Do pre-analysis plans protect against false discoveries?

Workshop on Pre-Analysis Plans for the Statistical Analysis of Large-Scale and Complex Datasets
British National Centre for Research Methods

Philip B. Stark

Department of Statistics, University of California, Berkeley

28 October 2021

In replication lies truth.

Why preregister?

- Reduce *P*-hacking
- Reduce file-drawer effect
- Improve reproducibility
- Allow scrutiny before it's too late

Costs

- Infrastructure, time, ...
- Makes analysis inflexible
- Can reduce power
- Encourages gaming
- Doesn't address core problems: multiplicity, selection/conditioning, (p)reproducibility

Issue 2.4, Fall 2020 ••• 2 more Published on Dec 16, 2020 DOI 10.1162/99608f92.fc62b261

Selective Inference: The Silent Killer of Replicability

by Yoav Benjamini

Published on Dec 16, 2020

last released
10 months ago

ABSTRACT

Replicability of results has been a gold standard in science and should remain so, but concerns about lack of it have increased in recent years. Transparency, good design, and reproducible computing and data analysis are prerequisites for replicability. Adopting appropriate statistical methodologies is another identified one, yet which methodologies can be used to enhance replicability of results from a single study remains controversial. Whereas the *p*-value and statistical significance are carrying most of the blame, this article argues that addressing selective inference is a missing statistical cornerstone of enhancing replicability. I review the manifestation of selective inference and the available ways to address it. I also discuss and demonstrate whether and how selective inference is addressed in many fields of science, including the attitude of leading scientific publications as expressed in their recent editorials. Most notably, selective inference is attended when the number of

Selective inference is focusing statistical inference on some findings that turned out to be of interest only after viewing the data. Without taking into consideration how selection affects the inference, the usual statistical guarantees offered by all statistical methods deteriorate. Since selection can take place only when facing many opportunities, the problem is sometimes called the multiplicity problem.

Both *invisible* & *evident*, even w/ preregistration.

- Selecting for emphasis, e.g., abstract, table or figure
- Emphasizing endpoints w/ small P-values or CIs that exclude "no effect"
- Selecting from the literature
- Failing to account for multiplicity or model selection in P-values & CIs

Controlling for multiplicity & selection

- Multiplicity adjustments to control FWER (Bonferroni, Sidak, Holm, ...)
- $\mathrm{FDR} := \mathbb{E}\frac{F}{1 \vee R}$
  - Benjamini & Hochberg, variants
  - Knockoffs
  - AdaPT
- $\mathrm{FCR} := \mathbb{E}\frac{N}{1 \vee M}$
- Conditional on selection: $\mathbb{P}\{\mathrm{CI}_i \ni \mu_i | i \text{ is selected}\} \geq 1 - \alpha$.
- Simultaneous over selected, ...

J. R. Statist. Soc. B (2018)
80, Part 4, pp. 649–679

AdaPT: an interactive procedure for multiple testing with side information

Lihua Lei and William Fithian

University of California, Berkeley, USA

[Received January 2017. Revised March 2018]

Summary. We consider the problem of multiple-hypothesis testing with generic side information: for each hypothesis H_i we observe both a p-value p_i and some predictor x_i encoding contextual information about the hypothesis. For large-scale problems, adaptively focusing power on the more promising hypotheses (those more likely to yield discoveries) can lead to much more powerful multiple-testing procedures. We propose a general iterative framework for this problem, the adaptive p-value thresholding procedure which we call AdaPT, which adaptively estimates a Bayes optimal p-value rejection threshold and controls the false discovery rate in finite samples. At each iteration of the procedure, the analyst proposes a rejection threshold and observes partially censored p-values, estimates the false discovery proportion below the threshold and proposes another threshold, until the estimated false discovery proportion is below α. Our procedure is adaptive in an unusually strong sense, permitting the analyst to use any statistical or machine learning method she chooses to estimate the optimal threshold, and to switch between different models at each iteration as information accrues. We demonstrate the favourable performance of AdaPT by comparing it with state of the art methods in five real applications and two simulation studies.

1. Introduction

1.1. Interactive data analysis

In classical statistics we assume that the question to be answered and the analysis to be used in answering the question are both fixed in advance of collecting the data. Many modern applications, however, involve extremely complex data sets that may be collected without any specific hypothesis in mind. Indeed, very often the express goal is to explore the data in search of insights that we may not have expected to find. A central challenge in modern statistics is to provide scientists with methods that are sufficiently flexible to enable exploration, but that nevertheless provide statistical guarantees for the conclusions that are eventually reported.

Selective inference methods blend exploratory and confirmatory analysis by allowing a search over the space of potentially interesting questions, while still guaranteeing control of an appropriate type I error rate such as a conditional error rate (e.g. Yekutieli (2012), Lee *et al.* (2016) and Fithian *et al.* (2014)), familywise error rate (e.g. Tukey (1994) and Berk *et al.* (2013)) or false discovery rate (FDR) (e.g. Benjamini and Hochberg (1995) and Barber and Candès (2015)). However, most selective inference methods require that the selection algorithm be specified in

Address for correspondence: Lihua Lei, Department of Statistics, University of California at Berkeley, 397 Evans Hall, Berkeley, CA 94720, USA.
E-mail: lihua.lei@berkeley.edu

 1369–7412/18/80649

advance, forcing a choice between either ignoring any difficult-to-formalize domain knowledge or sacrificing statistical validity guarantees.

Interactive data analysis methods relax the requirement of a predefined selection algorithm. Instead, they provide for an interactive analysis protocol between the analyst and the data, guaranteeing statistical validity as long as the protocol is followed. The two central questions in interactive data analysis are 'what did the analyst know and when did she know it?'. Previous

More generally, prior information could arise in more complex ways. For example, consider testing for association of 400000 single-nucleotide polymorphisms (SNPs) with each of 40 related diseases. If gene regulatory relationships are known, then we might expect SNPs near related genes to be associated (or not) with related diseases, but without knowing ahead of time which gene–disease pairs are promising. In a similar vein, Fortney *et al.* (2015) used prior knowledge of each SNP's associations with age-related diseases to focus their search for SNPs that are associated with longevity, leading to novel discoveries. Inspired by examples like this,

Preregistration of Preregistration evaluation 2016

Public registration ▾

Overview
Files
Wiki
Components 0
Links 0
Analytics
Comments 0

Study Information

Title

Do current study pre-registrations on the Open Science Framework limit opportunistic use of researcher degrees of freedom in the design, data collection, analyses and reporting of studies?

Research Questions

Attempts to replicate original research findings seem uncommon in psychology (Asendorpf et al., 2013; Mahoney, 1985; Makel, Plucker, & Hegarty, 2012; Sterling, 1959), although some argue that the number of replication studies are underestimated (Neuliep & Crandall, 1993a, 1993b). When replications do occur, they often produce weaker or no evidence for the original findings (Asendorpf et al., 2013; Open Science Collaboration, 2015). In recent years, many psychologists have expressed their concerns about the size and the gravity of this problem (Nosek & Lakens, 2015; Open Science Collaboration, 2015; Pashler & Harris, 2012; Spellman, 2015).

Contributors

Coosje Lisabet Sterre Veldkamp, David Thomas Mellor, Marjan Bakker, Marcel A.L.M. van Assen, Jelte Wicherts, Brian A. Nosek, How Hwee Ong, Elise Anne Victoire Crompvoets, and Courtney K. Soderberg

Registration type

Prereg Challenge

Date registered

September 14, 2016

Date created

September 2, 2016

Registered from

osf.io/emh59

Data collection procedures

We randomly selected a total of 106 registrations from two sets of pre-registrations:

- 53 public OSF Standard Pre-Data Collection Registrations
- 53 public COS Prereg Challenge registrations

Because the two types of pre-registrations are collected from different data bases and because they differ in their standard features, we had slightly different selection procedures for each type of pre-registration. Both are outlined in the following paragraphs.

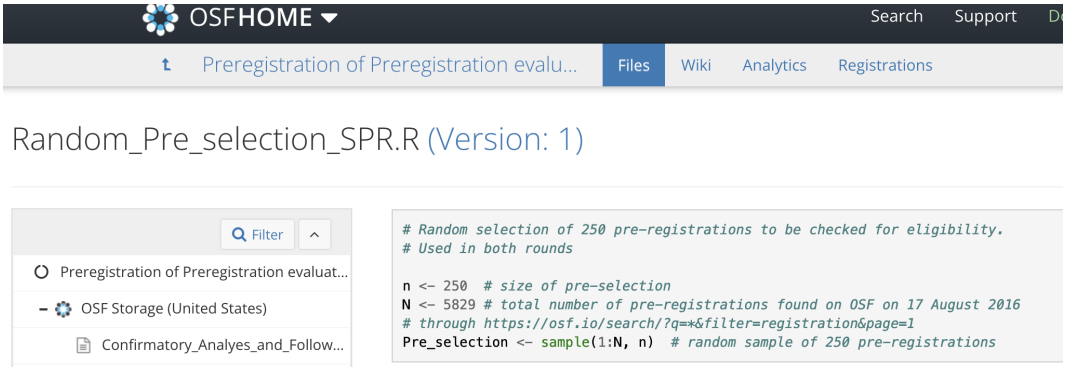
Protocols selection of pre-registrations

Public OSF-Standard Pre-Data Collection Registrations

The total number of pre-registrations listed on the Open Science Framework on Wednesday 17 August was determined by going to https://osf.io/search/?q=*&filter=registration&page=1 and viewing the total number of results for this search. Because we estimated that approximately 30% of the pre-registrations on the Open Science Framework would meet our inclusion criteria, we first selected a random set of 250 pre-registrations from the total of 5,829 results as a pre-selection. For this, we used the R (version 3.2.4) code in script 'Random_Pre_selection_SPR' which can be found in the component 'Scripts'.

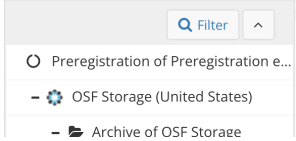
A group of researchers from the Meta-Research Center at the Tilburg School of Social and Behavioral Sciences (see metaresearch.nl) created a list of 34 degrees of freedom that researchers have in formulating hypotheses, and in designing, running, analyzing, and reporting of psychological research (Wicherts et al., under review). The list was composed in order to raise awareness of the risk of bias implicit in a lot of research designs in psychology and other fields, and in order to serve as a basis for the current study. Based on this list, we created a protocol to evaluate pre-registrations in terms of the extent to which they restrict 29 out of the 34 researcher degrees of freedom on the list. The reason that we only included 29 out of the 34 researcher degrees of freedom in this protocol in that five of the researcher degrees of freedom only concern the reporting phase of a study and therefore are not expected to be restricted by a pre-registration itself.

In the current study, we will apply our protocol to two sets of pre-registrations on the


OSFHOME
Search
Support
Preregistration of Preregistration evalu...
Files
Wiki
Analytics
Registrations
Random_Pre_selection_SPR.R (Version: 1)
Filter
Preregistration of Preregistration evaluat...
OSF Storage (United States)
Confirmatory_Analyes_and_Follow...
# Random selection of 250 pre-registrations to be checked for eligibility.
# Used in both rounds
n <- 250 # size of pre-selection
N <- 5829 # total number of pre-registrations found on OSF on 17 August 2016
# through https://osf.io/search/?q=*&filter=registration&page=1
Pre_selection <- sample(1:N, n) # random sample of 250 pre-registrations

Randomization_Pre_selection_PCR.R (Version: 1)

Download

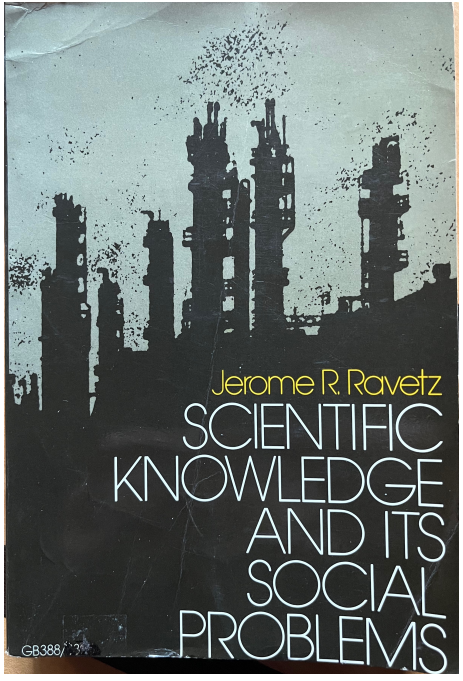



```r
# Randomization of order of 122 pre-registrations to be checked for eligibility.

n <- 122  # size of pre-selection
N <- 122 # total number of public Prereg Challenge registrations provided by
# the Centre of Open Science on August 16, 2016
Pre_selection <- sample(1:N, n)  # selection in random order
```

```r
# RDF T1
wilcox.test(data_non_imputed$T1[data_non_imputed$group==1],data_non_imputed$T1[data_non_imputed$group==0],
alternative = "two.sided", conf.int = TRUE, exact = F )
# means and standard deviations
# Standard Pre-Data Collection Registrations = DF Restriction Score DFRS_T1_SPR
mean(data_non_imputed$T1[data_non_imputed$group==0])
sd(data_non_imputed$T1[data_non_imputed$group==0])
# Prereg Challenge Registrations = DF Restriction Score DFRS_T1_PCR
mean(data_non_imputed$T1[data_non_imputed$group==1])
sd(data_non_imputed$T1[data_non_imputed$group==1])
# overall DF Restriction Score DF T1
mean(data_non_imputed$T1)

# RDF T2
wilcox.test(data_non_imputed$T2[data_non_imputed$group==1],data_non_imputed$T2[data_non_imputed$group==0],
alternative = "two.sided", conf.int = TRUE, exact = F )
# means and standard deviations
```


Jerome R. Ravetz
SCIENTIFIC KNOWLEDGE AND ITS SOCIAL PROBLEMS

INTRODUCTION

THE activity of modern natural science has transformed our knowledge and control of the world about us; but in the process it has also transformed itself; and it has created problems which natural science alone cannot solve. Modern society depends increasingly on industrial production based on the application of scientific results; but the production of these results has itself become a large and expensive industry; and the problems of managing that industry, and of controlling the effects of its products, are urgent and difficult. All this has happened so quickly within the past generation, that the new situation, and its implications, are only imperfectly understood. It opens up new possibilities for science and for human life, but it also presents new problems and dangers. For science itself, the analogies between the industrial production of material goods and that of scientific results have their uses, and also their hazards. As a product of a socially organized activity, scientific knowledge is very different from soap; and those who plan for science will neglect that difference at their peril. Also, the understanding and control of the effects of our science-based technology present problems for which neither the academic science of the past, nor the industrialized science of the present, possesses techniques or attitudes appropriate to their solution. The illusion that there is a natural science standing pure and separate from all involvement with society is disappearing rapidly; but it tends to be replaced by the vulgar reduction of science to a branch of commercial or military industry. Unless science itself is to be debased and corrupted, and its results used in a headlong rush to social and ecological catastrophe, there must be a renewed understanding of the very special sort of work, so delicate and so powerful, of scientific inquiry.

If we are to achieve the benefits of industrialized science, and avert its dangers, then both the common sense understanding of science and the disciplined philosophy of science will need to be modified and enriched. As they exist now, both have come down from periods when the conditions of work in science, and the practical and ideological problems encountered by its proponents,

Must build replication into the practice of science.

Currently not valued by the profession.

Benjamini proposal:

- required to replicate something you rely on in papers & proposals
- pre-register the replication study
- recognition for authors of studies others deem worthy of replication

Frontispiece of Fisher's *The Design of Experiments*:

I AM very sorry, *Pyrophilus*, that to the many (elsewhere enumerated) difficulties which you may meet with, and must therefore surmount, in the serious and effectual prosecution of experimental philosophy I must add one discouragement more, which will perhaps as much surprise as dishearten you; and it is, that besides that you will find (as we elsewhere mention) many of the experiments published by authors, or related to you by the persons you converse with, false and unsuccessful (besides this, I say), you will meet with several observations and experiments which, though communicated for true by candid authors or undistrusted eye-witnesses, or perhaps recommended by your own experience may, upon further trial, disappoint your expectation, either not at all succeeding constantly or at least varying much from what you expected.

ROBERT BOYLE, 1673, *Concerning the Unsuccessfulness of Experiments.*

experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate the whole of the possible effects of chance coincidence, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 trials is decidedly "significant," in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

Science may be described as the art of systematic over-simplification—the art of discerning what we may with advantage omit. —Karl Popper

NAOMI A. P. STARK

No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says **Philip Stark**.

From time to time over the past few years, I've politely refused requests to referee an article on the grounds that it lacks enough information for me to check the work. This can be a hard thing to explain.

Our lack of a precise vocabulary — in particular the fact that we don't have a word for 'you didn't tell me what you did in sufficient detail for me to check it' — contributes to the crisis of scientific reproducibility. In computational science, 'reproducible' often means that enough information is provided to allow a dedicated reader to repeat the calculations in the paper for herself. In biomedical disciplines, 'reproducible' often means that a different lab, starting the experiment from scratch, would get roughly the same experimental result.

In 1992, philosopher Karl Popper wrote: "Science may be described as the art of systematic oversimplification — the art of discerning what we may with advantage omit." What may be omitted depends on the discipline. Results that generalize to all universes (or perhaps do not even require a universe) are part of mathematics. Results that generalize to our Universe belong to physics. Results that generalize to all life on Earth underpin molecular biology. Results that generalize to all mice are murine biology. And results that hold only for a particular mouse in a particular lab in a particular experiment are arguably not science.

Communicating a scientific result requires enumerating, recording and reporting those

SCIENCE SHOULD BE 'SHOW ME', NOT 'TRUST ME'.

or analysis is preproducible if it has been described in adequate detail for others to undertake it. Preproducibility is a prerequisite for reproducibility, and the idea makes sense across disciplines.

The distinction between a preproducible scientific report and current common practice is like the difference between a partial list of ingredients and a recipe. To bake a good loaf of bread, it isn't enough to know that it contains flour. It isn't even enough to know that it contains flour, water, salt and yeast. The brand of flour might be omitted from the recipe with advantage, as might the day of the week on which the loaf was baked. But the ratio of ingredients, the operations, their timing and the temperature of the oven cannot.

Given preproducibility — a 'scientific recipe' — we can attempt to make a similar loaf of scientific bread. If we follow the recipe but do not get the same result, either the result is sensitive to small details that cannot be controlled, the result is incorrect or the recipe was not precise enough (things were omitted to disadvantage).

Depending on the discipline, preproducibility might require information about materials (including organisms and their care), instruments and procedures; experimental design; raw data at the instrument level; algorithms used to process the raw data; computational tools used in analyses, including any parameter settings or ad hoc choices; code, processed data and software build environments; or analyses that were tried and abandoned.

Peer review is hamstrung by lack of pre-

An experiment or analysis is preproducible if it has been described in adequate detail for others to undertake it.

STEVEN SHAPIN & SIMON SCHAFFER

LEVIATHAN AND THE AIR-PUMP

HOBBES, BOYLE, AND THE EXPERIMENTAL LIFE

ological Essays of 1661 were written to another nephew, Richard Jones; the *History of Colours* of 1664 was originally written to an unspecified friend.[74] The purpose of this form of communication was explicitly to proselytize. The *New Experiments* was published so "that the person I addressed them to might, without mistake, and with as little trouble as possible, be able to repeat such unusual experiments. . . ."[75] The *History of Colours* was designed "not barely to relate [the experiments], but . . . to teach a young gentleman to make them."[76] Boyle wished to encourage young gentlemen to "addict" themselves to experimental pursuits and thereby to multiply both experimental philosophers and experimental facts.

In Boyle's view, replication was rarely accomplished. When he came to publish the *Continuation of New Experiments* more than eight years after the original air-pump trials, Boyle admitted that, despite his care in communicating details of the engine and his procedures, there had been few successful replications.[77] This situation had not materially changed by the mid-1670s. In the seven or eight years after the *Continuation*, Boyle said that he had heard "of very few experiments made, either in the engine

I used, or in any other made after the model thereof." Boyle now expressed despair that these experiments would ever be replicated. He said that he was now even more willing "to set down divers things with their minute circumstances" because "probably many of these experiments would be never either re-examined by others, or reiterated by myself." Anyone who set about trying to replicate such experiments, Boyle said, "will find it no easy task."[78]

Prolixity and Iconography

The third way by which witnesses could be multiplied is far more important than the performance of experiments before direct witnesses or the facilitating of their replication: it is what we shall call *virtual witnessing*. The technology of virtual witnessing involves the production in a *reader's* mind of such an image of an experimental scene as obviates the necessity for either direct witness or replication.[79] Through virtual witnessing the multiplication of witnesses could be, in principle, unlimited. It was therefore the most powerful technology for constituting

intellectual collective had mutually to assure themselves and others that belief in an empirical experience was warranted. Matters of fact were the outcome of the process of having an empirical experience, warranting it to oneself, and assuring others that grounds for their belief were adequate. In that process a multiplication of the witnessing experience was fundamental. An experience, even of a rigidly controlled experimental performance, that one man alone witnessed was not adequate to make a matter of fact. If that experience could be extended to many, and in principle to all men, then the result could be constituted as a matter of fact. In this way, the matter of fact is to be seen as both an epistemological and a social category. The foundational item of experimental knowledge, and of what counted as properly grounded knowledge generally, was an artifact of communication and whatever social forms were deemed necessary to sustain and enhance communication.

of trust and assurance that the things had been done and done in the way claimed.

The technology of virtual witnessing was not different in kind to that used to facilitate actual replication. One could deploy the same linguistic resources in order to encourage the physical replication of experiments or to trigger in the reader's mind a naturalistic image of the experimental scene. Of course, actual replication was to be preferred, for this eliminated reliance upon testimony altogether. Yet, because of natural and legitimate suspicion among those who were neither direct witnesses nor replicators, a greater degree of assurance was required to produce assent in virtual witnesses. Boyle's literary technology was crafted to secure this assent.