AFTER STATISTICS REFORM: SHOULD WE STILL TEACH SIGNIFICANCE TESTING?

Tony Hak
Rotterdam School of Management
Erasmus University, Rotterdam, the Netherlands
thak@rsm.nl

In the longer term null hypothesis significance testing (NHST) will disappear because p-values are not informative and not replicable. As with any reform, the question can be asked whether we should continue to teach the procedures of abolished routines (i.e., NHST), not as a commendable practice but as a means of understanding what our (more or less statistically ignorant) predecessors did. Three arguments are discussed for ***not*** *teaching NHST in (introductory) undergraduate courses in inferential statistics: experience shows that NHST is too difficult for introductory courses; dichotomous thinking inherent to NHST is a cognitive obstacle for interpretation; and students can find relevant information in research reports without knowing NHST.*

STATISTICS REFORM

Although it has been known for a long time that null hypothesis significance testing (NHST) has many severe flaws, it is still the routine technique used by researchers for drawing conclusions from data in the social sciences. In the last decade, however, renewed recognition of the flaws of NHST has emerged, particularly in psychology. A number of journals in psychology are about to require that empirical results are reported as estimates, i.e. as effect sizes (ESs) with confidence intervals (CIs), or have already done so (see Eich, 2014).

In the discussion below it is assumed that there is a consensus among statisticians about the need of statistics reform in the social sciences, and no detailed argumentation for this reform will be provided. It will suffice here to just mention the main arguments provided in a large literature which is expertly summarized by Cumming in Chapter 2 ("From Null Hypothesis Significance Testing to Effect Sizes") of his book *Understanding the New Statistics* (Cumming, 2012):

- A *p*-value does not give additional information to the information that is already available before its computation, i.e., an effect size (ES) with its confidence interval (CI). Nuanced quantitative information (about an ES and its precision) is reduced to dichotomous information ("significant", "not significant").
- A *p*-value is subject to immense sampling variation and can take any value (0 to 1) in a routine research situation with medium effect size (e.g., Cohen's $d = 0.5$) and "normal" power (of about 0.5). Hence, a *p*-value is not replicable (See also Cumming, 2008).
- NHST itself yields inevitably "false negative" results because of its priority on avoiding "false positive" results. In actual practice, in which "negative" results have a considerably smaller chance to be published than "positive" findings, published results tend to overestimate true effects.
- In any field, results of multiple studies should be synthesized or "meta-analyzed". Outcomes of NHST cannot be meaningfully synthesized.

Note that each of these arguments, except the one about publication bias, pertains to NHST proper. A list of routinely occurring abuses and misuses of NHST outcomes would, additionally, include forms of violation of the preconditions for its use and of over-interpretation of its outcomes.

Current routine statistics practice (NHST) is not sustainable because its outcomes are at best uninformative (beyond the information already provided by estimation) and are at worst misleading or plain wrong. Statistics reform is inevitable and NHST will disappear in the longer term.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

THREE ARGUMENTS FOR NOT TEACHING NHST

It will not require much discussion that, after successful statistics reform, we should teach the "new" statistics (estimation and meta-analysis) in undergraduate courses as the preferred way to present and analyze empirical results. But it is less evident whether or not NHST (though not preferred as an analytic tool) should still be taught. Because estimation is already routinely taught as a preparation for the teaching of NHST, the necessary reform in teaching will not require the addition of new elements in current programs but rather the *removal of the current emphasis on NHST* or *the complete removal of the teaching of NHST* from the curriculum. It will be argued that NHST should *not* at all be taught in (introductory) undergraduate courses in inferential statistics. The main arguments are:

1. *NHST is very difficult to understand and hence it is very hard to teach successfully*

Tversky & Kahneman (1971) showed that many researchers did not appreciate the fact that any NHST outcome is subject to sampling variation and believed that a significant result obtained in one study almost guaranteed a significant result in a replication, even one with a smaller sample size. Is it then surprising that also our students do not understand what NHST outcomes do tell us and what they do not tell us? The principles and procedures of NHST are not well understood by undergraduate students who have successfully passed statistics courses in which NHST is taught. This can be illustrated with the following quotes from reviews of research papers written by students in one of the Master programs of the Rotterdam School of Management:

- *"This conclusion is convincing because it is significant"*
- *"Significant means that there is low probability that this could have happened by chance"*
- *"The difference cannot be explained by chance"*
- *"The p-value indicates the probability that the observed pattern occurs, given that the expected pattern is true"*
- *"Significant at 1% means that when the difference is larger than 1% from the expected average, the difference can be deemed significant"*
- *"There is a very unlikely probability (<1%) that there is an observation which disproves the null hypothesis"*

These quotes have been selected for inclusion in this paper because of their brevity, not because of their authors' exceptional ignorance about NHST. It is very hard indeed to find a comment on NHST in any student paper (an essay, a thesis) that is close to a correct characterization of NHST or its outcomes. No doubt, similar quotes and similar ignorance about NHST will be found in a study of any other population of students or researchers. Courses on NHST fail to achieve their self-stated objectives, assuming that these objectives include achieving a correct understanding of the aims, assumptions, and procedures of NHST as well as a proper interpretation of its outcomes.

What is the cause of this failure of NHST teaching? The following explanations could be tentatively offered:

- First, and above all other reasons, NHST itself is a complicated procedure. It requires students and researchers to understand that a *p*-value is attached to an outcome (an estimate) based on its location in (or relative to) an imaginary distribution of sample outcomes around the null. Although a small proportion of students is able to quote the phrase "*p* is the probability that this result (or a result farther from the null) is obtained under the assumption that null is true" (or a version of it), only very few of this minority can explain what this phrase means. When pressed, they almost always are confused and end up with a phrase in which *p* is interpreted as an indication of the likelihood of an effect (different from the null).
- Most students understand that "significance" has something to do with dealing with the effects of "chance" (as expressed in the quote "The difference cannot be explained by chance"), but they fail to relate the concept of "chance" to sampling error. They tend to think that NHST is aimed at cancelling out *any type of error* and that, hence, it is able to produce a result that is generalizable to non-observed populations, missing cases, the future, etc. (all assumed to be "missing at random"). They assume that a *p*-value is an outcome of procedures that "correct" for "chance". They cannot, cognitively, accept that *p*-values themselves are subject to sampling variation.

- Students have no real understanding of the concept of "probability sampling". Many of them confuse it with the concept of "representativeness" (meaning: a sample that gives a "true" result). This confusion is probably not confined to students. The belief that each sample is "representative" was also observed by Tversky & Kahneman (1971) among their colleagues.
- More fundamentally, students and researchers do not appreciate how "lumpy" chance (or randomness) is (Abelson's Law No.1; Abelson, 1995).
- There is also a heroic way of thinking about science in which scientific progress is achieved by designing and conducting a decisive single study (conducted by the single excellent researcher, who will be rewarded with a Nobel Prize). Replications are secondary. Their only purpose is to confirm the "original discovery" by generating the "same" result, i.e., the same *p*-value or, at least, the same "level" of significance.

Apart from the inherent complexity of NHST, most of these explanations for misunderstanding the principles of NHST are due to ignorance of the implications of sampling variation. Therefore, it might be argued that it is not fair to present these explanations here as evidence of a general failure of NHST teaching. Could not teaching the "new statistics" fail in the same way? This is not likely because a confidence interval itself is an acknowledgement and representation of the phenomenon of sampling variation! Hence, cognitively it is much less likely that an estimate with a confidence interval will be interpreted as having been "corrected" for "chance".

2. Dichotomous thinking inherent to NHST is a cognitive obstacle to interpretation

One might argue that there is no harm in adding a *p*-value to an estimate in a research report and, hence, that there is no harm in teaching NHST, additionally to teaching estimation. However, the mixed experience with statistics reform in clinical and epidemiological research suggests that a more radical change is needed. Dichotomous thinking in everyday talk about health behavior has increasingly been replaced by estimation thinking. Whereas in the past it would be discussed *whether* it is good or bad for your health to smoke or to drink coffee, or alcohol, now it is more likely that we discuss *how many* days or years of our life might be lost or won by the intake of, for instance, two glasses of wine per day on average. This change has been facilitated by a change in reporting about clinical research which to a large extent was driven by changes in reporting standards as required by journals. Reports of clinical trials and of studies in clinical epidemiology now usually report estimates and confidence intervals, in addition to *p*-values. However, as Fidler et al. (2004) have shown, and contrary to what one would expect, authors continue to discuss their results in terms of significance. Fidler et al. therefore concluded that "editors can lead researchers to confidence intervals, but can't make them think". This suggests that a successful statistics reform requires more than just requiring researchers to report confidence intervals in addition to *p*-values. It requires a cognitive change that should be reflected in how results are interpreted in the Discussion sections of published reports.

The stickiness of dichotomous thinking can also be illustrated with the results of a more recent study of Coulson et al. (2010). They presented estimates and confidence intervals obtained in two studies to a group of researchers in psychology and medicine, and asked them to compare the results of the two studies and to interpret the difference between them. It appeared that a considerable proportion of these researchers, first, used the information about the confidence intervals to make a decision about the significance of the results (in one study) or the non-significance of the results (of the other study) and, then, drew the incorrect conclusion that the results of the two studies were in conflict. Note that no NHST information was provided and that participants were not asked in any way to "test" or to use dichotomous thinking. The results of this study suggest that NHST thinking can (and often will) be used by those who are familiar with it.

The fact that it appears to be very difficult for researchers to break the habit of thinking in terms of "testing" is, as with every addiction, a good reason for avoiding that future researchers come into contact with it in the first place and, if contact cannot be avoided, for providing them with robust resistance mechanisms. The implication for statistics teaching is that students should, first, learn estimation as the preferred way of presenting and analyzing research information and that they get introduced to NHST, if at all, only after estimation has become their routine statistical practice.

3. Students can find information on effect sizes and confidence intervals in most research reports

It might be argued that students who have learnt to estimate (only) and who not have been introduced to NHST might not be able to read the (past and current) academic literature in which authors themselves routinely focus on the statistical significance of their results. It is obvious that someone who does not know NHST cannot correctly interpret outcomes of NHST practices. However, who cares? Because NHST outcomes are at best uninformative (beyond the information already provided by estimation) and are at worst misleading or plain wrong, nothing is lost by just ignoring the information that is related to NHST in a research report and by focusing only on the information that is provided about the observed effect size and its confidence interval. We can teach undergraduate students to find that information and to just ignore the "asterisks".

It is often argued that this approach is not feasible because many research reports do not contain sufficient information about the effect size and its confidence interval. This argument is flawed because it implies that the information that *is* provided (such as a "significance level", e.g. $p<0.05$) would be valuable even without the observed effect size, which we know is not the case. We cannot draw practical conclusions from evidence that suggests that there is an association between variables (or an "effect") if its size is not specified. We cannot meta-analyze (exact) *p*-values because they are neither replicable nor comparable between studies, and obviously we cannot meta-analyze *p*-value levels (such as $p<0.05$) either. Research reports that do not give information from which the effect size can be calculated must simply be dismissed as irrelevant.

Fortunately, most research reports that do not present and discuss the effect size that is observed in the study happen to provide information from which it can be calculated. Reports of experimental studies, for instance, quite often present information about observations in the various experimental groups (such as group means and standard deviations) from which the difference between group means and its confidence interval can be calculated (without reference to NHST). Even so, there are a (relatively small) proportion of research reports that give, e.g., an exact *p*-value and a standard error or group means only. Knowledge of NHST is required only in this group of reports for calculating an effect size and/or a confidence interval from the information that is given. If we do not teach NHST to students, these students will be handicapped in interpreting only this group of papers. This is a small price to pay for an immense gain, i.e. the capability to generate relevant statistical outcomes as well as to discover them in published research reports.

CONCLUSION

Three arguments have been discussed for teaching only estimation as the preferred form of statistical inference and for *not* teaching NHST, at least in beginner statistics courses:

(a) Students must be reminded all the time of the fact that they deal with phenomena that are subject to sampling variation. Confidence intervals do this reminding.
(b) Dichotomous thinking seems to be addictive and must be kept at a distance.
(c) The past record of empirical evidence will still to a large extent be accessible for students who have been only trained in estimation and meta-analysis. More importantly, these students will be able to read this past record much more critically than previous generations.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: L. Erlbaum Associates.

Coulson, M., Healy, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit: But do not guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement, 20*(1), 37-46.

Cumming, G. (2008). Replication and *p* Intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4): 286-300.

Cumming, G. (2012). *Understanding the new statistics*. New York / London: Routledge.

Eich, E. (2014). Business not as usual. *Psychological Science*. doi: 10.1177/0956797613512465.

Fidler, F., Thomason, N., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. Statistical reform lessons from medicine. *Psychological Science, 15*(2), 119-126.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105-110.