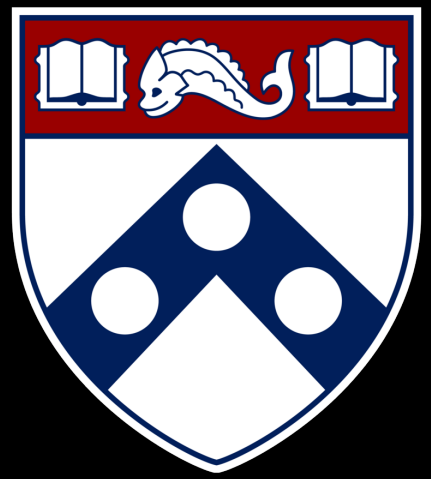
Empirical Security & Privacy, for Humans

UPenn CIS 7000-010

10/2/2025



Misuse of statistical analysis

Readings

Misuse, Misreporting, Misinterpretation of Statistical Methods in Usable Privacy and Security Papers

Jenny Tang
Carnegie Mellon University

Lujo Bauer
Carnegie Mellon University

Nicolas Christin
Carnegie Mellon University

Abstract

Null hypothesis significance testing (NHST) is commonly used in quantitative usable privacy and security studies. Many papers use results from statistical tests to assert whether effects or differences exist depending on the resulting *p*-value. We conduct a systematic review of papers published in 10 editions of the Symposium on Usable Privacy and Security over a span of 20 years to evaluate the field's use of NHST. We code statistical tests for potential statistical validity, reporting, or interpretation issues that may undermine assertions made in the 121 papers that use NHST. Most problematically, tests in 23% of papers inadequately account for non-independence between samples, leading to potentially invalid claims. 58% of papers lack information to verify whether an assertion is supported, such as imprecisely specifying the statistical test conducted. Many papers contain more minor statistical issues or report statistics in ways that deviate from best practice. We conclude with recommendations for statistical reporting and statistical thinking in the field.

1 Introduction

Statistical methods are often used in human-computer interaction research to support assertions about the presence (or absence) of an effect of scientific significance (e.g., some magnitude of difference) accompanied by a measure of statistic significance. Indeed, one of the most common refrains in statistical analysis is that a result is significant because

icance testing (NHST, also known as statistical significance testing)—that is, methods using *p*-values from inferential statistical tests as evidence to reject a null hypothesis—remains the dominant form of statistical analysis and evaluation [17]. However, simply dichotimizing results into "significant" and "non-significant" through their associated *p*-values without reporting other information is not in itself sufficient to convey the scientific importance of the claims, nor the richness and complexity of data collected from human subjects. This reliance on *p*-values to support assertions sometimes leads other information vital to understanding statistical and scientific significance to be omitted.

As a result, complete reliance on *p*-values is increasingly frowned upon, with some journals banning the reporting of *p*-values altogether [75, 81]. Most other current guidance is less drastic, and recommends using statistical hypothesis testing as a starting point and providing sufficient context (such as effect sizes, confidence intervals, and underlying data) to convey the scientific significance of the claims [2,13,49,59, 80,81]. We use this guidance to evaluate whether the scientific assertions made on the basis of NHST in usable privacy and security (UPS) are accompanied by sufficient reporting for readers to validate whether these assertions are supported by the information present in the paper. We focus on UPS as it is still a fairly young area, with evolving standards, features a considerable amount of quantitative research, and errors or misinterpretations can be detrimental to user safety in the digital world and beyond.

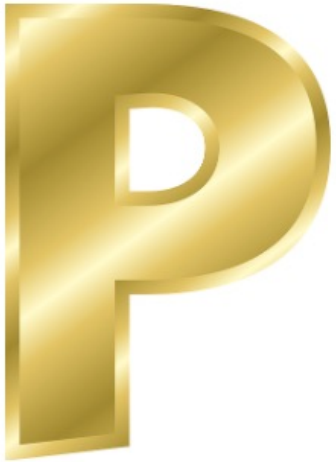
What is P Hacking: Methods & Best Practices

By Jim Frost — 2 Comments

P-Hacking Definition

P hacking is a set of statistical decisions and methodology choices during research that artificially produces statistically significant results. These decisions increase the probability of false positives —where the study indicates an effect exists when it actually does not. P-hacking is also known as data dredging, data fishing, and data snooping.

P hacking is the manipulation of data analysis until it produces statistically significant results, compromising the truthfulness of the findings. This problematic practice undermines the integrity of scientific research.



It occurs because high-impact journals strongly favor statistically significant results in today's scientific landscape. For researchers, publishing in these prestigious outlets is a career-boosting achievement. However, this prestige comes with pressure that can tempt researchers towards the perilous path of p-hacking.

P Hacking History

The term p-hacking was born during a crisis within the scientific community. Scientists were

Regression models: Charateristics

- Categorical vs. numeric variables
- Degrees of freedom
- Aikake Information Criterion (AIC): for judging relative quality of two models based on precision and parsimony
  - Prefers higher precision, and lower number of parameters

Regression models: Effect size

- Coefficient of determination (R^2), aka *effect size*: measures how well a model's predicted outcomes compare against real outcomes
  - Nagelkerke R^2 used for logistic regression
- Cohen's f^2 (*local effect size*): measures the impact of one parameter on the model effect size
- Cohen's d: quantifies the magnitude of a difference between two group means in terms of standard deviations

Statistical tests: Terms

- Alpha and p values of a statistical test, confidence intervals
- Statistical power, power analysis
- Type I vs. type II errors
- Paired sample test vs. a two-sample hypothesis tests, for comparing sampled measurements
  - t test, $\chi 2$ test, MW test, etc.
- One-sided vs. two-sided (“tailed”) null hypothesis test

P Hacking

- Jacks up the false positive rate (type I errors) by carrying out multiple tests on the same data
- How to do it
  - Test while you sample, stop when you get the result you want
  - Data tampering: Remove outliers
  - Change variables of results based on results
  - Too many hypothesis tests
  - Model selection based on p value
  - Cherry-picking the outcomes, failing to discuss nonsignificant results
- Bonferroni correction

Discussion points

- What did you think of the paper's methodology?
- Can you explain the results the paper found in terms of incentives? What are other explanations/causes?
- Despite exhibiting these issues, how might a paper nonetheless be providing value?
  - Is the value sufficient, most times?
- How might the review process change to help adjust these issues?
  - How to encourage correction even post publication?

How to do better?

Tea: A High-level Language and Runtime System for Automating Statistical Analysis

Eunice Jun[1], Maureen Daum[1], Jared Roesch[1], Sarah Chasins[2], Emery Berger[3,4], Rene Just[1], Katharina Reinecke[1]

[1]University of Washington, Seattle, WA {emjun, mdaum, jroesch, rjust, reinecke}@uw.edu

[2]University of California, Berkeley, CA schasins@cs.berkeley.edu

[3]University of Massachusetts Amherst, [4]Microsoft Research, Redmond, WA emery@cs.umass.edu

ABSTRACT

Though statistical analyses are centered on research questions and hypotheses, current statistical analysis tools are not. Users must first translate their hypotheses into specific statistical tests and then perform API calls with functions and parameters. To do so accurately requires that users have statistical expertise. To lower this barrier to valid, replicable statistical analysis, we introduce Tea, a high-level declarative language and runtime system. In Tea, users express their study design, any parametric assumptions, and their hypotheses. Tea compiles these high-level specifications into a constraint satisfaction problem that determines the set of valid statistical tests and then executes them to test the hypothesis. We evaluate Tea using a suite of statistical analyses drawn from popular tutorials. We show that Tea generally matches the choices of experts while automatically switching to non-parametric tests when parametric assumptions are not met. We simulate the effect of mistakes made by non-expert users and show that Tea automatically avoids both false negatives and false positives that could be produced by the application of incorrect statistical tests.

Author Keywords

statistical analysis; automated statistical analysis; declarative programming language; constraint-based system; data science; reproducibility; pre-registration

Since the development of modern statistical methods (e.g., Student's t-test, ANOVA, etc.), statisticians have acknowledged the difficulty of identifying which statistical tests people should use to answer their specific research questions. Almost a century later, choosing appropriate statistical tests for evaluating a hypothesis remains a challenge. As a consequence, errors in statistical analyses are common [26], especially given that data analysis has become a common task for people with little to no statistical expertise.

A wide variety of tools (such as SPSS [55], SAS [54], and JMP [52]), programming languages (e.g., R [53]), and libraries (including numpy [40], scipy [23], and statsmodels [45]), enable people to perform specific statistical tests, but they do not address the fundamental problem that users may not know which statistical test to perform and how to verify that specific assumptions about their data hold.

In fact, all of these tools place the burden of valid, replicable statistical analyses on the user and demand deep knowledge of statistics. Users not only have to identify their research questions, hypotheses, and domain assumptions, but also must select statistical tests for their hypotheses (e.g., Student's t-test or one-way ANOVA). For each statistical test, users must be aware of the statistical assumptions each test makes about the data (e.g., normality or equal variance between groups) and how to check for them, which requires additional statistical tests (e.g., Levene's test for equal variance), which themselves may demand further assumptions about the data. This cog-

Tisane: Authoring Statistical Models via Formal Reasoning from Conceptual and Data Relationships

Eunice Jun
emjun@cs.washington.edu
University of Washington
Seattle, Washington, USA

Audrey Seo
alseo@cs.washington.edu
University of Washington
Seattle, Washington, USA

Jeffrey Heer
jheer@cs.washington.edu
University of Washington
Seattle, Washington, USA

René Just
rjust@cs.washington.edu
University of Washington
Seattle, Washington, USA

ABSTRACT

Proper statistical modeling incorporates domain theory about how concepts relate and details of how data were measured. However, data analysts currently lack tool support for recording and reasoning about domain assumptions, data collection, and modeling choices in an integrated manner, leading to mistakes that can compromise scientific validity. For instance, generalized linear mixed-effects models (GLMMs) help answer complex research questions, but omitting random effects impairs the generalizability of results. To address this need, we present Tisane, a mixed-initiative system for authoring generalized linear models with and without mixed-effects. Tisane introduces a *study design specification language* for expressing and asking questions about relationships between variables. Tisane contributes an *interactive compilation* process that represents relationships in a graph, infers candidate statistical models, and asks follow-up questions to disambiguate user queries to construct a valid model. In case studies with three researchers, we find that Tisane helps them focus on their goals and assumptions while avoiding past mistakes.

KEYWORDS

statistical analysis; linear modeling; end-user programming; end-user elicitation; domain-specific language; transparent statistics; validity

ACM Reference Format:

Eunice Jun, Audrey Seo, Jeffrey Heer, and René Just. 2022. Tisane: Authoring Statistical Models via Formal Reasoning from Conceptual and Data Relationships. In *CHI Conference on Human Factors in Computing Systems (CHI*

1 INTRODUCTION

Statistical models play a critical role in how people evaluate data and make decisions. Policy makers rely on models to track disease, inform health recommendations, and allocate resources. Scientists use models to develop, evaluate, and compare theories. Journalists report on new findings in science, which individuals use to make decisions that impact their nutrition, finances, and other aspects of their lives. Faulty statistical models can lead to spurious estimations of disease spread, findings that do not generalize or reproduce, and a misinformed public. The challenge in developing accurate statistical models lies not in a lack of access to mathematical tools, of which there are many (e.g., R [63], Python [52], SPSS [58], and SAS [24]), but in accurately applying them in conjunction with domain theory, data collection, and statistical knowledge [26, 38].

There is a mismatch between the interfaces existing statistical tools provide and the needs of analysts, especially those who have domain knowledge but lack deep statistical expertise (e.g., many researchers). Current tools separate reasoning about domain theory, study design, and statistical models, but analysts need to reason about all three *together* in order to author accurate models [26]. For example, consider a researcher developing statistical models of hospital expenditure to inform public policy. They collect data about individual hospitals within counties. Based on their domain knowledge, they know that counties have different demographics and that hospitals in these counties have different funding sources (private vs. public), all of which influence hospital spending. To model county-level and hospital-level attributes, the researcher may author a generalized linear mixed-effects model (GLMM) that accounts for clustering within counties. But which variables should