A New Look at Evidence Testing and Measures of Error

Robert H Riffenburgh

Naval Medical Center San Diego
and
San Diego State University

July 2017

Abstract

Recently, considerable attention has been paid to the misuse of statistical testing outcomes, particularly p-values. I envision three stages to rectify this problem: conceptualization (providing a clear, unequivocal exposition of proper procedures), personalization (making this procedure accessible to the general non-statistician statistics-using public in a readily understandable form), and implementation (effecting the adoption of the second stage among the vast statistics-using public). The conceptualization (first) stage was partially addressed in 2016 by a policy statement from the American Statistical Association posing the process in traditional statistical terms. This paper attempts to start the personalization (second) stage with the goal of stimulating debate on the issue. Five components are proposed and discussed: Use terms to make the statistical discovery process meaningful to non-statistician statistics-users, concentrate on the effect of an experiment rather than a test, view a test result as a measure of belief in the effect, provide joint difference/equivalence tests, and provide a scaled ranking of beliefs relating p-values to practical real-life interpretations. The components are shown in a practical example. A template to help users to follow proper statistical discovery procedures is proposed.

Key words: Statistical inference, *p*-values, statistical tests, measures of error.

The Goal

Recently, considerable attention has been paid to the misuse of statistical testing outcomes, particularly p-values. I envision three stages to rectifying this problem: conceptualization (providing a clear, unequivocal exposition of proper procedures), personalization (making this procedure accessible to the general non-statistician statistics-using public in a readily understandable form), and implementation (effecting the adoption of the second stage among the vast statistics-using public). The conceptualization (first) stage was partially addressed in 2016 by a policy statement from the American Statistical Association posing the process in traditional statistical terms. Additional conceptualization is expected to develop at and from the October 2017 Symposium on Statistical Inference in Bethesda, MD.

This paper is an attempt to start the personalization (second) stage with the goal of stimulating debate on the issue and does not presume to provide new methodology.

An Example

Let me open with an example. To treat a fractured ankle, medical standard of care mandates pinning (implanting braces on fractured bones) by device 1. An investigator compares pinning device 1 with treatment by a new cheaper and more easily installed pinning device 2 in a randomized controlled trial with 30 patients treated by each device. The measure of success used for comparison is the distance—measured in inches—covered in a triple hop on the injured leg four months after repair (the longer the hop, the better the healing). The investigator hopes the new pinning device is superior, so chooses a difference test rather than an equivalence test. He submits H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$. Sample means and standard deviations were $m_1 = 33$, $s_1 = 4$; $m_2 = 35$, $s_2 = 6$. A two-sample t test yields $p = 0.134$ with confidence intervals CI_1 32-34 and CI_2 33-37 (rounded for clinical interpretation; values to 4 decimal places were used in calculations). H_0 cannot be rejected by the $p < 0.05$ rule, so the statistical conclusion is indeterminate. Statistical orthodoxy says there is no evidence to prefer one over the other. Since there is no evidence to change the current standard of care, the real life consequence defaults to a “use standard of care treatment 1” decision. What is wrong here?

The problems

Recently there has been extensive criticism of *p*-values, even to the point of urging that they not be used (Trafimow and Marks 2015). However, an abandonment of *p*-values follows faulty logic similar to saying that a misused tool should be abandoned rather than being used only correctly. The controversy became so intense that the American Statistical Association was motivated to publish an official policy statement in this regard (Wasserstein and Lazar 2016), a helpful first stage in addressing the misuse of statistical inference.

Non-statistician users of statistical methods (and even some statisticians) often misinterpret the statistical discovery process in the following ways:

Rigid adherence to misunderstood concepts. Users find it difficult to accept corrective statistical explanations because of their rigid adherence to previously ingrained concepts. If we try to explain the correct logic of the statistical discovery process, non-statisticians often nod politely, then return to the definitions and statistical procedures they’ve been taught.

Neglecting descriptive statistics. Users present descriptive statistics only as support for a statistical test rather than as primary evidence. In the statement of the example above, the difference between average hop distances for the two treatments is not even mentioned, although it is the primary outcome of the experiment.

Test result as indicator of efficacy. Too often, the test’s resulting *p*-value is taken as evaluating the experiment’s outcome, whereas the test is designed to give evidence on whether or not the experiment’s outcome is compatible with the experiment’s model. In

the example, the investigator focuses on the *p*-value as the outcome of the experiment. This faulty logic interprets the experiment's outcome as detectability rather than effect.

One-sided test of efficacy. Users employ a test that gives a one-sided decision—a defined outcome is unlikely to be compatible with a hypothesis of no difference versus the decision is indeterminate. When indeterminate, there is no evidence of effect, but this does not constitute evidence of no effect. In the example, in the absence of test evidence, the orthopedist must revert to standard of care and use the more costly and difficult-to-install surgical device.

Inflexible use of p-value. The cut-point on making a statistical discovery decision (using the *p*-value) is an almost universally accepted, firmly fixed, originally arbitrary value of 0.05. This results in a test outcome of $p = 0.049$ to be taken as "golden" but of $p = 0.051$ taken to show the experiment's outcome to be useless. In the example, the *p* is large enough not to be an issue, but we have all seen cases in which the rigid *p*-value cut point would be an issue.

The general issue of rigid employment of the Neyman-Pearson hypothesis testing logic to fit all applications has been addressed repeatedly, most thoroughly by Hurlbert & Lombardi (2009). The solutions suggested below are not new; what is new is the purpose, the framework in which these suggestions are posed, and the fabric woven from their combination.

Some Possible Solutions

Re: *Rigid misconceptualization.* Statisticians have tried again and again to explain these problems to users and how testing should be done properly, but the erroneous thinking has been so ingrained in the non-statistician user that explanations are little heeded. We need a rework of the basic definitions to jog the user out of scientific lethargy. We should retell the story of statistical discovery with new names for concepts so that users will see the story afresh in proper connotation and not revert to their ingrained misunderstanding. Equally important, these new names should be intuitively meaningful to users.

The measurement or rate or rank whose value is the issue of the experiment might be called the *experimental effect*, a term long used by statisticians. We must be led to ask first in a study what is the experimental effect and only then to ask about the ability to detect this effect.

Dealing with formal hypotheses has long been an anathema to non-statistician users, especially since the null hypothesis they are testing is usually opposite from the practical hypothesis they think likely. To shake users' thinking from erroneous patterns, we could rename *hypotheses* as *interpretations* and *hypothesis testing* as *evidence testing.* Thus we would be testing for evidence of an interpretation that the experimental effect is or is not important. This makes intuitive sense to the non-statistician investigator.

The concept of *significance* of a test result, too often confused with the non-technical concept of practical significance, is a statement of belief in the effect, not a measure of the effect itself, and might better be thought of as *detectability*.

The *p-value* could be accurately defined as the rate under a specified statistical model at which repeated samples drawn from the same population would show an effect at least as large as the observed one when that effect was absent. “Model”, obscure to many non-statisticians, includes at least the null and alternate hypotheses; sample size; and the assumptions of form of data distribution in the population being sampled, data independence, and data identicality. *p* is not an indicator of lack of effect, discussed in some detail below, but rather is a measure of doubt in the observed level of effect. To remove it from previous rigid concepts, we could rename *p* as the proportion of doubt in the observed effect. Similarly, the measure of doubt in observing the absence of an effect that is present could be denoted the proportion of doubt in a non-effect. For abbreviated reference, these could be named the *proportion of positive doubt*, or *PPD*, and the *proportion of negative doubt*, or *PND*.

Using these terms, we are able to explain the statistical discovery process in a way meaningful to the investigator and conveying the process simply:

> We estimate the experimental effect and assess its practical importance. Then we evaluate how strongly we may believe in this practical importance by evidence testing, augmented by consideration of sample size and sample characteristics. The believability in the experimental benefit’s detectability is indicated by the proportion of positive doubt in an observed effect (*PPD*) and the proportion of negative doubt (*PND*).

Re: *Neglecting descriptive statistics.* The key information in an experiment is in the value of the experimental effect itself, not how strongly we can believe in its detectability. We need to focus on presenting the experimental effect—after all, that’s why we undertook the experiment. We should not be shy about expressing an informed and experienced judgment of effect importance. As noted by Gelman (2013): “Epidemiologists and applied scientists in general have knowledge of the sizes of plausible effects and biases.”

We need to establish a cut point differentiating the effect’s importance from its unimportance. This value is often denoted Δ. An estimate of effect less than Δ would be an amount that could have arisen from any number of causes unrelated to the purpose of the experiment and obscuring any result too small to be of importance. To borrow a term from engineering applications, values less than Δ may arise from “machine error”.

In our fractured ankle example, a hop distance less than half the length of a foot (6 inches) may be due to differences in shoes, floor finish, subject motivation on that day, etc. We take Δ= 6 in. In the example, m_1 = 33, s_1 = 4; m_2 = 35, s_2 = 6, all in inches. Since the difference in hop distance is only 2 inches, considerably less than Δ, we judge that it is not clinically important.

Only after we have made this practical judgment, may we ask: How strongly may we believe in this practical judgment?

Re: *Test result as indicator of efficacy.* After our attention to the experimental effect value per se as the indicator of efficacy, we proceed to evidence testing.

A number of suggestions on novel ways to do evidence testing have been made, mostly providing some improvement but still imperfect. For example, Poole (1987) proposed using the *p*-value graphed on the odds ratio, Rosenthal & Rubin (1994) proposed a "counternull" statistic (a multiple of the effect size) combined with the *p*-value, Schweder & Hjort (2002) proposed confidence distributions in place of confidence intervals, Bender et al (2005) proposed confidence curves in place of confidence intervals (closely related to Poole's *p*-value graphs), Killeen (2005) proposed effect replication in place of significance testing, Demidenko (2016) proposed a "D-value" method using standard deviations in place of standard errors, and Noguchi (2016) proposed using "range-preserving" confidence intervals.

However, I suggest that the smallest variation on current practice will have the greatest likelihood of successfully reëducating statistics users and gaining their compliance. This smallest variation would be to leave historical and legitimate methods in place but to induce their proper use by requiring the use of new terms that invoke the right interpretations. With the new terms and revised explanation of statistical discovery, we will not find it difficult to maintain awareness that evidence testing provides only guidance in how strongly to believe in the putative efficacy.

From the example, we have decided that 2 inches is not a practically important difference. A *t* test on the difference in means from two independent samples yields the proportion of positive doubt in the effect, the *PPD*, of 0.134, too large a doubt to believe that 2 inches is evidence of a true difference. So we fail to detect a difference; we have no evidence regarding our outcome.

Re: *One-sided test of efficacy.* The study failed to show evidence of a difference. Is that the end of the story? No. We can now ask, is there evidence that there is no difference, i.e. can we test the data to learn if our decision of no difference is believable? Yes. We can make an equivalence test on the difference on means from the two samples and estimate the proportion of negative doubt, or *PND*.

Schuirmann (1987) introduced the two one-sided tests (TOST) for equivalence, but not a joint difference-and-equivalence test. Tryon (2001) and Tryon and Lewis (2008) proposed a joint difference and equivalence test for a two-sample *t* test, although their method relies on confidence intervals, yielding no indication of strength of difference or equivalence, and uses two confidence intervals on the respective means rather than one confidence interval on the difference between means. In this paper, I treat only the two-sample *t* test case, as the purpose is to codify a simpler approach to inference, not address methodology.

I propose that a joint difference and equivalence test be used routinely. In that case, there would exist four possible outcomes. Let us denote by δ the true but known difference in means, estimated by *d*. The null hypothesis would be H_0: $\delta \leq -\Delta$ or $\delta \geq \Delta$, and the alternate hypothesis would be H_1: $-\Delta < \delta < \Delta$. If we denote by *L* and *U* the lower and upper bounds of the confidence interval on δ, the four outcomes would be as follows.

(1) [$L<-\Delta$, $U<0$] or [$L>0$, $U>\Delta$] corresponding to evidence of difference and no evidence of equivalence.
(2) [$L\subset(-\Delta,0)$, $U\subset(0,\Delta)$] corresponding to evidence of equivalence and no evidence of difference.
(3) [$L<-\Delta$, $U>0$] or [$L<0$, $U>\Delta$] corresponding to indeterminacy, i.e. no evidence of equivalence or difference.
(4) [$L,U\subset(-\Delta,0)$] or [$L,U\subset(0,\Delta)$] corresponding to the degenerate case in which statistical variability is much smaller, i.e. measurement precision much larger, than the practical considerations giving rise to Δ.

In the example, $m_1 = 33$, $s_1 = 4$; $m_2 = 35$, $s_2 = 6$. The difference between means is $d = 2$, its $df = 58$, its standard deviation $s_d = 5.0990$, and its standard error $SE_d = 1.3166$. A confidence interval on δ is –4.6354 to 0.6354. The t statistics for H_0 are 3.0381 and 6.0763, yielding probability levels 0.0018 and < 0.0001. So the proportion of negative doubt (*PND*) would be reported as 0.002. At an $\alpha = 0.05$ level, H_0 is rejected in favor of H_1, showing evidence of equivalence. From a clinical decision standpoint, we can have reasonable belief in our decision that device 2 is not inferior in benefit to device 1 and the 2-inch improvement suggests that it could be slightly better. Equivalence mandates not that we default to device 1, but rather that we choose the device on a basis unrelated to subject performance. In this example, we would change clinical practice by choosing device 2 based on ease of surgical installation and cost.

Re: *Inflexible use of p-value.* Investigators using a difference test often find a *proportion positive doubt* (*p*-value) of, perhaps, 0.08 and say that, while there is “no statistical significance”, there is a “trend”. What they intend is that they did not satisfy the arbitrary requirement of *PPD* < 0.05, but they still believe there is practical evidence of a difference. We should not require that they “weasel word” their result, nor should we deny them further attention to a promising outcome.

Some attention has been given to the need for a more flexible relation of *PPD* to practical interpretation. Hurlbert and Lombardi (2009) note this need and Gelman & Robert (2014) recommend adjusting the significance level to the scenario. Boos and Stephanski (2011) note the use of *, **, and sometimes even *** to denote ordered categories of *p*-values but do not attach them to interpretive wording.

The most easily usable guide would be to scale the evidence levels as sort of a Likert scale with non-statistical descriptors that can be interpreted by anyone. Such a scale was hinted at by Goodman (1999) and proposed by Bland (2000). I propose such a scheme, shown as Table 1. This scheme is a little more liberal than those of Bland and Goodman because I base my values upon my experience.

Table 1.

Range of Doubt Level	How Strongly to Believe in an Effect
< 0.005	Very Strongly
0.005 to < 0.05	Strongly
0.05 to < 0.10	Moderately
0.10 to < 0.20	Weakly
≥ 0.20	Not at All

Using the scale in Table 1, the decision maker in the example would conclude that the *PND* = 0.002 is very strong evidence in favor of the clinical decision of no difference between the modes of treatment.

However, we also note that *PPD* = 0.134 would provide weak evidence for the 2-inch improvement by device 2. Thus there is very strong evidence that device 2 is not inferior to device 1 and weak evidence that it may be even better.

A Guide for the User

If we provide a sequence of steps to follow, it will guide and continually remind the user of the proper statistical discovery process. A possible guide appears as Table 2, exemplified for the broken ankle example as Table 3. As many columns could appear to the right as there are statistical questions to be addressed.

Table 2.

Issue to be Addressed		
Indicator of Benefit		
Value of Indicator		
Practical Interpretation		
Test on Indicator		
Proportion Positive Doubt		
Proportion Negative Doubt		
Consider sampling		
Evidence Shown		
Belief in Interpretation		

Table 3

Issue to be Addressed	Compare devices
Indicator of Benefit	Mean difference in hop distance
Value of Indicator	2 in.
Practical Interpretation	Device 2 not worse, maybe slightly better
Test on Indicator	*t* test
Proportion Positive Doubt	0.134
Proportion Negative Doubt	0.002
Consider sampling	Sample size moderate; data approx. normal & IID
Evidence Shown	Device 2 not worse, may be slightly better
Belief in Interpretation	Very strong in noninferiority, weak in better

Conclusion

To make statistical discovery easier to work with, to shake the research public from misconceptions, to fill gaps in the discovery process, and to reduce loss of research findings due to arbitrariness: We redefine terms to be meaningful to users; we focus on descriptors that are the purpose of the study; we interpret test results as believability, not efficacy; we use joint difference and equivalence testing; and we allow flexible measures of believability.

References

Bender, R. et al (2005). Tutorial: using confidence intervals in medical research, *Biometrical Journal*, 47, 237-247.

Bland, M. (2000). *An Introduction to Medical Statistics*, 3rd Ed. Oxford: Oxford University Press. Section 9.4.

Boos, D. D. and Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician*, 65(4), 213-221.

Demidenko, E. (2016). The *p*-value you can't buy, *American Statistician*, 70(1), 33-38.

Gelman, A. (2013). P values and statistical practice, *Epidemiology*, 24(1), 69-72.

Gelman, A. and Robert, C. P. (2014). Revised evidence for statistical standards. *Proc Nat Academy of Sciences*, 111(19), E1933.

Goodman, S.N. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor, *Ann Internal Medicine*, 130(12), 1005-1013.

Hurlbert, S. H. and Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Ann Zool Fennici*, 46, 2311-349.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests, *Psychological Science*, 16, 345-353.

Noguchi, K. and Marmolejo-Ramos, F. (2016). Assessing equality of means using the overlap of range-preserving confidence intervals, *American Statistician*, 70(4), 325-334.

Poole, C. (1987). Beyond the confidence interval, *American J Public Health*, 77, 195-199.

Rosenthal, R. and Rubin, D. B. (1994). The counternull size of an effect size: a new statistic, *Psychological Science*, 5, 329-334.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *J Pharmacokinetics and Biopharmaceutics*, **15**, 657-680.

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood, *Scandinavian J of Statistics*, 29, 309-322.

Trafimow, D. and Marks, M. (2015). Editorial in *Basic and Applied Social Psychology*, **37,** 1-2.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests, *Psychological Methods*, **6(4)**, 371-386.

Tryon, W. W. and Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor, *Psychological Methods*, **13(3)**, 272-277.

Wasserstein, R. L. and Lazar, N. A. (2016). ASA statement on statistical significance and *p*-values, *The American Statistician*, **70(2)**, 129-133.