



Southern African Association for Institutional Research

SAAIR 2015 Conference 2015

Reflecting on the ban of NHT: Life after 0,05

Dion van Zyl & Hanlie Liebenberg
Unisa

Why are we talking

The purpose of this session is to review and reflect on the use of NHSTP



Content

- Introduction
- Reviewing - Steps in hypothesis testing
- Basic and Applied Social Psychology on NHSTP
- General critique against NHSTP
- Why so deeply entrenched in our research practices?
- Myths of NHSTP
- A practical example
- Where to from here?
- Conclusion



Introduction

“Null-hypothesis significance testing (NHST) has long been the mainstay method of analyzing data and drawing inferences in psychology and many other disciplines.” (Cumming, 2014)

“Hundreds of papers and blogposts have been written about what some statisticians deride as ‘null hypothesis significance testing’. NHST deems whether the results of a data analysis are important on the basis of whether a summary statistic (such as a P value) has crossed a threshold.” (Leek & Peng, 2015)



Stats 101



My reaction to p-values



My reaction to p-values

20 years
later!



My confession...

- I am but a researcher /scholar that dwells in the world of statistics
- This presentation is based on selective reading – so not my own thinking
- To ignite your interest in this “world” that we either take for granted or knowingly avoid
- I take the plunge and scared as hell!!

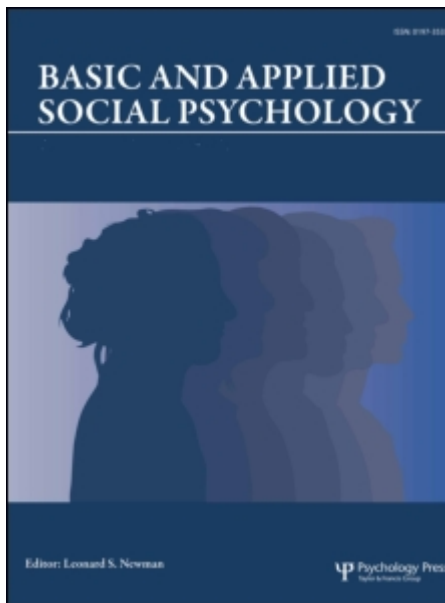


The four steps of NHSTP

1. **Develop H0** about some phenomenon or parameter (is generally opposite of the research hypothesis, which is what researcher truly believes and wants to demonstrate [alternative hypothesis] – this might be generated either inductively, from a study of observations already made, or deductively, deriving from theory) [H0 = no difference]
2. **Data collection** via experiment or sample – H0 often developed after data collection!)
3. **Statistical test of H0** conducted, which generates a p-value (set level of significance)
4. **Interpretation of p-value** - several interpretations of P often are made.

(Johnson, 1999)

Banning NHSTP



The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. **From now on, BASP is banning the NHSTP.**

12 Feb 2015

General critique against NHSTP

“Statistical hypothesis testing has received an enormous amount of criticism, and for a rather long time.” (Johnson, 1999)

“[NHSTP is]...no longer a sound or fruitful basis for statistical investigation.”

Clark (1963)

“...essential mindlessness in the conduct of research.” (Bakan, 1966)

Carver (1978) recommended that statistical significance testing should be eliminated; it is not only useless, it is also harmful because it is interpreted to mean something else.

General critique against NHSTP

“The main criticism is that NHST is **overrated**.” (Yatani, 2014)

“[NHSTP]...**does not tell us what we want to know**, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!” (Cohen, 1994)

“The current practice of focusing exclusively on a **dichotomous reject-nonreject decision strategy** of null hypothesis testing can actually impede scientific progress...” (Kirk, 2003)

General critique against NHSTP

“After review of the debate about NHST, I argue that the criticisms have sufficient merit to support the minimization or elimination of NHST in the behavioral sciences.” (Kline, 2004)

“Given the discourse, it is no surprise that some hailed as a victory the banning of NHST methods (and all of statistical inference) in the journal Basic and Applied Social Psychology in February.” (Leek & Peng, 2015)

Why so deeply entrenched?

“People are by nature good at **pattern recognition**... Our pattern recognition ability is so well-developed that sometimes we see *too much* meaning in otherwise random events... scientists may be extraordinarily good at pattern recognition, which also makes them subject to the potential error of seeing too much meaning in certain events. This seems to be true about NHST, because many common fallacies about it involve **exaggerating what can be inferred from statistical tests**.” (Kline, 2004)

“I suspect it’s **the seductive but misleading hints of importance** and certainty - even truth - in a statement that we’ve found a “statistically significant effect.” NHST decisions can be wrong, and every decent textbook warns that a statistically significant effect may be tiny and trivial. **But we so yearn for certainty that we take statistical significance as pretty close.**” (Cumming, 2014)

Dichotomous reject-nonreject decision strategy (Kirk, 2003)

Why so deeply entrenched?

Nester (1996) suggested several reasons:

1. They appear to be objective and exact
2. They are readily available and easily invoked in many commercial statistics packages
3. Everyone else seems to use them
4. Students, statisticians, and scientists are taught to use them
5. Some journal editors and thesis supervisors demand them (publication bias)

What's wrong with NHSTP?

Regarding confidence intervals, the problem is that, for example, a 95% confidence interval does not indicate that the parameter of interest has a 95% probability of being within the interval. Rather, it means merely that if an infinite number of samples were taken and confidence intervals computed, 95% of the confidence intervals would capture the population parameter.

(Editorial BASP, Trafimow & Marks, 2015)

Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

J. Neyman

Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 236, No. 767. (Aug. 30, 1937), pp. 333-380.

* Consider the variables (4) and assume that the form of their probability law (5) is known, that it involves the parameters $\theta_1, \theta_2, \dots, \theta_b$, which are constant (not random variables), and that the numerical values of these parameters are unknown. It is desired to estimate one of these parameters, say θ_1 . By this I shall mean that it is desired to define two functions $\bar{\theta}(E)$ and $\underline{\theta}(E) \leq \bar{\theta}(E)$, determined and single valued at any point E of the sample space, such that if E' is the sample point determined by observation, we can (1) calculate the corresponding values of $\underline{\theta}(E')$ and $\bar{\theta}(E')$, and (2) state that the true value of θ_1 , say θ_1^0 , is contained within the limits

$$\underline{\theta}(E') \leq \theta_1^0 \leq \bar{\theta}(E'), \quad (18)$$

It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results will tend to α .*

Consider now the case when a sample, E' , is already drawn and the calculations have given, say, $\underline{\theta}(E') = 1$ and $\bar{\theta}(E') = 2$. Can we say that in this particular case the probability of the true value of θ_1 falling between 1 and 2 is equal to α ?

The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made, that is except for the hypothetical and trivial ones

$$P\{1 \leq \theta_1^0 \leq 2\} = \begin{cases} 1 & \text{if } 1 \leq \theta_1^0 \leq 2 \\ 0 & \text{if either } \theta_1^0 < 1 \text{ or } 2 < \theta_1^0, \end{cases} \dots \quad (21)$$

which we have decided not to consider.

Some fallacies of NHSTP

Concerning misinterpretations (Kline*, 2004)

When the p-value is interpreted as the probability that the null hypothesis is true given the data [$p(H_0|D)$] → Inverse probability error (Cohen, 1994) (posterior probabilities in light of the data, and is probably what we would really like to know.

$p = \text{Pr}[\text{observed or more extreme data} | H_0]$ ✓ *[probability of horse winning on a rainy day vs. probability of raining when the horse wins]*

If H_0 is rejected, the p-value is the probability that this decision is wrong.

Only with sufficient replication could we discern whether a specific decision to reject H_0 was correct ✓

Complement of p is probability that a result will replicate under constant conditions.

Replication is more a matter of experimental design – and not one that could be answered by a single study and a statistical test ✓

Some fallacies of NHSTP

Mistaken conclusions after making a decision about the null hypothesis*

P-value is an indication or index value of magnitude of an effect; thus, low p-values indicate a large effect (magnitude fallacy).

Statistical tests and p-values measure sample size and effect size – if sample size is large, low p-values just confirm a large sample ✓

Failure to reject the null hypothesis means that the population effect size is zero.

(1) Basic tenet of science that absence of evidence is not evidence of absence; (2) Decision to fail to reject H_0 may be a Type II error – e.g. there maybe a real effect, but the study lacks sufficient power ✓

Failure to reject $H_0: \mu_1 = \mu_2$ means that the two populations are equivalent.

H_0 is usually false *a priori* ✓

Another myth...

Yatani (2014)

Threshold for p value

- Proposed by Fisher -> haven't changed it
- Probably reasonable choice, but there is no theoretical reasoning for it
- E.g. why is $p=0.04$ significant and $p=0.06$ is not? Why is $p=0.0499$ significant and $p=0.0501$ is not? If they have the equal sample sizes, the cases of $p=0.0499$ and $p=0.0501$ have almost the same effects.
- Marginally significant?
- Moving away or towards?

Practical example



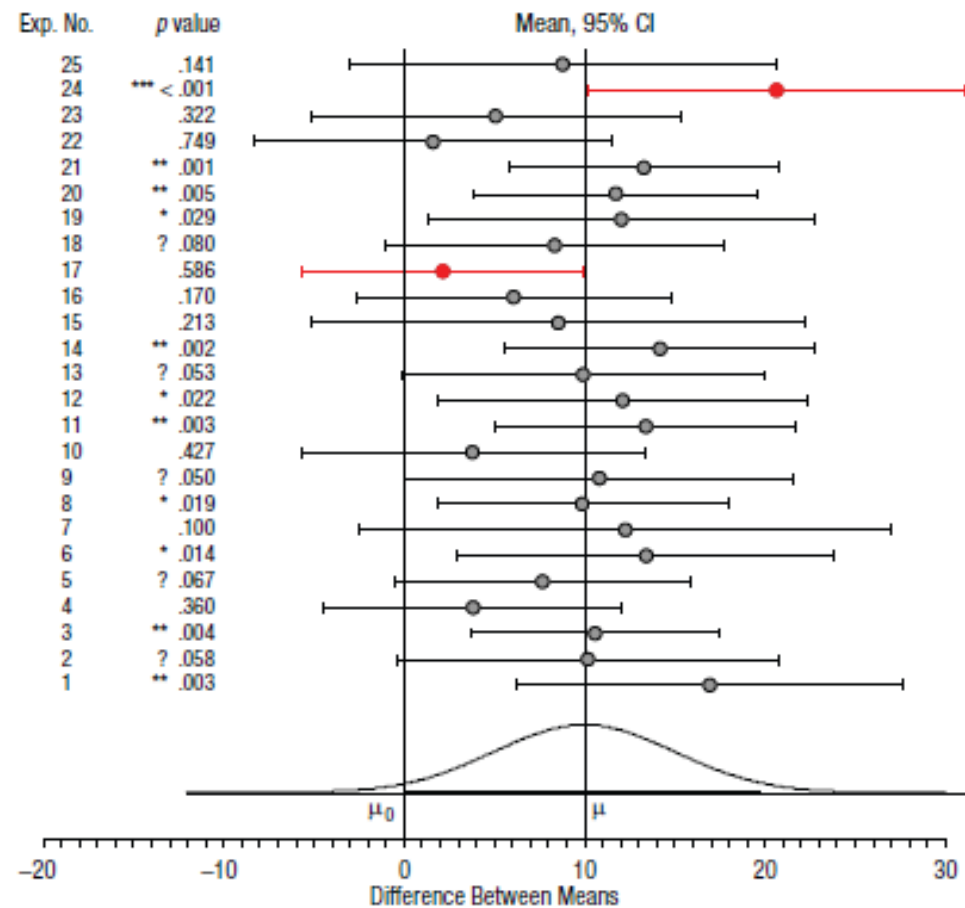


Fig. 1. Simulated results of 25 replications of an experiment (numbered at the left). Each experiment comprises two independent samples with an n of 32; the samples are from normally distributed populations with $\sigma = 20$ and means that differ by $\mu = 10$. For each experiment, the difference between the sample means (circle) and the 95% confidence interval (CI) for this difference are displayed. The p values in the list at the left are two-tailed, for a null hypothesis of zero difference, $\mu_0 = 0$, with σ assumed to be not known (* $.01 < p < .05$, ** $.001 < p < .01$, *** $p < .001$; a question mark indicates $.05 < p < .10$). The population effect size is 10, or Cohen's $d = 0.5$, which is conventionally considered a medium-sized effect. Mean differences whose CI does not capture μ are shown in red. The curve is the sampling distribution of the difference between the sample means; the heavy line spans 95% of the area under the curve.

Where to from here?

Is NHST bad?

So, you may want to ask “isn't NHST good and don't we even bother doing it?” or “should we stop believing the results of NHST?” My answer is we just need to be a little bit more careful to use it. If we can use NHST properly, it is a great tool to analyze your data if

1. you understand the meaning of the p value correctly (Myth 1, Myth 4),
2. you use the term “significant” properly (Myth 5),
3. your null hypothesis is appropriate, and
4. your research question can be answered by “yes” or “no”.

However, I also feel that the current HCI papers (and maybe reviewers, who are also us) overrate NHST a little too much. I have heard and seen that some papers got rejected simply because they couldn't find significant differences although they showed some interesting things. Again, being statistically significant is not necessarily being important in a practical setting (See Myth5 for more details).

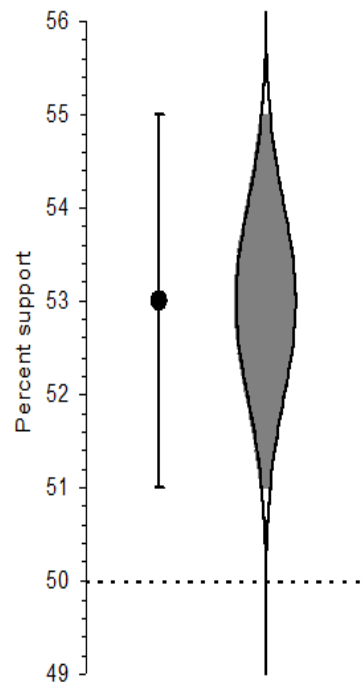
Thus, we need to be a little more careful when we are interpreting the results of NHST. Even if the authors cannot find statistical differences, there may be some other things that could benefit or inspire other researchers in that paper.

(Yatani, 2014)

Where to from here?

There's life beyond .05 (Cumming, 2014)

- Advocate estimation – effect sizes and confidence intervals (error bars)
- Meta-analysis
- Bayesian techniques



Frequentist vs Bayesian

Frequentists vs. Bayesian
Round 1

Parameters
fixed

Data
varies



Data
fixed

Parameters
Vary

Nate Silver



Where to from here?

Table 1. Twenty-Five Guidelines for Improving Psychological Research

1. Promote research integrity: (a) a public research literature that is complete and trustworthy and (b) ethical practice, including full and accurate reporting of research.
2. Understand, discuss, and help other researchers appreciate the challenges of (a) complete reporting, (b) avoiding selection and bias in data analysis, and (c) replicating studies.
3. Make sure that any study worth doing properly is reported, with full details.
4. Make clear the status of any result—whether it deserves the confidence that arises from a fully prespecified study or is to some extent speculative.
5. Carry out replication studies that can improve precision and test robustness, and studies that provide converging perspectives and investigate alternative explanations.
6. Build a cumulative quantitative discipline.
7. Whenever possible, adopt estimation thinking and avoid dichotomous thinking.
8. Remember that obtained results are one possibility from an infinite sequence.
9. Do not trust any p value.
10. Whenever possible, avoid using statistical significance or p values; simply omit any mention of null-hypothesis significance testing (NHST).
11. Move beyond NHST and use the most appropriate methods, whether estimation or other approaches.
12. Use knowledgeable judgment in context to interpret observed effect sizes (ESs).
13. Interpret your single confidence interval (CI), but bear in mind the dance. Your 95% CI just might be one of the 5% that miss. As Figure 1 illustrates, it might be red!
14. Prefer 95% CIs to SE bars. Routinely report 95% CIs, and use error bars to depict them in figures.

Where to from here?

15. If your ES of interest is a difference, use the CI on that difference for interpretation. Only in the case of independence can the separate CIs inform interpretation.
16. Consider interpreting ESs and CIs for preselected comparisons as an effective way to analyze results from randomized control trials and other multiway designs.
17. When appropriate, use the CIs on correlations and proportions, and their differences, for interpretation
18. Use small- or large-scale meta-analysis whenever that helps build a cumulative discipline.
19. Use a random-effects model for meta-analysis and, when possible, investigate potential moderators.
20. Publish results so as to facilitate their inclusion in future meta-analyses.
21. Make every effort to increase the informativeness of planned research.
22. If using NHST, consider and perhaps calculate power to guide planning.
23. Beware of any power statement that does not state an ES; do not use post hoc power.
24. Use a precision-for-planning analysis whenever that may be helpful.
25. Adopt an estimation perspective when considering issues of research integrity.

Where to from here?

An eight-step new-statistics strategy for research with integrity

The following eight steps highlight aspects of the research process that are especially relevant for achieving the changes discussed in this article.

1. **Formulate research questions in estimation terms.** To use estimation thinking, ask “How large is the effect?” or “To what extent . . . ?” Avoid dichotomous expressions such as “test the hypothesis of no difference” or “Is this treatment better?”
2. **Identify the ESs that will best answer the research questions.** If, for example, the question asks about the difference between two means, then that difference is the required ES, as illustrated in Figure 1. If the question asks how well a model describes some data, then the ES is a measure of goodness of fit.

Where to from here?

3. **Declare full details of the intended procedure and data analysis.** Prespecify as many aspects of your intended study as you can, including sample sizes. A fully prespecified study is best.
4. **After running the study, calculate point estimates and CIs for the chosen ESs.** For Experiment 1 in Figure 1, the estimated difference between the means is 16.9, 95% CI [6.1, 27.7]. (That is the APA format. From here on, I omit “95% CI,” so square brackets signal a 95% CI.)
5. **Make one or more figures, including CIs.** As in Figure 1, use error bars to depict 95% CIs.
6. **Interpret the ESs and CIs.** In writing up results, discuss the ES estimates, which are the main research outcome, and the CI lengths, which indicate precision. Consider theoretical and practical implications, in accord with the research aims.
7. **Use meta-analytic thinking throughout.** Think of any single study as building on past studies and leading to future studies. Present results to facilitate their inclusion in future meta-analyses. Use meta-analysis to integrate findings whenever appropriate.
8. **Report.** Make a full description of the research, preferably including the raw data, available to other researchers. This may be done via journal publication or posting to some enduring publicly available online repository (e.g., figshare, figshare.com; Open Science Framework, openscienceframework.org; Psych FileDrawer, psychfiledrawer.org). Be fully transparent about every step, including data analysis—and especially about any exploration or selection, which requires the corresponding results to be identified as speculative.

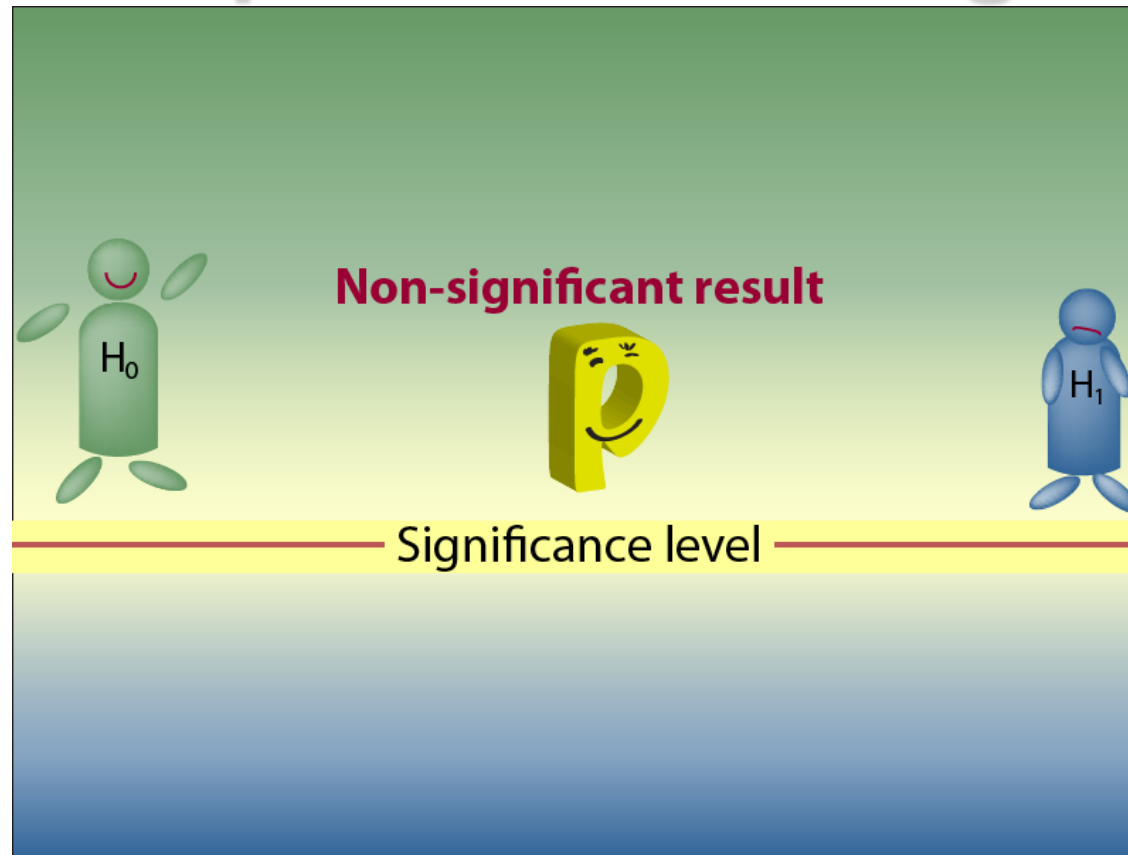
Statistics: P values are just the tip of the iceberg

“Such a ban will in fact have scant effect on the quality of published science. There are many stages to the design and analysis of a successful study... In practice, decisions that are made earlier in data analysis have a much greater impact on results — from experimental design to batch effects, lack of adjustment for confounding factors, or simple measurement error.”

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one.

Leek & Peng (28 April 2015, Nature)

Statistics: P values are just the tip of the iceberg



“The focus of research should be on . . . what data tell us about the magnitude of effects, the practical significance of effects, **and the steady accumulation of knowledge.**” (Kirk, 2003)

In conclusion

We are researchers that apply statistical analysis to answer questions about data (some specific and some not so specific)

...it is not about frequentist or Bayesian!!

It is the responsibility of the researcher to consider the most appropriate approach to answer these questions about data (... and not to blindly or religiously apply techniques – Irizarry, 2014)

One study does not provide all the answers

...both frequentist and Bayesian approaches have contributed towards insight

Map the world!



The scientist is not a person who gives the right answers, he is one who asks the right questions.

(Claude Levi-Strauss)

izquotes.com

Moneyball



"Adapt or Die"

It's unbelievable how much you don't know about
the game you've been playing all your life.

-Mickey Mantle