The Undetectable Difference: An Experimental Look at the "Problem" of *p*-Values

William M. Goodman, Ph.D.
University of Ontario Institute of Technology, 2000 Simcoe St. N., Oshawa, ON, Canada

Abstract

In the face of continuing assumptions by many scientists and journal editors that *p*-values provide a gold standard for inference, counter warnings are published periodically. But the core problem is not with *p*-values, *per se*. A finding that "*p*-value is less than α" could merely signal that a critical value has been exceeded. The question is why, when *estimating* a parameter, we provide a range (a confidence interval), but when testing a hypothesis about a parameter (e.g. μ = x) we proceed as if "=" entails *exact* equality of the parameter with x. That standard is hard to meet, and is not a standard expected for power calculations, where we are satisfied to reject H_0 if the result is merely "detectably" different from (exact) H_0. This paper explores, with resampling methods, the impacts on *p*-values, and alternatives, if the null hypothesis is defined as a thick or thin range of values. It also examines, empirically, the extent to which the *p*-value may or may not be a good predictor of the probability that H_0 is true, given the distribution of the data.

Key Words: p-value, significance, evidence, inference, hypothesis testing, detectable difference

1. Background and Introduction to the Problem

Expressed as neutrally as possible, the *p*-value is an output from a certain procedure—a hypothesis test—used conventionally to guide judgment with respect to a reference hypothesis (the "null hypothesis") as to which is more plausible: That (a) the hypothesis is true (or, at least, not false); or (b) the hypothesis is not true. The null hypothesis ("H_0") posits, explicitly, a particular value for a parameter of interest, such as the population mean of a variable; implicitly, the null hypothesis models the entire population distribution for the variable (e.g. Gaussian, with a variance that can be estimated from the sample). Data for a test are collected by sampling, without prior knowledge as to whether the parameter modeled in H_0 is factually true. If sample results differ from those given in the null hypothesis (e.g. if the sample mean differs from the population mean in H_0), then this difference in values can be expressed directly or (for parametric tests) in standardized form, taking into account the expected distribution of the sample statistic, when samples are drawn from H_0. The *p*-value, therefore, gives the conditional probability of obtaining a magnitude at least as large (with suitable sign) for the difference in values (absolute or standardized) between the actual and expected sample results, *if* the null-hypothesis assumptions were in fact true.

The preceding definition was expressed as impartially as possible—as just the (mechanical) output of "a certain" procedure. If you follow the mandated steps, and

report a suitable number, there is no problem to discuss. The problem, of course, lies in the *interpretation* and use of the *p*-value. The impression easily arises that a small *p*-value from an experiment suggests an outcome that was "unexpected", if not actually impossible, had the null hypothesis been true; and so, it might appear, the hypothesis is more likely to be false. This very interpretation is posted on a research hospital's web page, and their view is widely held and applied in the research community: "A p-value is a **measure of how much evidence we have against the null hypothesis**. ... The smaller the p-value, the more evidence we have against H_0." (Simon, 2007, emphasis included).

Simon's view draws on the apparent "surprisingness" of getting a small *p*-value. Berger and Selke (1987), however, speak for many who challenge the conventional perspective: To them, *p*-values and evidence are "irreconcilable." Directly stated by Marden (2000), "the *p*-value is not the probability that the null hypothesis is true". Empirically, a *p*-value can be small (i.e. conventionally taken as good evidence that H_0 is false), but yet if looked at from, say, from a Bayesian perspective, the posterior probability that H_0 is in fact false, given the sample data, may not look as convincing.

Some argue that evidence for, versus evidence against, a hypothesis should be distinguished. Donahue, for example, grants that small *p*-values may provide evidence against a null hypothesis, but when a large *p*-value does "not warrant rejection of the null," he objects to the *p*-value's use as "improper evidence for *accepting* the null hypothesis" (Donahue, 1999).

A quite different objection to the conventional use of *p*-values appears in a classic paper by William Roozeboom (1960). Though inclined to Bayesian approaches to assessing evidence (to what he calls calculating "inverse probability"), he does not outright reject legitimacy for *p*-values. Instead, what he opposes is the very model of the null-hypothesis significance test, to which, in current practice, *p*-value calculations are often conjoined. Science, says Roozeboom, does not proceed by binary decisions to simply accept versus reject a given starting hypothesis. Instead, continued experimentation should lead to progressive, informed changes in our degrees of belief in various hypotheses.

This wedge between the original conception of *p*-values by R.A. Fisher, on the one hand, and the Neyman-Pearson hypothesis test model, on the other, is driven further by Steven Goodman in a historical treatment of these ideas (Goodman, 1993). He claims that the two ideas' combination (into a view such a Simon's, quoted above) has been "improper"; and like Donahue and many others, he says that "the *p*-value substantially overstates the evidence against the null hypothesis". Similar points have appeared more recently (Hubbard & Armstrong, 2005; Hubbard & Bayarri, 2003; Ziliak and McCloskey, 2009).

One other aspect of the problem, underlying how *p*-values are conventionally used, appears in the literature. This is aspect is called by Berger and Delampady (1987) "the actual 'width' of H_0." The problem is that the *p*-value/hypothesis-test model proceeds as if the null parameter was an exact point value. For continuous data, it is mathematically impossible for a real parameter to equal a point null; and in some fields such as Psychology it may be unrealistic (and meaningless in practice) to formulate a widthless value for a research hypothesis (Berger & Sellke, 1987; Chow, 1988; Folger, 1989; Meehl, 1967; Nunnally, 1960). Curiously, the mathematical aspect is handled at times more like a side issue, than as central to discussion; for example, Berger and

Delampady's paper largely ignores the issue, then provides some formal calculations for how "wide" H_0 can become, without invalidating their main points of interest.

Two other trends in the literature should be noted. The papers by Meehl and Nunnally, cited above, brush against a larger literature that attempts to turn the focus from hypothesis testing (geared to *rejecting* hypotheses) to what are variously called, depending on context, model validation, equivalence testing, and tests of clinical noninferiority. (Examples include: Berger & Hsu, 1996; Evans, 2009; Robinson & Froese, 2004; Rogers et al, 1993; and Wellek, 2003). Although these topics are beyond the scope of the present paper, they do share a recognition of the "thickness" or "width" of H_0, so that evidence is consistent with H_0 if, in effect, it falls within a confidence interval that is centered by H_0.

A second important trend is closely intertwined with literature, cited above, that questions the use of *p*-values—but these other papers explicitly proffer Bayesian approaches as the alternative (Edwards et al, 1963; Cassella & Berger, 1987; Goodman, 1999). These particular arguments are not addressed explicitly in this paper, yet Bayesian concepts are implicit in the experiments that follow below. All the assumptions built into the experiments are in effect models of "prior probabilities" that will clearly impact the unfolding of the simulations. Moreover, assessment of the evidential impact of a *p*-value in terms of the "posterior probability" of H_0 is essentially what graphs such as Figure 4, below, attempt to enable.

2. The Problem Reconsidered—Experimentally

Many of the papers written about *p*-values tend to be formal and abstract. But as we know in teaching, a picture can be very helpful to enhancing learning and understanding. In the statistical context, a "picture" might be the graphical outcome of simulations. So, rather than pursue the arguments in Section 1 theoretically, the remainder of this paper will try a hands-on approach, using resampling-based simulation, to see how *p*-values actually get generated in action, and to see how, or if, they relate to the independently known truths or falsities of null hypotheses.

All simulations will address the "problematic inequality" shown in Figure 1, which lies at the heart of many of concerns about *p*-values expressed in the literature. In this process, the paper will focus on two particular issues:

Issue 1: The *thickness* of H_0.
Issue 2: The possibly *mismatched structure* of the inequality.

$$P([\text{The sample data have the obtained distribution}] \mid [H_0 \text{ is true}]) \neq P([H_0 \text{ is true}] \mid [\text{The sample data have the obtained distribution}])$$

Figure 1: The Problematic Inequality

The first (upper) term in the inequality expresses the *p*-value; namely the conditional probability of obtaining the sample distribution (or specific sample statistic) of interest, assuming that H_0 is true. The second (lower) term reverses the elements' positions in the conditional probability, and so appears to provide a measure of "evidence"—namely, the conditional (or "posterior") probability that H_0 is true, given the dataset actually obtained

as the sample. (The notation compares with Berger & Sellke's formulation for this term (1987b): "Pr($H_0|x$)".)

There are three basic ways that the inequality is problematic:

(1) If *p*-values were straightforward measures of evidence for H_0, then the two (upper and lower) terms in Figure 1 should be, if not equal, at least always closely correlated. (We will demonstrate that, in practice, Issue 1 (see next point) can affect both the magnitude and curve of the correlation between those two terms.)

(2) (***Issue 1: Thickness of H_0.***) If *p*-values can provide at least *some* evidence for H_0, then we have to determine the precision expected, when saying "H_0 is true." (We will demonstrate that how much deviation from exact equality is accepted as "good enough to satisfy 'H_0 is true'" can make a big difference.)

(3) (***Issue 2: Mismatched structure of the inequality***.) Even to try correlating the two terms in the Inequality requires that both terms mean the same thing by "probability" of X. (We will explore the impact of this potential incompatibility.)

3. Methodology

Three very similar, but independent, experiments were undertaken. All three followed identical, resampling-based procedures; their only differences are as noted in the third paragraph below Figure 2. In each pass for a given experiment (i.e. each loop of resampling), imagine that a scientist collects a random sample of 40 values from a population, and then tests the hypothesis H_0: $\mu = 100$. The scientist will not know the mean and standard deviation for the *real* population (see yellow boxes in Figure 2); but his or her sample will nonetheless be drawn from this real population, with its parameters—i.e., *not* from the null population. Throughout the experiment, it is *assumed* that the real population distribution is normal, and that σ for the real (sampled-from) population is the same as the (unknown) σ for H_0.

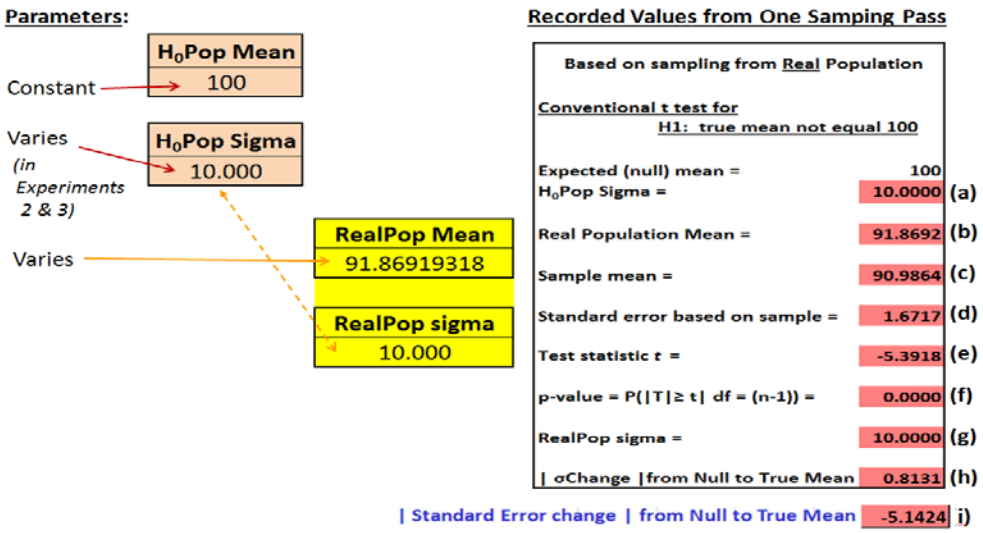

Figure 2: Values Recorded During One Pass in the Experiment

The highlighted values at the right of Figure 2 illustrate one set of possible, recorded outcomes for a single pass of an experiment. At the outset of each pass, the computer

randomly determines the "real" values for the population mean and population standard deviation. (See second paragraph, below, for more details on this selection.) These values are recorded in cells (b) and (g), respectively, in Figure 2. Random sampling of 40 values from a population having those particular parameters is simulated, and the sample's mean (in cell (c)) and standard deviation are calculated. The following are then calculated for the same sample, using conventional calculations: The standard error (in cell (d), based on the sample *s*, divided by $\sqrt{(n = 40)}$; the test statistic *t* (in cell(e), based on (sample mean minus H_0 mean), divided by standard error); and the (two-tail) *p*-value (in cell (f), based on the *t* statistic, and the cumulative distribution function for *t*).

Also recorded (in cells (h) and (i)) are indicators, for that experimental pass, of the *actual* distance between the null mean and the real mean. The units for the distance "standard error change" (in (i)) are analogous to the distance units "standard errors", effectively used for calculating the *t* statistic; namely, the distance from the null mean to the true mean is measured in the units: (true population σ) / ($\sqrt{n}$). This scale makes it intuitive, for example, in cell (i) of the figure, that in the pass illustrated, the real mean was objectively far from the null mean (well over 3 standard errors); so the small *p*-value for this case, in (f), is unsurprising. Note that all experiments are based on two-sided tests.

Two of the three, independent experiments consisted of 20,000 resample passes, each, as described above; a third experiment employed nested resampling, with 5000 passes in the main (outer) loop. As illustrated in cells (a) to (i) of Figure 2, results were recorded for each pass, making it possible to compare obtained *p*-values with the actual, corresponding truth (or near truth, or falsity) of H_0, for each of a large number of passes. In Experiment 1, the real population's sigma (and H_0's sigma) were both held constant, equal to "10", for all resample passes. In Experiments 2 and 3, the real sigma (and H_0's sigma) varied (together) randomly, from 4 to a maximum of 60 (i.e. the coefficient of variation could range, uniformly, from 4% to 60%). In all passes, the value set for the real mean could differ from the null mean by a random amount, between -3 to +3 times the true sigma value. Experiment 3 was more complex, to ensure that the preceding experiments' results were not an artefact of using *t* tests within each pass: Instead of determining the p-value for each main pass based on assuming a t-distributed sampling distribution, a nested, "inner loop" of 3000 repetitions was undertaken, for each "outer loop" pass, to determine the corresponding p-value by non-parametric resampling techniques. (The reduced number of repetitions, compared to Experiments 1 and 2, was used to limit the run time for the total 5000 × 3000 nested replications.)

4. Results

4.1 p-Values and the Truth of H_0

Figure 3 shows directly what many literature sources have tried to calculate or discuss speculatively: the actual relationship between *p*-values and the extent to which H_0 is true. Each point represents the outcome of one pass of Experiment 1, for which the real mean had some arbitrary value, as recorded, and then the *p*-value was calculated traditionally (in relation to null assumptions) based on a random sample drawn from the real population. Presumably, H_0 was more "true" when the standardized distance between the null parameter and the real parameter (for short, the "true distance from the null") approached zero. On that basis, the results shown are largely consistent with the claims of *p*-value proponents: *p*-values do get larger as H_0 gets truer, and do get smaller as H_0

gets more false (i.e. is more appropriate to reject). There is, however, a considerable variance around that correlation.

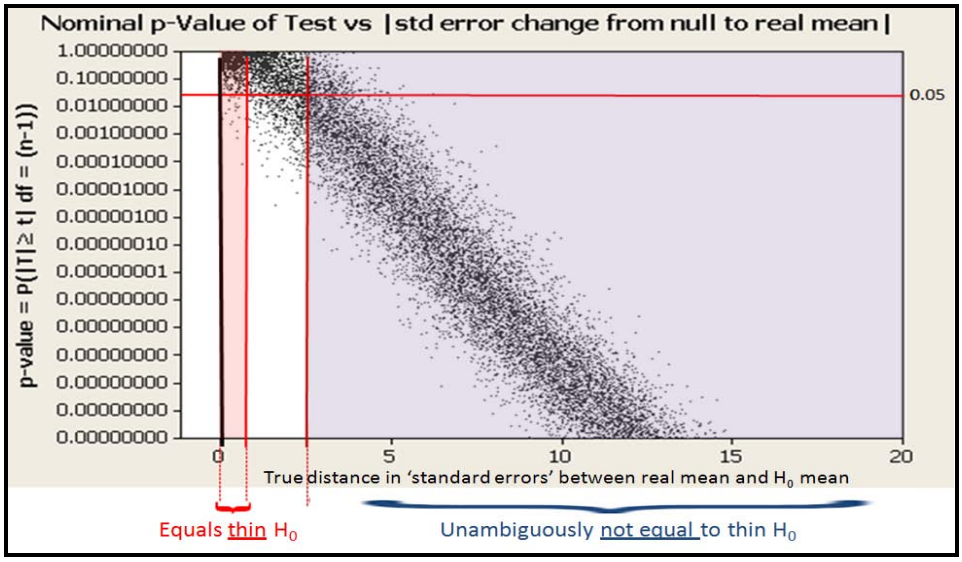

Figure 3: Relation Between *p*-Values and the True Distance of the Real Mean from the Null Mean

At the bottom of the figure, the concept of H_0's "thickness" is introduced. Just as we do not expect, in estimation, that a point estimate for a parameter will be exactly true; similarly, it is not realistic to expect in testing that the true parameter will *exactly* equal the null assumption for the parameter. If H_0 is conceived as "thin", then H_0 is true enough if there is at least a close, if not identical, match with the real parameter. Observe in the figure that when the thin H_0 is true enough in the above sense, then the *p*-value works essentially as advertised: In the red-tinted area of the graph, the true distance was consistent with a thin H_0 hypothesis; note that almost all the corresponding *p*-values are in the non-significant (e.g. larger than 0.05) range. In all cases where the true distance was in the range for a thin H_0, only occasionally (perhaps about α of the time) did the *p*-values by chance have deceptively small values.

Also, if H_0 is "thin", but the true distance from the real to the H_0 mean is *unambiguously* not close, then *p*-values also seem to behave as we'd like: In that region of the figure, tinted blue, observe that, in aggregate, virtually all the corresponding *p*-values were quite low (so H_0 would correctly be rejected as false), whereas for only a small proportion of cases (perhaps about β of them) did the *p*-values happen by chance to have deceptively large values.

Where the simplicity of using *p*-values to judge about H_0 breaks down is in the non-tinted, middle range for true distances, where the standardized true distance from the null is (for this simulation) roughly between 0.5 and 2.5. If a researcher has a "thick" model of H_0, that distance may seem unimportant—i.e. not worth detecting, and possibly designed to remain undetectable, when setting the power for the test. Considering that scenario, observe in the figure the large proportion of tests that would nonetheless reject a (true enough) thick null hypothesis, due to obtaining a misleadingly small *p*-value. This re-characterizes the oft-made complaint about *p*-values, that they can lead to rejecting more true null hypotheses than the nominal rejection-trigger α implies.

On the other hand, suppose the researcher has in mind a thin model for H_0—but it so happens the true distance is in the range of 1 to 2 standard errors. In such cases, the researcher wants the test to reject H_0, but a large proportion of the time, it won't. I.e. it lacks the power. It is noteworthy that while *p*-values are commonly discussed in relation to a reference value α for significance, we see that their performance is equally dependent on considerations of β, power, and the researcher's choice for detectable difference when designing for power.

4.2 The Problematic Inequality—Illustrated by Experiments

4.2.1 The Basic Model

An alternative way to represent the results of Experiment 1 is shown in Figure 4. The two axes directly correspond to the two terms of the Problematic Inequality (see Figure 1). For each pass (resample loop) of the experiment, a specific *p*-value was generated. The *p*-values are mapped on the *x* axis in the figure. The *y* axis shows the conditional probabilities for H_0 being true, given that the *p*-value specified on the *x* axis was generated in a pass of the experiment. The latter probabilities do not have *a priori* magnitudes, or even interpretations. (More on this point, in a following section.) Instead, the y–axis probabilities are frequentist approximations obtained as follows: The results of passes for which the generated *p*-values were closely similar were aggregated; then, for each such group, the proportion of cases in which, as it happened, the null mean was (approximately) true is interpreted as the desired, conditional probability. There are several curves in the figure because, as noted in the previous section, different standards are possible for counting H_0 as true.

The following example illustrates the intended interpretation of the figure. Consider the point labelled "(a)" in the upper right section. This conveys the following information: "Considering H_0 as true when the true, standardized distance from the null to the real mean is at most two standard errors, then for all recorded cases wherein the resulting *p*-value ≈ 0.05, the proportion of such cases for which H_0 was in fact true (by the given standard) was 0.7." In other words, for H_0 of this thickness, 70% of the time when the *p*-value is at the threshold of suggesting (as conventionally interpreted) that H_0 might be false, H_0 was in fact true.

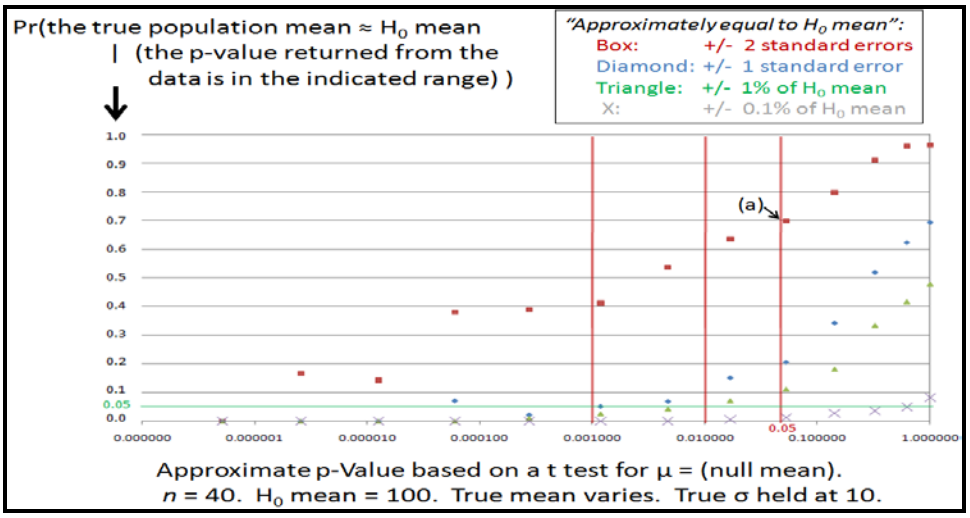

Figure 4: Experimental Illustration of the Figure 1 Inequality

We quickly see that for realistic thicknesses of H_0, the *p*-value seems to generally underestimate the true probability that H_0 is true. So if used as a trigger to reject the null hypothesis, the *p*-value does so too soon (i.e. it starts rejecting at *p*-values that are not yet low enough). Consider when the thickness of the null is just ±1% of the value of the null mean; at that thickness, you would need the power to detect that a true mean equal to 99 or 101 is significantly different from a null mean = 100. Even for this thin of an H_0 (see the triangle marker, two points below the point labelled "a"), over 10% of the time when the *p*-value ≈ 0.05, and so conventionally is at the boundary of suggesting that H_0 might be false, H_0 was really true, by that standard.

Notwithstanding the preceding, the figure refutes what some have suggested, that the *p*-value problem somehow can bottom out, under the right conditions, so the *p*-value gives the right answer about the *y* axis probability. To the contrary, we find that as H_0 gets thinner, there is no barrier to prevent the conditional probability for (H_0, given *p*-value) to fall lower than the *p*-value, itself. For example, in the figure: When the thickness of H_0 is just ±0.1% of the value of the null mean, the conditional probability for H_0, given a *p*-value ≈ 0.05, is only about 0.01. This makes sense: With H_0 so thin, the true mean rarely ever matches the null mean sufficiently—regardless of the *p*-value. This further illustrates, as well, why a no-thickness null is not realistic.

4.2.2 No Apparent Impact for Varying True Population σ

Figure 5 shows the effect of revising Experiment 1, in Experiment 2, with this one change: Just prior to each resampling pass, when the new real mean value is arbitrarily determined, a new true standard deviation value is determined as well. (In Experiment 1, recall, the true σ was always equal to 10; in experiment 2 the true σ can be, for a particular pass, any random value between 4 and 60.) For clarity, the Figure shows only the analyses for a thick H_0; a real mean is deemed "equal" to the null mean if the true distance between them is within ±2 standard errors.

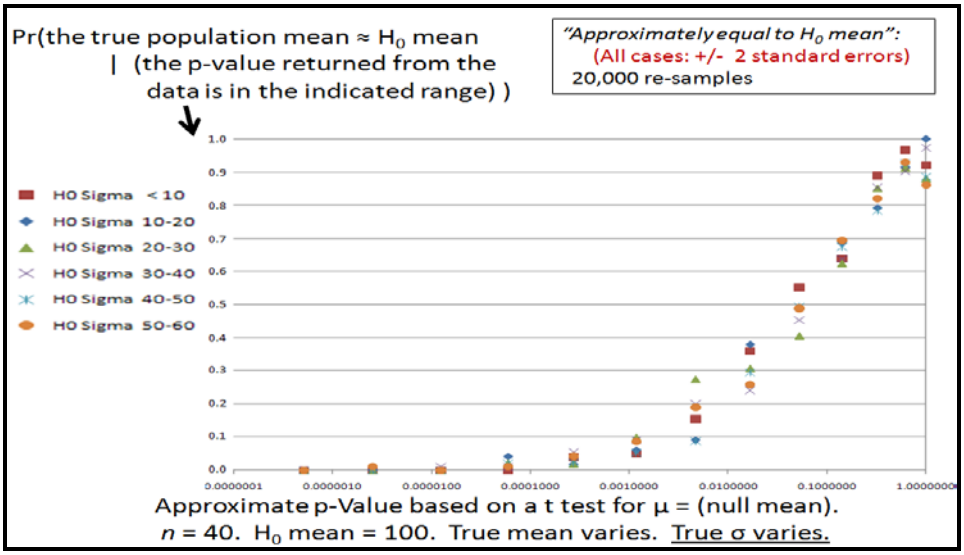

Figure 5: Supplemental Experiment—Varying the Real Population σ

The same basic effects as illustrated in Figure 4 appear to also apply when the real population standard deviation is not fixed at one particular value. While there is variance in exact locations of the points—both within this figure, for different ranges of σ, and

between this figure and Figure 4—there is no clear pattern of difference; for example, for *p*-values close to 0.001, the cases with the lowest true σ's had the lowest conditional probability for H_0, yet for *p*-values close to 0.08, the cases with the lowest true σ's had the highest conditional probability for H_0. Possibly, this variation is simply random noise; consider also that the number of cases in each *p*-value category, for calculating the conditional probabilities for H_0, are reduced here, due to the stratification of cases based on true σ's.

4.2.3 No Apparent Impact for Switching to Nonparametric Generation of p-values

Figure 6 shows the effect of revising Experiment 2 to Experiment 3. As above, each main pass (resampling loop) results in an obtained sample, taken from a population with a randomly re-set real mean and real standard deviation; and the distance δ is measured between the sample mean and the null hypothesis mean. The *p*-value (with respect to H_0) is determined not by a parametric test (using the *t* value), but by the following, non-parametric procedure: A nested loop of 3000 resamples is taken from the null hypothesized population, and the *p*-value (i.e.for that one pass of the outer loop) is the proportion of the nested resamples for which the magnitude of difference between the resample mean and the null mean is greater than or equal to |δ|. As in Figure 4, Figure 6 shows effects based on various selections for H_0 thickness.

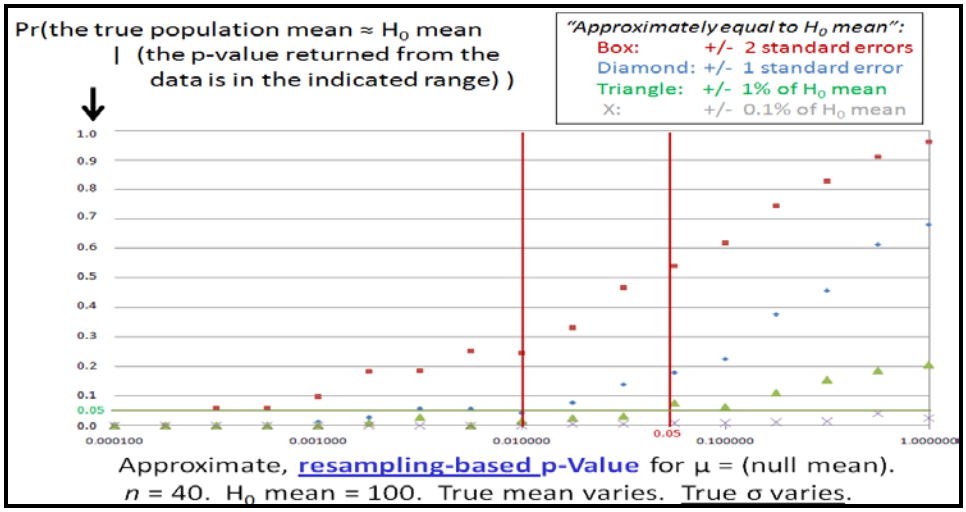

Figure 6: Supplemental Experiment—Non-Parametric Generation of the *p*-Value

Once more, we find the same basic pattern of results as occurred for the *t* test approach, in previous sections. It is clear that the *t* test is not, itself —either specifically, or in virtue of being parametric—causing the replicable pattern for these experiments.

4.3 Mismatched structure of the inequality

The results labelled **4.2.1-4.2.3**, above, concern the issue of H_0 "thickness." Figures 7 and 8 address the second issue that was mentioned in Section 2; namely whether both sides of the inequality in Figure 1 are really both "probabilities" in the same sense. Probabilities for *p*-values are not a problem; the *p*-value is a random variable, with a magnitude determined by the outcome of the hypothesis-test sampling. But the "posterior probability of H_0" is *not* a random variable in the true sense. The true mean has the value it has—*prior* to hypothesis testing; there is no randomness, and no probability. (Compare the confidence level, when estimating a parameter: This too is actually about the expected reliability of a procedure; it is not a probability for the value

of the parameter, itself, which pre-exists the procedure.) Figure 7 clearly demonstrates this situation:

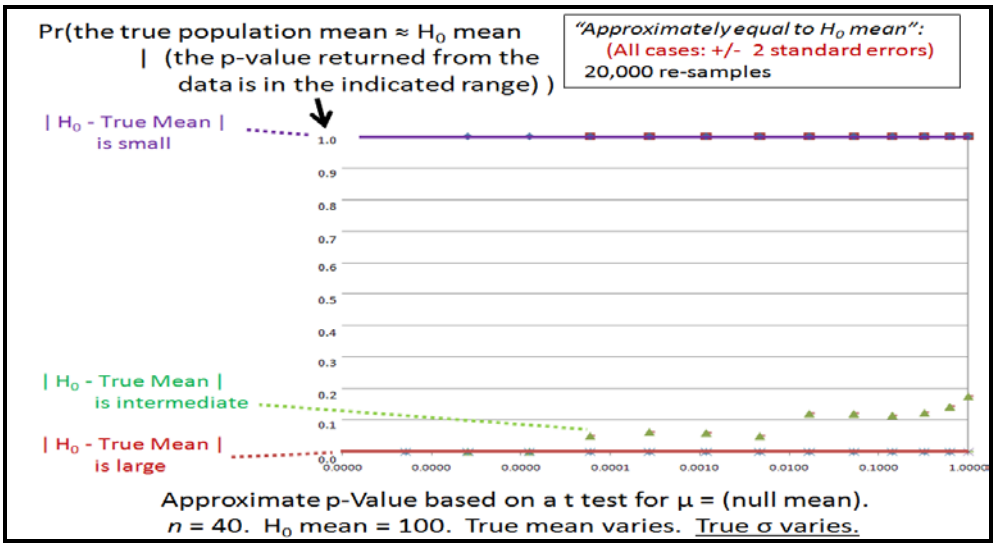

Figure 7: The Non-Random "Posterior Probability of H_0"

In the figure, where the true and H_0 means are very close, in relation to H_0's thickness, then the truth of H_0 is a fact, and its conditional (and absolute) probability = 1.00. When the true and H_0 means are far apart, in relation to H_0's thickness, then the falsity of H_0 is certain, and its conditional (and absolute) probability = 0.00. Observe in both cases that on occasion, the *p*-values calculated in different passes of the experiment were "all over the map" (i.e. small or large); but this had no bearing whatsoever on the certainties.

The points (in green) in the figure appear to be exceptions, but these are actually artefacts, due to aggregating cases where the distances between the real and null were known to be borderline. In this group, by chance, the output *p*-values could range from low to high—both for cases where the null and real means were (barely) "equal enough" and for cases where they were not (quite) equal enough.

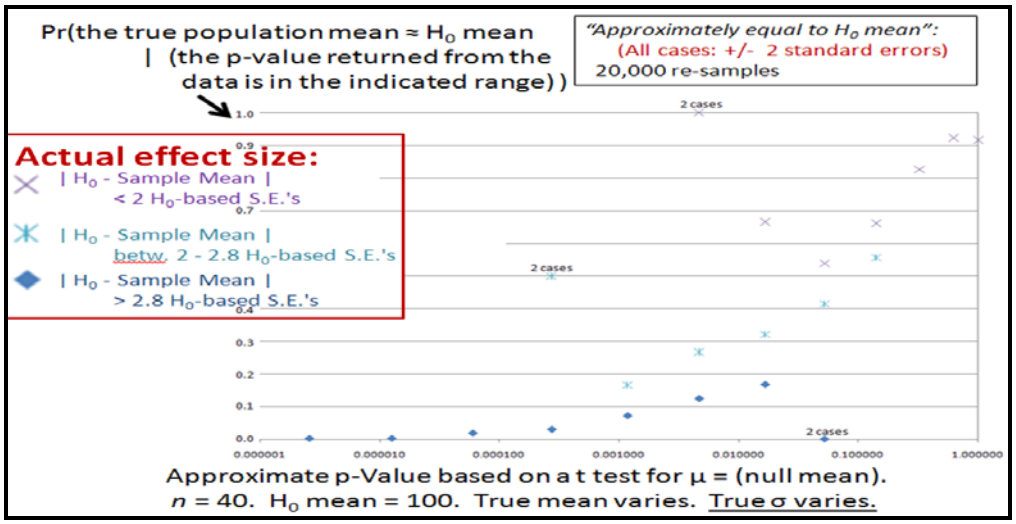

Figure 8: Effect Sizes and the "Posterior Probability of H_0"

In practice, of course, Figure 7 is not very helpful because we do not *know* the real value of the parameter. (If we did, there is nothing to test.) So, where Figure 7 refers to the true distance between the null and real means, Figure 8 uses the best proxy available: the effect size—in this case, the actual distance between the null and *sample mean*. Now, values on the vertical axis are (to the researcher) not certainties. Like confidence levels, the *y* values can be viewed as measures of expected reliability in this sense: If getting a *p*-value similar to a value on the *x* axis is used to trigger a rejection of the null hypothesis (of a specified thickness), then in what likely proportion of such cases would H_0 have in fact been true (and so rejected in error)?

This chart does not solve the *p*-value problem, but it gives a sense of when *p*-values might be informative, and when, to the contrary, they are more likely to be redundant—or, worse, misleading. If the effect size is not at least as large as the specified H_0 thickness (e.g. the detectable distance you used when calculating sample size), or, preferably, a bit larger, then the best guess is to stick with H_0 as likely true—regardless of what *p*-value you obtain. If, on the other hand, the effect size is quite a bit larger than the H_0 thickness, then rejecting H_0 is a safer—even if the *p*-value is not that persuasive.

5. Discussion and Conclusions

As Section 1 demonstrated, there have been many objections raised to using *p*-values for hypothesis tests, yet the convention of using them persists, as evidenced by their prominent inclusion in modern statistics packages. Large numbers of papers have been written on this subject, but most tend to be abstract and/or theoretical or historical. Some of these papers include simulations. Yet, surprisingly, no one has seemed to use simulations to try out, "first hand", what exactly it is that *p*-values do or don't do, in relation to the truth of a null hypothesis. This paper has taken that step; and in this section, we will review and summarize its findings.

5.1 There is a strong monotonic relationship between the order of magnitude of the p-value and the relative likelihood that H_0 is true.

This conclusion is strongly supported by Figure 3. As discussed in Section 4.3, this "likelihood" is not a probability of H_0, strictly speaking, because H_0 is not a random variable. Instead, Conclusion 5.1 is a statement about the long-term reliability of a strategy that would relate *p*-values to the unknown truth-values of null hypotheses. Figure 3 suggests that comparing two sizes of *p*-values (in orders of magnitude) that have been generated conventionally, generally the higher the order of magnitude for the *p*-values the greater the proportion of cases for which it happens that H_0 is really true.

5.2 The variance around the correlation pattern in Conclusion 5.1 is very large.

While Conclusion 5.1 seems supportive for the use of *p*-values, Conclusion 5.2 provides a serious check on that support: The standard error for the impressive regression line between the (log) *p*-value and the associated true distance in Figure 3 is 1.8 (orders of magnitude)—that is, the *p*-value that would be expected, based on the true standardized distance from the real mean from the null, could easily differ from the p-value actually obtained by over two orders of magnitude. If, for example, the *p*-value expected based on the true distance was 0.005, the actually computed *p*-value might range from 0.5 (or higher) down to 0.00005 (or lower). It would be hard to justify a conclusion from such imprecise findings.

5.3 The information in the p-value is highly dependent on the "thickness" of H_0

The thickness of H_0 might be determined based on practical or clinical considerations, regarding what difference from exact equality might truly matter. Logically, this could be the figure used when estimating a minimum sample size: When calculating the power for a test, one specifies a minimum distance $\pm\delta$ that needs to be detectable. The clear implication, once the test results arrive, is that if the observed effect size is *less than* $|\delta|$, then this small, non-differentiatable distance from the null should be counted as "equal enough", or "effectively equal".

Figure 4 demonstrates that any attempt to estimate a probability (or "likelihood") for H_0, given a *p*-value, is highly dependent on the posited "thickness" of H_0. In the example, the likelihood that H_0 is true, given a *p*-value of 0.05, might be anywhere from 70% (for the thickest null) to virtually zero (for a very thin null). There is no evidence of any lower bound for this likelihood. (At some thickness of H_0, presumably, the null's likelihood would be exactly equal to the *p*-value; but this paper does not examine where that point occurs, or whether it can be generalized.)

This sometimes dramatic swing of the *p*-value's possible impact or meaning can be viewed by relating the *p*-value to α. In hypothesis tests, α is supposed to represent the Type I error rate (i.e. the probability of mistakenly rejecting a true null hypothesis) that one is willing to accept, in following the testing procedures. Supposedly, by conventional models, one can ensure α by the algorithm of finding the *p*-value, and then rejecting H_0 only when the *p*-values reaches or falls below α. (E.g. if one's standard is $\alpha = 0.01$, do not reject H_0 for any *p*-value greater than 0.01.) In point of fact, the procedure described does *not* preserve the effective α for one's test. If using a *p*-value algorithm to decide whether or not to reject H_0, then (all else being equal):

a) For <u>thick</u> H_0's: (the *effective* α) > (nominal α);

b) For <u>very thin</u> H_0's: (the *effective* α) < (nominal α)

This follows because in (a), *p*-values lead to rejecting too many H_0's that are true, so the risks of Type I error (i.e. α) are greater than acknowledged; and in (b) *p*-values fail to reject virtually any H_0's (whether true or false), so the risks of rejecting true H_0's (i.e. α) is less than anticipated.

5.3.1 The reported effects seem independent of (a) the real value of σ (in relation to μ), and (b) the method used to obtain the p-values

These points were illustrated in Figures 5 and 6. Obviously, not all variations could be tested. Yet, the effects clearly persisted through a wide divergence of possible σ values, and from the parametric to the non-parametric procedures to generate the *p*-values.

5.3.2 p-values cannot tell you the "probability that (on this occasion) H_0 is true (or false)"

This is because *nothing* can tell you this. The real value of the parameter is not a random variable. At best, *p*-values can be used as part of a regimen that, over time, tends to be reliable when pointing to what *may be* true on particular occasions.

6. Recommendations

Based on the findings of this paper, the author would offer the following recommendations to those who are considering the use of *p*-values in the process of drawing, and/or stating, statistical conclusions:

1. Do not necessarily give up on *p*-values, but do keep clear on what they do—and do not—tell us, and under what conditions

2. At the very least, provide (or look for) this supplementary information:
   a. The actual effect size, *and*
   b. The thickness of H_0, i.e. the minimum difference that is detectable or, based on the context or purpose of the research, cared about.

p-values have the advantage of generally being readily accessible (e.g. through publication in articles, or by generating with software), and they do convey information; yet they cannot stand alone. Based on the results of experiments described in this paper, the author finds that combining the three elements named in the recommendations provides the best chance of "guesstimating" the net likelihood of H_0's being true. No one test can give certainty; but it is felt that using these guidelines will help to minimize error over the long term.

This closing example shows the intention of the guidelines that are recommended:

Suppose you run a multiple regression, and are examining the contributions of the independent variables to the overall regression. Most software will generate *p*-values, based on *t* tests, for each *x* variable coefficient. For each variable a null hypothesis is implied that its coefficient simply equals zero; i.e. that changes in *x* have no impact on the dependent variable. Before considering the size of the *p*-value, ask: What is a reasonable thickness for this coefficient's null hypothesis? As shown in Figure 4, small *p*-values can be easily obtained if the nulls are very thin (for example, if the mean of the dependent variable is 2000, yet you claim that an *x* coefficient equalling 1.2 is "not equal enough" to a null of zero). But inquire whether such small contributions to the regression would be practically meaningful; and whether the *t* test even has the power to discriminate differences that small? On the other hand, if the null coefficient has some reasonable thickness, *and* the coefficient actually obtained is notably larger than that baseline, then, as depicted with the diamond symbols in Figure 8, you may be onto something. Throw in a *p*-value that is very small (preferably, much lower than the usual "0.05"), and now (assuming no confounding, etc., are involved) the odds do look reasonable that the (thick) null hypothesis is probably false; i.e. the particular *x* variable is likely making an actual contribution to the regression.

This example shows the spirit of the recommendations. Although more sophisticated attempts to transform *p*-values into something better have been tried, and several are cited in the references, there is really no way to make the indeterminate (i.e. what is the true value of the population parameter) determinate—certainly not in a single sample. But by combining the three proposed elements in making one's assessment, a reasonable basis for future research or action can often be obtained.

Acknowledgements

My special thanks Milo Schield, who kindly invited me to present an earlier version of this paper in a Statistical Education section at the 2010 JSM, and who also provided helpful comments on my Proceedings version of the paper. Thanks also to those who attended my paper at the conference and who provided their comments and encouragement..

References

Bellhouse, D.R. (1993). Invited commentary: *p* values, hypothesis tests, and likelihood. *American Journal of Epidemiology*, 137, 497-499.

Berger, J.O. & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3), 317-352.

Berger, J.O. & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *Journal of the American Statistical Association*, 82(397), 112-122.

Berger, J.O. & Sellke, T. (1987). (Rejoinder to comments on "Testing a point null hypothesis: The irreconcilability of *p* values and evidence".) *Journal of the American Statistical Association*, 82(397), 135-139.

Berger, R.L. & Hsu, J.C. (1996). Bioequivalence trials, intersection—union tests and equivalence confidence sets. *Statistical Science*, 11, 283-319.

Brown, L.D. (2000). An essay on statistical decision theory. *Journal of the American Statistical Association*, 95(452), 1277-1281.

Browne, R.H. (2010). The *t*-test *p* value and its relationship to effect size and P(X>Y). *The American Statistician*, 64(1), 30-33.

Casella, G. & Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82(397), 106-111.

Chow, S.L. (1988). Significance test or effect size? *Psychological Bulletin*, 103(1), 105-110.

Chow, S.L. (1989). Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin*, 106(1), 161-165.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences, 2nd edition.* Hillsdale, NJ: Lawrence Erlbaum Associates

Donahue, R.M.J. (1999). A note on information seldom reported via the *p* value. *The American Statistician*, 53(4), 303-306.

Evans, S. (2009). Noninferiority clinical trials? *Chance*, 22(3), 56-62.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193-242.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin*, 106, 155-160.

Goodman, S.N. (1993a). *p* values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137(5), 485-496.

Goodman, S.N. (1993b). Author's response to "Invited commentary: *p* values, hypothesis tests, and likelihood." *American Journal of Epidemiology*, 137(5), 500-501.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine,* 130, 1005-1013.

Hoekstra, R., Kiers, H.A.L., Johnson, A., & Groenier, M. (2006). Problems when interpreting research results using only *p*-value and sample size. *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7)*; online, at URL: http://www.redeabe.org.br/ICOTS7/Proceedings/PDFs/ContributedPapers/C437.pdf

Hubbard, R. & Armstrong, J.S. (2005). Why we don't really know what "statistical significance" means: A major educational failure. Published online, at URL: http://marketing.wharton.upenn.edu/ideas/pdf/Armstrong/StatisticalSignificance.pdf

Hubbard, R. & Bayarri, M.J. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical teaching. *The American Statistician,* 57(3), 171-178.

Hung, H.M.J., O'Neill, R.T., Bauer, P., & Kohne, K. (1997). The behavior of the *p*-value when the alternative hypothesis is true. *Biometrics*, 53, 11-22.

Marden, J.I. (2000). Hypothesis testing: From *p* values to Bayes factors. *Journal of the American Statistical Association,* 95(452), 1316-1320.

Meehl, P.E. (1967). Theory testing in psychology and in physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, 20, 641-650.

Panagiotakos, D.B. (2008). The value of *p*-value in biomedical research. *The Open Cardiovascular Medicine Journal,* 2, 97-99.

Robinson, A.P., Duursma, R.A., & Marshall, J.D. (2005). A regression-based equivalence test for model validation: shifting the burden of proof. *Tree Physiology,* 25, 903-913.

Robinson, A.P. & Froese, R.E. (2004). Model validation using equivalence tests. *Ecological Modelling*, 176, 349-358.

Rogers, J.L., Howard, K.I., & Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin,* 113(3), 553-565.

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

Simon, S. (2007). What is a *p*-value? Published online; accessed 5/3/2010 at URL: http://childrenshospitalkc.com/stats/definitions/pvalue.htm

Wellek, S. (2003) *Testing Statistical Hypotheses of Equivalence.* Boca Raton, FL: Chapman & Hall/CRC.

Ziliak, S.T. & McCloskey, D.N.. (2009). The cult of statistical significance. *JSM Proceedings*, 2302-2316.