The pointlessness of the point null *p*-value

> *“Small wonder that students have trouble [understanding* p*-values]. They may be trying to think.”*[1]

Small wonder indeed. Deming was taking issue with the mangled definitions of the *p*-value that are littered throughout statistics textbooks. But textbooks may have no choice but to mangle them. It is hard to explain to students why they ought to calculate *p* if they are told it is the probability of observing a test statistic at least as extreme with respect to the null hypothesis as the one that was observed. Much easier if students are told (incorrectly) that *p* is the likelihood that the observations were due to chance.

I argue that the reason why the justification for the *p*-value is elusive to students is that there is no justification for *p* at all. *p* is in practice the probability of observing extreme results under the hypothesis that there is no difference at all between groups in the population with respect to some variable. This is never literally or approximately true of the actually-existing population from which the relevant sample is drawn. The best argument for point *p* is that it is equivalent to the likelihood that there is a casual relationship between variables under consideration. But that is rarely (if ever) the case. And if point null *p*-values do not say anything about whether there is a causal relationship between variables of interest, there is no reason to use them.

The point *p*-value

[1] W. Edwards Deming, "On Probability as a Basis for Action," *American Statistician* 29, no. 4 (1975): 150.

Researchers (including the aforementioned textbook writers) often define the *p*-value incorrectly, so it is worth clarifying the definition. The *p*-value is the probability of observing a test statistic that was at least as extreme as the test statistic that was actually observed under the null distribution. The null distribution is the sampling distribution of the test statistic under the null hypothesis, which is the hypothesis the researcher wishes to test. A test statistic is more extreme than what was observed if it is larger than the parameter specified by the null hypothesis.[2]

In practice, the null hypothesis has two other features. First of all, the null predicts that there is no difference between groups under observation.[3] Second, the null is almost always a point null. It predicts that the population parameters of the groups are *exactly* equal.

I will dwell a bit longer on the point null. Suppose a political scientist was interested in whether wealthier Australians were more or less likely to vote for the Australian Labor Party in the most recent federal election. The political scientist collects survey data on the total personal income of Australian voters and about how they voted. Following standard research practice, they conduct a significance test of the hypothesis that wealthy voters are as likely to vote Labor as not-so-wealthy voters. They of course do not test the hypothesis that the voting patterns of the wealthy and the not-so-wealthy are *about* the same. No, our political scientist only cares about the hypothesis that Australians who make $50 000 a year are exactly as likely (all-else-equal) to have voted Labor as

[2] Though this is only true for observations that are ordered. For discussion of this issue, see Jason Grossman, "Statistical Inference: From Data to Simple Hypotheses" (Australian National University, 2011), 191-6.

[3] Researchers often claim that the null hypothesis is by definition the hypothesis of no difference. But Fisher defined the null as whatever hypothesis the researcher wished to test. For a round-up of Fisher's statements on the subject, see Leonard J Savage, "On Rereading R. A. Fisher," *Annals of Statistics* 4, no. 3 (1976): 471-2.

Australians who make $100 000 a year. If, in the population, there is even one more ALP voter in the $50 000 group than there is in the $100 000 group, then the political scientist will reject the null if their sample is large enough.[4]

The point *p*-value is a curious statistic. The straightforward interpretation of the point null is that it describes a hypothetical parameter of the actual population (e.g. t=0). But in that case the point null is virtually always false. There is almost no chance at all that a discrete variable with many categories (e.g. GDP) will take on any value in particular, including the point null value. And the chance that a continuous variable will take on a particular value is *exactly* zero. Suppose that rejecting a null hypothesis was not the same as reckoning the null to be false. Even so, it is not clear why a researcher who was certain that a hypothesis was false would bother to test it at all. And yet this is what researchers appear to be doing when they calculate point *p*-values.

There is no mathematical reason why *p*-values have to relate to point nulls. The political scientist researching Labor voting could test the hypothesis, for example, that the relevant test statistic (say, t) was roughly equal to the null parameter (say, $-0.1 < t < 0.1$). There is a good chance that the population parameter would fall between these values. Of course, it is not as easy to use statistical software to calculate *p*-values for ‘small-interval nulls’ as it is to calculate point *p*-values. It is not impossible though. And some researchers have proposed computationally simple methods of calculating small-interval *p*.[5] I can only think of two reasons other why researchers would choose to calculate small-interval *p*-values rather than point *p*-values (aside from laziness). The first

[4] This problem is similar to Lindley's Paradox, which is well-known among theoretical statisticians. See D V Lindley, "A Statistical Paradox," *Biometrika* 33, no. 1/2 (1957).

[5] R C Serlin and D K Lapsley, "Rationality in Psychological Research: The Good-Enough Principle," *American Psychologist* 40, no. 1 (1985): 81-2.

reason, and the focus of this paper, is that there are other philosophical reasons why point *p* is informative. The second reason is that point *p*-value calculations are ‘uniform’ across studies and so are more universal or ‘objective’ than small-interval *p*-values. I will raise some worries about the ‘universality’ argument at the end of the paper.

It is worth mentioning that there are significance tests of point nulls that do not yield *p*-values. Bayesians for example often calculate ‘Bayes factors,’ the posterior of the point (or ‘sharp’) null hypothesis where the prior of the null is a half.[6] My arguments against point *p* probably apply to point null significance tests generally. But they may not. In any case, nearly all studies in social science and medicine include point *p*-values. If point *p* is unjustified, then many scientific ‘discoveries’ based on *p*-values are in jeopardy. And at least some may turn out not to be discoveries after all.

Bad arguments for point *p*

Point p *is roughly equal to small-interval* p

Most researchers would realise that the point null is never literally true of the actual population. Instead they would understand themselves to be investigating whether the point null is roughly true. The thought is that, if the value of point *p* is low, the value of nearly-point *p* would be almost as low. For two detailed versions of this argument, consult the edited collection of papers *What If There Were No Significance Tests?*[7]

[6] Robert E Kass and Adrian E Raftery, "Bayes Factors," *Journal of the American Statistical Association* 90, no. 430 (1995).

[7] Paul E Meehl, "The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions," in *What If There Were No Significance Tests?*, ed. Lisa L Harlow, Stanley A Mulaik, and James H Steiger (New York: Routledge, 1997); David Rindskopf, "Testing 'Small,' Not Null, Hypotheses: Classical and Bayesian Approaches," ibid.

There are at least two reasons to doubt that point *p* is equivalent to small-interval *p*. First of all, the claim that *p*-values for point nulls are equivalent to *p*-values for small-interval nulls is a mathematical claim. As such, it needs to be mathematically proven. To my knowledge, no proof has been published.

But grant that significance tests of point nulls get roughly the same results of significance tests for extremely-small-interval nulls (e.g. $-10^{-8} < t < 10^{-8}$). On its face, that’s at plausible claim. The second problem is that it is not clear how wide the interval could get before point *p* ceased to be a useful approximation of small-interval *p*. Is the *p*-value for the hypothesis that Labor voters earn the same income on average as Liberal voters close to the *p*-value for the hypothesis that the average Labor voter earns up to $50 less? What about up to $500 less? $5000? How would a researcher know without a mathematical proof?

Moreover, point *p* may be useful approximation of small-interval *p* for one hypothesis but not for another, even if the measurements for both hypotheses yielded the same test statistic. If a political scientist conducted a census that showed women were only a bit more likely (say, five per cent more likely) than other Australian voters to have voted Labor at the federal election, they might have accepted that gender affects voting behaviour. They may have been less inclined to accept that eye colour affects voting if Labor voters were five per cent more likely to have blue eyes. The interval for the small-interval null would be wider for the eye-colour study than for the gender study. Political scientists require less extreme evidence to be convinced that gender influences voting behaviour as there are good theoretical reasons to expect it would. There would be no reason to think that point *p* would in general approximate small-interval *p*. This is not to say point *p* circumvents this problem. Researchers also generally require extreme point *p*-values before they accept (or fail to reject) implausible hypotheses. What I am suggesting is that extreme point *p*-values will rarely approximate small-interval *p*-values as interval nulls vary in size. This is especially true for implausible alternative hypotheses where the interval nulls are wider and the *p*-values for a set of observations are (compared to point nulls) larger.

Defenders of point *p* might argue that the way to test wider-interval nulls is to shrink the critical region. Test statistics that are more extreme for point nulls are also more extreme for interval nulls. So, researchers might require that studies reject the null only if $p < 0.005$, or $p < 0.00000001$, or other small values of *p*. This just pushes the problem back. How small does point *p* if small-interval *p* is to be significant? And for intervals of what size? Researchers who want to argue that they have been testing small-intervals null rather than point nulls have their work cut out for them.

Point p *is the 'probability that the samples were drawn from different populations'*

I do not think this argument is worth much attention, but it is widely cited, so I will address it. According to Hagen, the point null says that the samples of the groups under consideration 'were drawn from the same population.'[8] Taken literally, this argument is obviously false. A population is just a collection of objects under consideration.[9] The group samples are always samples of a single population.

[8] Richard L Hagen, "In Praise of the Null Hypothesis Statistical Test," *American Psychologist* 52, no. 1 (1997): 20.

[9] Maurice G Kendall, *Advanced Theory of Statisics*, 5th ed., vol. 1 (London: C Griffin, 1987).

Perhaps Hagen understands a population to be a group of objects with a parameter. For Hagen, the point null predicts that samples from different groups will come from populations with identical parameters. In other words, the groups will not systematically differ. But what does ‘systematically differ’ mean here? It could mean that the groups in the finite population differ. This is however always true. Hagen could also take ‘systematically’ differ’ to mean that there is a non-random process that causes the groups differ in a variety of situations. In that case significant values of point *p* would count as evidence against the hypothesis that there were no systematic differences between groups of interest. This is not straightforwardly true. If it were, significant results would always prove that there were systematic differences the data. This is never the case.

One-tailed tests are not tests of point nulls

Perhaps only *p*-values derived from two-tailed tests are susceptible to objections to point *p*. One-tailed tests may be immune to these objections because they ostensibly test the hypothesis that the parameter is greater than or less than a particular value. The one-tailed null hypothesis is not a single value but rather a half-open set. And in most cases it is highly likely that the one-tailed null set will contain the parameter. So, it would make sense to calculate *p*-values for one-tailed tests as one-tailed nulls are likely to be true.

But one-tailed tests are in practice equivalent to tests of point nulls. The kinds of one-tailed tests that researchers actually conduct are identical to two-tailed tests except that they distribute the mass of the critical region on one side of the null sampling distribution rather than across both sides. The null distribution is always what the sampling distribution would be if the point null were true. This is a problem because it means that other members of the null set do not feature in the calculations. As Salsburg points out, one-tailed tests are only appropriate when effects only go in one direction.[10] Formally, one-tailed tests are only sensible where, for $H_0 = \{ \theta :] -\infty, \theta_0] \}$, $\Pr (\theta < \theta_0) = 0$. This is always literally false, for the same reasons that the point null in the two-tailed test is always literally false. There is always a chance, however small, that the parameter will be greater than or less than θ_0. Defenders of one-tailed *p* have to explain why it is useful to test a null hypothesis for which the null set only contains a single value, even though the null is never true in the actual population. They are faced with the same problem as defenders of two-tailed point *p*.

The superpopulation argument is better

The superpopulation argument says that the point *p*-value allows researchers to test the hypothesis that there is no interesting causal link between variables of interest. The argument starts from the premise that the point null hypothesis of no difference predicts that there will be no relationship between the variables under investigation in the ‘long-run,’ or in the ‘superpopulation.’

I will explicate the meanings of ‘long-run’ and ‘superpopulation’ presently. But I will start by explaining what it means for a set of processes to have an ‘interesting causal link’ and why the concept matters. Researchers tend to distinguish between causal and spurious relationships, or relationships that are 'due to chance' (whatever that means) and relationships that are not. There is for example a relationship between pool drownings and films Nicolas Cage has

[10] David Salsburg, "Use of Restricted Significance Tests in Clinical Trials: Beyond the One- Versus Two-Tailed Controversy," *Controlled Clinical Trials* 10 (1989).

appeared in from 1999-2009.[11] Researchers often call relationships like these non-causal or ‘due to chance.’ But there has to be something that caused this relationship. What researchers seem to mean when they say there is a causal relationship between some processes is that there is a causal relationship specified by the theory or model the researcher is testing. When they say that an effect isn't causal (like the case of drowning and Nicolas Cage) what they're really saying is that they don't know of any interesting theory that would explain the effect.

The statement that there is no interesting explanation for an observed relationship is equivalent to the statement that the observed relationship would all-else-equal fail to obtain in the 'long-run.' It obviously makes sense to talk about long-run probabilities with respect to infinite populations. When I roll a six-sided die arbitrarily many times to determine whether it is fair, I am testing the theory that the long-run probability of rolling each face is one-in-six. But it is less clear what it means to say, for example, that rich and poor voters are equally likely to vote Labor in the long-run. After all, there are only so many Australian elections that will occur. The sample of rich and poor is finite. Instead, the 'long-run probability' of an event occurring for a finite population should be interpreted as the probability of the event occurring in the 'superpopulation.' The superpopulation contains all possible populations (and therefore all possible samples), including the actual one, in which the researcher is interested.

Superpopulation modelling is an approach to model-building that assimilates finite population statistics with predictive statistics. It treats the finite

[11] Tyler Vigen, "Spurious Correlations," http://www.tylervigen.com/spurious-correlations.

population, the source of the sample under consideration, as a random sample from a superpopulation.[12] Other possible worlds that are relevantly similar to the actual one may have different populations in which the parameter is different (say, that lower-income earners were *less* likely to vote Labor). The point of superpopulation modelling is to make inferences about the hypothetical populations from a sample of the actual finite population.

If there is an interesting causal relationship between some variables, then the variables will correlate in the superpopulation (taking into account irrelevant factors). To see why this is so, consider what it would mean if there was a relationship between the variables in the finite population but not in the superpopulation. This would mean that certain facts obtained in the finite population that caused some variable to influence the others but which do not generally obtain across other populations in the superpopulation. Perhaps Nicolas Cage happened to appear in more movies in hotter years where people spent more time swimming in pools. The thought is that weather patterns could just as easily have been cooler without harming Cage’s employment prospects. In other words, the effect of Cage appearances on drownings will 'cancel out' across populations (or, ‘in the long run’). And when this happens, the point null will be true in the superpopulation. Where the point null is not true in the superpopulation, the processes under scrutiny are casually linked in similar circumstances to the ones in which we find ourselves. This is useful to know. It means that in future situations similar to the situation in which the data was collected, we can condition our behaviour on there being a causal link between the processes.

[12] T M F Smith, "The Foundations of Survey Sampling: A Review," *Journal of the Royal Statistical Society. Series A (General)* 139, no. 2 (1976): 189.

This is not to say that effects in the superpopulation are always interesting or that they always support the theory the researcher is testing. There may be other facts which generally obtain that the theory we are testing does not specify. For example, racial differences between high- and low-income earners may explain why low-income voters tend to vote for Labor. The way to solve this problem is to ensure that the model we're using to generate the relevant estimates is fully specified or that the experiment we're conducting controls for as many irrelevant factors as practicable.

Of course, researchers do not always need inferential statistics to determine whether 'noise' explains a correlation they have discovered in the finite population. It is for example hard to see what interesting model could explain the link between Cage's career and drownings. But for cases where it seems like an interesting model might explain a correlation, researchers need to investigate whether the correlation is also present in the superpopulation.

The best justification for tests of point nulls is that the point null is equivalent to the hypothesis that there is no relationship in the long-run between variables of interest. When a researcher rejects the point null and concludes that a relationship they observed could not have been 'due to chance,' the most charitable reading of their conclusion is that there is at least one interesting model that explains the relationship. Another way of saying this is that, all-else-equal, the relationship would have been observed in the superpopulation.

Where point p *comes in*

Point *p* may be the probability of observing extreme results where the sampling distribution of the superpopulation follows the point null distribution.

Significance tests becomes tests of the hypothesis that there is no difference between groups in the long run. It is not correct to say point *p* just is the probability of getting extreme results under the superpopulation point null. The sample under consideration is drawn from the finite population, not the superpopulation. The question is whether point *p* for the finite population is close enough to whatever would be point *p* for the superpopulation to justify inferences made from finite-population *p*. The answer is plausibly 'yes.'

Even Bayesians might like the superpopulation argument for point *p*. Bayesian significance tests (like the Bayes factor) calculate the posterior of the null in light of the observations under consideration. To my knowledge, there are two rigorous studies that investigate whether *p*-values and Bayesian procedures lead researchers to similar conclusions (or induce them to behave in similar ways). Jim Berger and Sellke find that *p*-values *underestimate* the likelihood that the null is true.[13] But, as Casella and Roger Berger point out, Berger and Sellke's are only true if the prior of the null is a half. They argue that lower priors may generate *p*-values that are closer to the posteriors of the null.[14] All four authors are far too generous to point *p* if it is understood to be a test of the point null for the finite population. The prior of the null in the finite population case is about zero. But Bayesian significance tests might approximate point *p* if point *p* is interpreted as the likelihood of extreme test statistics taken from the superpopulation. The prior of the point null for the superpopulation is just the prior likelihood that there is a causal relationship between variables of interest. And the likelihood that two variables are causally linked in an interesting way is plausibly quite high.

[13] James O Berger and Thomas Sellke, "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence," *Journal of the American Statistical Association* 82, no. 397 (1987).

[14] George Casella and Roger L Berger, "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," ibid.: 111.

Bayesians might not concede that *p*-values are worth calculating even in light of the superpopulation argument for point *p*. They may maintain that Bayesian statistics are better. But if the superpopulation argument goes through, Bayesians need not regard inferences based on *p*-values with great suspicion.

But the superpopulation argument isn't much better

The argument does not go through. At least it's not clear it does, for two reasons. First, there are issues with superpopulation modelling more generally that defenders of point *p* need to resolve. Superpopulation models require that the finite population was randomly sampled from the superpopulation. It is not clear that this is the case. Consider again the study on income and voting in Australia. What random processes caused this particular version of the Australian electorate to emerge in 2016? Sociological and economic processes that set up Australian society such that people who earn a certain amount understand their preferences in a particular way? Geological processes that cause the Australian continent to have a certain size and location with a particular endowment of natural resources? A finite population is only a random sample of a superpopulation the probability that it will be 'drawn' is known. But no one knows or could know the probability of drawing a particular finite population. Moreover, some kinds of populations may be more likely to get 'drawn' than others. It is for example unlikely that there would arise a version of Australian society where voters never took their material interests into consideration when they were voting.

Suppose the finite population was randomly sampled from the superpopulation. It is unlikely that point *p* for the finite population would be anywhere near point *p* for the superpopulation. The way to calculate superpopulation point *p* would involve estimating a multilevel model with the finite samples as level 1 units and the finite populations as level 2 units. It seems impossible to estimate this model as the variance of the superpopulation is unknown. Even if we did know the superpopulation variance for a particular problem, one would expect it to be extremely large. After all, the size of the sample drawn from the superpopulation is always one – the population in the world we live in. The total standard error of the finite population and of the superpopulation is going to be extremely high. So, the *p*-value for the multilevel model will be way higher than the *p*-value for the finite population model. Point *p* isn't even close to the probability of observing extreme results if there is an interesting effect in the data.

The problems facing the superpopulation argument for point *p* are problems facing superpopulation modelling in general. If researchers who have used superpopulation models in other contexts have resolved these issues, maybe they can be resolved in relation to point *p*. But researchers using point *p* should not assume that this is the case. What it means for a finite population to be 'randomly sampled' is at least partly a philosophical issue that involves grappling with the meaning of 'randomness.' Moreover, it is hard to see how the large-variance problem could be dissolved. It could be that multilevel superpopulation models are useful even though they have large standard errors. But the standard errors are so much larger for the superpopulation than for the finite population that point *p* is a poor estimate of the significance of the estimates in the superpopulation model.

In any case, suppose (optimistically) that the variance of the superpopulation model could be reduced such that the *p*-values for the finite population and superpopulation models were similar. Frequentists would be happy with the superpopulation argument. But Bayesians might not be. Recall that Berger and Sellke show that point *p* is higher than the posterior of the null for many reasonable priors distributing half mass under the null parameter.[15] What this suggests is that the prior of the null has to fall within a relatively narrow range of values (though it is not clear how narrow is the range). Casella and Berger counter that point *p* approximates the posterior of the null for a range of priors. But their conclusions only apply to one-tailed tests, including one-tailed tests where the effect may go in either direction. And, as discussed in previous section, one-tailed tests in practice assume that the effect can only go in one direction.

It is unlikely that enough priors of the superpopulation null will fall into the range of priors for which point *p* is sensible. The prior of the superpopulation null is the prior probability that there is an interesting causal relationship between variables of interest. But we do not have a general theory of causation that would yield the probability than any two processes in the universe were causally linked in an interesting way. And there is no use searching for one either. A researcher always knows more about processes under consideration than that they are processes. So, their priors will change accordingly. The trouble is, the priors for different null hypotheses will vary widely. The prior of the hypothesis that there is no interesting link between gender and voting is much lower the than the prior of the hypothesis that there is no interesting link between voting and eye-colour. The prior for gender and voting is also likely to

[15] James O Berger and Thomas Sellke, "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence," ibid., no. 397.

be different from the prior for the hypothesis that democracies don’t go to war with each other. Which prior is higher? How would a researcher decide? Even assuming the problems with superpopulation modelling are solvable, at best some *p*-values will turn out to match the posteriors of the superpopulation null. It would be immensely difficult to determine for which *p*-values this is the case. Bayesians might prefer to disregard conclusions based on *p*-values altogether.

Conclusion

If not point *p* then what else? The most obvious alternative to the point p-value is the small-interval *p*-value. Researchers could test interval nulls where the width of the interval corresponds to the ‘best practice’ in the relevant field. The trouble is, there is no best practice anywhere because interval nulls are not common anywhere. The alternative is for researchers to use their judgement to construct interval nulls. But one of the supposed virtues of the *p*-value is that it is universal. There is not much disagreement about how *p*-values ought to be calculated for most sets of observations among empirical researchers (though this is not to say there is agreement among statisticians). The universality of the *p*-value means that comparing results across studies is ostensibly straightforward.

Universal measures are also arguably less susceptible to corruption. One can imagine unscrupulous pharmaceutical companies stretching and shrinking intervals for null hypotheses until their study showed that their new drug significantly reduced the incidence of some ailment. I submit that we should not use tests that are completely uninformative, no matter how widely-used they are. It is also worth mentioning that tests of interval nulls are probably *harder* to corrupt than tests of point nulls. Corrupt researchers are generally interested in finding significant effects (e.g. of a new drug) and significant results are less likely for interval null tests. Of course, unscrupulous researchers may also prefer to use significance tests that are less likely to generate positives (for example, tobacco companies funding research that ‘shows’ there to be no statistically significant relationship between smoking and lung cancer). So, it is not clear whether small-interval *p* is more or less conducive to unscrupulous research than point *p*. I am just casting raising doubts about the importance of ‘universality’ in the selection of statistical techniques, at least with respect to point *p*.

And why stop at abandoning point *p*? Why not abandon *p* altogether? It is not hard to find other objections to point and interval *p*-values. And there are many other methods available that researchers could use to determine whether some effect is worth their attention. Bayes factors, confidence intervals, credible intervals, maximum likelihood estimation and many other techniques can all arguably fill the gap that *p*-values leave behind.

Having abandoned point *p,* researchers would also have to change how they read existing studies in frequentist disciplines. Ignore *p*-values. Pay attention to raw effect sizes and to sample sizes. Most studies at least contain confidence intervals, which are arguably a step up from *p*-values.[16] If a study does not show raw effect sizes, or if it claims that effects were worth paying attention to purely because they were statistically significant, then that study is not worth much (though that is probably true whatever the philosophical status of point *p*).

[16] Though confidence intervals have problems of their own. See James O Berger and Robert L Wolpert, *The Likelihood Principle*, ed. Shanti S Gupta, 2nd ed., Lecture Notes - Monograph Series (Institute of Mathematical Statistics, 1988), 5-6; Jason Grossman, "A Couple of Nasties Lurking in Evidence-Based Medicine," *Social Epistemology* 22, no. 4 (2008): 343-7.

Researchers who are particularly wedded to tradition might argue that *p*-values are still useful because they combine information about sample size, effect size and dispersion. But why combine these three characteristics of a set of observations in the way significance tests combine them? Why, for example, should a researcher studying a regression coefficient combine the effect size and the standard error estimate in the form of a *t*-score? Only if they are interested in calculating the probability of observing extreme *t*-scores under the null. I have argued that there is no reason to perform this calculation, at least with respect to point nulls.

There is no point calculating *p*-values for point nulls. Point *p* is probably nowhere near the likelihood of making extreme observations under the null hypothesis for the superpopulation. And I see no other reason to calculate it.

I recognise that this is conclusion is disconcerting. Medications have been prescribed, policies have been implemented and psychotherapies adopted because researchers used p-values to reject point nulls. I am contemplating that doctors, policymakers, psychologists and other members of frequentist professions have been basing decisions on a useless statistical test. But I see no way of avoiding this conclusion. It is clear enough why students and even practitioners struggle to grasp the point *p*-value. There is nothing to grasp.

Reference list

Berger, James O, and Thomas Sellke. "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence." *Journal of the American Statistical Association* 82, no. 397 (1987): 112-22.

Berger, James O, and Robert L Wolpert. *The Likelihood Principle*. Lecture Notes - Monograph Series. Edited by Shanti S Gupta. 2nd ed.: Institute of Mathematical Statistics, 1988.

Casella, George, and Roger L Berger. "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem." *Journal of the American Statistical Association* 82 (1987): 106-11.

Deming, W. Edwards. "On Probability as a Basis for Action." *American Statistician* 29, no. 4 (1975): 146-52.

Grossman, Jason. "A Couple of Nasties Lurking in Evidence-Based Medicine." *Social Epistemology* 22, no. 4 (2008): 333-52.

———. "Statistical Inference: From Data to Simple Hypotheses." Australian National University, 2011.

Hagen, Richard L. "In Praise of the Null Hypothesis Statistical Test." *American Psychologist* 52, no. 1 (1997): 15-24.

Kass, Robert E, and Adrian E Raftery. "Bayes Factors." *Journal of the American Statistical Association* 90, no. 430 (1995): 773-95.

Kendall, Maurice G. *Advanced Theory of Statisics*. 5th ed. Vol. 1, London: C Griffin, 1987.

Lindley, D V. "A Statistical Paradox." *Biometrika* 33, no. 1/2 (1957): 187-92.

Meehl, Paul E. "The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions." In *What If There Were No Significance Tests?*, edited by Lisa L Harlow, Stanley A Mulaik and James H Steiger, 353-82. New York: Routledge, 1997.

Rindskopf, David. "Testing 'Small,' Not Null, Hypotheses: Classical and Bayesian Approaches." In *What If There Were No Significance Tests?*, edited by Lisa L Harlow, Stanley A Mulaik and James H Steiger, 287-98. New York: Routledge, 1997.

Salsburg, David. "Use of Restricted Significance Tests in Clinical Trials: Beyond the One- Versus Two-Tailed Controversy." *Controlled Clinical Trials* 10 (1989): 71-82.

Savage, Leonard J. "On Rereading R. A. Fisher." *Annals of Statistics* 4, no. 3 (1976): 441-500.

Serlin, R C, and D K Lapsley. "Rationality in Psychological Research: The Good-Enough Principle." *American Psychologist* 40, no. 1 (1985): 73-83.

Smith, T M F. "The Foundations of Survey Sampling: A Review." *Journal of the Royal Statistical Society. Series A (General)* 139, no. 2 (1976): 183-204.

Vigen, Tyler. "Spurious Correlations." http://www.tylervigen.com/spurious-correlations.