SIGNIFICANCE IS NOT THE WHOLE STORY – DECISION MAKING IN HYPOTHESIS TESTING HAS TWO SORTS OF POSSIBLE ERRORS

James Nicholson
SMART Centre, Durham University, UK
j.r.nicholson@durham.ac.uk

Hypothesis testing has come under scrutiny, and attack, because of the way it is being misused. Much of the misuse seems to stem from a fundamental lack of understanding of some key principles within the methodology. In particular, losing sight of the fact that there are two possible wrong decisions. Where p-values are used, and computed by software, it is very difficult to maintain the perspective of the test as trying to identify shifts in parameters – because 'significance' has been viewed as the holy grail. The foundations for understanding hypothesis testing are undermined in some curricula where key aspects, such as the existence of two potential errors in the decision, are omitted. This paper will develop a pedagogical basis for teaching the logical foundations of hypothesis testing, and will provide links to electronic resources to support this approach.

INTRODUCTION

Simulations and other electronic resources have been around for a long time. Teachers with even a modest amount of classroom experience are capable of looking quite quickly at textbook or other paper based resources making a decision as to which resource they would prefer to use, and critically, are able to support their decision with some reasoning based on pedagogy. In comparison I think even several days would be insufficient time to make comparisons between electronic resources, and still the teachers would not have a preference which they could back up with a sound pedagogical argument. However, the understanding of many statistical concepts can be supported by visualisations and simulations.

One of the most contentious areas in statistical education currently is significance testing. Wasserstein and Lazar (2016) provide a very good background of the issues surrounding this controversy and the wider issues of *reproducibility* and *replicability*, which led to The ASA's Statement on *p*-Values: Context, Process and Purpose. Broadly speaking this was a list of don'ts, and it was a first step in a process, because there is fairly widespread agreement about the don'ts (what *not to do* with *p*-values). The American Statistician (2019) has produced an open-access special issue with 43 articles addressing the more difficult issue of the *do's* - what to do about the very hard problem of making decisions under uncertainty. Wasserstein, Schirm and Lazar (2019) provide a very accessible overview of the papers and their recommendations, which they summarise very helpfully in two sentences: *Accept uncertainty. Be thoughtful, open and modest*.

A lot (though not all) of the abuses of *p*-values seem to stem from losing sight of the fact that there are *two* possible wrong decisions in significance testing. The foundations for understanding hypothesis testing are undermined in some curricula where key aspects, such as the existence of two potential errors in the decision, are omitted. This paper will develop a pedagogical basis for teaching the logical foundations of hypothesis testing, and will provide links to electronic resources to support this approach.

DEVELOPING THE FOUNDATIONS OF HYPOTHESIS TESTING

A null hypothesis test of a parameter (NHT) requires both a null and an alternative hypothesis to be stated before any analysis is done (indeed, it should really be done before data is even collected – which would ensure that observations cannot influence the choice of hypotheses). The null hypothesis (H_0) must provide a sampling distribution for the statistic of interest and the alternative hypothesis (H_1) must identify what will constitute the most unusual outcomes if the null hypothesis is true. Non-parametric tests have a similar logical foundation, but this paper will consider the case of testing a Binomial parameter *p* because the illustration of the sampling distributions is relatively simple to grasp through simulations, and testing proportions is one of the most common applications of hypothesis testing.

One of the major issues in statistics today is whether or not hypothesis testing has outgrown its usefulness. Certainly it is misused horribly in very many places, and this is especially true if the *p*-

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

value approach is used where '*significant or not?*' is the only aspect which is addressed. The ASA statement on *p*-values is very helpful in teasing out many of the issues around why *p*-values are so badly misused, and provides some background understanding about why the use of language is extremely important.

In my teaching, I get students to construct critical regions under the null hypothesis first, and then consider where the observation lies – as a prelude to making a decision. Only after they are comfortable with this and competent in doing it correctly would I introduce the notion of *p*-value which is the statistic they will report when they use software to carry out tests.

The file '*Binomial test exploration'* (Nicholson, 2019) offers a number of different tabs which develop key ideas and concepts that are the foundations for understanding what is happening with the Binomial test. It can be downloaded from https://drive.google.com/open?id=1fnzGLZ14QY-ixDNWticLOqL2mjpttv34. There are some features in it which are worth noting as important aspects of working with any electronic graphical display or simulation:

- I have fixed the vertical scales so changes in the display are not distorted by the scales having changed.
- This is especially important when looking at the spreadsheets with $n = 150$ and wanting to compare with what happens when $n = 15$.

A couple of other points are worth making at this stage:

- This resource is intended to be used initially by the teacher in a classroom, with students spending time later exploring the behavior for themselves to build an intuitive understanding of the concepts which has a basis in experience. The definition of key terms such as Type I and Type II errors is expected to be reinforced by the teacher at appropriate points.
- Students should already be familiar with the Binomial distribution, and its applications, before introducing hypothesis testing.

Binomial distribution (n = 15)

This allows the user to explore the effect of changing the parameter in a Binomial. Figures 1 and 2 below show screenshots of the first tab in the file with different probabilities. The probability *p* is in cell B2, with a spinner which allows you to show values from a list between 0.1 and 0.9, in increments of 0.05. Limiting the values to this range means that no individual has a probability higher than 0.35, allowing the vertical scale to be fixed to provide the fixed frame of reference.

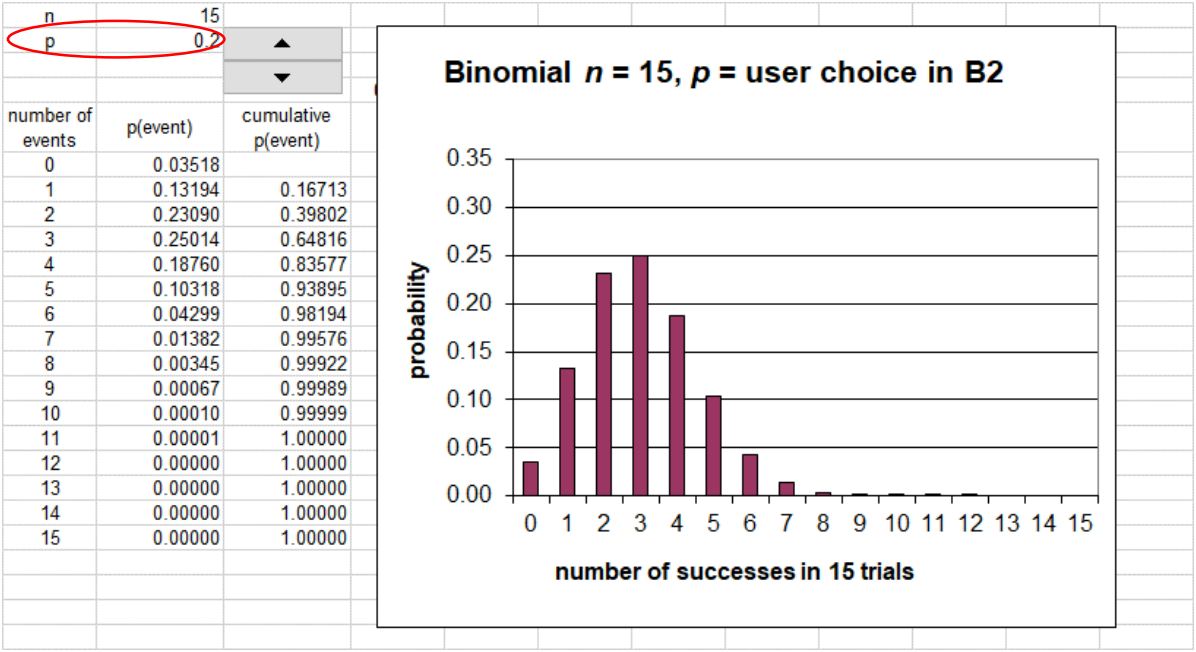
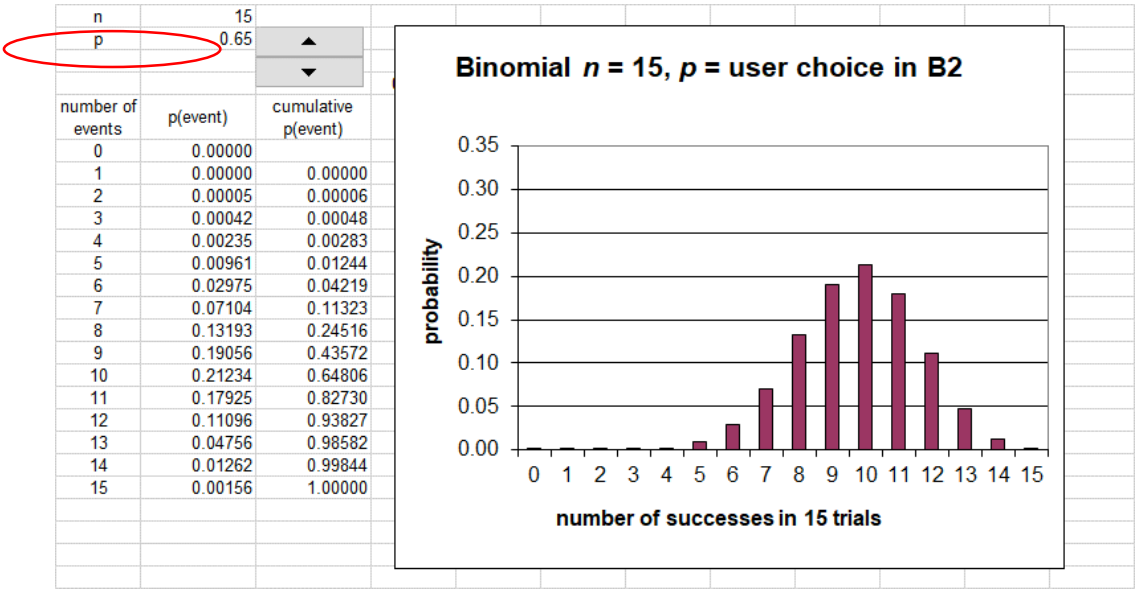


Figure 1: the Binomial distribution with $n = 15$, $p = 0.2$



n	15	
p	0.65	

number of events	p(event)	cumulative p(event)
0	0.00000	
1	0.00000	0.00000
2	0.00005	0.00006
3	0.00042	0.00048
4	0.00235	0.00283
5	0.00961	0.01244
6	0.02975	0.04219
7	0.07104	0.11323
8	0.13193	0.24516
9	0.19056	0.43572
10	0.21234	0.64806
11	0.17925	0.82730
12	0.11096	0.93827
13	0.04756	0.98582
14	0.01262	0.99844
15	0.00156	1.00000

Figure 2: the Binomial distribution with $n = 15$, $p = 0.65$

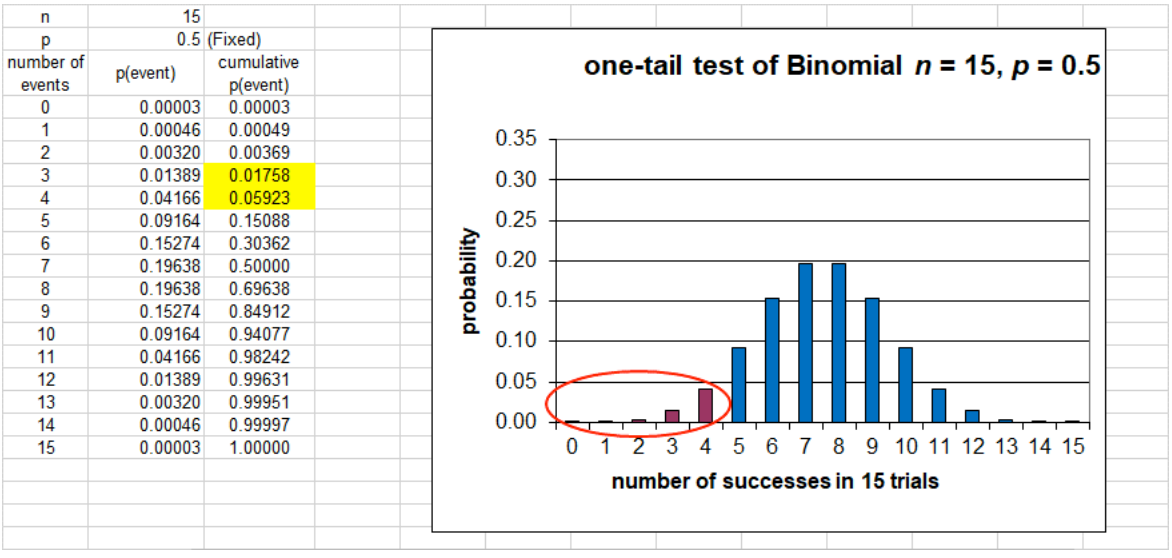
Many students rarely, if ever, see a full Binomial distribution. Most textbooks focus almost exclusively on the calculation of individual probabilities, so even this first tab showing how the distribution of probabilities changes as p changes for a fixed number of trials is important foundation in understanding what is happening.

One-tail test at 5% level of significance of $p = 0.5$ ($n = 15$)

Because the Binomial is a discrete distribution, a 'decision rule' to test $p = 0.5$ against $p < 0.5$ can not be exactly a 5% test - if you are to reject $p = 0.5$ if you see 3 or fewer successes in 15 trials then you have a 2% test (actually 1.76%), and if you decide to reject when you see 4 or fewer successes then it a 6% test (5.92%).

Figure 3 below has two cells (C7 and C8) high-lighted in yellow which show these cumulative probabilities We will use 4 or fewer as our criteria for deciding to reject the hypothesis that $p = 0.5$ on the basis of taking some observations.

Note: Cells B4 – C19 have a Binomial formula which use 0.5 as the parameter rather than a user defined value, because this is the distribution under the null hypothesis.



n	15	
p	0.5	(Fixed)

number of events	p(event)	cumulative p(event)
0	0.00003	0.00003
1	0.00046	0.00049
2	0.00320	0.00369
3	0.01389	0.01758
4	0.04166	0.05923
5	0.09164	0.15088
6	0.15274	0.30362
7	0.19638	0.50000
8	0.19638	0.69638
9	0.15274	0.84912
10	0.09164	0.94077
11	0.04166	0.98242
12	0.01389	0.99631
13	0.00320	0.99951
14	0.00046	0.99997
15	0.00003	1.00000

Figure 3: the one-tailed test of $p = 0.5$ for the Binomial distribution with $n = 15$

One-tail test exploration:

This tab tries to help generate an understanding of how effective a Binomial test with sample size 15 is in picking up a shift away from the hypothesized value of 0.5. The cells highlighted in purple represent the probabilities of the values which lie in the critical region - the bold red value is the cumulative probability - or the probability that a value in the critical region will be observed - leading to rejection of the null hypothesis. This is the *power of the test* for this value of the parameter as it is correctly rejecting the null hypothesis in a situation where p is < 0.5.

Altering the value of p using the up / down arrows will show just how far p has to move before there is a realistic chance of detecting the shift: you have to get down as far as 0.3 before it is better than evens chance of rejecting the null. Note that the spinner now uses increments of 0.01 and only runs from 0.49 to 0.1: these limits are chosen because 0.49 is the largest value (to 2 decimal places) which satisfies the alternative hypothesis, and 0.1 is still used as the lower limit so that fixing the vertical scale with a maximum of 0.35 allows all individual probabilities to be displayed.

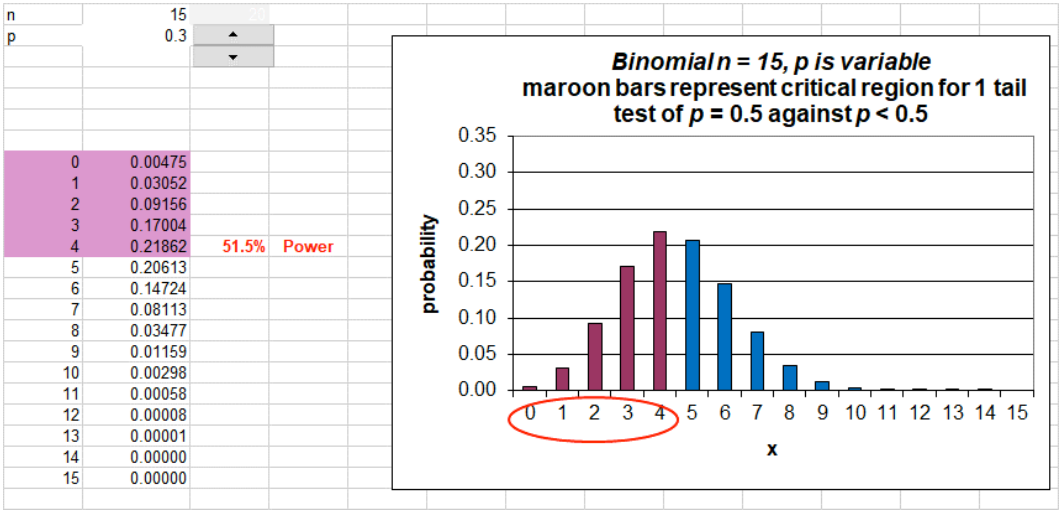


Figure 4: exploring the power of the one-tailed Binomial test with n = 15.

One-tail test simulation:

Where the previous tab showed the probability correctly rejecting the null hypothesis for different parameter values lying in the alternative hypothesis, this tab runs a simulation to demonstrate what is happening in a dynamic fashion. Cells P2-P1001 have 1000 random numbers, and Q2-Q1001 use these random numbers to simulate the outcome of a Binomial random variable with n = 15 and parameter equal to the user defined value in B2 (again this is limited to the range 0.1 – 0.9 so the vertical scale can be fixed with max at 0.35). The cells highlighted in purple represent the proportion of the outcomes of 1000 simulated tests which lie in the critical region - the bold red value in D6 is the total proportion of outcomes in the critical region - leading to rejection of the null hypothesis.

Altering the value of p using the up / down arrows will show just how far p has to move before there is a realistic chance of detecting the shift: you have to get down as far as 0.3 before it is better than evens chance of rejecting the null. Pressing F9 will allow you to take another set of 1000 tests with the same value of p, and altering the value of p automatically generates a new set of 1000 trials. There are then repeats of the first three of these sheets, but with n = 150 instead of 15. This should help the user understand why the larger sample tests are so much more effective in detecting shifts in the parameter value.

Binomial distribution (n = 150):

This now uses n = 150 where previously n was 15. Note that the vertical scale has been fixed to be the same as before, so the comparisons of probabilities that you see are real - the horizontal scale now includes many more values so the bars are much thinner, but the constant vertical scale means you make the correct comparison of probabilities. Now the spinner will allow you to go from 0.01 up to 0.99 without any individual outcome having a probability of more than 0.35.

One-tail test at 5% level of significance of p = 0.5 (n = 150):

As before, because the Binomial is a discrete distribution, the 'decision rule' to test $p = 0.5$ against $p < 0.5$ can not be exactly a 5% test - if you are to reject $p = 0.5$ if you see 64 or fewer successes in 150 trials then you have a 4% test (actually 4.3%), and if you decide to reject when you see 65 or fewer successes then it a 6% test (6.0%). We will use 64 or fewer as our criteria for deciding to reject the hypothesis that $p = 0.5$ on the basis of taking some observations (you can maybe just see the bars are shaded purple up to 64 and blue above that). I have reduced the row heights for a big block of low values (with negligible probabilities) so you can see the values contributing to the cumulative tail probability. The following tabs allow you to explore how effective this is in picking up shifts away from 0.5. Remember (or look back at the earlier tab) what happened with $n = 15$.

One-tail test exploration n = 150:

This allows you to explore how much more effective a test using 150 trials is in detecting a shift away from 0.5 than the test using 15 was. For example when $p = 0.4$, the power has increased to over 77% from under 22%; and when $p = 0.33$ the power is 99.5% compared to only 41.5% when $n = 15$ (see Figure 5).

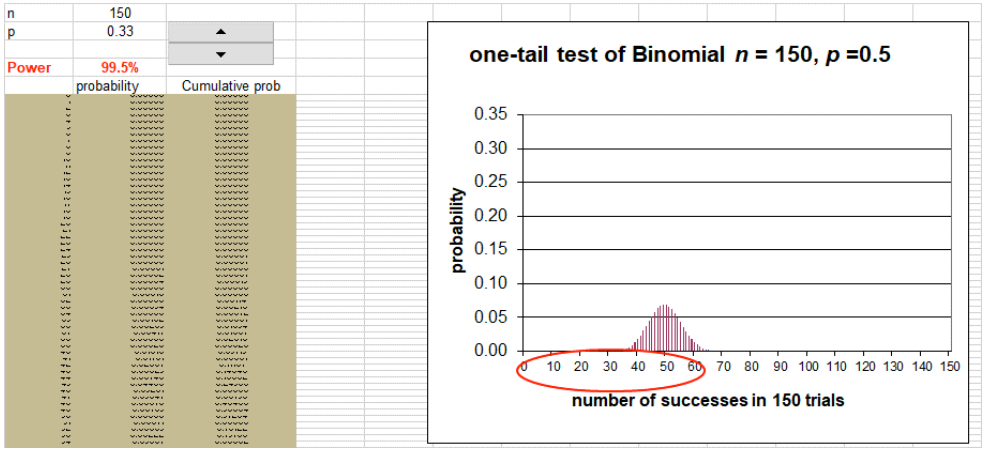


Figure 5: the one-tailed test of $p = 0.5$ for the Binomial distribution with $n = 150$

Two-tail test exploration:

This is similar to the one-tail exploration, but now you have to take the top and bottom tails as part of the critical region – so now the range of values you can explore is from 0.1 to 0.9.

Two-tail test simulation:

As with the exploration tab this counts the proportion lying in either tail, and again pressing F9 will allow you to take another set of 1000 tests with the same value of p.

Power function testing 'fair coin' with 15 trials:

This tab shows the power function for values between 0.1 and 0.9 for both the one-tail (red line) and two-tail (blue line) tests.

DISCUSSION

One of the statistical concepts that benefits most from visualisation and simulation is parameterised discrete probability distributions. In particular the binomial distribution where the parameters have extremely simple definitions, but it is extremely difficult to develop any intuition about the distribution if a student's only experience is in calculating the probability of individual outcomes (or sums of individual outcomes). This resource starts by providing a visualisation of the distribution for $n = 15$, This is not large in terms of hypothesis testing, but it is still larger than the sample sizes used in many assessment items in the UK. Nicholson and Ridgway (2018) reported that four of the thirteen tests in one set of assessments used sample with size 7, 8, 9 and 15. It is large enough for the effect of changing the parameter p on the distribution to be explored visually. Students

would already have explored this when they dealt with the binomial distribution as a topic in its own right (separate from hypothesis testing) and would have explored the effect of changing p for a variety of values of n.

Revisiting this exploration of the effect of changing p should have already given students some intuition in relation to how little the distribution changes when p changes a little. The second visualisation helps to make this explicit in the context of constructing a one-tailed test of $p = 0.5$: the colour coding of both the graph and the table in Figure 3 highlight the critical region for the test and students can see concrete evidence for how far the parameter p has to shift from 'fair' to stand a reasonable chance for the observed outcome to lie in the critical region. A key element in the resource is the comparison with the larger sample size in Figure 5 – for a fixed level of significance (or as comparable as possible with a discrete test) enlarging the sample size is the way that the decision-making process can be more effective in correctly identifying where there has been an important shift in the parameter p.

Hypothesis testing is a decision-making technique – under uncertainty. Any risk assessment should involve understanding both the likelihood of any errors which might occur, and the potential consequences of those errors. The reality is that much of the statistical practice that gave rise to the ASA's statements of *don'ts* and *do's* in this is the result of focusing at least primarily, and in many instances exclusively, on significance as though one of the two errors was the only one of any importance. The relationship between the two potential errors is inextricably interwoven with the sample size, and the "other evidence" Wasserstein et al. (2019) refer to in the do's for using hypothesis testing properly means that addressing this area is of critical importance early on in a pedagogically sound statistics curriculum.

It might be helpful here to articulate what the second type of error is in a specific context. If a company does not racially discriminate in its employment practices one might expect the proportion of ethnic minority employees fired would be similar to the proportion of minorities in the workforce. However, it is likely that the number of employees fired altogether will be small, giving only a small sample test. Even if the company does actually discriminate, the small sample may have a proportion of minorities fired which does not lie in the critical region, and the company will escape censure.

Nicholson and Ridgway (2018) offer a commentary on the new A level Mathematics curriculum introduced into the UK in 2017. Hypothesis testing is a substantive and compulsory element of that curriculum, testing Binomial and Normal distributions, and testing for correlation coefficients, but the content contains not a single reference to Type I and II errors, to effect size, or gives any indication that the size of the sample used in the test is of any relevance to the decision-making process. This is likely to produce a generation of young people who will perpetuate the poor statistical practice which gave rise to the ASA engagement with this issue.

While curriculum reform is a notoriously slow process, and continual change in curricula places a considerable burden on the teachers who have to implement the changes, it is to be hoped that the UK curriculum can be revisited in the near future to either move away from hypothesis testing as the main inferential tool, or, if it remains, then introducing sufficient other elements (such as Type I and II errors, effect size, minimum sample sizes needed to detect a particular effect size effectively) alongside significance that students will be introduced to a coherent and complete technique that can be used in decision-making.

REFERENCES

Nicholson, J., & Ridgway, J. (2018). Real-World Contexts in Statistics Components of UK Mathematics Examinations: Aiming Forward, Walking Backwards. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

Nicholson, J. (2019). Binomial test exploration. Downloaded from https://drive.google.com/open?id=1fnzGLZ14QY-ixDNWticLOqL2mjpttv34 30 April 2019.

Wasserstein, R., & Lazar, N. (2016). The ASA's statement on *p*-Values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a World Beyond "$p < 0.05$", *The American Statistician*, *73*: sup1, 1-19.