



Abandon Statistical Significance

BY BLAKELEY B. MCSHANE , DAVID GAL,
ANDREW GELMAN , CHRISTIAN ROBERT,
AND JENNIFER L. TACKETT

Presentation by Julian Naue and Emily Finne

Agenda

- Motivation
 - The Status Quo
 - Benjamin et al. (2018)
- Problems General to Null Hypothesis Significance Testing
 - Implausible Null Hypothesis
 - Misinterpretation of the p-Value
- Problems Specific to the Benjamin et al. (2018) Proposal
- Abandoning Statistical Significance
 - For Authors
 - For Editors and Reviewers
 - Abandoning Statistical Significance Outside Scientific Publishing
- Discussion

Authors

Blakeley B. McShane:

- Ph.D. and M.A. in Statistics and B.S. in Economics from the Wharton School of the University of Pennsylvania
- M.A. and B.A. in Mathematics from the College of Arts and Sciences of University of Pennsylvania
- faculty member in the Marketing Department of the Kellogg School of Management at Northwestern University

David Gal:

- BS, Computer Science, Penn State University, 1999
- MS, Management Science & Engineering, Stanford University, 2004
- Ph.D., Business Administration, Stanford University 2007
- Professor of Marketing, University of Illinois at Chicago, 2016-present

Andrew Gelman:

- He earned an S.B. in mathematics and in physics from MIT in 1986
- earned his Ph.D. in statistics from Harvard University in 1990

The status Quo

- 0.05 threshold relative to the sharp point null hypothesis of zero effect and zero systematic error
- The status quo is that $p < 0.05$ is deemed as strong evidence
 - To be published
 - To be taken serious
- $p < 0.05$ rule has been considered a safeguard against
 - noise-chasing
 - guarantor of replicability

Benjamin et al.

- redefine statistical significance
 - Change p-value threshold from 0.05 to 0.005

“changing the p-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance,”


-Benjamin et al. 2018



insufficient to overcome current difficulties with replication

Null hypothesis significance testing (NHST)

Abandon statistical significance, to drop the null hypothesis significance testing (NHST) paradigm.



p-value be demoted from its threshold screening role and instead, treated continuously

- just one among many pieces of evidence

- p-values should not be banned

- Just should **not** be thresholded

- Take no priority over other factors

- Seldom makes sense to calibrate evidence as function of p-values

Problems General to Null Hypothesis Significance Testing

- General Problems with NHST remain unresolved by Benjamin et al.
- Problems specific to Benjamin et al.
- Recommendation p-value be demoted from its threshold screening role
 - Even in big sample sizes
 - possibility of systematic bias and variation -> equivalent of small or unrepresentative samples
 - Estimates of every single study are generally noisy

systematic or nonsampling error which vary by field but
include measurement error

- problems with reliability and validity
- biased samples
- nonrandom treatment assignment
- Missingness
- Nonresponse
- failure of double-blinding
- Noncompliance
- confounding

systematic or nonsampling error which vary by field but include measurement error

- problems with reliability and validity
- biased samples
- nonrandom treatment assignment
- Missingness
- Nonresponse
- failure of double-blinding
- Noncompliance
- confounding

sharp point null hypothesis of zero effect and zero systematic error is highly problematic

- Effects are generally small and variable

systematic or nonsampling error which vary by field but include measurement error

- problems with reliability and validity
- biased samples
- nonrandom treatment assignment
- Missingness
- Nonresponse
- failure of double-blinding
- Noncompliance

– confounding

sharp point null hypothesis of zero effect and zero systematic error is highly problematic

- Effects are generally small and variable



Assumption of zero effect is false

measurements are generally noisy and systematically biased

Implausible Null Hypothesis

Problems are exacerbated per the status quo and the Benjamin et al. (2018) proposal

- Especially noisy estimates
 - are upwardly biased in magnitude
 - often of the wrong sign
- lexicographic decision rule -> tarnished literature
- smaller, less resource-intensive, noisier studies more likely to yield one or more statistically significant results
- Than fewer larger, more resource-intensive, better studies



encourages the former over the latter

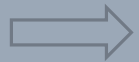
-features of the biomedical and social sciences

-for example, small and variable effects, systematic error, noisy measurements, a lexicographic decision rule for publication, and research practices

-make NHST and in particular the sharp point null hypothesis of zero effect and zero systematic error particularly poorly suited for these domains

Categorization of Evidence

- NHST dichotomization of evidence into
 - statistically significant
 - not statistically significant
- Sometimes trichotomization
 - marginally significant



0.05 threshold is arbitrary

“from an ontological viewpoint, there is no sharp line between a ‘significant’ and a ‘nonsignificant’ difference; significance in statistics...varies continuously between extremes”

-Rosnow and Rosenthal (1989)

Categorization of Evidence

- Treating p-value continuously than in a threshold is an improvement
- Go further: calibrate evidence as a function of the p-value because:
 1. p-value is defined relative to the generally implausible and uninteresting sharp point null hypothesis of zero effect and zero systematic error.

Categorization of Evidence

- Treating p-value continuously than in a threshold is an improvement
- Go further: calibrate evidence as a function of the p-value because:
 1. p-value is defined relative to the generally implausible and uninteresting sharp point null hypothesis of zero effect and zero systematic error.
 2. it is a poor measure of the evidence for or against a statistical hypothesis

Categorization of Evidence

- Treating p-value continuously than in a threshold is an improvement
- Go further: calibrate evidence as a function of the p-value because:
 1. p-value is defined relative to the generally implausible and uninteresting sharp point null hypothesis of zero effect and zero systematic error.
 2. it is a poor measure of the evidence for or against a statistical hypothesis
 3. it tests the hypothesis that one or more model parameters equal the tested values

Categorization of Evidence

- Treating p-value continuously than in a threshold is an improvement
- Go further: calibrate evidence as a function of the p-value because:
 1. p-value is defined relative to the generally implausible and uninteresting sharp point null hypothesis of zero effect and zero systematic error.
 2. it is a poor measure of the evidence for or against a statistical hypothesis
 3. it tests the hypothesis that one or more model parameters equal the tested values
- Small p-value
 - may be a problem with at least one assumption, without saying which one
- Large p-value
 - this particular test did not detect a problem

Erroneous Scientific Reasoning

- Rejection as positive or even definitive evidence in favor of some preferred alternative hypothesis
 - logical fallacy
- make scientific conclusions based on whether or not a p-value crosses the 0.05 threshold
- confuse statistical significance and practical importance
- believe a result with a p-value below 0.05 is evidence that a relationship is causal

Erroneous Scientific Reasoning

- Rejection as positive or even definitive evidence in favor of some preferred alternative hypothesis
 - logical fallacy
- make scientific conclusions based on whether or not a p-value crosses the 0.05 threshold
- confuse statistical significance and practical importance
- believe a result with a p-value below 0.05 is evidence that a relationship is causal



NHST encourages researchers to engage in dichotomous thinking

“the difference between ‘significant’ and ‘not significant’ is not itself statistically significant.”

-Gelman and Stern (2006)

Erroneous Scientific Reasoning

In medicine, epidemiology, cognitive science, psychology, and economics

- i. interpret p-values dichotomously rather than continuously, focusing solely on whether or not the p-value is below 0.05 rather than the magnitude of the p-value

Erroneous Scientific Reasoning

In medicine, epidemiology, cognitive science, psychology, and economics

- i. interpret p-values dichotomously rather than continuously, focusing solely on whether or not the p-value is below 0.05 rather than the magnitude of the p-value
- ii. fixate on p-values even when they are irrelevant, for example, when asked about descriptive statistics

Erroneous Scientific Reasoning

In medicine, epidemiology, cognitive science, psychology, and economics

- i. interpret p-values dichotomously rather than continuously, focusing solely on whether or not the p-value is below 0.05 rather than the magnitude of the p-value
- ii. fixate on p-values even when they are irrelevant, for example, when asked about descriptive statistics
- iii. ignore other evidence, for example, the magnitude of treatment differences

Misinterpretation of the p-Value

The p-value has often been misinterpreted as:

- I. The probability that the null hypothesis is true
- II. one minus the probability that the alternative hypothesis is true
- III. one minus the probability of replication

Example(Gigerenzer(2004))

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t-test and your result is significant ($t = 2.7$, d.f. = 18, $p = 0.01$). Please mark each of the statements below as “true” or “false.” “False” means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

1. You have absolutely disproved the null hypothesis (that is, there is no difference between the population means)

2. You have found the probability of the null hypothesis being true.

3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).

4. You can deduce the probability of the experimental hypothesis being true.

5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.

6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

example of research conducted on psychology professors, lecturers, teaching assistants, and students

1. Subjects were given result of a simple t-test of two independent means
 2. ($t = 2.7$, $df = 18$, $p = 0.01$)
 3. asked six true or false questions based on the result and designed to test common misinterpretations of the p-value
 4. **ALL** six of the statements were false
 5. study materials noted “several or none of the statements may be correct,”
 - i. none of the 45 students
 - ii. only four of the 39 professors and lectures who did not teach statistics
 - iii. only six of the 30 professors and lectures who did teach statistics
- marked **ALL** as false
- each group marked an average of 3.5, 4.0, and 4.1 as false

Problems Specific to the Benjamin et al. (2018) Proposal

1. Benjamin et al. (2018) propose the 0.005 threshold because:
 - i. „corresponds to Bayes factors between approximately 14 and 26”
 - ii. “would reduce the false positive rate to levels we judge to be reasonable.”

➡ little to no justification is provided for either of these choices of levels.

2. “restrict [their] recommendation to claims of discovery of new effects” problematic for two reasons
 - i. proposed policy is rendered entirely impractical
 - ii. the proposed policy would lead to incoherence

Problems Specific to the Benjamin et al. (2018) Proposal

1. Benjamin et al. (2018) propose the 0.005 threshold because:
 - i. „corresponds to Bayes factors between approximately 14 and 26”
 - ii. “would reduce the false positive rate to levels we judge to be reasonable.”

➡ little to no justification is provided for either of these choices of levels.

2. “restrict [their] recommendation to claims of discovery of new effects” problematic for two reasons
 - i. proposed policy is rendered entirely impractical
 - ii. the proposed policy would lead to incoherence

➡ Specifically, given one study with $p < 0.005$ and another with $p \in (0.005, 0.05)$, it would matter crucially which study was conducted first.

Problems Specific to the Benjamin et al. (2018) Proposal

- A study would be deemed a success under the Benjamin et al. (2018) proposal if the first study was the $p < 0.005$ study but a failure otherwise
- 3. uncorrected multiple comparisons—both actual and potential—are the norm in applied research
- 4. mathematical justification has come under no small amount of criticism
 - i. Consequently, the logic underlying the proposal to move to a lower p-value threshold avoids firmly confronting the nature of the issue
 - ii. tradeoff between Type I and Type II error
 - iii. tradeoff should depend on the costs, benefits, and probabilities of all outcomes

Problems Specific to the Benjamin et al. (2018) Proposal

the more stringent 0.005 threshold Pro and Contra

PRO:

- short term: could reduce the flow of low quality work
- medium term: motivate researchers to perform higher-quality work

CONTRA:

- lead to more overconfidence
- concomitant greater exaggeration of the effect sizes
- discounting of important findings



there are better approaches to statistical analyses than null hypothesis significance testing

Part II: Summary and recommendations

FOR AUTHORS, EDITORS AND REVIEWERS
& OUTSIDE SCIENTIFIC PUBLISHING

Abandoning Statistical Significance

– But what to do instead?

Authors' position:

- no quick fixes – it is not that simple!
- the proposed solutions discussed so far
 - Changing p-value threshold
 - Using confidence intervals instead (H0-representing value inside vs. outside)
 - Using Bayes factors

All suffer from the same issues:

- ▶ explicitly or implicitly use a threshold relative to implausible hypothesis
- ▶ purely statistical measures – no holistic view on evidence

Holistic view: Currently subordinate factors



General problem with statistical thresholds:
binary statement

- ▶ false promise of certainty offered by dichotomization
- ▶ Need more radical approach than just other statistical measures or thresholds

Subordinate factors:

- related prior evidence
- plausibility of mechanism
- study design
- data quality
- real world costs and benefits
- novelty of findings
- other factors that vary by research domain

Recommendations – for **authors**

1. Use the currently subordinate factors to motivate
 - data collection
 - statistical analysis
 - interpretation of results
 - writing
 - and related matters
2. Analyze and report all data and relevant results → instead of focusing on single comparisons exceeding a statistical threshold (like p-value)

Specific recommendation 1

Include in the manuscript a ***section*** that directly addresses how each of the currently subordinate factors

- prior evidence,
- plausibility of mechanism,
- study design and data quality,
- real world costs and benefits,
- novelty of finding,
- other factors specific to research domain

motivated decisions regarding

- data collection,
- statistical analysis,
- interpretation of results, and
- writing

in the context of the totality of the data and results.

How? - For example:

- Discuss study design in the context of subject-matter knowledge and expectations of effect sizes (as discussed by Gelman and Carlin (2014))
- Discuss the plausibility of the mechanism by
 - i. formalizing the hypothesized mechanism for the effect in question and explaining the various components of it,
 - ii. clarifying which components were measured and analyzed in the study, and
 - iii. discussing aspects of the results that support AND those that undermine the hypothesized mechanism

Specific recommendation 2

Analyze and report ALL data and relevant results!



= A fundamental principle of science!

BUT status quo is:

- results published only if $p < 0.05$
- focus on such results and to not report all relevant findings

HOW to do that? Instead of recipe-like guidance

Examples: Case study + Publications from different domains

- Clinical psychology (Tackett et al. 2014)
- Epidemiology (Gelman and Auerbach 2016a,b)
- Political science (Trangucci et al. 2018),
- Program evaluation (Mitchell et al. 2018),
- Social psychology and consumer behavior (McShane and Böckenholt 2017)

Recommendation – for **editors and reviewers**

- Evaluate papers with regard to statistical measures BUT ALSO the currently subordinate factors! → independent of statistical threshold reached
- Consider subordinate factors at all stages of the review process

Specific suggestions:

- ▶ Reviewers have to quantitatively evaluate each factor + overall quantitative evaluation of the strength of evidence (besides current mostly qualitative evaluations)
- ▶ Weigh quantitative evaluations by editors' or reviewers importance rating of each factor
- ▶ Editors could address each factor in decision letters for a more holistic view of the evidence

BUT: Do not Editors and Reviewers need clear decision rules?

- ▶ Statistical thresholds needed to decide if results are far enough from noise and worth publishing?
- ▶ Statistical thresholds as objective standards for evidence – protection against subjectivity and bias of editors and reviewers?

Authors argue: NO

- Even if threshold needed → no sense to base it on p-value
- p-value itself is not an objective standard → different model specifications and statistical tests may result in different p-values
- Many decisions on data protocols and analysis are subjective and can impact reported p-value
- No threshold screening rule needed: publication decisions are already made based on qualitative factors

 **MAIN POSITION: NO single number can eliminate subjectivity and personal biases!**

It is not that simple and dichotomous!

- Papers should be published if relevant to research question and if interpretation is accurate → even with $p > 0.05$ or CI including H_0 -value
- Should also be possible to publish a result with a $p < 0.001$ without interpreting this as truth of some favored H_1
- The p-value *IS* relevant to the question of how easily a result could be explained by a particular null model
- ▶ BUT this should not be the crucial factor in publication!



Results can be relevant to science or politics even if consistent with the null model
AND vice versa (reject a null model without offering anything of scientific interest or policy relevance)

Case study

TO ILLUSTRATE IMPLEMENTATION OF RECOMMENDATIONS

Hypothetical case study: Effects of sodium on blood pressure

Related prior evidence: Researchers consider evidence on blood pressure as marker of healthy arteries, role of sodium in blood flow etc. in publication

Plausible mechanism: To get rid of excess sodium the blood pressure has to be increased

Study design + data quality:

- How to recruit subjects?
- Randomization to high vs. low sodium diet?
- Or observational study with longitudinal tracking of annual checkups?
- Is data available from other studies?
- When and how often measure sodium and blood pressure?
- How to measure it? (can be quick and noisy or more accurate but expensive)

Hypothetical case study II

Researchers do a statistical analysis (no NHST and no thresholding of p-values)

Result A:

- $p = 0.001$ is found → supports hypothesis on the effect
- ▶ But can they conclude sodium is associated with/causes high blood pressure? (NHST paradigm would)
- ***It depends....*** on context and limitations in study design and data quality (for ex.: results may not generalize to other cultures)
- ***Causal interpretation*** depends on prior evidence, plausible mechanism, study design, and data quality – causality supported by
 - consistent and strong associations between sodium consumption and blood pressure
 - evidence from physiological studies and animal models consistent with a causal effect
 - randomization to different sodium levels

Hypothetical case study III

If *causal interpretation*:

Authors could consider *clinical significance* next → real world *costs and benefits*

► Depends on

- estimates of magnitude of the effects on blood pressure and other conditions
- on the uncertainty associated with effects
- but NOT on the p-value!

► AND depends on costs of potential interventions

Novelty of findings should be discussed in light of all these aspects

Hypothetical case study IV

Result B:

- $p = 0.2$ is found → consistent with H_0
- ▶ Can they conclude sodium is not associated with high blood pressure?

Again, this depends on the subordinate factors

Key points:

1. more **holistic view** of evidence: consider subordinate factors in any case and statistical measures along with these factors as one piece of evidence → independent of p-value or other threshold
2. Statistical measures are treated **continuously** in this view of the evidence
 - Lower p-value means continuously stronger evidence → regardless of the level of the p-value
 - Continuous evidence should be balanced with strengths and weaknesses of subordinate factors in assessing the level of support for a hypothesis

Hypothetical case study V

More complex analyses instead of p-value

For example:

- Multilevel model of association between sodium and blood pressure as function of additional (health and dietary, demographic, geography) variables
- Would give many estimates that vary based on other variables plus the uncertainty in these estimates
- ▶ Does NOT encourage binary statements about “an effect” or “no effect”
- Instead: accepting uncertainty and embracing variation in effects
- ▶ Tells a richer story about the association between sodium and blood pressure

Hypothetical case study VI

Editors and reviewers

Should consider the same factors in evaluating the submitted paper on sodium and blood pressure → evaluate:

- How does the paper fit in with and build upon related prior evidence?
- Is the mechanism plausible?
- Are the study design and data quality sufficient to justify the conclusions?
- What are the implications in terms of real world costs and benefits?
- How novel are the findings?
- How appropriate are the statistical analyses and how strong is the statistical support from these analyses (incl. p-value)?

Recommendations - **outside scientific publishing**

– Examples and specific recommendations

Similar issues with p-values in other areas of statistical decision making - *Examples:*

- **Neuroimaging:** voxelwise (voxel = 3-dimensional pixel) NHSTs to decide which results to take seriously → BETTER: reporting voxelwise estimates and uncertainty of brain activity changes
- **Medicine:** agencies use NHSTs to decide whether or not to approve new drugs
- **Policy analysis:** organizations use NHSTs to determine whether interventions are beneficial or not
- **Business:** managers use NHSTs to make binary decisions via A/B tests.
 - ▶ Cost-benefit calculations are superior to acontextual statistical thresholds. Thresholds implicitly express a particular tradeoff between Type I and Type II error - BUT this tradeoff should depend on the costs, benefits, and probabilities of all outcomes.
 - ▶ Nonstatistical thresholds are sometimes useful here (for example a firm may want to send an offer only to customers that yield a profit greater than a specific threshold)

Examples and specific recommendations II

Research projects: NHSTs used to decide how to go on based on preliminary findings

- ▶ No obvious cost-benefit calculation → for example, when comparing possible modes of actions of two drugs to treat some disease
- STILL: No need for statistical thresholds
- ▶ INSTEAD report ALL results: estimates, standard errors, CIs etc., even if inconclusive
- p-values for screening and decisions on which ideas or variables to pursue: thresholds simplify decisions BUT do not use data efficiently: p-values aren't connected to gains of pursuing specific lines of research or the probability that this will be successful
- ▶ INSTEAD: decisions should be based on a model of effect size distributions and variation
→ work with hypotheses of interest *directly* instead of reasoning from null model (*indirectly*)

Further (statistical) recommendations

In all these settings - when possible:

- Use more precise individual-level measurements
- Use within-person or longitudinal designs
- Increased consideration of models that use informative priors
- Models that feature varying treatment effects
- Multilevel (mixed-effect) or metaanalytical models

How to go on....

...to get to this more holistic approach and embrace uncertainty instead of relying on p-values or other statistical thresholds?

Possible problems with recommendations:

- ▶ Researchers often have to satisfy expectations of funding agencies in study design as well as editors and reviewers in publishing (and funding agencies can only fund and editors can only publish research which is submitted to them)
- ▶ This may often be more promising when relying on the traditional p-value paradigm than admitting uncertainties

Changes may need time

- discussed problems are difficult to solve

Improvements expected from applied and methodological research:

- Examples using improved methods demonstrate that it is possible to perform successful statistical analyses without binary decisions
 - ▶ Improved methods that move beyond NHST, f.e. include multilevel modeling, machine learning, statistical graphics, and other tools for analyzing and visualizing large amounts of data
- Theoretical work on statistical effects of selection based on statistical significance and other decision criteria
- Criticism of published work with overestimates of effect sizes or inappropriate claims of certainty
 - ▶ Change will likely require institutional reform with major modifications of current practices of funding agencies and editors and reviewers

Discussion

Authors' main point: Abandon statistical significance

from scientific publishing and overall statistical decision making

- ▶ not to completely “ban” p-values or other purely statistical measures
- ▶ Rather, do not threshold them and do not give priority over the currently subordinate factors



p-values are only one piece of information/evidence!

Are recommendations too radical?

Abandoning statistical significance may seem radical - BUT:

At least treating p-values (or other statistical measures) continuously is neither new, nor exceptional

- ▶ Early statisticians (e.g. Fisher, 1958; Pearson, Cox, 1977, 1982; Lehmann, 1993) already advocated for continuous treatment
- ▶ Others outside of statistics advocated for it (for example Eysenck, 1960, well known in psychology) as early as 1919 and again during the recent debate
- ▶ ALSO consistent with ASA (American Statistical Association) Statement on Statistical Significance and p-values:

“Principle 3: Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold” - Wasserstein and Lazar 2016


- ▶ BUT in direct opposition to the threshold-based status quo and Benjamin et al. (2018) proposal


Recommendations go beyond that

1. p-values or other purely statistical measures (thresholded or not) should not take priority over the subordinate factors → ALSO emphasized by ASA statement
2. Treating the p-value continuously rather than in a thresholded manner is better BUT recommendations go even further: it seldom makes sense to calibrate evidence as a function of the p-value or other purely statistical measures
3. Recommendations for authors, editors and reviewers given how to implement this proposal in publishing, and also in statistical decision making more broadly

Conclusion

Recommendations will not resolve the replication crisis → BUT they may push researchers

 **AWAY** from the pursuit of irrelevant statistical targets...
TOWARD understanding of theory, mechanism, and measurement

 And hopefully **BEYOND** the paradigm of routine “discovery” and binary statements...
TO a paradigm of continuous learning, accepting uncertainty and variation

Abandon Statistical Significance(?)

Time for
discussion and questions

Thank you for your
attention!