 samples of patients and re-ran the study for these samples we would find that the new treatment appears *less* efficacious in all of those samples. Moreover, we might have observed different results in our study sample had we randomized the patients to the two study arms differently. Perhaps in 100 such randomizations we would find that the new treatment appears more efficacious in only five randomizations, and the standard treatment more efficacious in the other 95. It is this kind of uncertainty due to randomness that the methods of inferential statistics are designed to account for.

2.2. Repeated sampling and p-values

Classical statistical inference is based on the idea of *repeated sampling*. This term is shorthand for repeating *ad infinitum* the entire process of drawing a sample from the population, carrying out the study, and computing the value of the statistic(s) to be used in inference. In our example this statistic might be the two-sample t statistic. We have a very good idea how the t and other statistics behave in repeated sampling when the null hypothesis is true—that is, when only randomness is at play. Like any statistic, the t statistic will "bounce around" from sample to sample, taking a different value each time, but we know it tends to take values closer to its mean more often than it takes values farther away. In addition, we know how frequently the t statistic takes values in any given range; that is, we know its *distribution*. Thus we can assess how unusual or extreme the observed value of our t statistic is by comparing it to the values we would expect to observe in repeated sampling—again, assuming the null is true. In fact, we can *quantify* how inconsistent the value of our t statistic is with the null by answering the following question: Assuming the null hypothesis is

true, how frequently in repeated sampling would we expect an outcome at least this inconsistent with the null?

This, of course, is the question we are answering when we compute the *p*-value. We can think of the *p*-value as a measure of surprise. If the null were true, how surprising would our outcome be? In our comparison of two treatments we would ask, "How often would we expect to see a difference in outcomes between study arms that is this large simply due to chance if, in fact, there were no difference whatsoever between the average effects these treatments would have in the population?" If, due solely to chance, we would expect to see an outcome at least as inconsistent with the null as our *t* statistic in ten samples out of 100 ($p = .10$), we may be somewhat surprised, but not fully convinced that anything other than randomness is at play. But if we would expect to see an outcome as extreme as our *t* statistic in only one sample out of 1000 ($p = .001$), we have stronger evidence that the null, which we assumed to be true in computing our *p*-value, is not true after all.

3. Errors in significance testing

The *p*-value was originally proposed as a rule of thumb for identifying findings worthy of further investigation, not a tool for making final judgments about whether findings were practically important, meaningful, or "significant" (Nuzzo, 2014; Wasserstein et al., 2019). It is useful for quantifying evidence against (never for) the null, but unfortunately hypothesis testing and interpretation of its results have fallen into widespread misuse and abuse, contributing to the replication/reproducibility crisis in science (e.g., see Gelman, 2015; Ioannidis, 2005; Open Science Collaboration, 2015). In 2015 the editors of *Basic and Applied Social Psychology* went so far as to ban *p*-values altogether, stating "we believe that the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research" (Trafimow, 2014; Trafimow and Marks, 2015). Readers interested in further discussion of issues related to *p*-values are referred to Nuzzo (2014) and Ioannidis (2019).

3.1. *Misinterpreting* p-*values*

One pervasive error in scientific research is assessing the importance of study outcomes based solely on tests of "statistical significance." The danger of mistaking "statistically significant" findings for practically meaningful findings has been highlighted many times, as has the importance of computing and reporting effect sizes along with hypothesis test results. Yet the message does not seem to have reached a large number of nursing researchers (Floyd, 2017; Gaskin and Happell, 2014; Polit, 2017). Recently the author reviewed a study where the researchers carried out extensive analyses to understand the association between two variables having a Pearson correlation of .12 with a *p*-value less than .05. Squaring this correlation for a better sense of the effect size, we find that we can account for less than 1.5% (.0144) of the variability in one variable using the information in the other. But the *p*-value for the test of this correlation decreases as the sample size increases, and in this case we are left with a finding that is "statistically significant" but means very little in practical terms.

Although the more common error is to mistake "statistically significant" findings for meaningful ones, we can err in the opposite direction by ignoring a potentially meaningful finding because its *p*-value is not less than the (arbitrarily chosen) cutoff value of .05. Researchers tend to pay more (though not enough) attention to avoiding false positives, but we also want to avoid missing potentially important findings. The challenge is to balance these two concerns given the costs associated with erring in each direction. The common practice of designing studies with statistical power of 80% theoretically reflects a willingness for roughly 20% of studies to fail due to a Type II error—a rate that seems surprisingly high. Perhaps this is further evidence that many researchers, not to mention funding agency personnel, have an incomplete grasp of significance testing. Most of us would be hesitant to invest a large sum of our money on a project with only an 80% chance of success.

Another pervasive error is interpreting a "non-significant" *p*-value as evidence *for* the null hypothesis. In our example study of 100 patients (50 per arm), suppose we revise the null to be that the proportion of patients in the population who would fully recover is the same for both treatments. If we find in our study sample that 60% of patients fully recover on the standard treatment and 70% fully recover on the new treatment, a test of independent proportions yields a *p*-value of .402 (see Fig. 1). If we observed these same recovery percentages in a smaller study of 10 patients per arm, the (Fisher exact test) *p*-value would equal 1.000—the maximum possible value. Yet the results of these studies, despite being nowhere near "statistical significance," could hardly be taken as evidence that the proportions of patients who recover on the two treatments are exactly equal; on the contrary, we have evidence in the form of a ten percentage point difference in recovery rates that they are *not* equal.

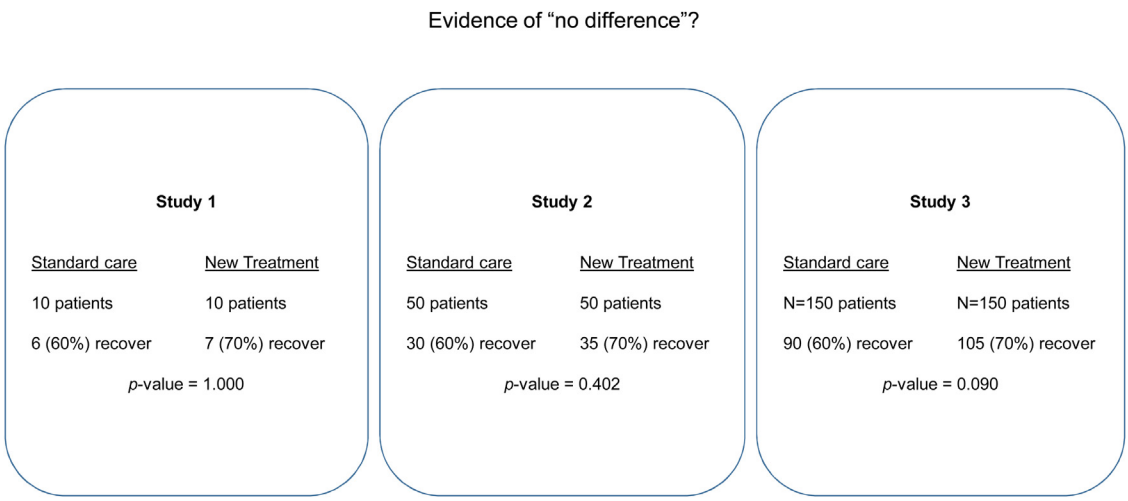

Fig. 1. Three hypothetical randomized trials.

In fact, we rarely observe evidence that two proportions or means are *exactly* equal in the population, or that a population correlation or regression coefficient is *exactly* zero. There is typically *some* evidence against the null, and *p*-values can be helpful in quantifying that evidence for a sense of whether it exceeds what we would expect due simply to chance. Thus, contrary to misconception, the *p*-value is not the probability that the null is true, nor should a non-significant *p*-value be taken as evidence for the null. In short, "no evidence of an effect" is not the same as "evidence of no effect." Or in our example, "insufficient evidence to conclude there is a difference" does not mean "evidence that there is no difference."

3.2. Misleading p-values

3.2.1. Multiple testing

Students in introductory statistics courses are often introduced to the issue of multiple testing and taught how to use the Bonferroni adjustment to keep the "experiment-wide error rate" (probability of one or more Type I errors) below the traditional .05 level. Multiple testing is a problem in this framework because we run the risk of a Type I error with each significance test, thereby increasing with every test the probability of committing at least one such error. The result is that the true experiment-wide error rate is not controlled at the advertised .05 level.

There is some controversy around the issue of multiple testing. It is not clear that the experiment-wide error rate is the appropriate error rate to control, nor is there consensus that *p*-values are the best method for testing hypotheses, or even that hypothesis testing should be the default framework for learning from data to begin with (Ioannidis, 2019; Thompson, 1998). For better or worse, however, the hypothesis testing framework remains the prevailing approach to statistical inference in nursing and other health-related fields. This is true both for studies involving well-specified hypotheses deduced from rigorous theoretical work, and for largely exploratory studies where hypothesis testing is not as well-suited due to the number of potential hypotheses and the challenge of specifying beforehand all the models and hypotheses to be examined.

Moreover, even when we are not making yes/no decisions to reject hypotheses and are not concerned with a study's experiment-wide error rate, subtle forms of multiple testing can result in findings with small *p*-values that are, in fact, meaningless results of randomness. In fact, methods that involve multiple testing, thereby allowing researchers to exploit chance, have been taught in statistics courses for decades and are now in routine practice. One such method is selecting explanatory variables for a model by first examining each candidate variable's correlation with the dependent variable. Another is stepwise regression modeling (see Harrell, 2015). In both cases the researcher is carrying out multiple hypothesis tests, implicitly if not explicitly, to arrive at a final model, typically without acknowledging the effects of multiple testing. With the first method, even if we only look at the magnitude of the correlations and not at their *p*-values, we are exploiting chance by selecting explanatory variables with the strongest association with the dependent variable *in this particular sample*, without accounting for the random variation in the sample correlations that we would expect to observe in repeated sampling; a correlation that appears meaningful in our sample may appear negligible in a different sample, and vice versa.

For both these methods, and for other methods of model selection involving comparison of multiple variables or models, the *p*-values computed for the final model are misleading because they do not reflect the randomness we encountered and took advantage of in the process of arriving at our final model. In practical terms, because we have followed the idiosyncracies of our unique sample to the (apparently) best possible model without considering the role of chance and quantifying our risk of being misled by it in each decision along the way, our findings are not as surprising (inconsistent with the null) as their *p*-values would lead us to believe, and they are less likely to be replicable in other samples. For the same reason, our confidence intervals will be too narrow, leading us to overestimate the precision of our parameter estimates, and our effect size estimates and R-squared value (if applicable) will be too large (Harrell, 2015). Like other forms of multiple testing, these model selection methods are not inherently wrong; the error is in their misuse. Recommendations are offered below.

3.2.2. Researcher degrees of freedom

The options available in model selection are an example of *researcher degrees of freedom*, a term used to describe the flexibility researchers have in carrying out a study, all the way from its design through the collection and analyses of data to the interpretation and reporting of findings (Simmons et al., 2011). Returning to our example study designed to compare two treatments, we must choose a target sample size, a method of recruitment, a dependent measure, a rule (or no rule) for identifying and excluding outliers, a statistical test for comparing the two study arms, etc. If we prespecify all these decisions, it is more likely that another researcher will be able to replicate our findings.

But suppose we try different options at a few points in the analysis process with an eye toward obtaining a "statistically significant"

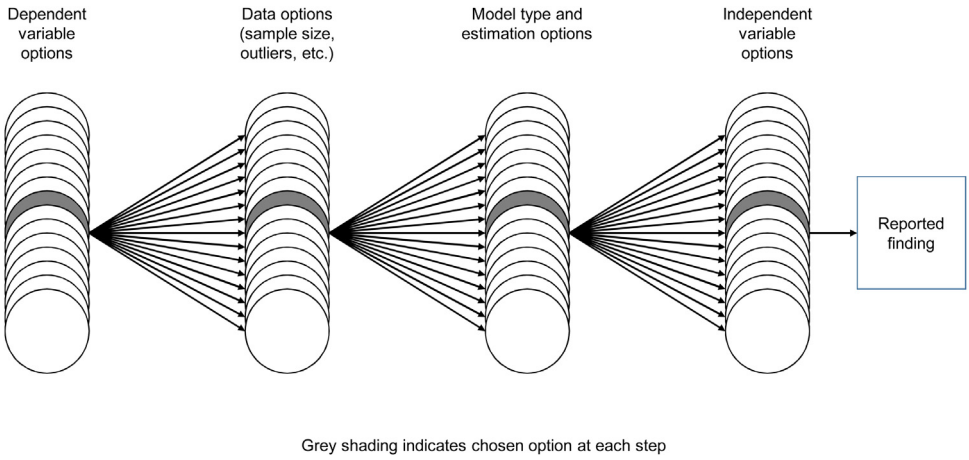

Fig. 2. Illustration of researcher degrees of freedom.

result, comparing different criteria for excluding outliers, for example, and two ways of measuring the outcome (improvement vs. percent improvement, say), and a parametric vs. a non-parametric test (see Fig. 2). Or if our initial finding is not "statistically significant," perhaps we consider re-running the analyses on a subsample with a particular symptomology, or adding covariates or interaction terms to our model, or bootstrapping. Each time we carry out a hypothesis test or examine the magnitude of our test statistic during this process of "fishing" for an apparently meaningful result, we are exploiting chance and undermining the validity of our final conclusions by multiple testing. The situation is akin to "proving" that a coin is unfair by flipping it repeatedly until we observe five heads in a row and then reporting the result as if the coin were flipped only those five times ($p = .031$).

Although it is not our intention to be dishonest in making use of our researcher degrees of freedom, and although we may convince ourselves that the apparently important finding that we eventually obtain is the product of objective decision-making (see Simmons et al., 2011), our final p-value will be misleading at best. If another researcher follows our final set of choices exactly in carrying out the same study on a new sample and fails to replicate our result, we should not be surprised. Nor should we be surprised if the treatment that appeared to be "statistically significantly" superior in our study turns out to be markedly inferior when introduced into clinical practice.

3.3. Randomization checks

In studies involving random assignment to one of two arms, it is not uncommon for authors to provide a table with descriptive statistics for the two arms on various participant characteristics (e.g., age, sex) and p-values for tests for differences on these variables. The idea is to ensure that the result of the randomization is a pair of arms that are not "statistically significantly different," increasing our confidence that any observed difference between arms on the outcome variable is due to their receiving different treatments. Conceptually, these tests are unworkable. Taking age as an example, if we test for a difference in average age between the two arms, the null hypothesis is that the mean ages of the two *populations* are equal. The problem is that we do not have two populations. Rather, we have a single population, we drew a sample from it, and we randomly divided the sample into two parts. Thus we know the null is true, in the sense that all participants were sampled from the same population with a single mean (not from two populations with different means). It follows that if we observe a small p-value, we know it is due solely to chance; it cannot be taken as evidence against the null because we know the null is true. And in several such tests, we should not be surprised to observe a small p-value simply due to chance.

We also learn very little when the result of such a test is a p-value that is not small. This, of course, is the result researchers generally want, to present as evidence that the two arms do not differ in any meaningful way. But as discussed above, a p-value is a measure of evidence *against* the null, and even a large p-value does not mean a null is true. In fact, there may be differences between the two arms that *are* important but do not yield a small p-value in the test for a difference between populations. Moreover, we already know the null is true, so we have little reason to subject it to hypothesis testing or to rely on such tests to detect important differences between the study arms.

4. Recommendations

4.1. Interpreting p-values

In reporting findings that appear "non-significant," researchers would do well to use precise language. For example, authors might state, "We found little evidence of an association," or "Evidence for an effect was limited," or "We observed a difference between arms, but it was too small to rule out chance as an explanation." In using statements like these, the idea is to consider not only the p-value, but also the direction, size, and any clinical or practical importance of the effect, and to avoid overstatements that would leave readers with the impression that the null is likely true. If we observe a non-zero effect with a p-value greater than .05 and simply report "There was no effect," it is tempting for readers, including journalists who report on science, to conclude "There is no effect," when in fact our analyses were set up only to find evidence *for* an effect, not to demonstrate a lack of one.

It is important to report confidence limits and effect sizes, and to assess whether effect sizes are large enough to be practically meaningful. It may be helpful to decide beforehand the smallest effect size that might be considered practically important to avoid the temptation of making too much of an effect that is too small to matter but has a low p-value. This could be part of a traditional power analysis, where the sample size needed to detect this smallest meaningful effect with high probability is computed in designing the study. Alternatively, we could use the smallest meaningful effect in deciding how narrow our confidence limits need to be and compute the sample size needed to achieve this level of precision.

If what we really want to know is whether the magnitude of an effect exceeds some threshold, not just whether it is different from zero, we can define the null hypothesis accordingly. This is the approach taken in *equivalence testing*, where the question is whether two treatments differ by more than some specified margin. In our motivating example, suppose the new treatment has fewer side effects in some patients, and we want to know whether the mean improvement for patients on this new treatment is comparable to the mean improvement for patients receiving the standard treatment. More precisely, suppose we want to know if the two treatments differ by more than .2 standard deviation (SD) on whatever measure we are using to assess improvement—a difference we judge to be clinically important. Rather than test a null hypothesis of zero difference between the two means, here we test the null hypothesis that the absolute difference in means is greater than .2 SD, hoping to find strong enough evidence against this null to conclude that the difference is no more than .2 SD.

A final recommendation regarding p-values, and hypothesis testing in general, is not to classify hypothesis test results as either "significant" or "non-significant." We can use p-values to quantify evidence against a null without categorizing them using an arbitrary threshold like .05 and adjudicating the importance of our findings on this basis. Wasserstein et al. (2019) write, "a declaration of 'statistical significance' has today become meaningless ... using bright-line rules for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making ... A label of statistical significance adds nothing to what is already conveyed by the value of p." McShane et al. offer the same recommendation in their paper entitled "Abandon Statistical Significance" (2017), as do Amrhein et al. (2019) and over 800 concurring signatories. These calls for reform are echoed by a group of 25 statisticians and quantitative methodologists in a recent guest editorial in *International Journal of Nursing Studies* (Hayat et al., 2019).

It should be noted that none of the authors above calls for banning p-values altogether, and that dropping the significant/non-significant dichotomy need not mean abandoning the entire hypothesis testing framework on which p-values are based. We can use the framework simply to quantify evidence against the null with a statistic (the p-value) that can take any value along the continuum from zero to one and stop there, rather than using the

hypothesis test as a pass/fail tests of "significance." In other words, the hypothesis test, like most of the tests familiar to us from school, should result in a *number*, not in an unnecessary yes/no decision. Scientific practice will not change overnight, to be sure, but this would be a step in the right direction.

4.2. Multiple testing

In the traditional framework of hypothesis tests involving a yes/no decision whether to reject the null, we can use the count of tests carried out and their *p*-values to apply an adjustment to control the experiment-wide Type I error rate. The well-known Bonferroni method is very conservative and can drastically reduce statistical power; the Holm-Bonferroni method (Holm, 1979) is always at least as powerful and easily implemented using a spreadsheet. A popular, generally more powerful alternative to these and other methods of controlling the experiment-wide Type I error rate is to control the *false discovery rate*—the proportion of hypotheses rejected that are false positives (Type I errors). The Benjamini-Hochberg procedure (see Hochberg and Benjamini, 1990) is a method for controlling the false discovery rate and, like the Holm-Bonferroni method, straightforward to apply using a spreadsheet program. More flexible extensions of this method are described by Storey (2003).

With methods of model selection that involve multiple testing, some kind of validation of the final model is required. One approach to validation is as follows: Suppose we begin by splitting our (sufficiently large) sample into two parts: a set of *training data* and a set of *testing data*. We then use the training data to carry out the entire model selection process, holding out the testing data for use in validating the final model. This simple approach allows us to avoid the problems described above. Whatever we gained by exploiting randomness in selecting the model will tend to be lost in fitting it to the testing data, and we will get more honest effect size estimates, confidence intervals, *p*-values, etc. The drawback of this approach, of course, is that we are not making full use of the data in model selection, potentially leading to a suboptimal choice for the final model, or in estimating model parameters, which means reduced precision, wider confidence limits, and lower statistical power. Harrell (2015) describes more efficient methods of validation, including *k*-fold cross-validation and bootstrap-based methods. But in many cases, we can avoid model selection altogether by simply including all the potential explanatory variables in one model and accepting the results.

4.3. Researcher degrees of freedom

Use of researcher degrees of freedom in analyses should be carefully restricted unless separate data are available for validation. It is best to pre-specify the analyses to be carried out to the extent feasible. Making this information publicly available in a written research protocol before any data analysis takes place is an excellent safeguard against abuse of researcher degrees of freedom.

This is not to say that every analysis decision must be made beforehand; certain decisions may require examining aspects of the data. For example, a skewed dependent variable may need to be transformed to meet a model assumption of Gaussian (Normal) residuals, and correlations between potential explanatory variables may need to be considered in deciding which to include in a model. But decisions like these should be based only on the relevant information (e.g., a plot of model residuals), not on whether a particular choice yields a "statistically significant" finding. A simple rule is that it is generally safe to examine the dependent variable in isolation (e.g., to check for outliers) and the set of explanatory variables in isolation (e.g., to assess multicollinearity), but one should not look at information regarding associations *between* the dependent and explanatory variables, including R-squared values and model fit statistics, in making analysis decisions without some kind of validation of the final model.

4.4. Randomization checks

In a randomized study, we hope to observe a difference in *outcomes* between the two arms that we can attribute to the difference in *treatment*. What we really want to know in a randomization check is whether there are differences between the arms—that is, between the two parts of the *sample* that we formed by randomizing participants—that might account for the observed difference in outcomes. In other words, we would like to assess the potential for *confounding*. Without resorting to hypothesis testing, the two arms can be compared using descriptive statistics, and their differences can be evaluated based on effect sizes. For example, the standardized mean difference can be computed for continuous variables, assuming their distribution is not too skewed.

It is possible, as a reviewer noted, to compute *p*-values for the tests of between-arm differences and treat them as *descriptive* statistics reflecting how extreme, in the sense of being unlikely to arise due to chance in repeated sampling, each difference is. Two caveats should be kept in mind with this approach: first, *p*-values are a function of sample size and should not be considered without also examining the effect size; and second, the *p*-values should be *evaluated*, not compared to an arbitrary threshold as if a hypothesis test is being carried out. After examining differences between arms, we can include variables on which the two arms differ as covariates in our statistical model.

4.5. Statistical expertise

A final recommendation the author would offer to nursing researchers is to continue developing their own statistical understanding and make use of the statistical expertise of others (for example, by collaborating with statisticians, whose perspective and knowledge can prove invaluable in designing studies and analyzing data appropriately). Admittedly, becoming a more knowledgeable user of statistics is sometimes easier said than done. Statistics is not always taught in an accessible way or with sufficient emphasis on conceptual understanding, and much like health websites can make medical diagnosis and treatment seem simple, easy-to-use statistical software can make statistics seem deceptively simple when, in fact, even those of us who practice statistics for a living can find it very challenging. Nevertheless, the more statistical expertise we can bring to bear on a project, whether from nursing researchers, statisticians, or both, the better the result will generally be.

5. Conclusions

Those involved in the education of nursing researchers would do well to ask what changes need to be made to raise the level of statistical practice in nursing research; Hayat et al. (2015) offer insights and recommendations. Although Bayesian methods are growing in popularity (see Lavine (1999) for a very brief introduction), for the foreseeable future classical methods of statistical inference based on the idea of repeated sampling will be the primary tools for quantifying randomness. The hypothesis testing framework, its limitations notwithstanding, can be helpful in ruling out chance as an explanation for effects observed in research studies. Nursing researchers who use quantitative methods and those who serve as journal reviewers and editors should understand it well.

Conflict of interest

None.

References

American Statistical Association, 2016. ASA statement on statistical significance and *p*-values. Am. Stat. 70 (2), 131–133.

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567, 3-5-307.

Floyd, J.A., 2017. A descriptive study of effect-size reporting in research reviews. J. Adv. Nurs. 73 (6), 1467–1481.

Gaskin, C.J., Happell, B., 2013. Power of mental health nursing research: a statistical analysis of studies. Int. J. Ment. Health Nurs. 22 (1), 69–75.

Gaskin, C.J., Happell, B., 2014. Power, effects, confidence, and significance: an investigation of statistical practices in nursing research. Int. J. Nurs. Stud. 51 (5), 795–806.

Gelman, A., 2015. Statistics and the crisis of scientific replication. Significance 12 (3), 23–25.

Harrell Jr, F.E., 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed. Springer-Verlag, New York.

Hayat, M.J., Higgins, M., Schwartz, T.A., Staggs, V.S., 2015. Statistical challenges in nursing education and research: an expert panel consensus. Nurse Educ. 40 (1), 21–25.

Hochberg, Y., Benjamini, Y., 1990. More powerful procedures for multiple significance testing. Stat. Med. 9, 811–818.

Hayat, M.J., Staggs, V.S., Schwartz, T.A., Higgins, M., Azuero, A., Budhathoki, C., Chadrasekhar, R., Cook, P., Cramer, E., Dietrich, M.S., Garnier-Villarreal, M., Hanlon, A., He, J., Hu, J., Kim, M., Mueller, M., Nolan, J.R., Perkhounkova, Y., Rothers, J., Schluck, G., Su, X., Templin, T.N., Weaver, M.T., Yang, Q., Ye, S., 2019. Moving nursing beyond $p < .05$. Int. J. Nurs. Stud. doi:http://dx.doi.org/10.1016/j.ijnurstu.2019.05.012.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Med. 2 (8), e124.

Ioannidis, J.P., 2019. What have we (not) learned from millions of scientific papers with P values? Am. Stat. 73 (sup1), 20–25.

Lang, T., Altman, D., 2016. Statistical analyses and methods in the published literature: the SAMPL guidelines. Med. Writ. 25, 31–36.

Lavine, M., 1999. What is Bayesian statistics and why everything else is wrong. J. Undergraduate Math. Appl. 20 (2), 165–174.

McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2017. Abandon Statistical Significance. arXiv Preprint arXiv:1709.07588. .

Nuzzo, R., 2014. Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506 (7487), 150–153.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251) aac4716.

Polit, D.F., 2017. Clinical significance in nursing research: a discussion and descriptive analysis. Int. J. Nurs. Stud. 73, 17–23.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366.

Storey, J.D., 2003. The positive false discovery rate: a Bayesian interpretation and the *q*-value. Ann. Stat. 31 (6), 2013–2035.

Thompson, J.R., 1998. Invited commentary: Re: "Multiple comparisons and related issues in the interpretation of epidemiologic data.". Am. J. Epidemiol. 147 (9), 801–806.

Trafimow, D., 2014. Editorial. Basic Appl. Soc. Psych. 36, 1–2.

Trafimow, D., Marks, M., 2015. Editorial. Basic Appl. Soc. Psych. 37, 1–2.

Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world Beyond "p <0.05.". Am. Stat. 73 (sup1), 1–19. doi:http://dx.doi.org/10.1080/00031305.2019.1583913.