Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don’t Expect Replication

Reading course - A world beyond $p < 0.05$

Kai Mandelkow, Elisabeth Spies, Swen Simon

January 2, 2021

Table of contents

Roadmap

	Adopting more holistic approaches
21.12.20	Billheimer – Predictive Inference and Scientific Reproducibility
04.01.21	Amrhein et al. – Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication
11.01.21	McShane et al. – Abandon Statistical Significance
18.01.21	Ziliak – How Large Are Your *G*-Values? Try Gosset's Guinnessometrics When a Little "*p*" Is Not Enough
25.01.21	van Dongen et. al – Multiple Perspectives on Inference for Two Simple Statistical Scenarios

Introduction

Introduction

- “crisis of unreplicable research” is not only about alleged replication failures
    - also nonreplication is often interpreted as a sign of bad science
- epidemic of misinterpretation of statistics:
    - leads to scientific misconduct
    - unfortunately, often common practice
        - e.g. selectively reporting of studies that were significant
- all results are uncertain
    - even those from the most rigorous studies

“No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon.”
(Sir Ronald Fisher 1937)

Introduction

- “scientific generalization is a broader question than mathematical description” (Boring 1919)
    - today students still indoctrinated with methods that claim to produce scientific generalizations from mathematical descriptions of isolated studies
    - such generalizations often fail to agree with those from other studies and thus statistical inference will fail to replicate
- a core problem is confounding statistics with reality
    - statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, models are extremely simplified compared to reality.

Introduction

- statistical results must mislead when communicated as representing the complex reality
  - not a problem of the model but of communication and interpretation
- we should use, communicate and teach inferential statistical methods as describing logical relations between assumptions and data
- not as a tool for providing generalizable inferences about universal populations.

The problem with hypotheses and the p-value

Inferences are not about hypotheses

- statistical models are sets of assumptions
- a model is a hypothesis how data could have been generated
- model matches reality to degree assumptions are met
    - starting from assumptions that we measured what we think, measurement errors were absent, sample was random sample, iid. of residuals,...
    - models imply countless assumptions about underlying reality

Inferences are not about hypotheses - Example

- a P-value refers not only to a hypothesis it claims to test (H0)
- a P-value refers to entire model
    - including other usually explicit assumptions
        - randomization of treatment
        - linearity of effects
        - …
    - plus implicit assumptions:
        - no measurement errors
        - …
- a small P-value is the net result of some combination of random variation and violations of model assumptions but does not indicate which assumption is violated!

Replication studies have a false negative problem

Variation

- varying assumption violations
- observed effect sizes can differ

$\Rightarrow$ variation from replication to replication

Replication studies have a false negative problem - Example

- 97 out of 100 psychological studies ($p \leq 0.05$)
- 35 replications had $p \leq 0.05$
- average power of 92%
  (= 89 of the 97 replicates were expected to have $p \leq 0.05$)

Replication studies have a false negative problem - Example

- 97 out of 100 psychological studies ($p \leq 0.05$)
- 35 replications had $p \leq 0.05$
- average power of 92%
  (= 89 of the 97 replicates were expected to have $p \leq 0.05$)

Explanation:

- results were selected for reporting
- original “significant” studies were mostly or entirely false positives.

62 replications with $p > 0.05$ = 64% of the 97 replications

Replication studies have a false negative problem - Example

- 97 “significant” original studies
- 70% (= 68) false H_0
- average power in non-null cases = 50% in the replication studies

Replication studies have a false negative problem - Example

- 97 "significant" original studies
- 70% (= 68) false H_0
- average power in non-null cases = 50% in the replication studies

$\Rightarrow$ expected "true positive" replications $0.50 \times 68 = 34$

$\Rightarrow$ expected "false positive" replications $= 0.05(97 - 68) = 1.45$

Replication studies have a false negative problem - Example

- 97 “significant” original studies
- 70% (= 68) false H_0
- average power in non-null cases = 50% in the replication studies

$\Rightarrow$ expected “true positive” replications $0.50 \times 68 = 34$

$\Rightarrow$ expected “false positive” replications $= 0.05(97 - 68) = 1.45$

$\Rightarrow$ total of $\sim$ 35 out of 97 replications having $p \leq 0.05$, as observed.

$\Rightarrow$ given selective reporting in the original studies, the observed 64% of the 97 replication attempts with $p > 0.05$ could have been expected even if only $97 - 68 = 29$ or 30% of H_0 were correct!

Replication studies have a false negative problem - Result

Selective reporting in the original studies

$\Rightarrow$ "nonsignificant" results in about two thirds of replications.

BUT

False-negative errors still present, even if

- no selective reporting
- only random variation present

Replication studies have a false negative problem - Example

- statistical power of 80%
- two conflicting studies
- replication's statistical power $\sim 100\%$

Replication studies have a false negative problem

Allow false-negative and false-positive errors

"Variation, and hence non replication, is the norm across honestly reported studies"
(Amrhein, Trafimow, Greenland, 2019)

Replication studies have a false negative problem

Take a look at prior information

BUT

Don't focus on estimates only!

Example: study imperfections

95% confidence interval $\neq$ 95% coverage of the true effect!

Overconfidence triggers Selection BIAS

Every reporting is biased!

Combination of studies

$\Rightarrow$ no guarantee of valid inferences

Overconfidence in statistical inference

$\Rightarrow$ result-selection bias

Overconfidence triggers Selection BIAS

Reduce selective reporting by providing all information:

- how the study was conducted
- what problems occurred
- what analysis methods were used
- detailed data tabulation and graphs
- complete reporting of results

"Move toward a greater acceptance of uncertainty and embracing of variation" (Gelman 2016)

Don’t blame the p-value - Example

- H_0 was correct
- ideal experimental conditions

$\Rightarrow$ P-value will vary uniformly between 0 and 1

- H_1 was correct
- ideal experimental conditions

$\Rightarrow$ P-value in the next sample will typically differ widely from our current sample

“The fickle P-value generates irreproducible results”
(Halsey et al. 2015)

Ban Statistical Tests?

Ban Tests?

Ban Tests?

Ban some practices

- temporarily
- in specific contexts

Ban Tests?

Ban some practices

- temporarily
- in specific contexts

$\rightarrow$ force researchers to learn how to analyze data in alternative ways

Ban Tests?

Ban some practices

- temporarily
- in specific contexts

$\rightarrow$ force researchers to learn how to analyze data in alternative ways

$\rightarrow$ may lead to misuse and abuse of other methods

Statistical Testing and Alcohol

Statistical Testing and Alcohol

- using statistical tests to force out inferences has become a *culturally ingrained habit*
- statistical testing often *gives the impression that complex decision can be oversimplified without negative consequences*
- researchers become in a sense *addicted* to such oversimplification

Long live no King

Banish the King?

Banish the King?

- fixed-cutoff hypothesis testing has been king for over 80 years
- it might still be useful in some cases
    - e.g. quality control
- abandon it in favor of data description and direct presentation of precise P-values for scientific inferences
- that also includes P-values for alternative hypotheses
- applies to any other statistical criterion as well

Long live no King

- evidence need to be weighted against or in favor of a scientific hypothesis
- statistical tests cannot suffice for that
- could even be destructive if degraded into a binary decision
- especially when results are sensitive to doubtful assumptions
    - e.g. absence of measurement-error dependencies

Options going forward

What comes next?

What comes next?

- most common proposal: replace hypothesis tests with interval estimates
- classical confidence interval is nothing more than a summary of dichotomized hypothesis tests
  $\rightarrow$ does not solve the core psychological problems

Empire of diversity

- empire of ”statistical significance” reached it’s dominance with the spread of cutoffs for testing
- relic of past era
- there is no substitute for accepting methodologic diversity
- careful assessment of uncertainty as the core motivation for statistical practice

The replacement for hypothesis testing

"Don't look for a magic alternative to NHST (Null Hypothesis Significance Testing), some other objective mechanical ritual to replace it. It doesn't exist"
(Cohen 1994)

- what needs to change is not necessarily the statistical methods we use
- but how we select our results for interpretation and publication and what conclusions we draw
  - as mentioned, every selection criterion would introduce BIAS

The replacement for hypothesis testing

Use the following steps to extent the feasible:

- target results for publication and interpretation before data are collected
- before analyzing data (and preferably before collecting them) make an analysis plan
- emphasize and interpret estimates rather than tests
- when reporting statistics, give their precise values rather than mere inequalities

The replacement for hypothesis testing

- do not use words “significant” or “confidence” to describe scientific results
- acknowledge that statistical results describe relations between assumptions and the data in the study and that scientific generalization from a single study is unwarranted
- openly and fully report detailed methods, materials, procedures data and analysis scripts

Example

Consider a study by Brown et al (2017) who reported that, "in utero serotonergic antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child"

- based on an estimated hazard rate ratio (HR) of 1.61
- a 95% confident interval of [0.997,2.59]
- as it is often the case, the authors misused the CI as a hypothesis test
- claimed to have demonstrated no association because lower limit was slowly below no association (HR=1)
- ignoring that the upper limit exceeded 2.59

Example

A more correct summary of the results would have been:

- the estimate of the hazard rate ratio was 1.61 and thus exposure could be associated with autism
- however, possible hazard rate ratios that are highly compatible with the data given the model ranged from [0.997,2.59]
- this could be followed by a discussion of why the authors seem to think the exposure effect might be irrelevant despite the association

Example

- had the authors found an interval [1.003,2.59] the reporting should have been the same
- even with an interval [0.900,2.59] the description of the results should largely be the same – The point estimate would still be a HR well above 1, indicating a possible positive association.

Anything goes?

- what do we conclude from a study like Brown et al (2017)?
- if we interpret [1.003,2.59] and [0.997,2.59] in the same way, does that mean that the floodgates of “anything goes” are wide open?
- everything should be published in some form
- publish if whatever you measured made sense before you obtained the data because it was connected in a potentially useful way to some research question
- even if after doing the study it appears the measure did not make sense or the methods were faulty

Anything goes?

- however, the floodgates should be closed for drawing conclusions from virtually any single study
- for example: because they found a CI that barely included the null value, Brown et al reported a conflict with previously observed associations that were nearly in the same size (HR rates about 1.7)
- goal of easy entry into meta-analyses

Abandon statistical inference

- no suggestion to completely abandon inference from our data to a larger population
- inference must be scientific rather than statistical
- all statistical methods require subjective choices, so there is no objective decision machine for automated scientific inference
- we must make the inferences and so claims about a larger population will always be uncertain

When can we be confident that we know something?

Abandon statistical inference

When can we be confident that we know something?

- a successful theory is one that survives decades of scrutiny
- if every study claims to provide decisive results, there will be ever more replication failures, which in turn will further undermine public confidence in science
- decision makers must act based on cumulative knowledge (not rely solely on single studies or even single lines of research)

Advice to researchers and journalists

If we are researchers...

- don't claim that the statistics indicate that there is no effect
  - even if the data remain consistent with a zero effect, they remain consistent with many other effects as well
  - lots of additional hypotheses outside the interval estimate will also be compatible with our data (due to methodologic limitations that we have not modeled)
  - almost never will we have found absolutely no effect
- "perfectly compatible" with one hypothesis, or model, does not mean that all other hypotheses, or models, are refuted

If we are researchers...

- remember the "dance of the confidence intervals"
- treat statistics as descriptions of the relation of the model to the data rather than as statements about the correctness of the model
- a small P-value is just a warning signal that the current model could have a problem

If we are researchers...

- science includes learning about assumption violations, then addressing those violations and improving the performance of our models about reality
- be honest and thorough in the description and discussion of our methods and of our data
- for **journal editors**: consider “results blind evaluation” of manuscripts

If we are scientific writers and journalists...

- continue writing about isolated experiments and replication
- if we think we found a good study, or a bad study, we may report it
- don't be impressed by what researchers say is surprising

If we are scientific writers and journalists...

- continue writing about isolated experiments and replication
- if we think we found a good study, or a bad study, we may report it
- don't be impressed by what researchers say is surprising
    - surprising results are often products of data dredging or random error

If we are scientific writers and journalists...

- continue writing about isolated experiments and replication
- if we think we found a good study, or a bad study, we may report it
- don't be impressed by what researchers say is surprising
    - surprising results are often products of data dredging or random error
      $\rightarrow$ makes them less reproducible
      $\rightarrow$ often do not point to general scientific discoveries

If we are scientific writers and journalists...

- continue writing about isolated experiments and replication
- if we think we found a good study, or a bad study, we may report it
- don't be impressed by what researchers say is surprising
    - surprising results are often products of data dredging or random error
      $\rightarrow$ makes them less reproducible
      $\rightarrow$ often do not point to general scientific discoveries
    - may still be valuable because they lead to new insights about study problems and violations of assumptions

If we are scientific writers and journalists...

- consider asking for the most boring results rather than what was surprising or unsurprising
    - those that were shown several times before and thus seem to be most trustworthy
- try looking for signs of overconfidence
    - we proved/ disproved
    - there was no effect/ association/ difference
    - our study confirms/ validates/ invalidates/ refutes previous results

Conclusion

Conclusion

- we generally cannot decide whether a result from a single study can be generalized
- important role for statistics in research is the summary and accumulation of information
- if replications do not find the same results, this is not necessarily a crisis, but is part of a natural process by which science evolves
- goal of scientific methodology: direct this evolution toward ever more accurate descriptions of the world and how it works, not toward ever more publication of inferences, conclusions, or decisions

Additional Information

A Descriptive View of P-values and Posterior Probabilities

Model Test - Fisherian P-value

- data-generating model M
- test statistic T
- observed value t

- test gave back p = 0.04

$\Rightarrow P(T \geq t|M) = 0.04$

A Descriptive View of P-values and Posterior Probabilities

$P(x|M)$ (x observed data)
$\Rightarrow P(x, M) = P(x|M)P(M)$

posterior P(M|x) becomes a deduction from the observed data x and the full model P(x,M)

Problem: choice of prior P(M)

References

1. Amrhein,V., Trafimow, D, Greenland, S. (2019), ”Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don’t Expect Replication”
2. Cohen, J. (1994), “The Earth Is Round (p < .05),” American Psychologist, 49, 997–1003. [265,266]
3. Halsey,L.G., Curran-Everett,D., Vowler,S.L., and Drummond, G. B. (2015), “The Fickle P-value Generates Irreproducible Results,” Nature Methods, 12, 179–185. [264]
4. Gelman, A. (2016), “The Problems with P-values are not Just with P-values,” The American Statistician, Supplemental Material to the ASA Statement on P-values and Statistical Significance. [264]
5. Fisher, R. A. (1937), The Design of Experiments (2nd ed.), Edinburgh: Oliver and Boyd. [262,264]
6. Boring, E. G. (1919), “Mathematical vs. Scientific Significance,” Psychological Bulletin, 16, 335–338. [262]
7. Brown,H. K.,Ray,J.G.,Wilton,A.S., Lunsky,Y., Gomes, T.,andVigod, S.N. (2017), “Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children,” JAMA: Journal of the American Medical Association, 317, 1544–1552. [266,267]

Discussion