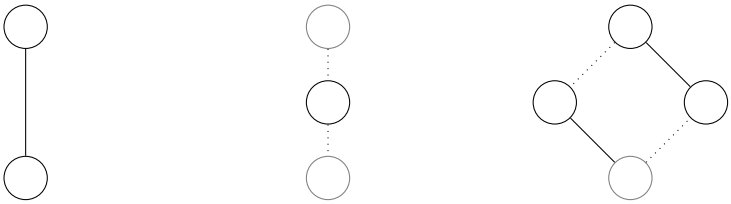
Towards a Measure of `code::proof`

A *toolchain walkthrough* for computationally developing a statistical estimator

Charles T. Gray



A thesis presented for the degree of
Doctor of Philosophy

Department of Mathematics and Statistics
School of Engineering and Mathematical Sciences
La Trobe University
Australia
November 2020

Towards a Measure of `code::proof`

A *toolchain walkthrough* for computationally developing a statistical estimator

Charles T. Gray

Abstract

Methods for generating and sharing data herald a technological revolution in scientific practice. The study of statistical methodology is increasingly steeped in simulation of random data with particular characteristics of interest to the scientist. Various broad statements about better data management and code curation emphasising transparency, accessibility, and extensibility are surfacing. However, *how* to implement these best practices are left to the researcher to determine for their discipline, and with their chosen scientific tools. This toolchain gap is the study of this dissertation, a contribution to the emerging literature addressing computational practice in the age of data. Taking a case study of developing a statistical estimator for meta-analysis of medians, this manuscript provides a toolchain walkthrough, from the computational structure, to simulation algorithms and data visualisation.

She sang, of course, "*M'ama!*" and not "he loves me," since an unalterable and unquestioned law of the musical world required that the German text of French operas sung by Swedish artists should be translated into Italian for the clearer understanding of English-speaking audiences. – *The Age of Innocence,* Edith Wharton, 1920.

Contents

List of Figures

List of Tables

Chapter 1

Preface

Towards a measure of `code::proof`

1.1 Structure of this manuscript

This doctoral project is not a conclusion, but an apprenticeship in computational science. This collection of five published or potential manuscripts demonstrate apprenticeship in the relatively new fields of research data engineering and interdisciplinary computational metascience. The first two chapters were published in the proceedings of the La Trobe Research School of Statistics and Data Science [38, 39] and are presented as published. The subsequent chapters are close to publication and sufficient to form the overarching story of this manuscript, a toolchain walkthrough for computationally developing a statistical estimator. The penultimate Chapter 6, is an example of metascientific questions that arise from doing interdisciplinary work. The final chapter, Foibles and Limitations, reflects on the strengths and weaknesses of this interdisciplinary undertaking.

Chapters 2 to 5 comprise the main content of the thesis, a collection of published or potentially publishable scientific essays. Chapters 2 and 3 explain why we should, and how to do so, adopt a computationally reproducible research workflow. The next two chapters provide an example research compendia of code, analysis, and mathematical statistics. Chapter 5 is thematically corollary, and extends into logic and philosophy of science to ask what other questions arise from interdisciplinary work such as this.

1.2 Nomenclature

As a dissertation on, amongst other things, research software engineering, there is much discussion of code and programming. Packages and the functions they contain are distinguished thus:

- `code`
- `package::`
- `::function`
- `package::function`

1.3 Assumed knowledge

This manuscript is accessible to any graduate-level student in data science. In the interests of brevity, and wishing to dive deep into discussion of reproducibility and simulations, there is much that is taken as given knowledge. There is an assumed understanding of the fundamental principles of mathematical statistics, such as found in [4]. Code examples begin from a working understanding of fundamental data science tools and `tidyverse::` syntax [40]. Finally, an understanding of meta-analysis [10] and `metafor::` [99].

1.4 Code suppression

In effort to stay true to the spirit of *toolchain walkthrough*, the theme of opinionated documentation of scientific workflow, central to this thesis, an effort has been made to present reproducible code.

However, in the interests of brevity, code is at times hidden from output, and only for the purposes of document formatting. For example, column header strings to be parsed by R into TeX are routinely omitted.

All code, including that which generates the formatting can be found on the associated GitHub repository[1].

[1] https://github.com/softloud/measureofcodeproof

1.5 Open source code

McElreath's `rethinking::`[2] package is not published on The Comprehensive R Archive Network (CRAN). His comments in *Statistical Rethinking*, serve just as well for the source code accompanying this dissertation.

> 'Note that **`rethinking::`** is not on the CRAN package archive, at least not yet. You'll always be able to perform a simple internet search and figure out the current installation instructions for the most recent version of the **`rethinking::`** package. If you encounter any bugs while using the package, you can check
>
> `github.com/rmcelreath/rethinking`
>
> to see if a solution is already posted. If not, you can leave a bug report and be notified when a solution becomes available. In addition, all of the source code for the package is found there, in case you aspire to do some tinkering of your own. Feel free to 'fork' the package and bend it to your will.'

Publishing on CRAN is a worthy goal, however, producing software is a different gambit to producing research software. All code accompanying this dissertation is provided in various open source repositories. Publication of software in and of itself, as distinct from the software created for analysis is not the focus of this manuscript. Rather, the focus here is to provide reproducible, accessible, and (aspirationally) interoperable code *for data analysis*.

To this latter aims, the R scripts supporting the work in this manuscript, are provided in packaged format (as described in Chapters 2 and 3), via online repositories:

- `simeta::` (https://github.com/softloud/simeta)
- `varameta::` (https://github.com/softloud/varameta)
- `parameterpal::`(https://softloud.github.io/parameterpal/)

This dissertation is positioned at the intersection of software engineering, computational research, and mathematical science. The code is offered in much the same spirit as McElreath's **`rethinking::`**. The software here does not necessarily meet all requirements for CRAN, demonstrative of a running theme of this dissertation is that computational research has different engineering requirements than commercial software. The purpose of packaging the code is not to create software, but to provide accessible, interoperable, and reproducible research compendia.

[2] https://github.com/rmcelreath/rethinking

Thus this software is provided open source, and as is, to be forked and explored for solving different problems by other scientists, as well as my future self.

1.6 Authorship

Required statement of authorship: *Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution. This work was supported by an Australian Government Research Training Program Scholarship.*

November 15, 2020.

This dissertation was written, researched, with analyses coded by myself, Charles T. Gray. Many chapters list co-authors, who have provided guidenace or insight. However, the writing and code was primarily, or entirely, produced by myself.

For example, Hannah Fraser was asked to read the Mathematistry chapter, that is, Chapter 6 and edit for framing of questionable research practices in ecology. However, she did not contribute to the mathematical arguments. Hien Nguyen and Dani Navarro provided sounding boards for the mathematical arguments, thus earning their co-authorships, however, all proofs and arguments were written by myself.

The most notable point of another's contribution to this dissertation is the concept of deriving the parameter for meta-analysis medians, presented in Chapter 4, provided by my advisor Luke Prendergast. He suggested exploring a estimator for the variance of the sample median, where quartiles are derived as described. Luke provided an example coverage probability simulation, from which the resulting analysis software was modelled.

Aside from this, co-authorship indicates minor editing and comments, not writing or programming. Thus, whilst others have contributed, I am first author on all manuscripts contained herein.

1.7 Acknowledgements

First, foremost, and above all, I thank my husband, Dr Alexander C. Gray, for his unwavering support in my career transition from piano teacher to whatever I am now.

I appreciate the patience and guidance of my advisors: Dr Hien Nguyen, Dr Hannah Fraser, Dr Luke Prendergast, and Dr Emily Kothe.

Particular thanks, too, to Dr Matthew Grainger, my academic compass, thank you for always being available to help me find north.

There have been many in the academic community that I have consulted in the making of this interdisciplinary dissertation; I appreciate the time taken by Kerrie Mengersen, Kate Smith-Miles, Mark Padgham, Gavin Stewart, W. Kyle Hamilton, Emily Riederer, and others in the open science community in consultation on particular sections in this manuscript.

Thanks, too, to the code reviewers I've consulted along the way: Adam Gruer, J. D. Long, James Goldie, and Heather Turner. It is the nature of open science that I may well have failed to acknowledge someone's contribution in code review, trust that this is inadvertent, a hazard of open science that we share, have conversations at unconferences, code together, and forget who suggested what where. Through the open science community, I've learned many skills, and believe that sharing research code as reproducible and extensible, is an important component of any research pipeline.

Finally, my `grrls`: Iris van Rooij, Susan D'Agostino, Laura Ación, Berna Devezer, and Danielle Navarro. No power in the 'verse can stop you [3], fierce ladies. Thank you for lending your fire when I needed it.

[3] In the space western *Firefly* (aired 2002-2003), River foreshadows her gifts by prodigiously gunning down her foes, remarking, 'No power in the 'verse can stop me' [29].

Chapter 2

Truth, Proof, and Reproducibility

There's no counter-attack for the codeless

Charles T. Gray and Ben Marwick

Abstract

Current concerns about reproducibility in many research communities can be traced back to a high value placed on empirical reproducibility of the physical details of scientific experiments and observations. For example, the detailed descriptions by 17th century scientist Robert Boyle of his vacuum pump experiments are often held to be the ideal of reproducibility as a cornerstone of scientific practice. Victoria Stodden has claimed that the computer is an analog for Boyle's pump – another kind of scientific instrument that needs detailed descriptions of how it generates results. In the place of Boyle's hand-written notes, we now expect code in open source programming languages to be available to enable others to reproduce and extend computational experiments. In this paper we show that there is another genealogy for reproducibility, starting at least from Euclid, in the production of proofs in mathematics. Proofs have a distinctive quality of being necessarily reproducible, and are the cornerstone of mathematical science. However, the task of the modern mathematical scientist has drifted from that of blackboard rhetorician, where the craft of proof reigned, to a scientific workflow that now more closely resembles that of an experimental scientist. So, what is proof in modern mathematics? And, if proof is unattainable in other fields, what is due scientific diligence in a computational experimental environment? How do we measure truth in the context of uncertainty? Adopting a manner of Lakatosian conversant conjecture between two mathematicians, we examine how proof informs our practice of computational statistical inquiry. We propose that a reorientation of mathematical science is necessary so that its reproducibility can be readily assessed.

Keywords: Meta-research · Reproducibility · Mathematics.

In David Auburn's Pulitzer prize-winning 2000 play *Proof*, a young mathematician, Catherine, struggles to prove to another mathematician, Hal, that her argument is not a reproduction

of the intellectual work of her deceased father, a professor [2]. Her handwriting similar to her father's, there is no way to discern her proof from his. But if Catherine were a computational scientist, we would have a very different story. We reimagine Hal challenging Catherine for different mathematical questions and the reproducibility of her solutions. We consider simple to complex mathematical questions that can be answered at the blackboard, and then consider the scenario where Catherine must use a combination of mathematical and computational tools to answer a question in mathematical science. Via these scenarios, we question to what extent proof methodology continues to inform our choices as mathematical scientists become as much research software engineers[1] as they are mathematicians.

Mathematical science is the compendium of research that binds the Catherine's methodology of work indistinguishably from her father's. However, in computational science, we not only do not have a common language in the traditional sense, with programming languages such as Python, R, and C++ performing overlapping tasks, but our research workflows comprise tools and platforms and operating systems, such as Linux or Windows, as well. Many inadvertent reasons conspire so that scientists are arriving at similar problems with different approaches to data management and version control. Code scripts, arguably the most immediately analogous to mathematical proof, are but one of the many components that make up the outputs of computational science.

If Catherine were a contemporary computational mathematician, she would not only struggle to reproduce another person's work, but she would likely struggle to reproduce her own. She may be overwhelmed by the diversity of research outputs [15], and find that she needs to rewrite her work to unpick what she did with specific computational functions under specific software package releases. The language of mathematical science has changed from something we write, to something we collect. In order to diligently answer scientific questions computationally, the mathematician must now consider her work within that of a research compendium. In this paper we ask: how can we extend the certainty afforded by a mathematical proof further down the research workflow into the 'mangle of practice' [81]? We show that communities of researchers in many scientific disciplines have converged on a toolkit that borrows heavily from software engineering to robustly provides many points to verify certainty, from transparency via version control, to stress testing of algorithms. We focus on unit testing as a strong measure

[1] We might argue here we employ the term *research software engineer* (RSE) as Katz and McHenry would define *Super RSE*s, developers who 'work with and support researchers, and also work in teams of RSEs who research and develop their own software, support it, grow it, sustain it, etc.' [53]. Or choose the more ambiguous Research Software Engineers Association definition of RSEs as people in academia who 'combine expertise in programming with an intricate understanding of research' [113].

of certainty.

2.1 The technological shift in mathematical inquiry

The task of a mathematical scientist in the pre-computer age was largely that of a blackboard rhetorician, where the craft of proof reigned. For a proof such as that featured in Auburn's play, the argument can often be included in the article, or as a supplementary file. This allows the reader to fully reproduce the author's reasoning, by tracing the flow of argument through the notation. As computers have become ubiquitous in research, mathematical scientists have seen their workflow shift to one that now more closely resembles that of a generic scientist, concerned with diligent analysis of observational and experimental data, mediated by computers [80]. But the answer to the question of what constitutes a diligent attempt to answer a scientific question examined in a computationally intensive analysis, is unclear, and remains defined by the era of the blackboard mathematician.

So, what is proof in mathematics, when experimental and computer-assisted methods are common? And, beyond mathematics, in fields where literal proofs are unattainable, what counts as an equivalent form of scientific certainty in a computational experimental environment? How do we measure truth in the context of uncertainty? Among the histories of science we can trace three efforts to tackle these questions. First is the empirical effort, most prominently represented by Robert Boyle (1627-1691), known for his vacuum pump experiments [88]. Boyle documented his experiments in such detail and to an extent that was uncommon at the time. He was motivated by a rejection of the secrecy common in science at his time, and by a belief in the importance of written communication of experimental expertise (as a supplement to direct witnessing of experimental procedures). Boyle's distinctive approach of extensive documentation is often cited by modern advocates of computational reproducibility [93]. Making computer code openly available to the research community is argued to be the modern equivalent of Boyle's exhaustive reporting of his equipment, materials, and procedures [59].

A second effort to firming up certainty in scientific work, concerned with statistical integrity, can be traced at least as far back as Charles Babbage (1791-1871), mathematician and inventor of some of the first mechanical computers. In his 1830 book 'Reflections on the Decline of Science in England, and on Some of Its Causes' he criticised some of his contemporaries, characterising them as 'trimmers' and 'cooks' [41]. Trimmers, he wrote, were guilty of smoothing of irregularities to make the data look extremely accurate and precise. Cooks retained only those results that fit their theory and discarded the rest [70]. These practices are now called

data-dredging, or p-hacking, where data are manipulated or removed from an analysis until a desirable effect or p-value is obtained [44].

A third effort follows the history of formal logic through to the time when an equivalence between philosophical logic and computation was noted. This observation is called the Curry-Howard isomorphism or the proofs-as-programs interpretation. First stated in 1959, this correspondence proposed that proofs in some areas of mathematics, such as type theory, are exactly programs from a particular programming language [92]. The bridging concepts come from intuitionistic logic and typed lambda calculi, which have lead to the design of computational formal proof management systems such as the Coq language. This language is designed to write mathematical definitions, execute algorithms and theorems, and check proofs [6]. This correspondence has not been extensively discussed in the context of reproducibility, but we believe it has relevance and is motivating beyond mathematics. Our view is that this logic-programming correspondence can be extended in a relaxed way beyond mathematics in proofs to scientific claims in general, such that computational languages can express those claims in ways that can establish a high degree of certainty.

Questions of confidence in scientific results are far from restricted to the domains of mathematics or computers; indeed, science is undergoing a broad reexamination under what is categorised as a crisis of inference [28]. How we reproduce scientific results is being examined across a range of disciplines [16, 101]. An early answer to some of these questions is that authors should make available the code that generated the results in their paper [27, 94]. These recommendations mark the emergence of a concern for computational reproducibility in mathematics. This paper extends this argument for computational reproducibility further into the workflow of modern statistical inquiry, expanding and drawing on solutions proposed by methods that privilege computational reproducibility.

Systemic problems are now being recognised in the practice of conventional applied statistics, with a tendency towards *dichotomania* [1] that reduces complex and nuanced questions to Boolean statements of `TRUE` or `FALSE`. This has diluted the trust the can be placed in scientific results, and led to a crisis of replication, where results can not easily be reproduced [28] and questionable research practices [31] proliferate.

As the conventions of statistics are called into question, it stands to reason that the research practices of the discipline of statistics itself require examination. For those practicing statistical computing, a conversation is emerging about what constitutes best practice [110]. But best practice may be unrealistic, especially for those applying statistics from fields where their background has afforded limited computational training. And thus the question is becoming

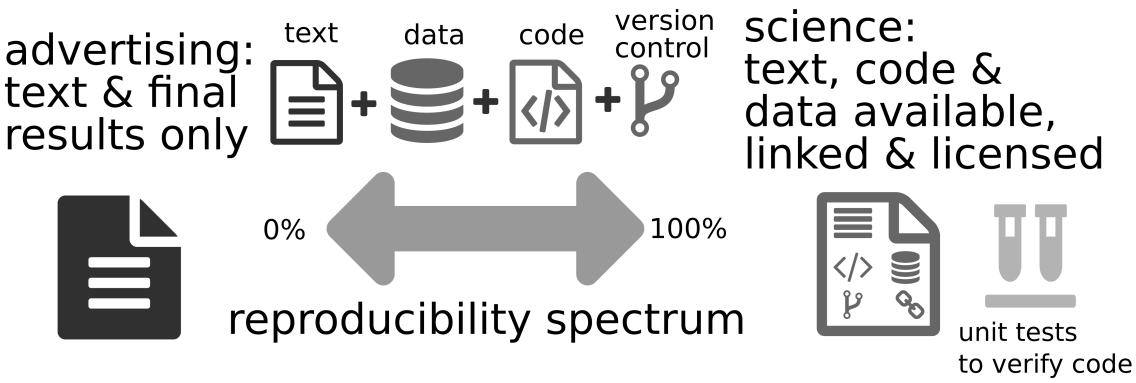


Figure 2.1: We propose updating this spectrum of reproducibility [66] with unit tests for data analysis. In addition to the advertising, the **formal** scientific argument put forward, many *informal* and traditionally hidden scientific outputs comprise the compendium of research that produces the results. Given the underutilised nature of unit tests, we suggest there is further work to be done to facilitate the adoption of *good enough* [111] research software engineering practices for answering mathematical questions computationally. The informal components of mathematical research compendium are shaded grey. This figure has been adapted with permission [86] and is licensed under CC-BY 2.0.

reframed in terms of *good enough* standards [111] we can reasonably request of statistical practitioners. By extension, we must reconsider how we prepare students in data-analytic degree programs.

Proofs, derivations, verification, all form the work of mathematics. How do we make mathematical arguments in a computational[2] environment? In constructing mathematical arguments, we posit that we require an additional core element: unit testing for data analysis. We propose an expansion of the spectrum of reproducibility, Figure 2.1, to include unit testing for data analytic algorithms facilitated by a tool such as `testthat::` [106], for answering mathematical research questions computationally. In order to motivate this practice, we turn to the purest of sciences, mathematical proof.

[2]We focus in this manuscript on R packages, but the reader is invited to consider these as examples rather than definitive guidance. The same arguments hold for other languages, such as Python, and associated tools.

2.2 Truth in mathematics

The titular proof [2] of Auburn's play is a mathematical argument, a formalised essay in mathematical science. The creator of the proof, Catherine, is questioned by Hal, who is capable of following the argument; that is, Hal can *replicate* an approximation of the type of thought process that leads to a *reproduction* of the argument presented in the proof.

In Figure 2.1, we have coloured the components, black **formal** argument, and grey *informal* work, of mathematics Hal would need to reproduce the proof. In order to verify the results, Hal would need to follow the formal argument, to understand what was written in the proof, but also need to do informal work, to understand the links between concepts for verification.

Hal would come to the problem with a different background and education to Catherine. Although work is necessary for the verification of the results, the reproduction of the reasoning, the work required would be different for Hal and Catherine, based on their respective relevant preparation. However, the language of mathematics carries enough uniformity that Hal can fill in the work he requires to understand the result, from reasoning and mathematical texts. If Catherine were asking a mathematical question computationally, the presentation of the results carries not millennia of development of methodology, as does the noble craft of mathematics, but less than a century of frequently disconnected developments separated by disparate disciplines.

We begin with traditional mathematics and end with answering questions in computational mathematics. To this aim, we adopt, in the manner of Lakatos' *Proofs and Refutations*' conversant conjecture, scenarios between Hal and Catherine, where Hal challenges Catherine over her authorship of the proof. In each scenario, we imagine the challenge would play out for different ways of answering mathematical questions. We argue the thinking work of mathematical science is not as immediately inferable in a computational experimental environment, and that the roots of mathematical science in proof lead to an overconfidence that science is as readily reproducible as a proof.

2.2.1 Prove it!

Let us suppose Catherine claimed she could demonstrate a property about the order[3] on natural numbers, $\mathbb{N} = \{1, 2, 3, \dots\}$, the counting numbers.

The order on a set of numbers is **dense** if, for any two numbers we can find a number in

[3]Let P be a set. An *order* on P is a binary relation $\leqslant$ on P such that, for all $x, y, z \in P$: we have $x \leqslant x$; with $x \leqslant y$ and $y \leqslant x$ imply $x = y$; and, finally, $x \leqslant y$ and $y \leqslant z$ imply $x \leqslant z$. We then say $\leqslant$ is reflexive, antisymmetric, and transitive, for each of these properties, respectively [22].

between. More formally, we say an ordered set P is dense if, for all $x < y$ in P, there exists z in P such that $x < z < y$.

Catherine presents the following argument that the order on $\mathbb{N}$ is not dense. In this case she chooses a type of *indirect* proof, an *existence* proof [14], where she presents a counterexample demonstrating that the density property is not true for all cases for $\mathbb{N}$.

Proof. **The order on $\mathbb{N}$ is not dense.** Let us, in the spirit of Lewis Carroll[4], be contrary and suppose, by way of contradiction, that the order on $\mathbb{N}$, *is* dense. Then for any two numbers, x and y, in $\mathbb{N}$ such that $x < y$, I should be able to find a distinct number z in $\mathbb{N}$ between them, that is $x < z < y$. But, consider the numbers 3 and 4. Let $x = 3$ and $y = 4$, then $x < y$. There is no distinct number, z, that exists between x and y. Since this rule must be true for any two numbers $x < y$ in the order to be dense, we have shown the order on natural numbers $\mathbb{N}$ is not dense. □

A standard way to prove something is *not* true, is to assume it *is* true, and derive a contradiction [23]. Arguably, this reasoning goes to the heart of the problem of *dichotomania* lamented by 800 scientists in a recent protest paper about the misinterpretation of statistics in *Nature* [1]. A null hypothesis test of a difference between two groups will assume the opposite of what we suspect is true; we believe there to be a difference between two groups and take a sample from each of the groups and perform a test. This test assumes there is no difference, null, between the two groups and that any observed differences in sampling are due to random chance. The calculation returned, the p-value, is the likelihood we would observe the difference under those null assumptions. Crucially, the calculation returned is probabilistic, a number between 0 and 1, not a `TRUE` or `FALSE`, the logic of a proof by contradiction. The logic does not apply to a situation where, within a single group of people, some people might be resistant to treatment, and some might not be, say, and we have estimated a likelihood of the efficacy of the treatment. Dichotomania is the common misinterpretation of a probabilistic response in a dichotomous framework; scientists are unwittingly framing null hypothesis significance testing in terms of a proof by contradiction.

In order to illustrate our central point, we now turn to a direct argument, rather than the indirect approach of contradiction, in order to examine the process of the making of a proof. In both the case of the direct, and indirect proofs, however, Hal could challenge Catherine, as he did in the play.

[4]Lewis Carroll, author of *Alice in Wonderland*, is a writing pseudonym used by Charles Lutwidge Dogson, born in 1832, who taught mathematics at Christ Church, Oxford [17].

> "Your dad might have written it and explained it to you later. I'm not saying he did, I'm just saying there's no proof that you wrote this" [2].

2.2.2 The steps in the making of a proof

Let us now suppose Catherine's proof instead demonstrated a density property on the order on the real numbers,

$$\mathbb{R} = \{\ldots, -3, \ldots, -3.3, \ldots, 0, \ldots, 1, \ldots, 100.23, \ldots\},$$

i.e., the whole numbers, and the decimals between them. Catherine claims the order on $\mathbb{R}$ is dense, which is to say, if we choose any two distinct numbers in the real numbers, we can find a distinct number between them.

Catherine would construct her proof in the manner laid out in the introductory monograph *When is a Proof?* [23], in Table 2.1, provided to undergraduate mathematics majors at La Trobe University. These steps comprise **formal** and *informal* mathematical work, showing that mathematical *work* comprises more than the *advertising*, as it is labelled in the reproducibility spectrum presented in Figure 2.1. In the case of pure mathematics, the advertising would be the paper that outlines the proof, the formal mathematical argument, but the informal work is left out.

Catherine presents the following proof to Hal to show the order on real numbers, $\mathbb{R}$, is dense.

Proof. The order on $\mathbb{R}$ is dense. Let $x < y$ in $\mathbb{R}$. Let[5] $z := (x + y)/2$. To see that $x < z < y$, we begin with $x < y$, so, $x + x < x + y$ and $x + y < y + y$, which gives,

[5] In mathematics, we read := as 'be defined as', $\implies$ as 'implies', and $<$ as 'less than but not equal to'.

Table 2.1: The steps in the making of a proof from Brian A. Davey's primer, *When is a Proof?* [23]. The formal steps that contribute to the final proof are in **bold**, the hidden informal work, in *italics*. These steps are summarised in terms of $p \implies q$ in the final column of the table.

Step -1	**Translate the statement to be proved into ordinary English and look up appropriate definitions.**	
Step 0	**Write down what you are asked to prove. Where appropriate, isolate the assumptions, p, and the conclusion, q.**	$p \implies q$
Step 1	**Write down the assumptions, p: "Let ... "**	**Assume** p**.**
Step 2	**Expand Step 1 by writing out definitions: "i.e., ... "**	**Define** p**.**
Step 3	*Write down the conclusion, q, which is to be proved: "To prove: ... "*	*State* q*.*
Step 4	*Expand Step 3 by writing out definitions: "i.e., ... "*	*Define* q*.*
Step 5	*Use your head: do some algebraic manipulations, draw a diagram, try to find the relationship between the assumptions and the conclusion.*	*Work.*
Step 6	**Rewrite your exploration from Steps 3, 4 and 5 into a proof. Justify each statement in your proof.**	**Formalise work**
Step 7	**The last line of the proof.**	**"Hence q."**

$$
\begin{aligned}
& x+x < x+y < y+y \\
\implies & \frac{x+x}{2} < \frac{x+y}{2} < \frac{y+y}{2} \\
\implies & \frac{2x}{2} < \frac{x+y}{2} < \frac{2y}{2} \\
\implies & x < \frac{x+y}{2} < y \\
\implies & x < z < y,
\end{aligned}
$$

since $z = (x+y)/2$, as required. □

Catherine presents the formal proof, the science that in Figure 2.1 is described as the advertising, a subcomponent, of the compendium of research she created in order to arrive at this argument. Hal wishes to verify the results and investigate whether Catherine merely reproduced her father's reasoning. In the case of proof, what is published is the formal argument, but as the steps in Table 2.1, this is not all of what makes a proof. We could think of the steps

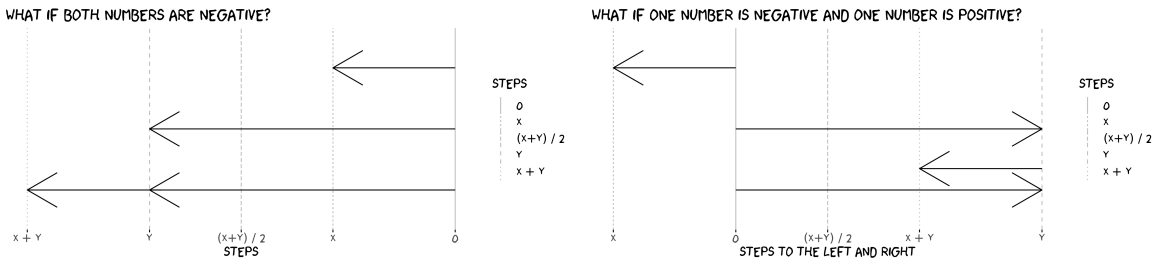


Figure 2.2: On the left, Hal might begin to verify his understanding of + by first considering the case where both numbers are negative, $x, y < 0$. In this case, we might think of + as combining x steps to the left with y steps to the left. The halfway point $(x + y)/2$, falls in the middle of the two arrows laid side by side, which also falls between where the two ends of the arrows fall. On the right, Hal considers the case where $x < 0$, $y > 0$ and $|x| < |y|$. Here $x + y$ can be thought of as y steps to the right and then x steps to the left. Again, the halfway point $(x + y)/2$ falls halfway between the tips of the two arrows above.

presented in Table 2.1 in terms of a mathematical statement $p \implies q$, which we read as p *implies* q, as given in the final column of the table. We now revisit the proof Catherine offered in terms of these steps.

We begin, step 0; we state what we wish to prove, $p \implies q$, in plain English. We wish to show the real numbers, $\mathbb{R}$, are dense; i.e., for all $x < y$ in $\mathbb{R}$, there exists z such that $x < z < y$.

Step 1, we **assume** p is true. We assume we have two distinct numbers x and y in $\mathbb{R}$ with $x < y$; i.e., x is less than y, and x is not equal to y. Step 2, nothing to define as we are familiar with $<$ and $\mathbb{R}$.

Step 3, we *state* what we wish to prove, q; the order on $\mathbb{R}$ is dense. Step 4, i.e., we need to show there exists z in $\mathbb{R}$ such that $x < z < y$. Now, Catherine has offered a solution $z := (x + y)/2$ that Hal wishes to verify.

Step 5, Suppose Hal asks, what if both x and y are negative numbers? Is it still true that $x < z < y$? Hal might verify his understanding of + by thinking about positive and negative numbers as steps taken to the left or the right. In Figure 2.2, Hal considers the case where both numbers are negative, $x, y < 0$. In this case, we have x steps to left, and y steps to the left, which we imagine as arrows of appropriate length. If we lay both arrows end to end, we see the number of combined steps to the left. If we consider the half-way point of x and y laid beside each other, $(x + y)/2$, we see this falls between where the arrow heads of x and y fall.

Now Hal can flip the arrows in the opposite directions to construct an argument for if both numbers were positive, $x, y > 0$.

But then Hal asks in Figure 2.2, what if one number were positive and one number were negative? Is $(x + y)/2$ still halfway between? Let us assume, as mathematicians say, without loss of generality that the magnitude of x is strictly less than y, that is $|x| < |y|$, the number of steps in x is less than the number of steps of y. Hal now considered where one would end up if one took y steps to the right and then x steps to left. He checks that he does not need to consider two cases, as he would end up in the same place if he took x steps to the left and then y steps to the right. Again, $(x + y)/2$ falls between where he would start and where he would end.

Now Hal has verified his understanding of $+$, which may or may not be the way that Catherine arrived at her result, but after this work he is capable of fully reproducing the mathematical result presented. He reads the proof Catherine has provided, and verifies Steps 6, and Step 7. Catherine has proved that the order on $\mathbb{R}$ is dense. With this proof, as with the proof presented in Section 2.2.1, Hal cannot disqualify the possibility that Catherine merely reproduced her father's work.

Even in these relatively simple proofs, Step 5, the informal work of verification and understanding vastly outweighs what goes into the formal proof. But these toy examples belie a process of redefinition and re-examination, as illustrated in the discussion within a hypothetical mathematics classroom that forms the narrative of Lakatos' *Proofs and Refutations* [58]. We now move to a recently published proof to illustrate this process of redefinition.

In the combat conditions of new mathematics

Suppose, now, that Catherine's proof were for the theorem pertaining to quasi-primal algebras, presented in the recent publication 'The homomorphism lattice induced by a finite algebra' [24] in *Order*, a mathematics journal devoted to 'original research on the theory and application of ordered sets'. In addition to the informal work demonstrated by the proof that the order on $\mathbb{R}$ is dense, the making of this proof involved a redefinition of the result proved, through a process writing several proofs. In terms of Table 2.1, initially a result was considered, $p \implies q$. A proof was written for this result. At this point the mathematicians realised, however, that the converse could be shown, that is, $q \implies p$. And so, a proof was generated for a new result, $p \iff q$. In the case of this proof, the act of writing the proof itself redefined the result in question. In the combat conditions of new mathematics, the process of writing a proof is doing mathematical science, and involves a great deal more work than is presented in the advertising of the science.

Hal may require graduate-level knowledge of abstract algebra to reproduce this proof, but as a professional mathematician, this is not a great leap. More challenging the proof may be, but the process of reproduction would be similar. Even if this were the proof, Hal would not know if Catherine merely reproduced, as he did, her father's proof.

But what if Catherine were posing her mathematical question computationally? Would Hal be able to reproduce her results?

2.2.3 Is computational mathematics mired in proof methodology?

When we are exploring and answering mathematical questions in a computational environment, we consider some aspects of our work to be **formal** and some *informal*. But in omitting the greyed *informal* work in Figure 2.1, are we still approaching compendia of research from the perspective of a blackboard mathematician?

Given we use statistics in most science, arguably most scientific questions are posed, to some extent, mathematically. The output format, a published paper, remains similar to mathematics of the pre-computer age. But the informal work of answering mathematical questions has changed significantly. Now that much work is done computationally, there are multiple research outputs that comprise the compendium of science that produces the published paper.

Let us now suppose that Catherine had a statistical estimator for a population parameter of interest. That is, Catherine has an equation that, given some data, she can approximate some value about the population, such as an overall average. Let us further suppose, as is increasingly common, that she does not have a closed-form solution, meaning she cannot write out a mathematical argument in the traditional sense. Instead, she demonstrates the estimator's performance through simulation studies.

Now suppose Hal challenges Catherine to prove that she created the science that produced the paper. Given what is on the piece of paper, how can Hal know that Catherine's code does what she said it does? It is unclear what assumptions were made, about, say, sample size and distribution. How can Hal verify her results? Through adopting research software engineering principles, Catherine can facilitate a process akin to proofs and refutations, the redefinition described in the Section 2.2.2, The combat conditions of new mathematics. The process of redefinition is transcribed by version control, but further to this, the software itself provides a modular framework, such as a theorem in mathematics, for future work to scaffold and extend. New software can be developed that either extends, or redefines the existing software. One analogous way this is occurring is in the rise of metapackages, such as `tidyverse::` [107] and

`metaverse::` [103], that collect software to solve particular problems in an opinionated [79] manner, that guide the end-user to what the creators consider to be good enough practice. This is analogous to classes of mathematics, such as group theory or analysis, that collect results, theorems, that rely upon each other, and where certain underlying assumptions, such as the *Axiom of Choice*[6], are made. Indeed, as Martin-Löf proposed a shift in terminology from computer science to computing science, they make the following remark.

> It has made programming an activity akin in rigour and beauty to that of proving mathematical theorems [64].

How are contemporary researchers answering mathematical questions? Alex Hayes, current maintainer and one of the many authors of `broom::` [85], an open source R package that amalgamates hundreds of contributions towards providing a suite of tools that *tidily*[7] [42] extract statistical model information from R algorithms, recently noted the underdeveloped nature of the implementation of statistical algorithms [43]:

> In practice, most people end up writing a reference implementation and checking that the reference implementation closely matches the pseudocode of their algorithm. Then they declare this implementation correct. How trustworthy this approach is depends on the clarity of the connection between the algorithm pseudocode and the reference implementation.

This is not to carp upon diligent scientists; we need to do far more to support the software engineering principles we expect from those who answer mathematical questions computationally [76]. Mathematicians are trained to provide enough work such that the hidden steps illustrated in italics in Table 2.1 can be reproduced by their target audience. The detail of mathematical work shown is tempered for level of the audience, but the same process described in bold in Table 2.1 is the same. But, does the workflow Alex describes above equip the target audience with enough information such that they can understand all the details of the entire argument put forward?

[6] Turning to the bible of algebra, *Lattices and Order* [22], we learn the *Axiom of Choice* 'asserts that it is possible to find a map which picks one element from each member of a family of non-empty sets'.

[7] From Wickham's *Tidy data* [42], we describe data as *tidy* if

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Table 2.2: Percentage of R packages in repositories that have unit tests included. These results are from Jim Hester's presentation on covr:: in September 2016 [46].

Repository	Tests	Total	
CRAN	2091	9772	21%
Bioconductor	449	1258	36%
rOpenSci	84	146	58%
	2624	11,176	24%

Code has the appearance of being highly logical, it's easy to assume it's infallible; and whilst the logic of the code is robust, the pipeline that carries the algorithm to implementation may be susceptible to compromising factors, with typos being just one example of inadvertent error.

Because code appears so logical, we assume it is analogous to proof for our intended audience to follow. But we were trained to leave out the informal messy thinking work associated with mathematics; trusting the formal argument provides enough information to verify and reproduce the mathematics. Does our code do what we think it does? In addition to providing the research outputs in the spectrum of reproducibility, Figure 2.1, we posit mathematical science should adopt the software development practice of unit testing, to ensure the mathematical results can be verified and reproduced.

2.3 Testing

Testing is the software engineering tool that is provides a key piece of the correspondence between scientific claim and programming. Just as the Curry-Howard isomorphism expresses proofs-as-programs to link mathematics and programming, we argue that tests are are the link between scientific claims more generally and programming. In a test the researcher isolates a scientifically meaningful part of their code, and creates a witness so that others can easily see that the code does what the researcher intends it to do. In this section we consider a 'vital' [104] research output, testing, that it is unlikely the mathematical scientist has been trained in. There are many such under-formalised skills represented in Figure 2.1[8]. In 2016, a quarter of packages on R package archives CRAN, Bioconducter, and rOpenSci, included tests, a repository by repository breakdown of this is shown in Table 2.2.

Now, Hayes advises people against using untested software [43]. It is alarming that, by this

[8]Indeed, the natural consequence of questioning how we practice mathematical science is how we train the next generation of practitioners. Important, however this may be, this is beyond the scope of this manuscript.

logic, we would be **insane** to use *three quarters* of packages available. But Hayes continues, 'You have two jobs. The first job is to write correct code. The second job is to convince users that you have written correct code' [43]. The disconnect here suggests a failure to communicate broadly the importance of testing of algorithms in the dissemination of research. As researchers, we believe our science is as reproducible as a traditional mathematical proof; however, the growing literature of the replication crisis demonstrates we have not succeeded in rendering our science reproducible.

rOpenSci's review system recommends using the `covr::` [47] package to measure how the code behaves with different expected outputs. From the creator of `covr::`, we obtain the following definition of test coverage.

> Test coverage is the proportion of the source code that is executed when running these tests [47].

2.3.1 What is a test?

Tests demonstrations that a given input produces an expected output. They are grouped contextually in a file; the context being a certain aspect of the algorithm that should be tested [104]. An example of a context for a test is the question, does a given function return the expected result for different inputs? Each test comprises a collection of expectations. Each expectation runs a function or functions from the package, and checks the returned output is as expected. In this case, we have a test for the `expect_equal` function: one expectation checks the function successfully runs when given equal inputs, and another expectation checks that the function fails when passed two non-equal inputs.

An example test from the `testthat::` [106] contains two expectations.

```
  test_that("basically principles of equality hold", {
  expect_success(expect_equal(1, 1))
  expect_failure(expect_equal(1, 2))
})
```

2.3.2 How good are we at *good enough* testing?

A response to the replication crisis has been to examine *questionable research practices* [31], frequently borne of tradition and convention within different disciplines, deviate from evidence-

based best-practice research methodology. We suggest it is a questionable research practice to draw conclusions about the efficacy of statistical estimators from untested code.

Given only a quarter of R packages have unit tests associated with them, we are falling short of best practice in scientific computing [110]. In a recent assessment of what constitutes *good enough* practice in scientific computing [111], unit testing was not included. However, for mathematical science, where the algorithms implemented and the code written is often complex, we suggest that unit testing should be considered good enough practice, in spite of the additional learning curve. With the backdrop of the replication crisis, it is crucial we have confidence in the algorithms we implement.

2.3.3 Analysis of testing code in R packages

So, what packages have tests? We provide a preliminary analysis of tests in CRAN packages in Figure 2.3. The code and data used to generate the results presented here are openly available at https://github.com/softloud/proof.

We provide analysis for packages associated with CRAN task view [114], opinionated [79] collections of R packages that are relevant to a particular type of statistical analysis, maintained voluntarily by experts in their respective fields [114]. CRAN task views provide a convenient taxonomy of R packages for a preliminary exploratory analysis of patterns of test use among R package authors.

Packages listed in a task view are may be interpreted by users as more stable and trustworthy than other packages, because they have passed some kind of inspection by maintainer of the task view who listed the package (however the review and curation process is not open or documented). And yet, even amongst the 4105 packages associated with task views, 1524 packages were without tests; 37 per cent of packages associated with CRAN task view were without tests.

The proportion of task view packages with tests has fallen over the last decade. This does not seem surprising given the uptake of R amongst communities of researchers in applied sciences with little formal programming and computer science training, such as psychology and ecology.

Figure 2.4 shows that there is wide variation in test coverage. Even the largest and fastest growing CRAN task views have very different proportions of packages with tests (Survival, about 0.23, compared to Web Technologies about 0.66). We find few clear patterns in the presence of tests over time, between different CRAN task views, and with metadata such as the number of authors, the size of the package and the centrality of the package (as measured by

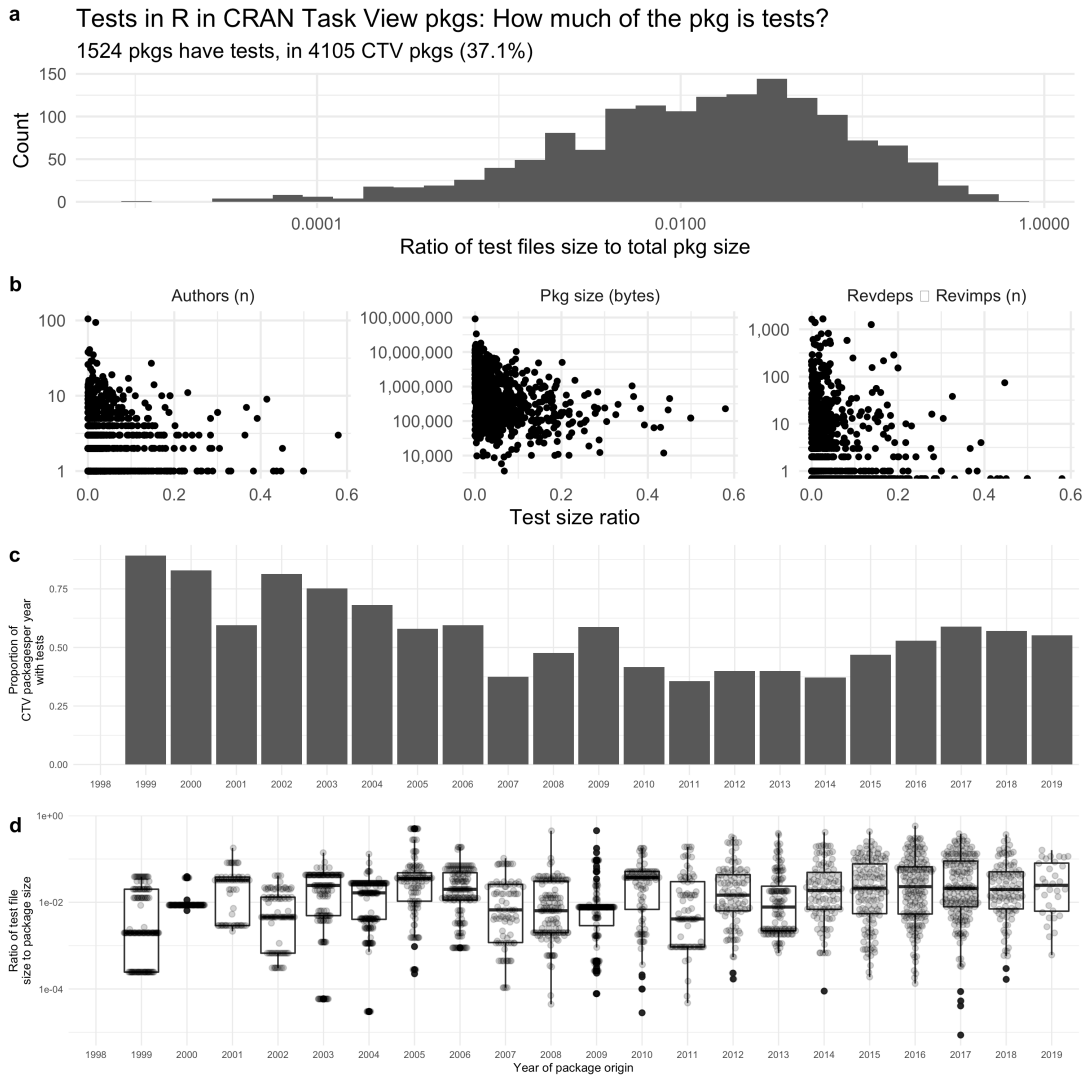


Figure 2.3: This panel shows some basic details of tests in R packages listed in CRAN task views [114]. The measure of interest, *test size ratio*, was calculated by dividing the test file size with the overall package source file size from the unofficial CRAN mirror on GitHub. This is a rough indicator of test coverage, future work should consider more precise metrics such as those produced by the covr:: package. a) the distribution of the ratio of test file size to total package size, test size ratio. b) scatter plots demonstrate the relationship between test size ratio and number of authors, overall package size, and number of packages imported and calling the package, respectively. c) the proportion of all task view packages that contain tests over time. d) boxplot detailing the distribution of file size ratio over time.

the union of the number of reverse dependencies and reverse imports). Based on these data, we suggest there is much work to be done in developing methods and opinionated tools that guide users towards good enough practices.

2.4 Tempered uncertainty and computational proof

It's easy to lie with statistics, but it's even easier without them [71]. In a computational experimental setting, we often cannot achieve the satisfying precision offered by a proof. We can, however, adopt good enough practices in sharing and testing code to increase confidence in our scientific conclusions. Given the prevalence of generalised linear models, we can think of the practice of much science as the interpretation of

$$y \approx bx,$$

where: x represents what we know about the data; y, the observed response of interest that we wish to investigate how it responds to x; and b, the *how* it responds, approximated unknown. It may not be possible to provide the rigour of a closed-form mathematical solution, but we can aim to temper the uncertainty, and bolster confidence, in computational arguments via automated testing, version control, and other computational outputs.

We suggest there is much work to be done in developing good enough practices [111] we can ask mathematical scientists to adopt. For example, we do not have a chance to discuss in this manuscript the role of markdown and html reporting in reproducible science. Indeed, the question of good enough practice can be posed for each research output. Less than offering answers, this manuscript seeks more to suggest there is a rich line of inquiry [76] in the relationship between scientific truth, mathematical proof, and computational reproducibility and rigour.

2.4.1 Coda

Returning to Catherine and Hal from Auburn's *Proof* [2], we can now imagine her as computational mathematician who provides a compendium of reproducible research. To demonstrate the rigour of her computational work, she would provide unit tests for the algorithms she had implemented. Catherine would share her work openly via her GitHub or similar repository, where the development of her ideas would be timestamped and recorded. The structure of her research compendium of would be automatically standardised via a tool such as `rrtools::` [65].

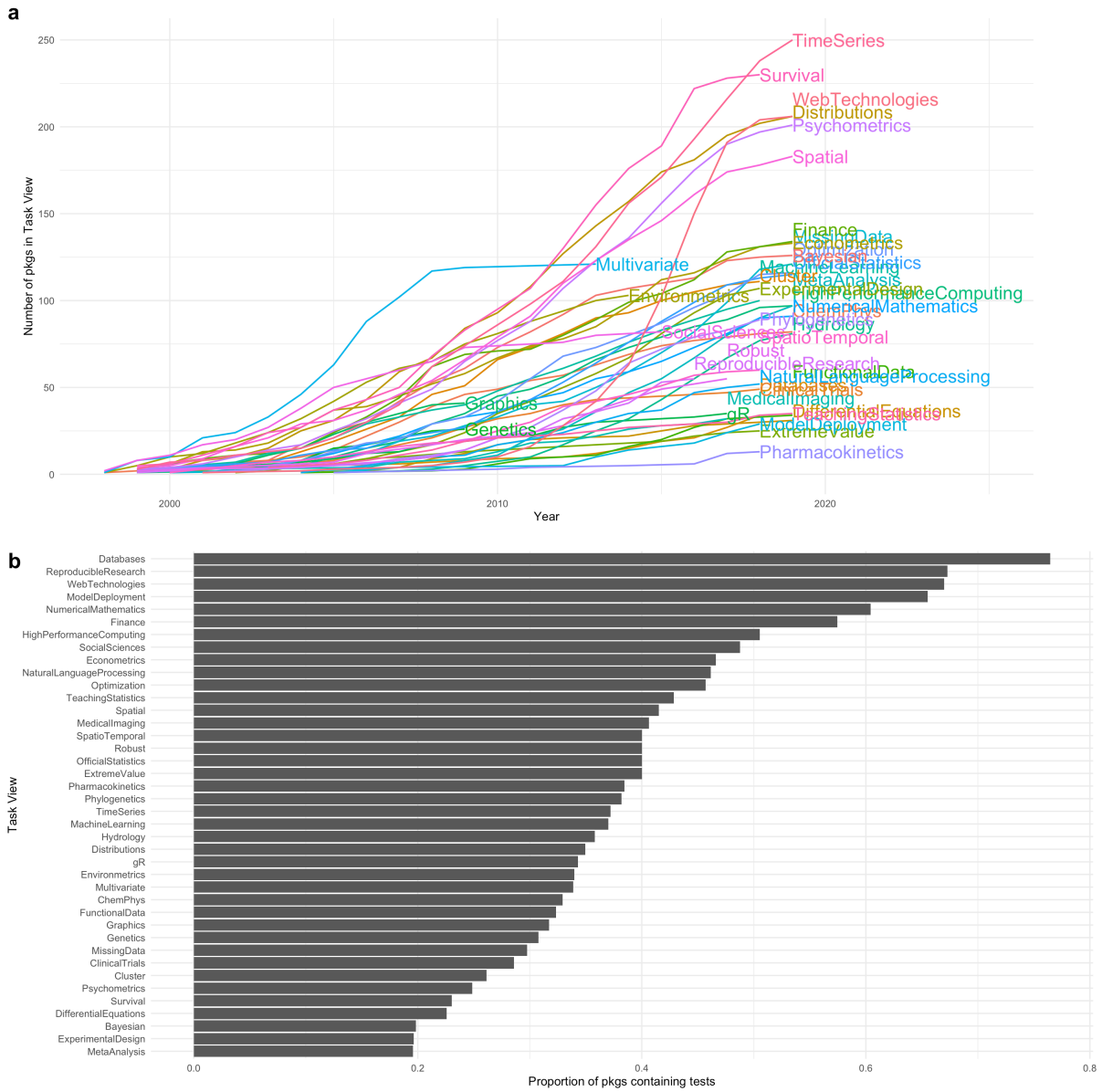


Figure 2.4: a) shows the change in the number of packages in each CRAN task view over time. b) shows the proportion of packages in each CRAN task view that have tests.

At publication, her compendium would be deposited on a trustworthy, DOI-issuing repository for others to link to and cite.

And she would feel safe asking questions about good enough practice [111], and how to avoid questionable research practices [31], because there is an understanding in the community that no one is trained in all these things, so we are all always learning.

There would be no struggle, as there was in Auburn's play, to show that the mathematician who created these research outputs was Catherine. But that wouldn't matter - she and Hal would be having far too much fun collaborating on the next question.

Chapter 3

`code::proof`

Prepare for most *weather conditions*

Charles T. Gray

Abstract

Computational tools for data analysis are being released daily on repositories such as the Comprehensive R Archive Network. How we integrate these tools to solve a problem in research is increasingly complex and requiring frequent updates. To mitigate these *Kafkaesque* computational challenges in research, this manuscript proposes *toolchain walkthrough*, an opinionated documentation of a scientific workflow. As a practical complement to our proof-based argument for reproducible data analysis, here we focus on the practicality of setting up reproducible research compendia, with unit tests, as a measure of `code::proof`, confidence in computational algorithms.

Keywords: Meta-research · Metaprogramming · Statistical computing.

3.1 The Kafkaesque dystopia of DevOps

In Franz Kafka's 1925 novel *The Trial* [52], the fictional character Josef K. is prosecuted for crimes that are not clear, in proceedings brought forth by an unidentified authority. For the diligent scientist attempting to answer a mathematical question computationally, such as measuring the efficacy of a statistical estimator via simulation, the process of implementing a scientific workflow to achieve this aim can be a *Kafkaesque* tour of computational tools and systems. The scientist may feel as if they are locked in a dystopia, tested repeatedly for practices in

which they have not been trained, such as shell scripts and computational architecture. Whilst there are detailed guides for specific computational tools, it is hard to tell what is still relevant, as code frequently slides into obsolescence [83], and identify the optimal place to begin [111]. Significant cultural barriers continue to exist in programming fora; for example, only one in seventeen contributors to Stack Overflow[1] identify as women [30].

For many an unfortunate scientist, the dystopian experience is not confined to the *DevOps*, the developmental operations of preparation for the implementation of an algorithm [50]. Just as Josef K. was tried multiple times, the labours of the scientist attempting to answer a mathematical question computationally have only just begun. Analogous to how a string will knot with mathematical predictability when jostled [5], an algorithm will reliably require *debugging*, the process of identifying and correcting code, either to incorporate a new feature, or to correct an error. This scientist finds themselves part of the first generation of *research software engineers* (RSEs), who use computational tools in discipline-specific research practices [113]. By virtue of pioneering, RSEs are inadvertently cast as meta-researchers[2], developing new methodologies for scientific technologies that hitherto did not exist [53]. With the aim of mitigating the dystopia of DevOps and debugging for RSEs, this manuscript proposes a *toolchain walkthrough*, an *opinionated* [79] documentation of a scientific workflow, towards a measure of `code::proof`, a *good enough* [111] effort to provide computational confidence through reproducible research compendia with unit tests.

3.2 Toolchain walkthrough

We define a *toolchain* as a collection of computational tools and commands that forms a scientific workflow to achieve a specific research objective, such as test the efficacy of a statistical estimator in a particular context. The term *walkthrough*, we borrow from video game terminology [20], and is defined as a guide for other players of the game. Various walkthrough formats exist to optimise the narrative enjoyment of the gamer. For example, the Universal Hint System [98] interface provides the gamer with ever more revealing hints without spoiling other parts of the game. Next generation walkthroughs see in-game modifiers, in games such as World of

[1] Stack Overflow (https://stackoverflow.com/) is forum for asking tightly scoped programming questions.

[2] Visit the discussion on meta-research and RSEs on the research compendium associated with this manuscript as an example of why this paper, and its companion [39], have so many acknowledgements. Canonical literature is not yet established in the field of RSE, and thus leaders of RSE projects, such as Alex Hayes' maintenance of the `broom::` [85]. This has propelled Hayes rapidly to the level of expert, by virtue of the pioneering collaborative structure of the package, where hundreds of statistical modellers contribute integrated code.

Warcraft, where these provide an option for on-screen boss-specific warnings [91].

We define *toolchain walkthrough* as an *opinionated* [79] documentation of a scientific workflow, where opinionated is a term appropriated from software engineering that acknowledges that software guides the user to certain choices. In this manuscript, we describe a workflow for building a research compendium that is opinionated in privileging reproducibility. As with the hint systems of gaming, a workflow can and must be tailored to the skill and background of the user. Thus toolchain walkthroughs can be extended and adapted for different disciplines.

Toolchain walkthroughs have not only intrinsic value in terms of solving the intended research problem, but also extrinsic value, pedagogically and from a developmental perspective. Frequently those who are undertaking research software engineering on statistical projects are not the most senior member of the team; in the case of university faculty, these are often also lecturers and service teachers. There is value in seeing the minutiae of what the footsoldiers of research development undertake and how they instruct others. This can inform as to what skillsets are required in graduate courses, or are required for those who wish to optimise scientific workflow for researchers. Much of what is being implemented right now, in workflows recommended in texts[3] such as *R Packages* [104] and *Advanced R* [105], is being adopted from existing software engineering principles. Toolchain walkthroughs can contribute to the literature on the adoption of these procedures in a research context, in addition to programming fora and blog posts.

Blog posts and programming fora, as well as printed texts, are inevitably bound for obsolescence [83]. Vignettes, tool-specific long-form documentation [104], focus on one tool in the chain. As a counterpoint to the inadvertently implied redundancy of the academic manuscript in the theoretical companion manuscript [39], here we consider if the ephemerality of most-recent publications, and the chronological nature of academic publishing, may serve the breakneck speed of research development. The toolchain walkthrough provides a documentation of a specific scientific workflow constructed by an expert, or expert in training, in the field. Indeed an expert in training is perhaps best placed, as by virtue of inexperience must research in order to solve the problem. The challenge above, say, the standard one might expect from a blog post, is to provide a *good enough*[4] [111] effort to avoid questionable research practices [31] that privilege, say, convention over optimal scientific methodology.

[3] As in the companion manuscript [39], we focus on R packages, but the reader is invited to consider these as examples rather than definitive guidance. The same arguments hold for other languages, such as Python, and associated tools.

[4] As opposed to ofttimes unattainable or impractical *best practices* [110] in scientific computing.

3.3 Two research compendia case studies

For concrete examples of the benefits of adopting software research engineering principals in mathematical science, we consider two in-development research compendia, `varameta::` and `simeta::`. The primary purpose of these packages is to provide a comparative analysis of estimators for the variance of the sample median when quartiles are provided, rather than a measure of standard deviation, within the meta-analytic context. However, by structuring the packages as such, rather than within a single script file, there is scope for solving similar problems.

3.3.1 The `varameta::` package; a comparative analysis

In contemporary meta-analytic computational tools, such as the R package `metafor::` [99], a measure of both an effect and its variance are required to estimate the population parameters of interest.

However, not all studies report a variance of effect; particularly when scientists suspect an underlying asymmetry in the distribution of the observed data, prompting them to report quartiles, rather than sample standard deviations. One solution to this is to approximate estimators for mean and variance from quartiles [7, 49, 102]. We wish to explore the comparative efficacy of an estimator for the variance of the sample median derived from the estimator of [87]:

$$\operatorname{var}(m) \approx \frac{1}{4nf(\nu)^2}$$

where m denotes the sample median, n the sample size, ν the population median, and f the population probability density function.

However, in an experimental setting, we do not know the true distribution, nor the true population median. Thu, our method proposes that we assume a distribution, and estimate the parameters of that characterized the assumed distribution from the sample size and sample quartiles. We provide estimators derived for different distributions, to assess the efficacy of this analysis framework. One of which is the exponential distribution, which this manuscript will focus on.

If we assume that f is an exponential probability density function, with unknown rate parameter λ, then we can estimate this rate parameter via the sample median. Since the true

median is given by $\log 2/\lambda$, we can estimate the rate parameter,

$$\lambda \approx \log 2/m, \tag{3.1}$$

via the sample median, m.

Each proposed estimator requires a different set of reported values as inputs and different calculations. It is notable that a most optimal estimation method for the problem above is generally unknown. For example, in the comparative analysis Wan et al. [102], it was shown that the performance of different estimators varied with the simulated sample sizes.

Thus, there is merit to providing not only the practical functionality of our proposed solution, but also the existing solutions. By structuring this comparative analysis as a reproducible research compendium we achieve practical improvements on a self-contained computational script file. Via `roxygen::`ised [109] documentation, estimators are provided in a modular fashion, with a devoted script file for each estimator that is easily sourced from the package environment. In addition to the advantage of debugging a single script file, the comparative analysis also serves a practical purpose, providing a characterisation of the functionality of each estimator.

To compare these estimators for the variance of the sample median, we undertook *coverage probability* simulations. Here, the coverage probability refers to the probability that the true parameter of interest falls within its constructed confidence interval. In order to do so, we require simulated meta-analytic data, which has the added complexity of a random effect that governs the variation *between* studies. To solve this with confidence in the implementation of computational algorithms and mathematical derivations, we structure this as a package. In addition to building `code::proof`, by separating the simulation component, we begin to develop a computational solution to not only solving this problem, but the testing of `any` estimator for the variance of the sample mean or median.

3.3.2 The `simeta::` package

A *coverage probability* simulation repeats several trials with the same simulation meta-parameters where the differing factor is the random sampling of data. In order to separate simulation meta-parameters from trial-level parameters, and delineate this algorithm, we begin by considering a single trial from a standard coverage probability simulation.

3.3.3 Coverage probability simulation

Each trial draws a random sample, for example `rnorm(n = 100, mean = 3, sd = 0.2)` will produce 100 values drawn randomly from a normal distribution with mean 3 and standard deviation 0.2. From this sample, we calculate summary statistics. Using these summary statistics, we can compute an estimate of the parameter of interest $\hat{\nu}$, and its variance $\hat{\gamma}$. With these estimates, we can produce a $(1-\alpha) \times 100\%$ confidence interval $\hat{\nu} \pm z_{1-\alpha/2}\sqrt{\hat{\gamma}}$, where $z_a = \Phi^{-1}(a)$ is the ath quantile of the standard normal distribution, and Φ is the standard normal distribution function. Given we set the parameters for the random sample drawn, we know the true parameter, ν. Thus we can ask, does ν fall within the confidence interval produced? We summarise the steps of a trial as an algorithm:

1. Draw a random sample from the distribution that is characterised by the parameter of interest, ν;
2. Calculate summary statistics from the random sample;
3. Calculate an estimate of ν from the summary statistics;
4. Construct a confidence interval using the parameter estimate;
5. Check if ν falls within the confidence interval.

A coverage probability simulation performs multiple trials and returns the proportion $p \in [0, 1]$ of confidence intervals for ν that contain the generative parameter value.

3.3.4 Simulating meta-analysis data

For a meta-analysis simulation, however, these steps are significantly more involved. And with this complexity, as we shall see, nesting, of the algorithm, the advantages of the package structure begin to become apparent. In a single script file, it is hard to find at which step of the algorithm that the code has failed. In addition to human error introduced into code, there are also practical considerations. For example, the random effects maximum likelihood model, `method = REML`, employed by `metafor::rma` [99] does not always converge on estimates for the effect and its variance, in which case a fixed effects model, `method = FE`, can be employed to produce parameter estimates.

The other point of complexity is in the sampling of meta-analytic data. As meta-analytic data is a collection of summary statistics for K studies of control and intervention samples, the first step of a coverage probability simulation trial,

1. Draw a random sample from the distribution that is characterised by the parameter of interest, ν,

requires several substeps. For the kth ($k \in \{1, \ldots, K\}$) study, we assume there is variation γ_k associated with that study, and, in particular, the control, with parameter ν_k^C, and intervention, with parameter ν_k^I, samples with ratio, $\rho = \nu_k^C / \nu_k^I$.

Let us consider a practical example from the estimators provided in the comparative analysis, `varameta::`. Our estimator of interest is the variance of the log-ratio of sample medians for control, ν^C, and intervention, ν^I groups. Since our focus is on building the research compendium to undertake this analysis, rather than the estimators in question, we will take the simplest case, where there is one parameter λ associated with the distribution of interest. Let us assume an underlying exponential distribution: Exponential(λ).

At the simulation level, which is to say, across all trials, we set λ, the parameter of the distribution of interest. Also at the simulation level, we define a ratio $\rho := \nu^C / \nu^I$ of interest for the population medians, where $\rho = 1$ would indicate no true difference between control and intervention groups. We assume that the log-ratio of sample medians $\log(m_k^I / m_k^C)$ for the kth study, can be characterised in terms of the log-ratio of populations medians $\log(\nu^C / \nu^I)$, with some error $\gamma \sim N(0, \tau^2)$ association with that study, as well as sampling error, $\varepsilon \sim N(0, \sigma^2)$,

$$\log(m_k^I / m_k^C) = \log(\nu^I / \nu^C) + \gamma_k + \varepsilon_k.$$

Since the underlying distribution is exponential, we need to find λ_k^J for $J \in \{C, I\}$ in order to sample n values $x_1, \ldots, x_n \sim \text{Exponential}(\lambda_k^J)$. We also know the median of the exponential distribution with rate parameter λ is given by $\log 2 / \lambda$. Then, assuming the sampling error will be attained through the random computational process, we have

$$\begin{aligned}
& \log(m_k^I / m_k^C) = \log(\nu^I / \nu^C) + \gamma_k \\
\implies & \log(\lambda_k^C) - \log(\lambda_k^I) = \log(\lambda^C) - \log(\lambda^I) + \gamma_k \\
\implies & \log(\lambda_k^C) - \log(\lambda_k^I) = (\log(\lambda^C) + \gamma_k/2) - (\log(\lambda_k^I) - \gamma_k/2)
\end{aligned}$$

If we then split the random effect associated with the variation between studies γ_k equally,

and divide the terms by experimental group $J \in \{C, I\}$, we obtain the following system for the control C and intervention I groups' kth parameter, λ_k^J.

$$\lambda_k^C = \lambda^C \exp(\gamma_k/2)$$
$$\lambda_k^I = \lambda^I \exp(-\gamma_k/2)$$

1. Draw a measure of variation for the kth study from $N(0, \tau^2)$ and calculate λ_I from fixed values, the ratio of medians, ρ, and the control group's rate parameter λ_C;

2. Calculate the rate parameters for the control, λ_k^C, and intervention, λ_k^I, groups for the kth study;

3. Draw a random samples of size n_k^J from Exponential(λ_k^J), for $J \in \{C, I\}$.

The sample size n_k^J for the Jth group of the kth study can also be sampled, by assuming $N_k := n_k^C + n_k^I$ and drawing N_k from a uniform distribution Uniform(a, b), where the minimum a, and maximum b, reflect knowledge about the domain of interest. The proportion of N_k given to n_k^I can be drawn from a beta distribution. But we shall omit the derivations of these sampling distributions, in the interests of brevity.

In the sampling steps that have been outlined, there are random values drawn, but there are also set simulation-level parameters. We may wish to see how our estimator performs for different numbers of studies, K, different expected variability between the studies, τ^2, and whether or not there is a difference between the control and intervention groups, ρ.

And finally, if we consider other distributions, with a mix of symmetric, say, normal or Cauchy distribution, and asymmetric, say, exponential or log-normal, we require different derivations for the sampling parameters.

3.3.5 Complexity and formalised analysis structures

Via the modular nature of a research compendium R package, we can separate each layer of the algorithm into functions. We can produce automated *unit tests* for these functions that, at the very least, check that each component of the algorithm returns an output of expected type. We cannot automate the mathematical derivations, but we can produce an algorithm structure that provides far more computational confidence in implementation than a single script file in which the entire algorithm is nested.

However, structuring an analyses in research compendia is more challenging than simply coding directly into a `.R` script. Thus, there is benefit to outlining the computational workflow. We now turn to the practical *toolchain walkthrough* for establishing these analyses as research compendia. We may not be able to prepare for all errors, but we can aim to weather *most* problems that arise in the computational implementation of mathematical algorithms.

3.4 Research compendia toolchain walkthrough

We now aim to provide a practical guide to computational research compendia for the comparative analysis, `varameta::`, and the simulation algorithm, `simeta::`, that supports it. As this is a first effort at a toolchain walkthrough, there will likely be aspects that are overlooked or underdeveloped.

3.4.1 DevOps

The DevOps section of this toolchain walkthrough aims to cover computational tools, why they were chosen, as well as some guidance as to how to source them.

Intended audience.

A toolchain walkthrough is a documentation of a specific scientific workflow created by a scientist who utilised this workflow for research. We begin by identifying the audience targeted who may benefit from detailing the minutiae of this process. We do not seek to generalise, but rather to provide a workflow that reflects the author's knowledge of good enough practices in scientific computing for this task, optimised for efficiency, scientific rigour, and, in the spirit of the gaming walkthrough: *fun*.

This toolchain walkthrough assumes an R user whose expertise is not primarily in computing, but rather a researcher who employs R for analysis in a discipline such as statistics, psychology, archaeology, or ecology. We make an effort to cover some of the less familiar aspects of computational workflow, such as shell commands, that might be considered trivial to a formally trained computer scientist.

Although many R users have gaps in their formal computational science education, researchers who utilise R are often implementing complex algorithms, such as the one outlined in Section 3.3.2, which describes the simulation of meta-analysis data for coverage probability simulation.

Burn it down.

This section only applies for work that has already begun. However, this is often the case for the development of a scientific project. We frequently have work that begin as small scripts, that develop in complexity and requirements.

In recognition of the ofttimes overwhelming density of resources, we list a few bash shell commands here that are particularly useful for moving files around when setting up an analysis as a research compendium. We enclose user input in <> and describe the utility of the command after \#. A directory is colloquially referred to as a folder. These can be executed from a terminal.

```
. # here
.. # up one
cd <directory path> # change location of .
ls -a # list files in .
cp <file> <toplace> # copy
mv <file> <toplace> # move or rename
rm -rf <directory> # remove directory and its contents
locate <partoffilename> # find a file
mkdir <directory> # create a directory
```

How to code.

The R software environment can be downloaded from R: The R Project for Statistical Computing. There are several excellent resources for getting started with programming with R. We list an opinionated selection here, chosen for clarity and enjoyment, all of which are freely available online:

- Learning Statistics with R by Danielle Navarro [73],
- R for Data Science by Grolemund Garrett and Hadley Wickham [40],
- R Cookbook by J.D. Long and Paul Teetor [60].

We now assume a working knowledge of the R programming language, as the intended audience of this toolchain workflow are researchers who have a working level of programming proficiency in R.

Where to code.

In this toolchain walkthrough, we emphasise cross-platform open-source software. There is, of course, the immediate benefit of accessibility. Furthermore, open-source invites an evolutionary development community where many can contribute small solutions that integrate to solve larger problems. RStudio is an integrated development environment for writing in the statistical language R. RStudio is *cross-platform* in that it can be installed on Windows, Macintosh, and Linux operating systems. There are many further advantages to this widely-used environment. For example, the `citr::` add-in [3] modifies RStudio to enable a connection to the open-source reference manager Zotero. Another example is the `datapasta::` [68] add-in that enables copy-paste of tables into R-formatted script.

3.4.2 Create compendium architecture

As `varameta::` is a research compendium containing comparative analyses and `simeta::` a package to provide simulation tools, the creation process for these two compendia are different.

We make use of two R packages, `rrtools::` [65] and `usethis::` [108], to assist in automating these tasks.

Compendiumise `varameta::`.

1. Open RStudio and close project via the toolbar File menu,
2. In the Console, set the working directory to desired location; e.g.,

```
> getwd()
[1] "/home/charles"
> setwd("Documents/repos/")
> getwd()
[1] "/home/charles/Documents/repos",
```

3. and `rrtools::use\_compendium("varameta")`,
4. and update `DESCRIPTION` file with author, title, etc.,
5. Create analysis file structure with `rrtools::use\_analysis()`.

For `varameta::`, we will have several reproducible documents that will form the basis of the analysis, as well as figures to contribute to the associated publication. The final step above automates the creation of a directory structure for a paper, figures, data, and templates.

Compendiumise `simeta::`.

In this case, the file structure is less involved, however the testing structure will be need to be considerably more robust because of the complexity of the simulation algorithm described in Section 3.3.2:

1. Create a package with `usethis::create\_package()`,
2. Switch to the package directory with `usethis::project\_activate()`.

3.4.3 Common steps across both packages

1. Set open source licence, with

   ```
   usethis::use\_mit\_license(name = "Charles Gray");
   ```

   this 'simple and permissive' choice of licence [108] serves the purpose of a comparative analysis of estimators,
2. Set up documentation for functions with `usethis::use\_roxygen\_md()`,
3. Set up data for internal datasets and examples with `usethis::use\_data()`.

Connecting to GitHub.

There are benefits to implementing a version control system, such as via the Git language and GitHub online repository archive, beyond the ability to trace work back to an earlier iteration [15]. The added benefit, arguably even greater benefit, is that of collaborative science. Storing work on GitHub allows for instantaneous sharing of code and analyses, and collaborative work with advanced project planning features, enabling other scientists to make very specific comments on work in progress.

Data ethics and further considerations.

In the case of `varameta::`'s estimators for meta-analysing medians, and `simeta::` for simulating meta-analysis estimators, there are no ethics in data considerations beyond ensuring contributors are recognised and credited for their work by time of publication. For some disciplines, sharing geographic locations might be an ethical consideration, say, in preventing fossil hunters from exploiting palaeontology sites [48]. Personal details, must, too be considered, that might inadvertently identify people and violate privacy considerations. Furthermore, various allowances might need to be made for institutional workflow. We note these here as a possible considerations, but as our case studies do not have such requirements, we now consider our research compendia instantiated.

However, as this algorithm has significant complexity, we need to include unit tests to provide confidence in our results , as we argue in the companion computational metamathematics manuscript [39], which motivates the practical steps laid out here.

3.5 Testing

We now expand in a practical sense on unit testing, which, in the theoretical companion manuscript, we describe 'the software engineering tool that provides a key piece of the correspondence between scientific claim and programming' [39]. It is in this manuscript that we sought to answer the question: why test? In this toolchain walkthrough, we will focus on the practical implementation of first unit tests.

3.5.1 What is a test?

Tests are collected in contexts. Each test comprises congruous expectation functions.

In the *head* of the 'bug hunt' context (under `context("bug hunt")`), we find the loading of packages. A seed is then set for reproducibility of errors. The first test, `"metasim runs for different n"`, tests the `simeta::metasim()` function for different orders of magnitude of `trials`. As each trial samples new data, this is the most direct way to test the scalability of the function for large datasets. We then follow up with a test that checks that the exponential distribution can be passed to all levels in the algorithm.

```
context("bug hunt")
```

```
set.seed(38)
library(tidyverse)
library(metasim)

test_that("metasim runs for different n", {
  expect_is(metasim(), 'data.frame')
  expect_is(metasim(trials = 100) , "data.frame")
  # expect_is(metasim(trials = 1000) , "data.frame")
})

test_that("exponential is parsed throughout", {

  # check sample
  expect_equal(
  sim_sample(10, rdist = "exp",
    par = list(rate = 3)) %>% length, 10)
  # check samples
  ...
```

3.5.2 Non-empty thing of expected type

Simply asking '*does a function produce the expected output?*', induces a surprising number of considerations. To illustrate this, we return to our case studies.

Testing a collection of estimators in `varameta::`.

In the interests of mathematical and computational brevity, we focus on one distributional example: the simple case of the exponential distribution, which is characterised by a single parameter. We return to the estimator of the rate $\hat{\lambda} := \log 2/m$ derived for the exponential distribution, as discussed in Section 3.3.4 and defined in Equation (3.1), explicitly coded in R.

```
function(n, median) {

  # Estimate parameters.
  lambda <- log(2) / median
```

```
  # Approximate the standard error of the sample median.
  1 / (2 * sqrt(n) * dexp(median, rate = lambda))

}
```

We create a context file, **tests/testthat/test-exponential.R** and provide a short context description in the first line of the script.

```
  context("exponential estimator")
```

As a starting point, we can write unit tests to automate a check that this function returns non-empty thing of expected type. We arbitrarily choose values, a sample size of 10, and a proposed sample median of 4, for instance. The function should return a numeric `double` value, and should be positive.

```
  test_that("non-empty thing of expected type, for fixed values", {

    # returns numeric
    expect_type(g_exp(10, 4), "double")

    # returns positive number
    expect_gt(g_exp(10, 4), 0)

  })
```

In addition to choosing explicit values, we can also randomly sample the sample size `n`, and sample median `m`. To ensure reproducibility of these testing results on any machine, we set a random seed, passing `set.seed` an arbitrary numeric value.

```
set.seed(39) # ensures reproducibility of test results

# sample fuzz testing parameters
n <- sample(seq(2, 100), 1)
m <- runif(1, 1, 100)
```

We can then use these random *fuzz* values [55] to produce analogous unit tests for non-empty thing of expected type.

```
test_that("non-empty thing of expected type, for random values", {
  expect_type(g_exp(n, m), "double")
  expect_gt(g_exp(n, m), 0)
})
```

We can extend these tests to cover expected input errors. For example, we wish this function to fail when passed negative numbers. The sample size cannot be less than or equal to 0, and due to the logarithm, the function only works for positive sample medians. Here, we include the fixed and randomised values in the same test.

```
  test_that("negative numbers throw an error", {
  expect_error(g_exp(-3, 4))
  expect_error(g_exp(3, -4))
  # with fuzz testing
  expect_error(g_exp(-n, m))
  expect_error(g_exp(n, -m))
})
```

Running all tests in a context tells us if the function is behaving as expected. The more tests we write, the more confidence we will have that our function behaves as we intended it to.

```
==> Testing R file using 'testthat'

Loading varameta
  |  OK F W S | Context
  |   8       | exponential estimator

  Results
OK:       8
Failed:   0
Warnings: 0
Skipped:  0
```

```
Test complete
```

There is a tradeoff with tests, in terms of time taken by updating the tests themselves. Here a test requires updating from an expected output of a numeric vector, to a dataframe. The function that is being tested.

```
==> Testing R file using 'testthat'

Loading simeta
  |  OK F W S | Context
  |   6 1     | bug hunt [7.1 s]

test-bug-hunt.R:21: failure: exponential is parsed throughout
sim_stats(rdist = "exp", par = list(rate = 3)) inherits from
`tbl_df/tbl/data.frame` not `numeric`.


  Results
Duration: 7.1 s

OK:       6
Failed:   1
Warnings: 0
Skipped:  0

Test complete
```

Testing a nested algorithm in `simeta::`.

Our other case study provides an example of a nested algorithm. In addition to ensuring each function returns a non-empty thing of expected type, we can automate checks that the functions form a toolchain. In the first place, it is helpful to know that our functions continue to form a toolchain under default settings.

We begin by setting our context. In this case, as we are running our functions on default settings, we do not require randomly sampled fuzz parmeters.

```
context("default pipeline")
```

We now check that the algorithm runs 'upwards', by running a test from most granular function in the algorithm to most nested. We could write a similiarly inverted test, from most nested function, downwards to most granular.

```
test_that("work upwards through algorithm", {
  expect_is(sim_n(), "data.frame")
  expect_gt(sim_n() %>% nrow(), 1)
  # sim_df calls sim_n
  expect_is(sim_df(), "data.frame")
  expect_is(sim_stats(), "data.frame")
  # metasim calls metatrial
  expect_is(metatrial(), "data.frame")
  expect_is(singletrial(), "data.frame") # alternate trial
  expect_is(metasim(trials = 3), "data.frame")
  # metasims calls sim_df & metasim
  expect_is(metasims(
    single_study = FALSE,
    trials = 3,
    progress = FALSE
  ),
  "sim_ma")
})
```

Now, if this test fails, we will know the combination of functions fails at some point in the nested algorithm. We follow this upwards test with a series of small tests for each function set to defaults to identify at which point in the pipeline where the algorithm fails, if the 'work upwards' test fails.

```
# test each component on defaults

test_that("sim_n", {
  expect_is(sim_n(), "data.frame")
})
```

```
  test_that("sim_df", {
    expect_is(sim_df(), "data.frame")
  })

  test_that("metatrial", {
    # metasim calls metatrial
    expect_is(metatrial(), "data.frame")
  })

  test_that("singletrial", {
    expect_is(singletrial(), "data.frame") # alternate trial
  })

  test_that("metasim", {
    expect_is(metasim(trials =  3), "data.frame")
  })

  test_that("metasims", {
    expect_is(metasims(
      single_study = FALSE,
      trials = 3,
      progress = FALSE
    ),
    "list")
  })
```

And we can now run all tests, for a starting point of automating checks that our algorithm runs on default settings.

```
==> Testing R file using 'testthat'

Loading simeta
  |  OK F W S | Context
```

```
  |  14       | default pipeline [28.7 s]

  Results
Duration: 28.7 s

OK:       14
Failed:   0
Warnings: 0
Skipped:  0

Test complete
```

To demonstrate how informative testing can be in identifying where an algorithm breaks, we now modify the `simeta::metasim` function to return a character string, `"error"`. Testing the default pipeline reveals where the algorithm is broken. Debugging is where the advantage of testing is exposed, and thus, arguably the requirement for testing increases with complexity of algorithm. Detailed output have been omitted for brevity.

```
==> Testing R file using 'testthat'

Loading simeta
  |  OK F W S | Context
  |  10 4     | default pipeline [32.3 s]

test-default-pipeline.R:12: failure: work upwards through algorithm
metasim(trials = 3) inherits from `character` not `data.frame`.

test-default-pipeline.R:14: error: work upwards through algorithm
Argument 1 must have names
...

test-default-pipeline.R:43: failure: metasim
metasim(trials = 3) inherits from `character` not `data.frame`.

test-default-pipeline.R:47: error: metasims
```

```
Argument 1 must have names
...


  Results
Duration: 32.3 s

OK:       10
Failed:   4
Warnings: 0
Skipped:  0

Test complete
```

From this output, we can see not only where the algorithm fails, but also what other functions fail because of a reliance on the elements that have failed.

3.5.3 Test-driven development

As we build new features into our package, such as checking that the single-trial setting works in the simulation function from `simeta::`, we can focus on a writing new tests that ensure our feature works within the ecosystem of our algorithm as expected. We can develop our algorithm from a testing setting, rather than focusing on rewriting functions and script files.

Another overview check that we can incorporate is from the `covr::` package [47]. Using `covr::package\_coverage()`, we can check what proportion of lines of code have been tested in each function.

For the `varmeta::` package, at the time of writing, we have the following test coverage.

```
varameta Coverage: 90.00%
R/g_cauchy.R: 44.44%
R/g_norm.R: 71.43%
R/hozo_se.R: 92.31%
R/bland_mean.R: 100.00%
R/bland_se.R: 100.00%
R/effect_se.R: 100.00%
```

```
R/g_exp.R: 100.00%
R/g_lnorm.R: 100.00%
R/hozo_mean.R: 100.00%
R/wan_mean_C1.R: 100.00%
R/wan_mean_C2.R: 100.00%
R/wan_mean_C3.R: 100.00%
R/wan_se_C1.R: 100.00%
R/wan_se_C2.R: 100.00%
R/wan_se_C3.R: 100.00%
```

This is enables us to target specific functions that may require further testing. Testing lines of code is somewhat a blunt instrument, as we are not ensuring tests for every combination of inputs. However, test coverage is still an informative measure of software reliability. For example, here we see not all code in the `g\_*` estimators have been checked.

These notes on testing are not intended to be comprehensive, but only aim to give the user an starting point for the initialisation of summarising an analysis in a reproducible research compendia, with an informative level of automated checks. Given only one quarter of packages on the largest R package repository CRAN have unit tests at all [39], it is arguable that there is much further scope for discussion and development with respect to the adoption of automated tests in reproducible research compendia.

3.6 Prepare for *most* weather conditions

Computational proof may be unachievable, however, a measure of `code::proof` can be attained by structuring research compendia in a standardised reproducible format, such as produced by `rrtools::` [65]. Perhaps we cannot prove our software in the traditional mathematical sense [39]. However, we could consider building confidence in the mathematics that we implement computationally, like waterproofing our shoes. If we step in a big enough puddle, our feet are still going to get wet, but at least we have prepared to weather *most* of the problems associated with the implementation of statistical algorithms.

Chapter 4

Meta-analysis of Medians

Estimating the variance of the sample median

Charles T. Gray, Luke Prendergast, Emily Kothe, and Hien Nguyen

4.1 Medians pose a problem in meta-analyses

Software tools for meta-analysis, such as Cochrane's RevMan [62], newly superseded by the cloud-based RevMan Web [63] or the R package `metafor::` [99], require estimates of both effect and variance of that effect. However, the sample variance for the reported effect of interest is not always available. When the reported statistics are medians, with the measure of spread commonly provided in the form of quartiles, as opposed to the required variance of the effect of interest. This leads to the omission of studies that report medians from the meta-analysis. In this manuscript we present a method for estimating the variance of the sample median so that studies reporting medians may be included in meta-analyses.

This manuscript is a component of the research compendium created to solve this problem. In this case, the research compendium comprises not just a manuscript, but a pair of software packages, `varameta::`, which translates estimators presented in this manuscript to code, and `simeta::`, for simulating meta-analysis data, to see how the estimators in `varameta::` perform.

Chapter 5 breaks down the computational components of the simulation, as well as the underlying derivations. In this chapter, we explore the theoretical underpinnings of the estimators provided in `varameta::`. The lens through which we discuss the problem of medians in this dissertation is reproducible computation, so, in addition to intrinsic questions regarding

meta-analysing medians, this manuscript considers on the ideas presented in the companion papers [39, 38] that ruminate, with this analysis as a case study, on why and how we may build reproducible research compendia.

4.1.1 What's the problem?

Skewed data is often summarised by reporting the median and either the interquartile range or range. Quartiles might be provided as an interval of two measures or a difference of quartiles. While this may be useful in a descriptive single-study sense, the lack of reported estimator variability poses a challenge in the context of meta-analysis. Software for performing meta-analyses, such as the widely-used R package `metafor::` [99], require an estimate of the variance of the reported effects to conduct the meta-analysis under the assumed model

$$\widehat{\delta}_k = \delta + \gamma_k + \varepsilon_k, \tag{4.1}$$

where $\widehat{\delta}_k$ is the estimated effect from the kth study, δ is the population effect of interest, ε_k is the error allowing for sampling variability and $\gamma_k \sim N(0, \tau^2)$ is the random effect to allow for differences in the true effects between studies. Given the estimated effects for K studies, all assumed to be normally distributed with a known (or estimated) variance, a meta-analysis can be carried out to estimate δ and the random effect variance τ^2. Our focus is on meta-analysis of three different effects involving the median. The first is simply the median itself when there is only one group of interest in each study. The second is the difference of two medians when there is two groups to be compared within each study (such as a case and control group). The third, which may be more suitable than the difference in medians when measurements of scale differ between studies, is the ratio of medians. For more on meta-analysis see, e.g., [9] and [56].

4.1.2 Why propose a new method?

In this paper we propose a method for meta-analyses of studies whose effects are reported in the form of median and interquartile range or range, that is, in the form of quartiles. The previously proposed method of [49], and extensions by [7, 102], solve this problem by estimating the mean and standard deviation from the provided summary statistics. For some applications there may be two noticeable drawbacks to this approach.

Disadvantage 1. *The methods to convert to a mean and standard deviation perform well when the underlying distribution is symmetric, and in some cases more specifically when it is*

normal. However, results have shown that performance can be poor in the presence of skew (e.g., see [89]).

Disadvantage 2. *Those who initially published the summary measures, may have chosen to report medians and ranges because they had decided that moment-based measures such as the mean were not suitable descriptors.*

In the presence of underlying skewed distributions, both Disadvantages 1 and 2 may cause a real threat to the validity of any inference following conversion from medians to means. Indeed, and the choice of whether to report mean or median is both random and not independent of the data. Meta-regression is one possibility for combining means and medians, however, the robustness of this approach would require consideration that is beyond the scope of this manuscript. This underscores the arguments made in Chapters 2 and 3, if the raw data for each study were available, the meta-analyst could extract whatever sample statistics they consider best to answer the question at hand.

The method that we propose can be easily adapted to both single-study and meta-analysis contexts. To illustrate this problem, we begin with an example meta-analysis from medical research. We then briefly touch on how our method contributes to the existing solutions for this problem. In Section 4.4, we define our estimator for the variance of the sample median and consider alternatives. We show how this estimator can be used in meta-analysis. Simulation results are provided in Section 4.5.1 that assess the performance of our estimator in both the single-study and meta-analysis setting. Finally, in Section 4.6.1, we return to the motivating example, discussed in Section 4.2, to demonstrate how our method can be applied. Concluding remarks are provided in Section 4.6.1.

4.2 A motivating example

To motivate our method, we detail an example of the variety of summary statistics that can arise in meta-analyses. We shall return to this example in Section 4.6.1 to see how our method facilitates meta-analysis of all studies, rather than just the three studies originally included, which reported means and standard deviations.

We choose notations similar to those used by Wan *et al.* [102]. Define: a, the minimum value; q_1, the first quartile; m, the median; q_3, the third quartile; b, the maximum value; n, the sample size. IQR denotes the interquantile range and this may be reported as an interval, i.e. (q_1, q_3), or a width, i.e. $q_3 - q_1$. We also let $\overline{x}$ and s denote the sample mean and sample standard

deviation, respectively. Later, we either subscript or superscript such measures appropriately to identify different groups within studies, and in the example below this equates to, e.g., m^C denoting the median for a control group and m^I for the intervention group.

As an example of how studies collated in meta-analyses report these summary statistics differently, consider the dataset presented in Table 4.1, taken from a systematic review of d-dimer in pre-eclampsia [82]. Ideally, one would want to perform a meta-analysis using the effects from all studies. However, estimator variance is only reported for three of the seven studies presented with the remaining studies reporting medians and ranges.

Table 4.1: Data from a meta-analysis of d-dimer levels in pre-eclampsia presented by [82], measures of location and scale are varied: there are means and standard deviations; medians and interquartile ranges; quartiles; and medians and ranges. The types of estimates reported are listed in the final column denoted 'reported'.

		Control group			Intervention group			
study	year	location	scale	n^C	location	scale	n^I	**reported**
Dusse	2003	1146.6	311.2	28	1263.8	411.9	43	$\boldsymbol{\overline{x}, s}$
Schjtlein	1997	1390.0	559.0	97	1545.0	849.5	200	$\boldsymbol{\overline{x}, s}$
Terao	1991	221.52	179.9	80	347.87	460.5	13	$\boldsymbol{\overline{x}, s}$
Catarino	2008	538.2	(391.2, 822.8)	42	448.5	(313.0, 1091.3)	44	$\boldsymbol{m, (q_1, q_3)}$
Bellart	1998	545.0	225.0	65	2090.0	1800.0	12	$\boldsymbol{m}$, **IQR**
Heilmann	2007	1149.0	456.0	33	1623.6	932.9	111	$\boldsymbol{m}$, **IQR**
He	1997	183.0	(110.0, 340.0)	24	315.0	(145.0, 1150.0)	30	$\boldsymbol{m, (a, b)}$

Three studies (first authors Dusse, Schjtlein, and Terao) detailed in the table provide the sample mean, $\overline{x}$, and standard deviation, s. Two studies (Bellart and Heilmann) provide the sample median, m, and interquartile range, IQR. One study (Catrino) provides the sample median, as well as the first and third quartiles, q_1 and q_3. Finally, one more study (He) provides the sample median and the minimum a and maximum b observed values. All studies provide the sample size n and their respective estimates of location and scale for both the control and the pre-eclamptic groups.

In order to perform a meta-analysis via conventional methods, we require, at minimum, the studies' effect estimates, associated variances, and sample sizes. In the original analysis, replicated in Figure 4.1, the study's authors omitted the four studies that reported medians, providing an incomplete summary of the available evidence.

While full access to the raw data of each study would enable researchers to calculate the

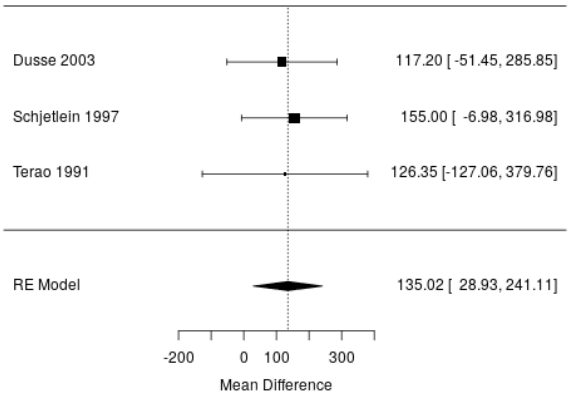


Figure 4.1: A replication of a meta-analysis [82] of studies in Table 4.1 that report the mean and standard deviation.

necessary sample variance for each study, there are many practical reasons, such as the time it would take to gather the data, that reduce the practicality of this approach, a point that is well made by others [7, e.g., p. 57]. Since only three studies presented in Table 4.1 report sample variance, Pinheiro *et al.*'s meta-analysis was restricted to these three datasets [82]. This paper provides a method that allows meta-analyses to be performed over studies reporting a variety of summary statistics, such as those outlined in Table 4.1.

4.3 Existing solutions to this problem

A potential solution is offered by Hozo *et al.* [49], who suggest estimating the mean and standard deviation from a reported median, minimum and maximum, as well as sample size, denoted $C_1 := \{a, m, b; n\}$ in Table 4.2. This provides a way to calculate the variance of the effect, as required by contemporary meta-analysis tools, although there are some limitations. For example, C_1 does not cover all cases of reported medians. In the example meta-analysis given in Table 4.1, there is only applicable study (He 1997) for this method.

Bland extends on Hozo *et al.*'s solution, but for the set $C_2 := \{a, q_1, m, q_2, b; n\}$ where

the minimum, maximum, median, as well as first and third quartiles are reported [7]. Wan *et al.* improve on Hozo and Bland's solutions, as well as providing a solution for the set $C_3 := \{q_1, m, q_3; n\}$ where the interquartile range is provided as an interval along with the median [102]. A nice review of the methods, including an improvement, can be found in [89]. However, it is noted that underlying normality appears to be the motivation for all methods and below we detail some limitations.

Table 4.2: This table is a rephrasing and detail of Table 3 from Wan *et al.* [102], and guides the simulation discussed in Section 4.5.1. Here we have details of various estimators for the sample mean and sample standard deviation, for different data sets. These estimators are defined in terms of sample summary statistics: minimum a, maximum b, median m, first quartile q_1, third quartile q_3, and sample size n. Source indicates the first author of the paper the equations are found in. The sample mean and sample standard deviation estimators are presented by columns $\overline{X}$ and S, respectively. The column C presents the sample summary statistics required as parameters in the coupled estimators.

Source	$\overline{X}$	S		C
Hozo [49]	$a + 2m + b/4$	$S \approx \begin{cases} \left[(b-a)^2 + (a-2m+b)^2/4\right]^{\frac{1}{2}}/\sqrt{12} & n \leqslant 15 \\ b - a/4 & 15 < n \leqslant 70 \\ b - a/6 & n > 70. \end{cases}$		$C_1 = \{a, m, b; n\}$
Bland [7]	$(a + 2q_1 + 2m + 2q_3 + b)/8$	$[(a^2 + 2q_1^2 + 2m^2 + 2q_3 + b^2)/16 +$ $(aq_1 + q_1m + mq_3 + q_3b)/8 -$ $(q + 2q_1 + 2m + 2q_3 + b)^2/64]^{\frac{1}{2}}$		$C_2 = \{a, q_1, m, q_3, b; n\}$
Wan [102]	$(a + 2m + b)/4$	$b - a/2\Phi^{-1}(n - 0.375/n + 0.25)$		$C_1 = \{a, m, b; n\}$
Wan [102]	$(a + 2q_1 + 2m + 2q_3 + b)/8$	$(b-a)/4\Phi^{-1}(n - 0.375/n + 0.25) +$ $(q_3 - q_1)/[4\Phi^{-1}(0.75n - 0.1215/n + 0.25)]$		$C_2 = \{a, q_1, m, q_3, b; n\}$
Wan [102]	$(q_1 + m + q_3)/3$	$(q_3 - q_1)/[2\Phi^{-1}(0.75n - 0.125)/(n + 0.25)]$		$C_3 = \{q_1, m, q_3; n\}$

Firstly, note that the summary statistics sets $\{a, m, b; n\}$, $\{a, q_1, m, q_2, b; n\}$, and $\{q_1, m, q_3; n\}$ do not cover all of the presentations of summary statistics seen in Table 4.1. Thus, even if Pinheiro *et al.* had access to all methods, the meta-analysers would still have work ahead of them to include all studies presented here.

Secondly, and more importantly, to convert medians and interquartile ranges (or ranges) to means and standard deviations ignores the implicit information conveyed by the reported summary statistics; that is, that the study's authors perceived asymmetry in the data, given the median was chosen as the measure of interest. Our motivation is to provide a solution that enables meta-analyses to retain this information and, in addition, provide a method of comparing the studies that reported means with the studies that reported medians.

4.4 Estimating the variance of the sample median

Before detailing our proposed solution for estimating the variance of the sample median under meta-anlaytic conditions, we provide expressions for approximations of the variance of a single median, a difference in two independent median estimators and the log ratio of two medians, when the underlying distribution and true median is known. It is these expressions we adapt in order to estimate these approximations.

Consider a population median denoted ν with corresponding estimator M, taken to be the middle order statistic from a sample with n observations. Let f denote the probability density function for the underlying population. Then the median estimator, M, is asymptotically normal with approximate variance (see, e.g. Ch.7 of [21])

$$\mathrm{Var}(M) \approx \frac{1}{n} \cdot \frac{1}{4\left[f(\nu)\right]^2}. \tag{4.2}$$

With this approximated variance, we can then extend to the variance of the difference and the variance of the ratio of two sample medians. For the difference of two sample medians, we have, assuming that the estimators are independent,

$$\mathrm{Var}(M_1 - M_2) = \mathrm{Var}(M_1) + \mathrm{Var}(M_2). \tag{4.3}$$

Using the delta method [77], the variance of the log ratio of two sample medians is given by

$$\mathrm{Var}\left[\log\left(\frac{M_1}{M_2}\right)\right] \approx \frac{\mathrm{Var}(M_1)}{\nu_1^2} + \frac{\mathrm{Var}(M_2)}{\nu_2^2}. \tag{4.4}$$

In practice we do not know the true population median ν, nor the true population density f, so estimates are required. It is common to only have access to the sample median and interquartile range (or range) from a single study. Or, in the case of the comparison of two samples, we may have two sample medians and associated interquartile ranges (or ranges).

However, as we shall explore in Section 4.4.2, both the log-normal and the normal densities provide surprisingly close approximations of the true densities evaluated at the median. In this paper, we propose the following adaptation of Equation (4.2)

$$\mathcal{V}(M) := \frac{1}{4n\left[g\left(M;\hat{\boldsymbol{\theta}}\right)\right]^2}, \tag{4.5}$$

where g is a pre-specified density and $\hat{\boldsymbol{\theta}}$ is a vector of parameter estimates for g where the estimates arise from the limited information in the reported median and interquartile range (or range).

Remark 1. *The choice of g does not need to be similar to the true underlying distribution. Instead, it only need be close to the density evaluated at the median. It turns out that there are excellent choices for g that approximate many unimodal densities evaluated at the median; that is, for appropriately chosen $\boldsymbol{\theta}$, $g(\nu;\boldsymbol{\theta}) \approx f(\nu)$ for many unimodal densities, f.*

We now derive each of these parameter sets, for the normal, log-normal, exponential, and Cauchy distributions, before comparing the estimators derived in Section 4.4.2.

4.4.1 Approximating the variance of the median from limited information

Given that the true population density f of Equation (4.2) is unknown, we propose replacing f with a nominated density g whose parameters are estimated, $\hat{\boldsymbol{\theta}}$, from the information available, evaluated at the sample median, M. In doing so we obtain our approximated variance in (4.5) by choosing a suitable $g\left(M;\hat{\boldsymbol{\theta}}\right)$.

Using the normal distribution

For the normal density with parameters μ and σ, the quantile function is $G^{-1}(p) = \mu + \sigma\Phi^{-1}(p)$ where Φ is the standard normal cumulative distribution function. Using the symmetry of Φ and assuming the interquartile range has been reported, we know that the true interquartile range is given by $2\sigma\Phi^{-1}(0.25)$. Thus we have estimators $\hat{\mu} := M$ and

$$\hat{\sigma}^{(1)} := \frac{\text{IQR}}{2\Phi^{-1}(0.25)}.$$

If the sample range is reported, as was the case with one study in the motivating example provided in Table 4.1 then we need a different estimate of σ. For $x_{[i]}$ denoting the ith order statistic for a sample of size n, $x_{[i]}$ is an estimate to approximately the $n^{-1}(i-0.5)$th population quantile. In particular, the maximum, or nth order statistic $x_{[n]}$, is an estimate to approximately

the $[(n-0.5)/n]$th population quantile. Thus, by a similar argument, we have

$$\hat{\sigma}^{(2)} := \frac{x_{[n]} - x_{[1]}}{2\Phi^{-1}\left[(n-0.5)/n\right]},$$

where $x_{[n]} - x_{[1]}$ is simply the reported range.

From above, if we choose the normal density for g, then $\hat{\boldsymbol{\theta}} = [\hat{\mu}, \hat{\sigma}^{(i)}]$ $(i = 1, 2)$ depending on whether the interquartile range or range is reported.

Using the log-normal distribution

If we were to choose the log-normal density with parameters μ and σ, then since the true median of a log-normal density is given by e^{μ}, we have an estimator for μ, given by

$$\hat{\mu} := \log(M).$$

We obtain our estimator for σ similarly when the interquartile range is reported. We know that the true interquartile range of the log-normal density is given by $G^{-1}\left(\frac{3}{4}\right) - G^{-1}\left(\frac{1}{4}\right)$ where G is the cumulative distribution function for the log-normal density. We have the associated quantile function

$$G^{-1}(p; \mu, \sigma) = \exp(\sigma\Phi^{-1}(p) + \mu).$$

Using this information, along with the symmetry of Φ and our estimate $\hat{\mu}$, we have

$$\hat{\sigma}^{(1)} := \frac{1}{\Phi^{-1}\left(\frac{3}{4}\right)} \log\left(\frac{\mathrm{IQR}e^{-\hat{\mu}} \pm \sqrt{\mathrm{IQR}^2 e^{-2\hat{\mu}} + 4}}{2}\right).$$

By similar argument to the derivations for the normal density in Section 4.4.1, if the range is reported then our estimate to σ is

$$\hat{\sigma}^{(2)} := \frac{1}{\Phi^{-1}\left(\frac{n-\frac{1}{2}}{n}\right)} \log\left[\frac{(x_{[n]} - x_{[1]})e^{-\hat{\mu}} \pm \sqrt{(x_{[n]} - x_{[1]})^2 e^{-2\hat{\mu}} + 4}}{2}\right].$$

Again, if we choose the log-normal density for g, then $\hat{\boldsymbol{\theta}} = [\hat{\mu}, \hat{\sigma}^{(i)}]$ $(i = 1, 2)$ depending on whether the interquartile range or range is reported.

Using the exponential distribution

For the exponential density, we need only to estimate the rate parameter λ. Since the true median of the exponential density is given by $\log(2)/\lambda$, we can estimate $\hat{\lambda} := \log(2)/M$. Here, $\hat{\boldsymbol{\theta}}$ takes the single parameter estimate $\hat{\lambda}$.

Using the Cauchy distribution

We need to estimate two parameters for the Cauchy density: a location parameter η and a scale parameter θ. From the quantile function for the Cauchy distribution $G^{-1}(p) = \eta + \theta \tan[\pi(p - 0.5)]$, we know that the true median is the location parameter η and that the interquartile range is equal to 2θ. Hence, we can estimate $\hat{\eta} := M$ and if the interquartile range is reported $\hat{\theta}^{(1)} := \text{IQR}/2$. If the range is reported then similar to previous arguments, we can estimate

$$\hat{\theta}^{(2)} = \frac{x_{[n]} - x_{[1]}}{2 \tan\left[\pi\left(\frac{n-0.5}{n} - \frac{1}{2}\right)\right]}.$$

When using the Cauchy, we then have $\hat{\boldsymbol{\theta}} = [\hat{\eta}, \hat{\theta}^{(i)}]$ $(i = 1, 2)$ depending on which range is reported.

4.4.2 Comparison between the four choices of g

To choose a distribution to inform g in Equation (4.5), we compared various choices of g for Equation (4.5). For the true median, ν, and true distribution f, we calculated the ratio of the approximated variance, given by Equation (4.5) with Equation (4.2) the true median ν evaluated by the true distribution f,

$$\rho := \frac{\mathcal{V}(\nu; g)}{\widehat{\text{Var}}(\nu; f)} \qquad \frac{(4.5)}{(4.2)}.$$

This was calculated for various distributions, visualised in Figure 4.2. The ratios for a choice of exponential for g in $\mathcal{V}$ were omitted as ρ took values greater than 3, in some cases, performing far worse than the other choices for g explored.

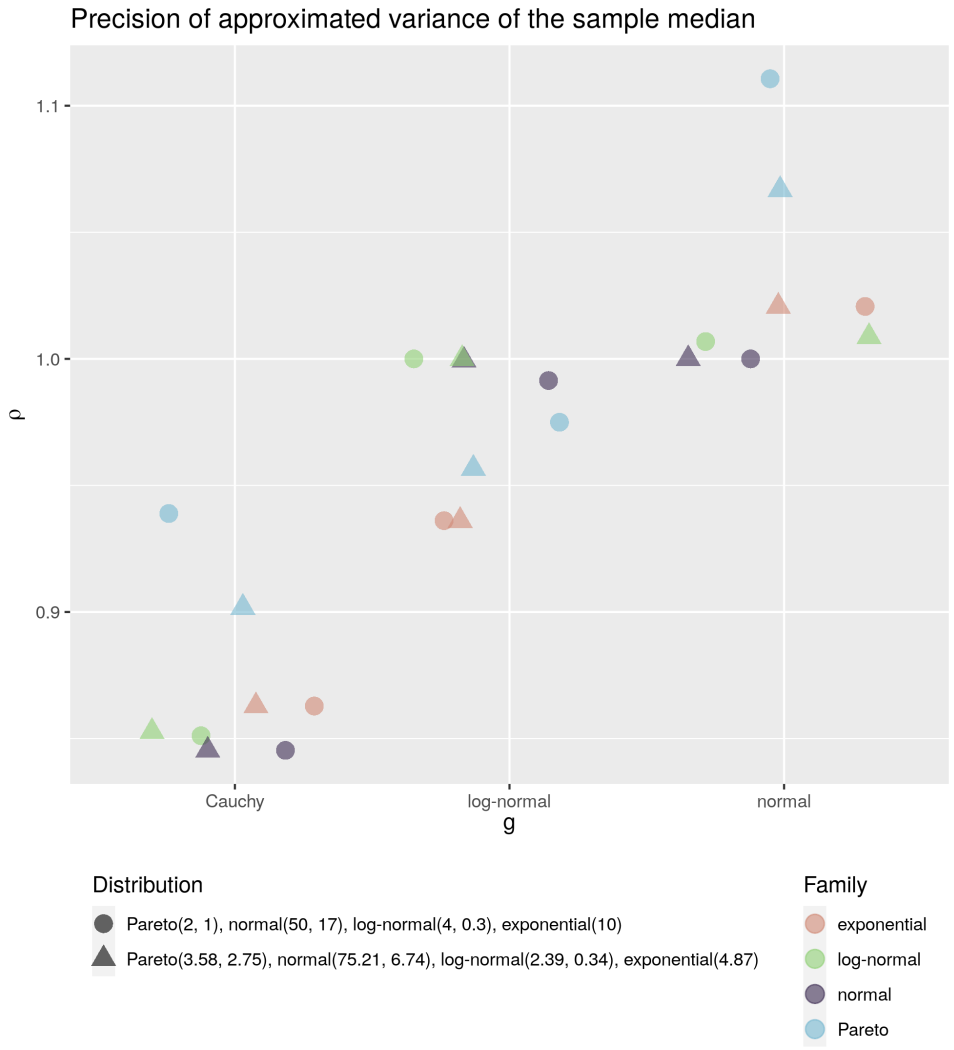


Figure 4.2: Each point represents a precision measure of approximated variance of the sample median: the ratio, $\rho := \mathcal{V}(\nu; g)/\widehat{\mathrm{Var}}(\nu; f)$, of approximated variances of the sample median. This is calculated for various distributions, indicated by colour and shape. The numerator, $\mathcal{V}(\nu; g)$, evaluates the median, ν, with the density, g, indicated by the x axis, into Equation (4.5), and approximates the parameters of g, as described Section 4.4. The denominator, $\widehat{\mathrm{Var}}(\nu; f)$, evaluates the true density, f, at the median, ν, with true parameters, as given by Equation (4.2). The parameters in the distribution labels have been rounded. In this horizontally-jittered plot, a small amount of horizontal random displacement is applied, so that points with the same value of ρ are easily discerned. Colours and shapes have been applied to facilitate between and within distributional family comparisons.

Using a Cauchy substitution as a choice of g underestimates $\widehat{\text{Var}}$, and the normal density overestimates. The log-normal sits between these results with most $\mathcal{V}$ within (0.95, 1) of $\widehat{\text{Var}}$; for this reason, the log-normal is chosen, and the variance of the sample median may be defined, adapting Equation (4.2) using the log-normal, so that $\mathcal{V}(M; \text{log-normal})$ provides the best estimator for an estimate for the approximation of the variance of the sample median.

Definition 1 (An estimator for the variance of the sample median)**.** We define an estimator $\mathcal{V}$ for the variance of the sample median M in terms of M, the quartiles reported, S, and sample size n,

$$\mathcal{V}(M, S, n) := \begin{cases} \frac{1}{\Phi^{-1}\left(\frac{3}{4}\right)} \log\left[\frac{\text{IQR} e^{-\hat{\log}(M)} \pm \sqrt{\text{IQR}^2 e^{-2\hat{\log}(M)} + 4}}{2}\right] & S = \text{IQR}, \\ \frac{1}{\Phi^{-1}\left(\frac{n-\frac{1}{2}}{n}\right)} \log\left[\frac{(x_{[n]} - x_{[1]}) e^{-\hat{\log}(M)} \pm \sqrt{(x_{[n]} - x_{[1]})^2 e^{-2\hat{\log}(M)} + 4}}{2}\right] & S = \{x_{[1]}, x_{[n]}\}, \end{cases}$$

where IQR denotes the interquartile range, and $\{x_{[1]}, x_{[n]}\}$ denotes the minimum and maximum values, the range, reported as an interval or a difference.

4.5 Performance of estimator in coverage probability simulations

Now that we have defined an estimator for meta-analysing medians, we explore the efficacy of this estimator under simulation, for different numbers of studies, distributions, and different assumptions about variation between studies and efficacy of intervention.

4.5.1 Coverage probability simulation

The approach we adopt for exploring the efficacy of a statistical estimator is to simulate **coverage probabilities**. In a coverage probability simulation, each trial randomly generates data from known parameters, calculates estimates of interest, and then produces a confidence interval based on that estimator. The trial is recorded as successful if the true parameter of interest falls within the confidence interval.

Enumerating these steps provides an algorithm for performing a coverage probability **trial**, one instance of a simulation.

1. Draw a random sample from the distribution that is characterised by the parameter of

interest, ν;

2. Calculate summary statistics from the random sample;
3. Calculate an estimate of ν from the summary statistics;
4. Construct a confidence interval using the parameter estimate;
5. Check if ν falls within the confidence interval.

In particular, we calculate a $(1-\alpha)\times 100\%$ confidence interval $\hat{\nu} \pm z_{1-\alpha/2}\sqrt{\hat{\gamma}}$, where $z_a = \Phi^{-1}(a)$ is the ath quantile of the standard normal distribution, and Φ is the standard normal distribution function

We intend **simulation** to be understood as the results of all trials. The **coverage** of the simulation is the proportion of trials for which the confidence interval contains the true parameter value.

The derivation of simulation meta-parameters and details of computational implementation are provided in Chapter 5.

4.5.2 Simulation results

Figures 4.3 and 4.4 summarise the results of simulations comprising 1000 trials, for different values of study heterogeneity, equal or unequal ratios between intervention and control, with data sampled from symmetric and asymmetric distributions.

Figure 4.3 shows desirable coverage for the estimator for several symmetric and asymmetric distributions. As the number of studies increase, the coverage increases. The coverage is comparable for no study heterogeneity, up to 0.4 study heterogeneity. Not surprisingly, the confidence intervals for log-normal estimates are outliers, as they are measured on a different scale. Mean confidence interval is reasonably consistent for all meta-parameters. In conclusion, we believe these simulation results demonstrate this is an effective estimator for the variance of the sample median for meta-analysis.

4.6 Meta-analysis of medians

For our motivating problem, meta-analysis of medians, we have followed a **toolchain walk-through** [38] for computationally developing a statistical estimator. This process took us from

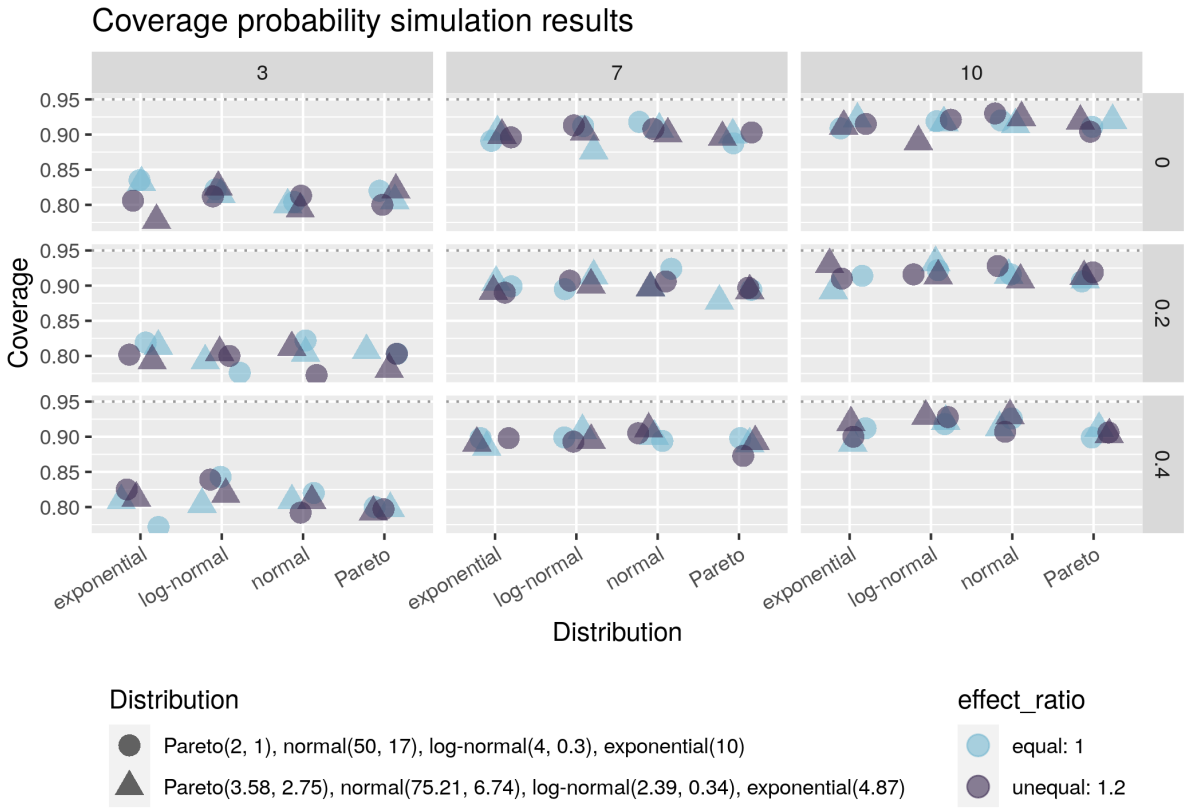


Figure 4.3: Simulation results presented as a coverage probability plot.

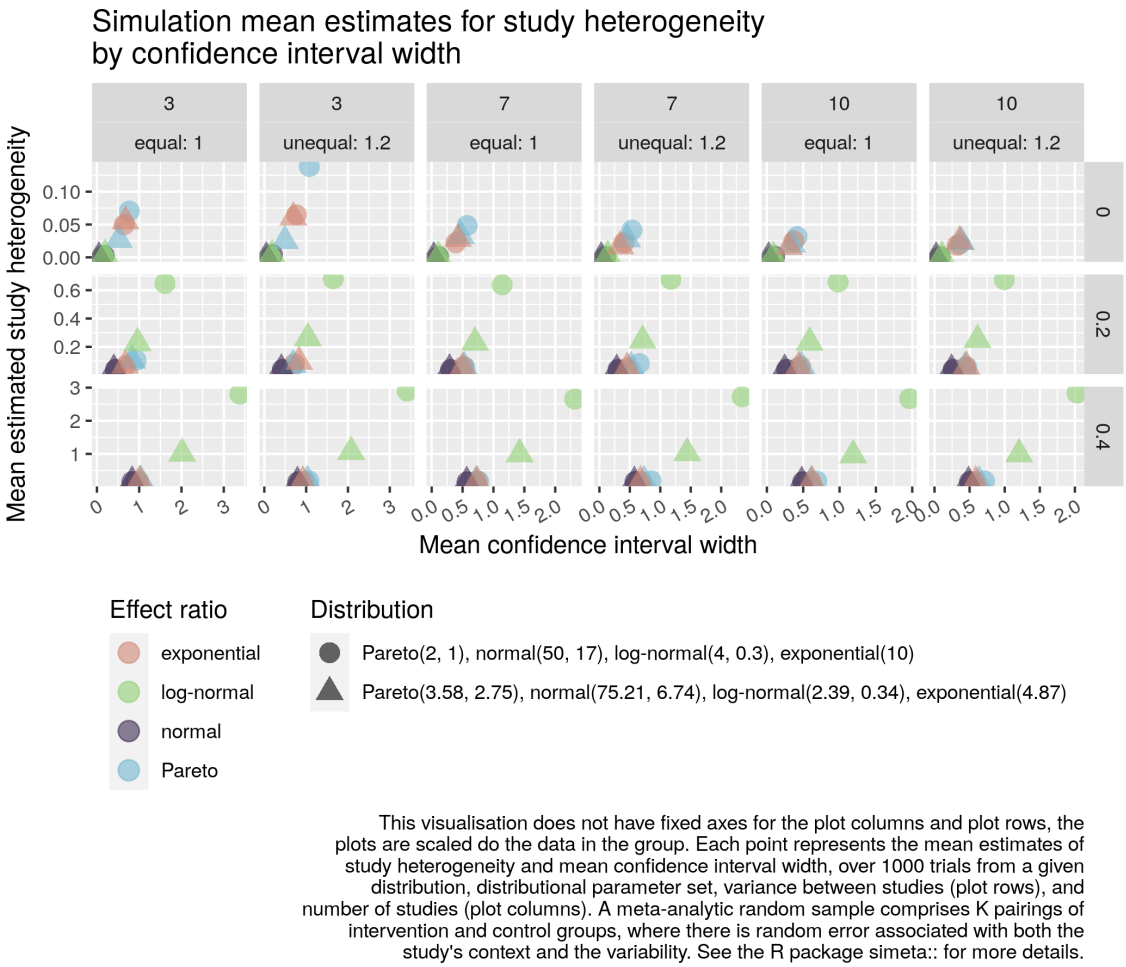


Figure 4.4: Bias of estimator by confidence interval width.

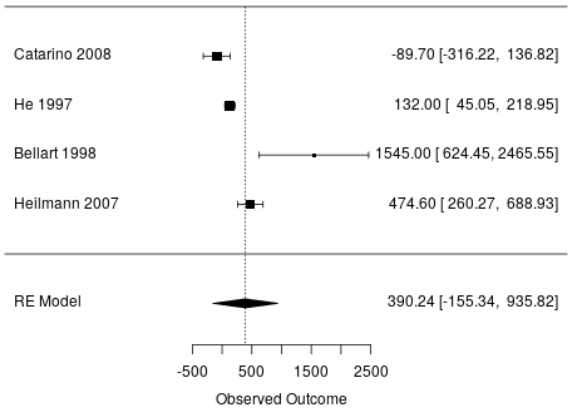


Figure 4.5: Using the estimator for the variance of the sample median, we can meta-analyse the median difference in the studies omitted (Table 4.1) from Pinheiro *et al.*'s meta-analysis of means, replicated in Figure 4.1

mathematical derivations, to estimator functions provided by **research compendia** [39] including analysis software `varameta::` and **coverage probability** simulations provided by the package `simeta::` to explore the efficacy of this estimator under different sampling conditions.

Through a case study, this manuscript raises the question of the value of examining research software engineering methodology. Beyond exploring the efficacy of an estimator for use in meta-analysis of medians, this dissertation has extrinsic value in the context of rapidly evolving statistical tools for simulation and analysis; structure of the analysis, the workflow, is of research merit in its own right [34]. We begin by revisiting our motivating example, and turn to meta-research observations from this analysis.

4.6.1 Revisiting the motivating example

With a method for incorporating medians, we revisit the motivating example discussed in Section 4.2. We can now compare the results of Figure 4.5's meta-analysis of means with meta-analysis of studies that reported means, shown in Figure 4.5.

This approach could potentially be extended to meta-regression of means and medians,

however this was not included as it seems inadvisable for asymmetric data.

4.6.2 Components of research for computational science

This manuscript derives mathematical estimators and presents simulation results of programmed instantiations of the algorithmic solution to the meta-analysis of medians. Rapid advances in the adoption software engineering provide an auxiliary context of meta-research; these are explored in companion manuscripts that describe the theoretical underpinnings of reproducible computing [39] and the practical steps in preparing this analysis as a reproducible research compendium [38]. Which is to say, meta-research questions from this project have generated more products of research than the question of meta-analysing medians itself.

This manuscript is one product of a research question. However, against the backdrop of technological revolutions in data collection and code sharing, contemporary researchers face a constant challenge of upskilling in computational tools, in addition to the challenges of the discipline in which the researcher is working. This manuscript used two packages, `varameta::` and `simeta::` to perform analysis; it is but one research component of the compendium of research to assess estimators for meta-analysis of medians. This splintered approach has the advantage that each component can exist for a different utility, individually, but together form a compendium of research.

We next derive the calculations and code required to produce Figures 4.3 and 4.4, before extending on meta-research questions.

Chapter 5

The `simeta::` Package

Extensible meta-analysis simulation

5.1 Basic usage

The `simeta::` package is an extensible set of tools for producing the simulations provided shown in Chapter 4, Figures 4.3 and 4.4. These tools are modularised, so that solutions are readily extractable and code accessible for extension and adaptation. It is this structure we explore in this chapter.

Install `simeta::` from GitHub.

```
# install simeta from gh
devtools::install_github("softloud/simeta")
```

We now load the package and run `::metasims` to computationally produce the simulation results presented in Chapter 4. The summary results of these simulations is provided by `::coverage_plot`, shown in Figure 4.3, providing a scatterplot of **coverage probability**, the number of trials in which the confidence interval produced by meta-analysis on randomly-generated data contain the true value. These are calculated for a variety of distributions, number of studies, ratio of true effects, and variation between studies. The code for this is shown in Figure 5.1. In this section, we will explore the code and derivations underpinning the code provided in the wrapper function `::metasims`.

In an effort to achieve some degree of computational reproducibility described in the opening

```
# packages
library(simeta)

## Loading required package: actuar
##
## Attaching package: 'actuar'
## The following object is masked from 'package:grDevices':
##
##     cm

# so these results are reproducible
set.seed(40)

# create coverage figure
cov_plot <-
  sims %>% # produced by default ::metasims
  coverage_plot()

# save figure
ggsave("coverage.png", plot = cov_plot)

## Saving 7 x 6 in image

# create bias figure
bias_plot <-
  sims %>%
  variance_plot()

# save bias figure
ggsave("bias.png", plot = bias_plot)

## Saving 7 x 6 in image
```

Figure 5.1: This code generates Figures 4.3 and 4.4 in Chapter 4.

chapters, this chapter is a reproducible .Rnw document that comprises code chunks and text that produces the exact same results when recompiled.

We now describe the motivation for this software, and the modularised elements, along with the underpinning mathematical derivations.

5.2 Motivation

This software was developed to support the analysis of the estimator for the variance of the sample median explored in Chapter 4. As the analysis developed, however, opportunities for developing the code to answer other questions presented themselves. Thus this software has been created with a particular focus on the extensibility [95], that is, how well code provided for a scientific analysis can be extended and adapted for other scientific questions. In this chapter, we explore the modularised components that produce the simulation results of Chapter 4.

5.3 Overview of codeflow

The codeflow underpinning `simeta::` comprises several modular components for simulation of meta-analysis data, from sample size generation, to meta-analysis, and summary visualisations. The code is modularised such that each piece of code performs one specific task. This section describes the way these components fit together. This is particularly important for the extensibility of this code, so that other researcher-developers can make use of any of the solutions provided in this collection of scripts. As a doctoral thesis, this is the work of an apprentice scientist. Arguably, demonstrating extensibility is now a core component of any computational doctoral work, as this is now considered best practice in scientific programming [95]. In addition, and not least, is the debugging benefit of testing to see if each code component performs the task assigned. The core function of `simeta::` is `::metasims`, which is the bridge between the back end (that is, not intended for the end user) of the code, and the reporting functions.

This is a tidy-structured algorithm [42], wherein each row denotes a simulation, and each column denotes a simulation meta-parameter. `::metasims` begins by using `::sim_df` (Table 5.1) to produce a table of simulation metaparameters, differentiated from parameters applied in each simulation. Based on user-chosen inputs, or, if the user defines no inputs, on the default metaparameters, `::metasims` constructs a table of simulation metaparameters, wherein each row denotes a simulation. And where by **simulation**, we mean randomly generating data and performing statistical analyses thereon, reproducing this `trials` (a user-specified input) times.

In Section 5.4, the derivations and codeflow for generating sample sizes are provided, and for sampling data in Section 5.5.

`::sim_df` generates sample sizes for each study, and allocates a proportion of the total sample sizes for the intervention group. Section 5.4 provides a toolchain walkthrough and derivation of how the sample sizes are arrived at.

`::metatrial` generates a sample meta-analaysis dataset, produces meta-analysis results and checks to see if the true log-ratio of intervention and control median fall within the estimated confidence interval for the log-ratio of medians. Section 5.5 provides a toolchain walkthrough and derivation of how the meta-analysis samples are generated.

With each row of the simulation metaparameters produced by `::sim_df`, a simulation of the specified number of trials (1000 trials in Figure 4.3), with `::metasim`, shown in Table 5.2, which runs `::metatrial` for the number of specified `trials` and summarises what proportion of trials were successful, that is, whose confidence interavals contained the true log-ratio of medians.

Summary statistics are produced to inform on the results and reported back in a table. This table is then appended with the simulation results, shown in Table 5.3, and visualised using `::coverage_plot`, as shown in Figure 4.3.

5.4 Simulating meta-analysis sample sizes

Meta-analytic samples vary in different ways. There are the number of studies, K, and variation τ^2 between them. There are small cohorts and large cohorts. And the intervention groups vary in proportion of total sample size. In this section, we focus on sample size generation for meta-analysis studies with control and intervention cohorts.

5.4.1 Codeflow

The `simeta::` package provides a means of producing a randomly-generated dataset of sample sizes. The `::sim_n` provides a method of producing a dataset of meta-analysis sample sizes, based on a user's expectations of minimum, maximum, and proportion of intervention cohort from the total in the control and intervention groups.

The software defaults to 3 studies, a minimum sample size of 20, a maximum sample size of 200. As a default, equal intervention and control groups are expected, with some allowance made for small variations, for, say, if a few people drop out of a study. Using `::fn_fmls` from

Table 5.1: Simulation metaparameters, where each row is a simulation, and each column is a parameter option for that column-header's variable. The first column, k, denotes the number of studies and τ^2 the variation between studies. The distribution and distributional parameters are provided in the next two columns, a simulation id, and finally, the true value of interest, in this case, the median. For example, note that where the mean of the normal distribution is 2, the median is 2, as expected. The parameters with one decimal place were fixed and the others were randomly sampled.

```
sim_df() %>%
  # select first 30 rows
  head(15) %>%
  # sample size dataframe variable omitted for brevity
  select(-n, -sim_id) %>%
  # format for pdf
  simeta_table_tex(
    col.names = cnames_simdf,
    escape = FALSE
  )
```

k	τ^2	ρ	distribution	parameters
3	0	1	pareto	2, 1
3	0	1	norm	50, 17
3	0	1	lnorm	4.0, 0.3
3	0	1	exp	10
3	0	1	pareto	3.576119, 2.745808
3	0	1	norm	75.209383, 6.739041
3	0	1	lnorm	2.3900182, 0.3383603
3	0	1	exp	4.86717
7	0	1	pareto	2, 1
7	0	1	norm	50, 17
7	0	1	lnorm	4.0, 0.3
7	0	1	exp	10
7	0	1	pareto	3.576119, 2.745808
7	0	1	norm	75.209383, 6.739041
7	0	1	lnorm	2.3900182, 0.3383603

Table 5.2: This simulation function takes the arguments of the rows of the simulation metaparameter dataframe produced by ::sim_df within the wrapper metasimulation function ::metasims.

```
metasim(
  tau_sq = 0.2,
  effect_ratio = 1.1,
  rdist = "lnorm",
  par = list(1,2),
  trials = 100
) %>%
  simeta_table_tex(
    col.names = cnames_metasim,
    escape = FALSE
  )
```

τ^2	width	bias	coverage	successful	id
0.15	1.12	0.24	0.82	100	simulation1

rlang::, we can extract the default values and display in Table 5.4.

In keeping with tidy data principles [42], the dataset is produced with one observation, that is, sample size, per row. However, meta-analyses conducted in R frequently make use of the metafor:: package, and a wide format, shown in Table 5.5, where each row is a study with intervention and control listed in columns, may be easier to work with for this context.

Smaller cohorts may be assumed. Here we will generate a five-study dataset of sample sizes that are small cohorts, say, less than 30, and an expected cohort of ten per cent for the intervention group. In Table 5.6, small data are generated in long form.

```
beta_par(
  proportion = 0.7,
  error = 0.2
)

## $alpha
## [1] 36.05
##
```

Table 5.3: Simulation results, by simulation metaparameters. Each row is a simulation, wherein the columns denote metaparameters (distribution, variation between studies, etc.), and simulation results associated with those metaparameters.

```
sims %>%
  # choose simulation results table
  pluck("results") %>%
  head(20) %>%
  # sample size dataframe variable omitted for brevity
  select(-n,-sim_results) %>%
  simeta_table_tex(
    col.names = sims_results_cols,
    escape = FALSE)
```

$\hat{\tau}^2$	ci width	bias	coverage	success	id	k	τ^2	ρ	dist'n	parameters
0.07	0.77	0.18	0.82	1000	sim 1	3	0	1	pareto	2, 1
0.00	0.17	0.04	0.80	1000	sim 2	3	0	1	norm	50, 17
0.00	0.14	0.03	0.82	1000	sim 3	3	0	1	lnorm	4.0, 0.3
0.05	0.65	0.14	0.84	1000	sim 4	3	0	1	exp	10
0.03	0.52	0.12	0.81	1000	sim 5	3	0	1	pareto	3.576119, 2.745808
0.00	0.05	0.01	0.80	1000	sim 6	3	0	1	norm	75.209383, 6.739041
0.00	0.19	0.04	0.82	1000	sim 7	3	0	1	lnorm	2.3900182, 0.3383603
0.05	0.68	0.15	0.83	998	sim 8	3	0	1	exp	4.86717
0.05	0.57	0.13	0.89	996	sim 9	7	0	1	pareto	2, 1
0.00	0.11	0.02	0.92	1000	sim 10	7	0	1	norm	50, 17
0.00	0.11	0.02	0.91	1000	sim 11	7	0	1	lnorm	4.0, 0.3
0.02	0.39	0.08	0.89	1000	sim 12	7	0	1	exp	10
0.03	0.47	0.10	0.90	996	sim 13	7	0	1	pareto	3.576119, 2.745808
0.00	0.03	0.01	0.91	1000	sim 14	7	0	1	norm	75.209383, 6.739041
0.00	0.11	0.02	0.88	1000	sim 15	7	0	1	lnorm	2.3900182, 0.3383603
0.03	0.43	0.09	0.90	1000	sim 16	7	0	1	exp	4.86717
0.03	0.42	0.08	0.91	1000	sim 17	10	0	1	pareto	2, 1
0.00	0.11	0.02	0.92	1000	sim 18	10	0	1	norm	50, 17
0.00	0.08	0.02	0.92	1000	sim 19	10	0	1	lnorm	4.0, 0.3
0.02	0.37	0.08	0.91	998	sim 20	10	0	1	exp	10

Table 5.4: The arguments of ::sim_n and their defaults. These are the values a user can change: the number of studies, k, is set to 3; the minimum sample size min_n is set to 20; the maximum sample size defaults to 200; the proportion of total sample size allocated to the intervention group is assumed to be 0.5 with 90 per cent of sample sizes falling within 0.1. The wide function toggles whether the control and intervention arms are columns or in separate rows.

```
# extract the arguments of sim_n
(fn_fmls(sim_n)) %>%
  # converts to list
  map(1) %>% {
    # construct a table of args
    tibble(argument = names(.),
           default = .)
  } %>%
  mutate(# convert list variable to string
    default = map_chr(default, toString)) %>%
  # display as table
  simeta_table_tex()
```

argument	default
k	3
min_n	20
max_n	200
prop	0.5
prop_error	0.1
wide	FALSE

Table 5.5: Wide format of simulated meta-analysis sample sizes.

```
sim_n(wide = TRUE) %>%
  simeta_table_tex()
```

study	intervention	control
Vëantur_1982	33	34
Déagol_1998	39	33
Amarië_1960	75	75

Table 5.6: Example of a randomly-generated small-cohort dataset.

```r
# generate table
sim_n(
  k = 5,
  min_n = 4,
  max_n = 30,
  prop = 0.1,
  prop_error = 0.01,
  wide = FALSE
) %>%
  simeta_table_tex()
```

study	group	n
rimë_1977	control	9
Lóni_1954	control	12
Gwindor_1969	control	24
Hundad_2013	control	26
Nob_1999	control	11
rimë_1977	intervention	1
Lóni_1954	intervention	1
Gwindor_1969	intervention	3
Hundad_2013	intervention	3
Nob_1999	intervention	1

```
zeta_plot(0.3,0.1)
```

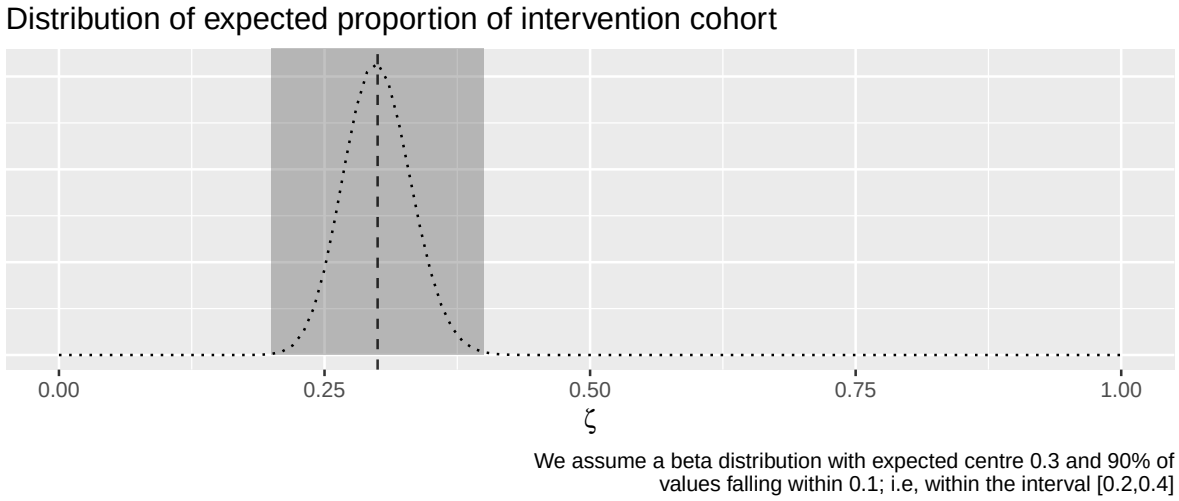


Figure 5.2: Plot of the distribution of proportion allocated to intervention from a total sample size comprising control and intervention groups.

```
## $beta
## [1] 15.45
```

Equipped with code the provides the beta parameters, we the call this, in a function that generates an intervention proportion based on sample sized, expected value and standard deviance of the proportion.

```
intervention_proportion(
  n = 5,
  proportion = 0.1,
  error = 0.005
)

## [1] 0.09965155 0.09926290 0.09902813 0.10151700 0.10059654
```

::zeta_plot, shown in Figure 5.2, shows the distribution of the proportion allocated to intervention group.

The next section provides the derivation that underlie this code.

5.4.2 Derivations for generating sample sizes

Total sample sizes, of control and treatment groups are generated from user-defined parameters: minimum, a, and maximum, b; and k, the number of studies. k total sample sizes, N_k, are randomly sampled from uniform(a, b). For each N_k, a proportion is allocated to the intervention group, and the rest to control. The user specifies the expected proportion of the intervention cohort, ζ, and the error associated, expressed in terms of an expection of 90 per cent of proportions falling within θ. ζ is sampled from a beta distribution, with the distributional parameters α and β derived as follows.

We assume the proportion ζ follows a beta distribution, that is, $\zeta \sim \text{beta}(\alpha, \beta)$, with expected value, $E(\zeta) = \tilde{\zeta}$, and that ninety per cent of ζ falling within τ of ζ. Then, Chebyshev's inequality [1] provides

$$P(|\zeta - \tilde{\zeta}| \geqslant \tau) \leqslant 0.1$$

where $\tau = k\sigma$, and σ denotes the standard deviation of ζ, and $k > 0$ [4]. Furthermore, the righthand side of the inequality is given in terms of k, so we have $k^{-2} = 0.1$. This gives

$$\tilde{\tau} = \sqrt{10}\sigma.$$

We now apply these assumptions to the definitions of the mean and variance of the beta distribution to obtain the parameters required to randomly sample the proportion of the total sample size given to the intervention cohort.

Since $\zeta \sim \text{beta}(\alpha, \beta)$, we have

$$\tilde{\zeta} = \frac{\alpha}{\alpha + \beta} \implies \beta = \alpha/\tilde{\zeta} - \alpha.$$

and, to find α, we combine this with the variance,

$$\begin{aligned}
& \frac{\tilde{\tau}^2}{10} &=& \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} && \text{as } \sigma = \tilde{\tau}/\sqrt{10} \\
\implies & \frac{\tilde{\tau}^2}{10} &=& \frac{\alpha(\alpha/\tilde{\zeta}-\alpha)}{(\alpha+\alpha/\tilde{\zeta}-\alpha)^2(\alpha+\alpha/\tilde{\zeta}-\alpha+1)} && \text{as } \beta = \alpha/\tilde{\zeta} - \alpha \\
\implies & \frac{\tilde{\tau}^2}{10} &=& \frac{(1/\tilde{\zeta}-1)\zeta^2}{\alpha/\tilde{\zeta}+1} \\
\implies & \alpha/\zeta + 1 &=& 10\tilde{\zeta}^2/\tilde{\tau}^2(1/\tilde{\zeta} - 1) \\
\implies & \alpha &=& \tilde{\zeta}[10\tilde{\zeta}^2/\tau^2[1/\tilde{\zeta} - 1] - 1] \quad ,
\end{aligned}$$

[1] This calculation relies on the questionable assumption that the inequality can be simplified to an equality. Through sharing this work open source an improved solution was created for specific software for this calculation, parameterpal::(https://softloud.github.io/parameterpal/).

if we assume we may divide the random effect for the study γ_k equally between both arms.

And this is in a form that is relatively easy to program. Leaving it in this format, meant that upon returning to code, it was relatively straightforward to rederive the result for transcription into the manuscript, from this piece of code.

```
beta_par(
  proportion = 0.4,
  error = 0.05
)

## $alpha
## [1] 383.6
##
## $beta
## [1] 575.4
```

5.5 Simulating meta-analysis data

This section mirrors the structure of Section 5.4, in first outlining the codeflow, that is, the toolchain walkthrough, of simulating meta-analysis data, and then delving into the theory underpinning these simulations.

5.5.1 Codeflow

Once we have sample sizes, we can randomly sample data from distributions specified by the user, currently implemented for the normal, exponential, Pareto, and log-normal distributions, and provide sample effect measures, with variance.

`::sim_stats` provides a means of randomly generating a meta-analysis dataset of effect, sample sizes, and effect spread, shown in Table 5.7. This piece of code extends directly from `::sim_n`, discussed in Section 5.7. For purely aesthetic reasons, we round the digits in Table 5.7 to two decimal places.

This function creates a random effect for each study in the dataset, and samples with `::sim_sample` from the distribution adjusted via that parameter. Here it is a sample of 18 values from a normal distribution with mean 20 and standard deviation 0.2. This study deviates

Table 5.7: Default output of ::sim_stats, assuming: an underlying normal distribution with mean 50 and standard deviation 0.2; variation of 0.4 between studies; and a true effect ratio, intervention over control, of 1.2.

```
sim_stats() %>%
  simeta_table_tex(digits = 2)
```

study	group	effect	effect_spread	n
Argonui_2000	control	39.43	0.31	75
Argonui_2000	intervention	76.07	0.28	70
Elfwine_2013	control	77.26	0.20	12
Elfwine_2013	intervention	38.78	0.37	10
Yavanna_1970	control	57.12	0.31	69
Yavanna_1970	intervention	52.53	0.25	61

by 0.2 from the overall true effect across all studies. In this case, a control (unadjusted) sample is provided.

```
sim_sample(
  n = 18,
  this_study_error = 0.2,
  rdist = "norm",
  par = list(mean = 20, sd = 0.2),
  control = TRUE
)

##  [1] 15.97560 16.11579 16.72661 16.10583 16.54135 16.45139 16.27503 16.36559
##  [9] 16.63843 16.15508 16.22943 16.29320 16.26190 16.32932 16.59718 16.35684
## [17] 16.23745 16.25456
```

To sample from the intervention group of a study, we set the control argument to FALSE, and provide a ratio of intervention true effect, effect_ratio, to control of 1.2, so that

$$\frac{\text{intervention true effect}}{\text{control true effect}} = 1.2.$$

```
sim_sample(
  n = 18,
  this_study_error = 0.2,
  rdist = "norm",
  par = list(mean = 20, sd = 0.2),
  control = FALSE,
  effect_ratio = 1.2
)

##  [1] 29.26291 29.33009 29.41119 29.18646 29.19458 29.31168 29.35614 29.40245
##  [9] 29.18717 29.54784 29.34734 29.44249 29.45987 29.74727 29.71541 29.51767
## [17] 28.91137 29.23608
```

For each study, these samples are summarised in quartiles. The `this_study_error` is sampled as a metaparemeter set in `::sim_stats`, Table 5.7 which generates each study's error by sampling from a normal distribution with variance τ^2.

`::sim_stats` calls `::sim_sample`, which produces a sample for each arm of each study, and returns the `effect` of interest, the sample median, and the `effect_spread`. So that each number in the corresponding columns of Table 5.7.

```
sim_sample(
  n = 5,
  this_study_error = 0.4,
  rdist = "pareto",
  par = list(2, 3),
  control = FALSE,
  effect_ratio = 1.4
)

## [1] 0.1026971 4.2486377 0.8726473 1.4324320 1.0580498
```

For four distributions, log-normal, normal, Pareto, and exponential, data can be sampled with the parameters for the sampling derived in the following section.

5.5.2 Derivations for meta-analysis data generation

We wish to generate data that mimic the structure of meta-analysis observations. In this section, we extend in detail on the method described briefly in Section 3.3.3.

Assumption 1. We assume the log-ratio of sample medians $\log(m_k^I/m_k^C)$ for the kth study can be expressed in terms of the log-ratio of population median, $\log(\nu_k^I/\nu_k^C)$, with normally distributed variation associated with the kth study, and sampling error. So,

$$\log(m_k^I/m_k^C) = \log(\nu^I/\nu^C) + \gamma_k + \varepsilon_k \tag{5.1}$$

where $\gamma_k \sim \text{normal}(0, \tau^2)$ denotes variation associated with the kth study and $\varepsilon_k \sim \text{normal}(0, \sigma^2)$ denotes the sampling error.

We take as known the ratio, ρ, of intervention, ν_I, and control median, ν_C. And thus we have $\nu_I = \rho/\nu_C$.

The first derivation is a restatement of the result provided in Chapter 3, which is then expanded to other distributions of interest.

For simplicity, and arguably a limitation of this analysis, we shall assume one parameter is dependent on the study, and all other parameters are fixed across all studies. We take the parameter for the control group and other parameters as known.

Another limitation is the following assumption, made for purely technical reasons, to make the sampling code easier to write, and the derivations simpler.

Assumption 2. The random effect γ_k may be divided evenly between both arms.

Assuming an exponential distribution

If we wish to sample $x_1, \ldots, x_n \sim \text{exponential}(\lambda_k^J)$, for $k \in K$ studies and $J \in \{C, I\}$ arms of each study, we must derive the rate parameter, λ_k^J. We take as given values, λ_C, and $\nu_I = \rho/\nu_C$, the ratio of the intervention and control median.

The median ν of an exponential distribution with rate parameter λ is $\log 2/\lambda$. So, we have

$$\nu_I = \rho\nu_C \implies \log 2/\lambda^I = \rho \log 2/\lambda_C \implies \lambda_I = \lambda_C/\rho. \tag{5.2}$$

Now, taking our assumption about the log-ratio of sample medians, we have

$$
\begin{aligned}
& \log(m_k^I/m_k^C) &&= \log(\nu^I/\nu^C) + \gamma_k && \text{Assumption 1} \\
\implies & \log((\log 2/\lambda_k^I)/(\log 2/\lambda_k^C)) &&= \log((\log 2/\lambda^I)/(\log 2/\lambda^C)) + \gamma_k/2 \cdot 2 && \\
& && \text{as } \nu_J = \log 2/\lambda^J \text{ for } J \in \{C, I\} && \\
\implies & \log \lambda_k^C - \log \lambda_k^I &&= \log \lambda^C - \log \lambda^I + \gamma_k/2 \cdot 2 && \\
\implies & \lambda_k^C &&= \lambda_C \exp(\gamma_k/2) && \text{and} \\
& \lambda_k^I &&= \lambda^I \exp(-\gamma_k/2) && v \\
\implies & \lambda_k^C &&= \lambda_C \exp(\gamma_k/2) && \text{and} \\
& \lambda_k^I &&= \lambda_C/\rho \exp(-\gamma_k/2). && (5.2)
\end{aligned}
$$

Assuming a log-normal distribution

If we wish to sample $x_1, \ldots, x_n \sim \text{log-normal}(\mu_k^J, \sigma^2)$, for $k \in K$ studies and $J \in \{C, I\}$ arms of each study, we must derive parameters μ_k^J and σ_k^J.

We fix three simulation metaparameters. The second parameter of the log-normal distribution, σ, is assumed to be the same across all studies, and arms, such that $\sigma_k^J = \sigma^J$. The first parameter for the control arm is fixed, so we take μ_k^C as known, and we make an assumption $\rho := \nu^I/\nu^C$ about the ratio of the true intervention median, ν^I, and control median, ν^C.

From Equation (1), we have,

$$\log(m_k^I/m_k^C) = \log(\nu^I/\nu^C) + \gamma_k,$$

which is in terms of the true medians of the control and arm distributions. Since we are assuming a log-normal distribution, we know the median is provided in terms of the first rate parameter, μ, such that $\nu = \exp(\mu)$. Thus to find μ^I in terms of known values μ^C and ρ, we have,

$$\rho = \nu^I/\nu^C \implies \nu^I = \rho\nu^C \implies \exp(\mu^I) = \exp(\mu^C)\rho \implies \mu^I = \mu^C + \log(\rho), \tag{5.3}$$

which we can use to find the parameters of interest, μ_k^J.

$$\begin{aligned}
\log(m_k^I/m_k^C) &= \log(\nu^I/\nu^C) + \gamma_k && \text{Assumption 1} \\
\log(\mu_k^I/\mu_k^C) &= \log(\mu^I/\mu^C) + \gamma_k && \text{as } m_k^J = \mu_k^J \text{ and } \nu^J = \mu^J \\
\implies \log\mu_k^I - \log\mu_k^C &= \log\mu^I - \log\mu^C + 2\gamma_k/2 \\
\implies \log\mu_k^I &= \log\mu^I + \gamma_k/2 \\
\log\mu_k^C &= \log\mu^C - \gamma_k/2 && \text{Assumption 2} \\
\implies \mu_k^I &= \mu^I\exp(\gamma_k/2) && \text{and} \\
\mu_k^C &= \mu^C\exp(-\gamma_k/2) \\
\implies \mu_k^I &= (\mu^C + \log\rho)\exp(\gamma_k/2) && (5.3), \text{ and} \\
\mu_k^C &= \mu^C\exp(-\gamma_k/2)
\end{aligned}$$

Assuming a normal distribution

If we wish to sample $x_1, \ldots, x_n \sim \text{normal}(\mu_k^J, \sigma^2)$, then we need to find μ_k^J for $J \in \{C, I\}$ arms of the study, and $k \in K$ number of studies. We begin with known values, σ, the same variance across all studies, μ^C the centre parameter of the control arm, and $\rho := \nu^I/\nu^C$, the ratio of the intervention median and the control median. The random effect associated with the kth study's error, $\gamma_k \sim \text{normal}(0, \tau^2)$, is sampled for a given value of the simulation metaparmeter, τ^2.

Now, the median of the normal distribution is the parameter of centrality, that is, for a normal distribution with measure of centrality, μ, the median is μ. Then,

$$\nu^I = \rho\nu^C \implies \mu^I = \rho\mu^C, \tag{5.4}$$

as $\nu^C = \mu^C$. Then,

$$\begin{aligned}
\log(m_k^I/m_k^C) &= \log(\nu^I/\nu^C) + \gamma_k && \text{Assumption 1} \\
\implies \log(\mu_k^I/\mu_k^C) &= \log(\mu^I/\mu^C) + 2\cdot\gamma_k/2 \\
\implies \log\mu_k^I - \log\mu_k^C &= \log\mu^I - \log\mu^C + 2\cdot\gamma_k/2 \\
\implies \log\mu_k^I &= \log\mu^I + \gamma_k/2 && \text{and} \\
\log\mu_k^C &= \log\mu^C + \gamma_k/2 && \text{Assumption 2} \\
\implies \mu_k^I &= \rho\mu^C\exp(\gamma_k/2) && (5.4), \text{ and} \\
\mu_k^C &= \mu^C\exp(-\gamma_k/2).
\end{aligned}$$

Assuming a Pareto distribution

If we assume a Pareto II distribution, such that, for $\alpha > 0$, $\sigma > 0$, and $x > 0$, we have

$$f(x) = \alpha/\sigma(1 + x/\sigma)^{-1(\alpha+1)}.$$

We write $x_1, \ldots x_n \sim \text{ParetoII}(\alpha, \sigma_{\text{k}}^{\text{J}})$, with shape parameter α, and scale parameter σ, for the kth study and $J \in \{C, I\}$.

We take as user-defined input, the shape parameter, α, and the control group's scale parameter, σ^C, and the ratio between intervention and control medians. We need to find the kth study's, Jth group's, scale parameter σ_k^J.

Since $x \sim \text{ParetoII}(\alpha, \sigma_k^J)$, we have the median ν^J for the arm $J \in \{C, I\}$,

$$\nu^J = \sigma^J(\sqrt[\alpha]{2} - 1), \tag{5.5}$$

So,

$$\rho = \nu^I/\nu^C \implies \nu^I = \rho\nu^C \implies \sigma^I(\sqrt[\alpha]{2} - 1) = \rho\sigma^C(\sqrt[\alpha]{2} - 1) \implies \sigma^I = \rho\sigma^C. \tag{5.6}$$

Then,

$$\begin{array}{rrcll}
 & \log(m_k^I/m_k^C) & = & \log(\nu^I/\nu^C) + \gamma_k & \text{Assumption 1} \\
 & \log\left(\frac{\sigma_k^I(\sqrt[\alpha]{2}-1)}{\sigma_k^C(\sqrt[\alpha]{2}-1)}\right) & = & \log\left(\frac{\sigma^I(\sqrt[\alpha]{2}-1)}{\sigma^C(\sqrt[\alpha]{2}-1)}\right) + \gamma_k/2 & (5.5) \\
\implies & \log\sigma_k^I - \log\sigma_k^C & = & \log\sigma^I - \log\sigma^C + \gamma_k/2 & \\
\implies & \log\sigma_k^I & = & \log\sigma^I + \gamma_k/2 & \text{and} \\
 & \log\sigma_k^C & = & \log\sigma^C - \gamma_k/2 & \text{Assumption 2} \\
\implies & \sigma_k^I & = & \sigma^I\exp(\gamma_k/2) & \text{and} \\
 & \sigma_k^C & = & \sigma^C\exp(-\gamma_k/2) & \\
\implies & \sigma_k^I & = & \rho\sigma^C\exp(\gamma_k/2) & (5.6), \text{ and} \\
 & \sigma_k^C & = & \sigma^C\exp(-\gamma_k/2). &
\end{array}$$

5.6 Coverage probability simulation

The `::metasims` function provides a means of performing `trials`, specified by user, number of simulations. Each simulation fixes a set of meta-analysis sample sizes as described in Section 5.5. For a chosen distribution, data are randomly sampled and summary statistics calculated.

A meta-analysis is conducted via `metafor::rma`, which yields a confidence interval for the effect and its variance. The **coverage probaiblity** is proportion of trials for which the true effect falls within the confidence interval.

Each row of the output coverage probability summary table represents a simulation, shown in Table 5.3. In this tidy format [42], where each row is a simulation's observed summary statistics, and each column represents a simulation metaparameter, the output lends itself well to `tidyverse::` visualisation.

In Figure 4.3 we show one way to summarise these data, faceted by variation between studies, number of studies.

5.7 Extensibility

A key feature of the structure of this collection of analysis code, and, indeed, this manuscript, is its modularity. Perhaps not all of the code is useful for another researcher, perhaps, for example, the sample size generation is useful for simulation analyses with more covariates of interest. Extensibility is now being recognised as an integral component of scientific computation [95]. Analogous to *scope* in replication, researchers can take a more active role in future usage their algorithms, by providing accessible, reproducible, and extensible code.

Fraser *et al.* discuss how replication relies on clearly defining the conditions under which it would be expected a subsequent experiment to yield the same results [32]. Thus, the onus of replication is not only on those who perform the secondary experiment, but also on the initial researchers to make the scope of the experiment explicit, to inform subsequent replication.

Similarly, in scientific computation, if we do not provide extensible, accessible code, we make it more difficult for future research to expand on and investigate our findings. Lengthy script files with hidden package dependencies can be a headache for those attempting to make use of our computational work. By providing transparent, reproducible code, in a modularised format, we facilitate extending on our scientific work. Beyond reproducibility, there is much to be gained from structuring and documenting analysis code.

Chapter 6

The Order of Mathematistry

Queering metascience with mathematics

Charles T. Gray, Hien Nguyen, Hannah Fraser, and Danielle Navarro

Abstract

Out of the set of reasonable pairings between methodologies and the scientific claims that they assess, what proportion of claims and methodology pairings are sufficiently strongly linked so as to provide a meaningful scientific answer to the claim? What is the scientific benefit in a methodological intervention, such as preregistration or reproducible computing, when the link between method and claim posed is weak? This manuscript constructs an order on the set of partitions of reasonable pairings of scientific claim and methodology to queer questions about psychological methodology, through a lens of mathematics. Within an order-theoretic framework, cardinality and density formalise scientific methodology so that we may discuss in which context specific methodological tools, such as preregistration or reproducible computing, provide meaningful evidence of the quality of scientific claims.

Keywords: Meta-research · Reproducibility · Mathematics

In the 1966 novel *Wide Sargasso Sea* [84], Jean Rhys retells Charlotte Brontë's *Jane Eyre* [13] from the point of view of the seemingly-mad woman in the attic [35]. By re-examining the narrative from the perspective of the woman silenced in Brontë's text, confronting questions are raised about the conventions of society. Rhys makes a compelling case for structural inequities and oppression, not madness, driving the character's actions; she *queers* the lens of the narrative from a white, patriarchal perspective, to the lived experience of a woman of the colonies in nineteenth-century British society. Rhys' text reveals the intersectionality of minoritised people, and we employ the term queer in this, rather than the explicit sexual orientation sense.

Phrases such as *science has shown* imply a certainty we are rarely afforded in the practice of science. Closer perhaps, to say *science provides evidence of*. The language that science uses to make these statements is statistics, and by its very nature, inference involves a degree of uncertainty. Although mathematics follows logic as no other science does, the application of mathematical calculataions are not necessarily so robust. Box's term **mathematistry** [12] can be adapted for model selection procedures to 'describe using formal tools to define a statistical problem that differs from the scientific one, solving the redefined problem, and declaring the scientific concern addressed' [72]. Arguably, all inferential statistics involves a degree of mathematistry, but the better applications seek to minimise the mathematistry of the calculations. In this manuscript, we shall use order theory to formalise a question from psychological science, posed in the manuscript, 'Is Preregistration Worthwhile?'. To this end, we construct an order of mathematistry.

By queering, instead of reinventing the wheel, we aim to further metascientific objectives, that is, to evolve and develop by learning from other disciplines. We confront questions about the limitations of conventional discipline-specific methodology, that so easily trap us in questionable practice [31]. One such example of evolution is the adoption of software engineering practices in research for scientific reproducibility[1]. The implicit politics and ethics of data science are poorly articulated [54]; metascientific queerings provide one means by which we may explore the intersectionality of computational science.

6.1 An order-theoretic approach to the question: 'Is Preregistration Worthwhile?'

A recent publication, 'Is Preregistration Worthwhile?', raised the question of the efficacy of preregistration when there is weak link between the statistical methodology to the scientific question of interest [97]. The Centre for Open Science (COS) defines **preregistration** as 'specifying' the research plan in advance [78]. That is, establish the manner in which data will be collected, what statistical models will be used, etc., *before* the experiment is undertaken. The benefit argued is that preregistration promotes transparaency and restricts the ability of researchers to engage in inadvertent methodological errors, frequently borne of discipline-specific convention, **questionable research practices** [31, 61], and delineates between confirmatory and exploratory research.

[1]The topic of which this manuscript forms a doctoral oeuvre [38, 39] thereof.

As COS produced the Reproducibility Project in psychology, which brought much-needed exposure to the lack of trustworthiness in scientific results [19], we might consider their resources as canonical current metascience practice. In what Makel describes as the 'most telling example of the lack of replicability' in science [61], the Reproducibility Project coordinated replications by 270 researchers, who found that out of 100 studies published in major psychological journals, only 39 per cent of these studies replicated their (mostly) significant results [19]. The push for preregistration is a direct response to this abysmal replication rate. Nosek, from the Centre for Open Science, with others, argue that 'Preregistration is Hard, And Worthwhile' [75], from a psychology metascience perspective.

Rethinking the conceptual framework of preregistration through a lens of mathematics, as opposed to psychology, raises questions about the space for which preregistration is defined, and draws parallels with other methodological interventions advocated to address the problems identified by COS' Reproducibility Project, such as computational reproducibility. For which statistical methodologies is preregistration worthwhile? Is the approach restricted to psychology? What about other sciences that share similar methodological practices? How should this be adopted in the craft of statistical computing and software design?

We begin with a brief overview of a recent publication in psychological science, 'Is Preregistration Worthwhile?', highlighting what is of specific interest to this manuscript, and follow with the meta-research objectives of this application of a conceptual framework from discrete mathematics in the context of psychological methodology.

6.1.1 Is preregistration redundant, at best?

Szollosi *et al.*'s manuscript questions the utility of preregistration, in the context of weak statistical inference, as opposed to other considerations, such as transparency in science [97]. They identify and respond to two arguments commonly put forward by proponents of preregistration.

Firstly, that preregistration reduces the potential for selectively including covariates to maximise statistical significance. This is but one of the many statistical model modifications sometimes referred to as ***p*-hacking**, **cherry-picking**, or more generally and collectively as **questionable research practice (QRP)** [31, 51, 61]. Questionable research practice is a family of activities that increase the chances of reporting spurious results and make research appear more reliable than is warranted.

In a modelling context, this includes changing modeling approach, model structure, parameterisation, or evaluation motivated by influencing the results of the model (either what it

finds or how informative it appears), rather than producing a model that is most appropriate for the dataset [36]. These practices are incentivised by an entrenched system of publish or perish in academia. Researchers are placed under enormous pressure to engage in QRP in order to maximise scientific successes and inflate publication records.

> Like airbrushing a photograph, QRPs may smooth unsightly blemishes to create a false sense of desirable appearance [61].

The prevalence of QRP is more about our scientific values as a culture, than calling out individual researchers for following long-standing conventions diligently. The second argument, advanced by proponents of preregistration, according to Szollosi *et al.*, is a weaker argument than the first: that preregistration encourages more reflective statistical choices in model selection and construction. The authors contend neither of these arguments hold when there is a weak link between the theory and practice:

> The diagonisticity of statistical tests depend entirely on how well statistical models map onto underlying theories, and so improving statistical techniques does little improve theories ***when the mapping is weak*** [97].

If we put aside the question as to whether preregistration is, itself, *worthwhile*, which is to say, despite the 'firestorm' that erupted when 'Is Preregistration Worthwhile?' was preprinted [74], we observe *all are in agreement* that methodological problems are widespread, and that we need to develop more robust scientific practices, particularly in an age of big data.

There is more that unites metascience than divides.

We aim to consider the statements above in the *lingua franca* of science: mathematics. From the perspective of mathematical science, if we are speaking of unity, it leads directly, as we shall explore in Section 6.4, to a union of sets.

There are many readings and possible responses to Szollosi *et al.*'s paper, even if only to expand on the ideas. In this manuscript, we shall restrict ourselves to an examination of how we might mathematically define what the authors mean when they say, '***when the mapping is weak***'.

Figure 6.1, reproduced from Wagenmakers' 'Breakdown' of Szollosi *et al.*'s paper, presents a nuanced way of thinking about bad, *wonky*, statistics and good enough practices in inference, that is, *sound* statistics. Instead of a dichotomy, crudely categorising scientific practice as

exploratory research ⟷ confirmatory research

sound statistics ⟷ wonky statistics

Figure 6.1: This is a reproduction of a detail from an image by Dirk-Jan Hoek in the blog post, 'A Breakdown of 'Preregistregration is Redundant, at Best"[2] [100]. It suggests we can think of weak and strong science applications of methodology as a spectrum. Also, a similar continuum between exploratory and confirmatory claims. It a visual representation of how informative statistical practice is, as not binary, but a continuous variable. However, by overlaying the two spectra, an implicit link is made between exploratory and weak application which does not allow for good methodology within exploratory research. By using the language of mathematics to describe the problem, we allow for more nuance, and capture more possibilities, than can be described in a two-dimensional figure.

arbitrarily good and bad, let us, as suggested in Figure 6.1, think of good enough scientific practice on a spectrum.

The conclusion of this manuscript takes us to Wagenmakers' observation that there are many methodological interventions to consider, in addition to preregistration, many of which may be at least as important.

> I view preregistration as one possible crutch (there are others) that fixes a small but essential part of the complete statistical inference process [100].

Disciplines use different approaches and have different intentions and it shouldn't be surprising when the solutions they offer to methodological problems also differ. Here we focus on defining a nomenclature to describe the relative merits of research methods that may or may not be relevant across many disciplines.

6.1.2 Questions about 'Is Preregistration Worthwhile?'

In this manuscript we use order theory to formalise the difference, described in 'Is Preregistration Worthwhile?', between the intrinsic measure of a mathematical estimator's efficacy, and it's value and trustworthiness in answering the scientific question of interest. In particular, *how* this is measured; how can we compare the utility of methodological interventions? How might we describe '***when the mapping is weak***' mathematically? In other words, an estimator may be nothing but mathematistry, obscuring the weakness of the statistical procedure behind mathematical complexity, and not furthering our understanding of the scientific question at hand.

There is a growth in the literature regarding preregistration, as there is emergence in reproducible computing in research, suggesting that these methodological interventions have great impact in particular contexts. But what are those contexts?

In Section 6.2, we question how we might measure how well methodologies answer scientific questions. Section 6.3 then uses the measure defined for the **order of mathematistry** so that we can compare the utility of different scientific methodologies under said measure, by class of measure. The first question posed is one of cardinality, in Section 6.4; how many pairings of claim and methodology are impacted in a scientifically meaningful way by preregistration? The second question posed, in Section 6.5, asks about how blurry the boundary, that is, mathematically continuous, as opposed to discrete, between how well methodologies inform scientific claims.

Perhaps heuristics are only effective at measuring the utility of a scientific methodology for a small subset of possible pairings of claim and methodology. Section 6.6 poses questions the self-limiting of any given heuristic of *good enough* [110], that is, realistically implementable best practices in science, as well as non-trivial cases. In Section 6.7, we consider other ways that we might have approached queering this particular question. In Section 6.8, we conclude by reflecting on the role queered scientific manuscripts across disciplines might play in furthering open metascience.

6.1.3 Why choose order theory?

Regardless of heuristic, we are discussing a value judgement of good enough practices [111] in science. We are considering which scientific practices are better, or **greater**, which is to say, $>$, mathematically. We ascribe a value to the efficacy of a scientific methodology when we say its output furthers, to borrow the parlance of Devezer *et al.*'s mathematical model, the 'process of scientific discovery' [26], or if the methodology is nothing more than mathematistry.

We shall formalise this in Section 6.3, so, for now, we foreshadow what we mean by **order** in this manuscript, intuitively in Figure 6.2.

Our purpose is to formalise a value judgement, a measure of from pure mathematistry to provable mathematical truths.

And a value judgement is, at heart, an order.

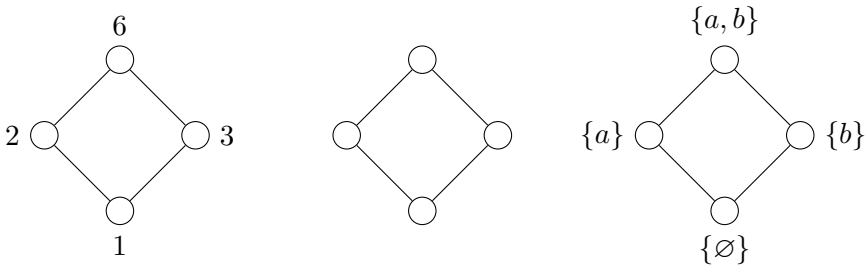


Figure 6.2: Order theorists study and describe the structure of ordering systems. Yes, the numbers 1, 2, 3, and so on, are ordered. But there are more complex ordering systems possible. On the left, we have the numbers that divide six in a *Hasse* diagram (where we read order ascending) [22]. In this order, $x \leqslant y$ if and only if x divides y. On the righthand side, we have the powerset (the set of all subsets) of two elments, ordered by subset. In the centre, we have the structure as an order theorist might think of it; the numbers that divide 6, ordered by division, is isomorphic (the same, up to a change of labels) to the powerset of two elements, ordered by subset. Note that in both cases, we have two elements that are in parallel, they are neither less than nor greater than the other, but do have relationships to the same elements: 2 does not divide 3, nor 3 divide 2; and $\{a\}$ is not a subset of $\{b\}$, nor vice versa. The central nodes are below the top, and above the bottom, in all three diagrams.

6.2 Measuring mathematistry

Let C denote the set of all possible scientific claims for which we might provide evidence of, using a scientific method or procedure.

Let M denote the set of all possible scientific methods that can be used to provide evidence of scientific claims. We are deliberately vague here, where scientific method may comprise a model, a method, such as preregistration, or a combination of statistical and procedural interventions. Any procedure that, at least in some contexts, furthers what Devezer *et al.* describe as the 'process of scientific discovery' [26]. We may consider statistical models in this nomenclature, at any time, by taking the subset S of methodologies that involve a statistical procedure. In this manuscript, we take reproducible computing and preregistration, in particular, and touch on the more complex case of questionable research practices in ecological models, in Section 6.6.2, to demonstrate the challenges of modelling through mathematistry.

Consider the set of ordered pairs $(c, m) \in C \times M$, that each describe a scientific claim presented in a published manuscript, thought of in terms of claim $c \in C$ and method $m \in M$. Let $\mathcal{X} \subseteq C \times M$ denote the subset $\mathcal{X}$ of reasonable pairings $C \times M$ of claims and methods.

The measure of mathematistry can be thought of as a perfect truth or not, as in a mathem-

atical proof, or a spectrum, as illustrated in Figure 6.1, or ordering more complex, from weak to strong applications. In the latter, the weakest applications are nothing but mathematistry, and in the strongest, the mathematics does not obscure the truth with mathematistry. Indeed, as we shall explore in Section 6.6.2, we may construct yet more complex orderings of mathematistry. First, however, we must define both the measure of mathematistry and its order.

6.2.1 Heuristics of mathematistry

It is worth pausing to question what characterises a strong application of method m from possible methods, M, to further, to borrow Devezer *et al.*'s terminology [26], the process of scientific discovery of claim c from possible claims, C. We will not seek to argue the merits of any particular measure, but to allow our heuristic to follow any possible ordering system. We consider $h(c, m)$ to be a measure of the strength of the mapping between method m and the scientific claim c; a measure from nonsense and mathematistry, obscuring of the results posed by answering a seemingly similar question, at one end, to truth at the other. All statistical inference has, arguably, a measure of mathematistry; thus it is what lies between nonsense and truth that is the study of this manuscript.

There are many interpretations of the term **heuristic** [18], and here we invoke it to represent the rule of thumb by which we denote science as good or bad. In particular, we might extend on the spectrum suggested in Figure 6.1, as opposed to a reductive dichotomy of good or bad science. Thus, we might think of four cases of how a heuristic might behave:

$$h : C \times M \to \begin{cases} \{0, 1\} & \text{if heuristic categorises as effective or not;} \\ [0, 1] & \text{if heuristic measures efficacy on a spectrum;} \\ H & \text{if heuristic categorises efficacy otherwise,} \\ \varnothing & \text{if the heuristic does not apply to the pairing.} \end{cases}$$

where h is some categorisation of $C \times M$ ordered by $\geqslant$ under h in some way from not effective, an abundance of mathematistry, to effective, very little mathematistry. A visual interpretation of these is shown in Figure 6.3 using *Hasse* diagrams, read from bottom to top, from minimal elements, uninformative mathematistry, to maximal elements, no mathematistry. From the left, we have the dichotomous case, then the spectrum, and, on the right, a more complex set of relationships.

We define our heuristic according to the commonalities we identify between all types of

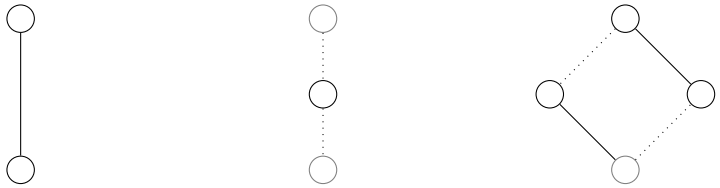


Figure 6.3: We can think of **heuristics of mathematistry** in terms of three forms: dichotomous, as Wagenmakers observes is too reductive; spectrum, as suggested by Wagenmakers [100]; or some combination of continuous or categorical value systems that allow comparison systems. In the language or order theory, we can abstract and describe all cases. The solid lines symbolise a **covering** connection, where the element is immediately above or below the connected element under heuristic h; think the natural numbers, $1, 2, 3, \ldots$, which we denote $\mathbb{N}$. The dotted lines indicate a continuum upon which there are infinitely many points; think real numbers $..., -\pi, \cdots -1, \ldots, 0, \ldots e, ...$, which we denote $\mathbb{R}$. We allow systems that have a true top $\top$ and bottom $\bot$, such as the interval $[0, 1]$. Also allowed are those that have theoretical limits not included in the order, coloured in grey, such as $(0, 1)$, as illustrated in the central *Hasse* diagram. Mathematics provides us with tools to ask questions of all possible heuristics.

heuristics.

Definition 2. *Heuristic of mathematistry.* A **heuristic h of mathematistry** measures the degree to which a method $m \in M$ obscures, through calculations and scientific procedure, the process of furthering scientific discovery towards claim $c \in C$. We denote $\mathbb{H}$ to be the set of all possible heuristics of mathematistry. We define the value $h(c, m)$ as the **measure of mathematistry** of pairing (c, m) under heuristic h.

6.2.2 Characterising heuristics of mathematistry

We are interested in heuristics that measure how well a method furthers what Devezer *et al.* describe as a process of scientific discovery [26]. The integrity of the design of a statistical model will impact on how much it furthers the process of scientific discovery. However, it is yet more nuanced still. A study may be perfectly designed and still not further the process of scientific discovery to the claim or question of interest.

In Grainger *et al.*'s recent paper on **research waste** [37], cumulative meta-analysis, assessing meta-analyses over time to see if further research will make a difference to the estimation of effect, is suggested to assess if conducting a study will contribute meaningfully to the field, or if it will be a waste of research resources. In this case, we may have an experiment performed rigorously, but if this does not further scientific discovery to the population effect, based on

the existing literature, then we arguably have close to the greatest measure of mathematistry. Mathematical calculations have been performed, but they do nought to further the process of scientific discovery.

Another concern may be the required computational power. Kwisthout *et al.*, point out that, for scientific claims $c_1, c_2 \in C$, an algorithm $m \in M$ may be computationally feasible for c_1, but not c_2 [57]. Another consideration is the limitations of computation in the questions that can be effectively asked computationally [8]. These are but examples of confounding factors that might affect the efficacy with which a given method might further the process of scientific discovery towards a claim.

Let us consider an example measure to see how its mathematistry changes when the claim paired to it is changed. Suppose a data scientist at a school wished to know how many autistic students n are in their cohort N, to inform how resources should be allocated. We will denote this question c_1.

They may extract the students' files as a table and filter on an autism indicator counting how many rows they have n, compared to how many rows total N. We will denote this method $m \in M$.

What is the mathematistry of this calculation? It is $\top$, with no mathematistry. The population the data scientist wished to know about was entirely known.

Now suppose the principal asked the data scientist to project what the proportion of the cohort will be autistic students over the next ten years. We denote this question as c_2. In this case, the method employed can only approximate the parameter of interest.

Does method m carry as much certainty for question c_1 and c_2? No, (c_1, m) is a population parameter, so is truth, $\top_h$, the minimum measure of h. In the other case, (c_2, m), it is an uncertain estimate of a parameter. Which is to say, under some heuristic h, we must have

$$\top_h = 0 = h(c_1, m) < h(c_2, m).$$

A heuristic must differentiate pairings of claims and methods by how well the method furthers the process of scientific discovery. With this, we now extend on Definition 2 to clarify the set of possible heuristics, $\mathbb{H}$.

Definition 3. Let $\mathbb{H}$ denote the set of all heuristics of mathematistry. For a heuristic h to be a member of $\mathbb{H}$ there must exist a scientific method m, and two distinct scientific claims we

might reasonably pair m with, c_1 and c_2, such that

$$h(c_1, m) > h(c_2, m)$$

or, conversely, there must exist distinct methods, m_1 and m_2, such that, for a claim c, we have

$$h(c, m_1) > h(c, m_2).$$

6.3 The order of mathematistry

In this section we define a binary relation $\to_h$ and show it is a quasi-order on pairings of claim and methodology, and that $\to_h$ is an order when we partition the reasonable pairings $\mathcal{X}$ of claim and method, $C \times M$, by the same relation $\to_h$. This construction follows Hell and Nešetřil's methodology for ordering of equivalence classes of the category of directed graphs [45], which was recently extended to the category of finite algebras [24].

Definition 4. Let $(c_1, m_1) \to_h (c_2, m_2)$ if and only if $h(c_1, m_1) \geqslant h(c_2, m_2)$ under heuristic h of mathematistry.

So, we consider (c_1, m_1) to be less than (c_2, m_2) if the measure of mathematistry of (c_1, m_1) is greater than (c_2, m_2). To be considered an order, $\to_h$ must satisfy three properties [22].

Definition 5. A binary relation $\to$ on set P is an **order** if, for all $x, y, z \in P$, we have

(i) $x \to x$,

(ii) $x \to y$ and $y \to x$ implies $x = y$,

(iii) $x \to y$ and $y \to z$ imply $x \to z$.

We refer to these properties as (i) **reflexivity**, (ii), **antisymmetry**, and (iii) **transitivity**, respectively.

When a binary relation satisfies (i) reflexivity and (iii) transivity, but not (ii) antisymmetry, we say it is a **quasi-order**.

Lemma 1. The relation $\to_h$ is a quasi-order on $\mathcal{X}$.

Proof. We will show the relation $\to_h$ satisfies reflexivity and transitivity, but not antisymmetry. Let $(c_1, m_1), (c_2, m_2), (c_3, m_3)$ denote paired claims and methodologies from reasonable pairings, $\mathcal{X}$, of scientific claim and methodology to assess that claim, a subset of $C \times M$.

To show reflexivity, we observe $(c_1, m_1) \to_h (c_1, m_1)$, as $h(c_1, m_1) = h(c_1, m_1)$, so $h(c_1, m_1) \geqslant h(c_1, m_1)$.

For transitivity, let us assume $(c_1, m_1) \to_h (c_2, m_2)$ and $(c_2, m_2) \to_h (c_3, m_3)$. Then, we have $(c_1, m_1) \geqslant (c_2, m_2)$ and $h(c_2, m_2) \geqslant h(c_3, m_3$. Which gives $h(c_1, m_1) \geqslant h(c_3, m_3)$ $(\star)$. Thus, $(c_1, m_1) \to_h (c_3, m_3)$.

But if we assume $(c_1, m_1) \to_h (c_2, m_2)$ and $(c_2, m_2) \to_h (c_1, m_1)$, we cannot assume $c_1 = c_2$ and $m_1 = m_2$, as more than one paired claim and methodology may have the same measure under the heuristic h.

Since we have satisfied reflexivity and transitivity, but not antisymmetry, we conclude $\to_h$ is a quasi-order on $\mathcal{X}$. □

Note that in the above proof we require, at $(\star)$, the heuristic h to be ordered by $\geqslant$.

Corollary 1. The heuristic h of mathematistry is an ordered set on the universe, that is, the set that is ordered, H,

$$h := \langle H; \geqslant \rangle.$$

If, however, we define an equivalence class, and partitioning of $\mathcal{X}$ by that equivalence relation, then we shall see $\to_h$ is an order relation in that space.

Definition 6. We define the equivalence class $[[c, m]]_h$ on $\mathcal{X} \subseteq C \times M$ under h, a measure of the mathematistry of method, m, from possible methods, M, in furthering the process of scientific discovery towards the claim c, from possible heuristics $\mathbb{H}$ of mathematistry,

$$[[c, m]]_h := \{(x, y) \in \mathcal{X} \mid h(x, y) = h(c, m)\}$$

Which is to say, we consider $[[c, m]]_h$ to be the equivalence class of things that take the value $h(c, m)$ under heuristic $h \in \mathbb{H}$ of mathematistry.

Definition 7. Let $\mathfrak{X}_h := \mathcal{X}/_{\to_h}$ denote the set of equivalence classes $[[c, m]]_h$ of $\mathcal{X}$ partitioned by heuristic h of mathematistry, chosen from possible heuristics of mathematistry, $\mathbb{H}$, of strength of evidence in scientific claim $c \in C$ provided by procedure m in M, which includes all statistical forms of estimation.

When we consider the application of the relation $\to_h$ in the space $\mathfrak{X}_h$, we can now more fully characterise the relation.

Theorem 1. The relation $\to_h$ is an order on $\mathfrak{X}_h$.

Proof. Let $[[c_1, m_1]], [[c_2, m_2]], [[c_3, m_3]]$ in $\mathfrak{X}_h$. According to Definition 5, we must show to_h satisfies reflexivity, transitivity, *and* antisymmetry.

Since $(c_1, m_1) \to_h (c_1, m_1)$, by Lemma 1, we have $[[c_1, m_1]]_h \to_h [[c_1, m_1]]_h$. So, $\to_h$ is reflexive.

For transitivity, let us assume $[[c_1, m_1]]_h \to_h [[c_2, m_2]]_h$ and $[[c_2, m_2]]_h \to_h [[c_3, m_3]]_h$. We wish to show $[[c_1, m_1]] \to [[c_3, m_3]]$. Then $(c_1, m_1) \to_h (c_2, m_2)$ and $(c_2, m_2) \to_h (c_3, m_3)$, so $(c_1, m_1) \to_h (c_3, m_3$, by Lemma 1. So, we have $[[c_1, m_1]]_h \to_h [[c_3, m_3]]_h$, as required.

Finally, we show that when considered acting on equivalence classes, $\to_h$ satisfies antisymmetry. Let us assume $[[c_1, m_1]]_h \to_h [[c_2, m_2]]_h$ and $[[c_2, m_2]]_h \to_h [[c_1, m_1]]_h$, then we have $h(c_1, m_1) \geqslant h(c_2, m_2)$ and $h(c_2, m_2) \geqslant h(c_1, m_1)$. Since h is antisymmetric, by Corollary 1, we have $h(c_1, m_1) = h(c_2, m_2)$. Thus $[[c_1, m_1]]_h \equiv [[c_2, m_2]]_h$. □

And so we may define our order of interest.

Definition 8. We refer to

$$\langle \mathfrak{X}; \to \rangle_h := \langle \mathcal{X}/\to_h; \to_h \rangle$$

as the **order of mathematistry** under heurisic $h \in \mathbb{H}$.

6.4 A question of cardinality

Figure 6.4 attempts, if a trifle ambitiously, to summarise the motivating question of this manuscript. Consider the pairings of claim and scientific method, where the method includes a given intervention, say, preregistration or reproducible computing. What proportion of these pairings furthers the process of scientific discovery such that the pairing's mathematistry is lifted above some predefined optimal level $\tilde{h}$ of mathematistry? This question motivated the order theoretic constructed provided in Definition 8; that is, the order of mathematistry was created to articulate this question.

In this section, we attempt to articulate, using the order of mathematistry, this motivating question of cardinality. That is to say, what is the size of the set of paired claim and methodologies for which preregistration, or reproducible computing, will make a meaningful scientific difference?

Figure 6.4 provides a Hassse diagram of the central question of the manuscript 'Is Preregistration Worthwhile?' [97]. We now present this claim mathematically.

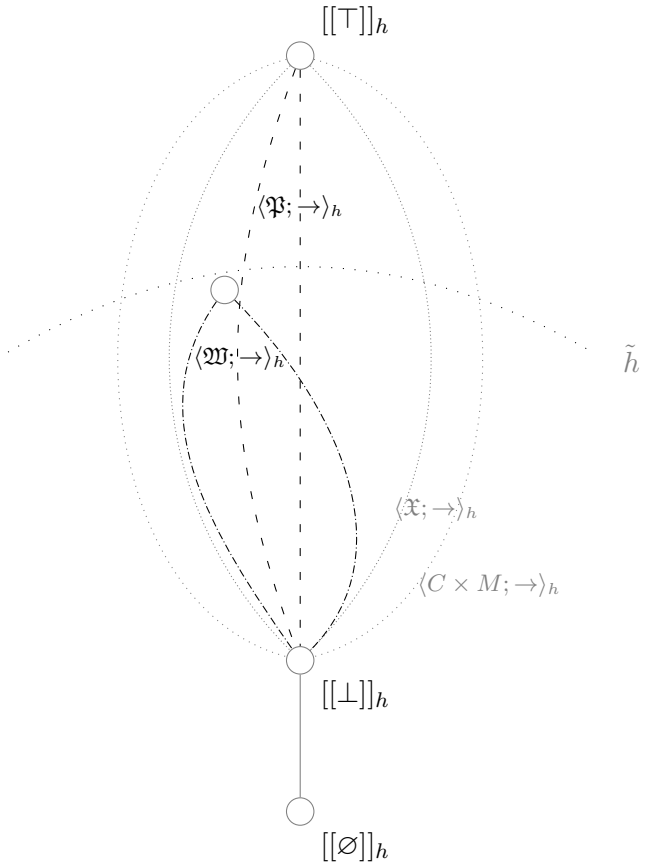


Figure 6.4: We use ····· to depict with the order of mathematistry, $\langle C \times M; \to \rangle_h$, on all pairings of claim and methodology. The order on reasonable pairings, $\langle \mathfrak{X}; \to \rangle_h$, depicted by ····· , are a proper subset of possible pairings, as there are combinations that are nonsensical, such as applying longitudinal methodology to data that has no properties of longitudinal data. We highlight two suborders of mathematistry in black. We use - - - to show the order of mathematistry on pairings whose methodology contains an intervention of interest, $\langle \mathfrak{P}; \to \rangle_h$, say, reproducible computing or preregistration. And -·-·- for the order of mathematistry where there is a weak application of theory to the given scientific question, $\langle \mathfrak{W}; \to \rangle_h$. An arbitrary band ····· represents the measure, $\tilde{h}$, above which is considered an acceptable measure of mathematistry under heuristic h. Not all preregistered scientific endeavours have a strong link between theory and application [97], so $\langle \mathfrak{W}; \to \rangle_h$ overlaps $\langle \mathfrak{P}; \to \rangle_h$. Is the proportion of $\langle \mathfrak{P}; \to \rangle_h$ that does not overlap $\langle \mathfrak{W}; \to \rangle_h$ so small that most of $\langle \mathfrak{P}; \to \rangle_h$ does not achieve $\tilde{h}$?

Claim 1. Let h be a heuristic of mathematistry, chosen from possible heuristics of mathematistry, $\mathbb{H}$. Let $\langle \mathfrak{P}; \to \rangle_h$ denote the order of mathematistry on reasonable pairings of claim and methodology whose methodology include an intervention of interest, for example, reproducible computing or preregistration. That is, if $\mathcal{P} := C \times P$, where $P \subset M$, denotes the methodologies with the intervention of interest, paired reasonably with appropriate claims. Then the pairings considered in $\mathcal{P} \subset \mathcal{X} \subset C \times M$ generate the suborder $\langle \mathfrak{P}; \to \rangle_h$. Let $\langle \mathfrak{W}; \to \rangle_h$ denote the order of mathematistry where there is a weak application of theory to the claim in question. These are both suborders in the order of mathematistry on reasonable pairings, $\langle \mathfrak{X}; \to \rangle_h$. Let $\tilde{h}$ denote the measure of mathematistry under h above which the pairing sufficiently furthers scientific discovery.

We then ask if the proportion of $\langle \mathfrak{P}; \to \rangle_h$ that overlaps, that is, intersects, with $\langle \mathfrak{W}; \to \rangle_h$ is much greater than $\langle \mathfrak{P}; \to \rangle_h$ where there is a strong application of theory to claim. That is,

$$\langle \mathfrak{P}; \to \rangle_h \cap \langle \mathfrak{W}; \to \rangle_h \gg \langle \mathfrak{P}; \to \rangle_h \backslash \langle \mathfrak{W}; \to \rangle_h$$

and, as such, leads to a larger set of pairings for which the minimal heuristic of mathematistry is not achieved. That is, if we assume that by *weakly applied*, we mean sufficiently minimal mathematistry to further the process of scientific discovery, we have

$$(c, m) \in [[c, m]]_h \text{ and } [[c, m]]_h \in \mathfrak{W} \implies h(c, m) > \tilde{h}.$$

Then only a small proportion of pairings of claim and methodology will further the process of scientific discovery by imposing this intervention. So, for any heuristic, $h \in \mathbb{H}$, and any methodological intervention $P \subset M$, ordered by heuristic h, we have,

$$\{(c, m) \in \langle \mathfrak{P}; \to \rangle_h \mid h(c, m) > \tilde{h}\} \gg \{(c, m) \in \langle \mathfrak{P}; \to \rangle_h \mid h(c, m) \leqslant \tilde{h}\}.$$

in the order of mathematistry $\langle \mathfrak{P}; \to \rangle_h$ induced by h on the methodological intervention that defines $\langle \mathfrak{P}; \to \rangle_h$.

6.5 A question of density

In addition to asking what subspaces affected meaningfully by methodological intervention, such as reproducible computing, or preregistration, as discussed in the previous section, we might also ask if this space is finite or contains infinite subspaces. Which we might ask, in the

context of mathematistry, is there a clear delineation between two classes of models where one has a greater order of mathematistry than the other? More formally, we might ask if the order of mathematistry is **dense**.

Definition 9. An ordered set P is **dense** if, for $x < y$ in P, there exists z such that $x < z < y$ [22].

One observation we might make is that there is a strict delineation between $[[\bot]]_h$ and $[[\varnothing]]_h$; which is to say, $[[\bot]]_h$ **covers** $[[\varnothing]]_h$, as shown in Figure 6.5.

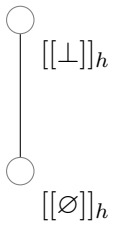


Figure 6.5: Detail from Figure 6.4 showing the set of pairings from $C \times M$ that have measure 0 under heuristic $h \in \mathbb{H}$ covering (with nothing between) the set of pairings from $C \times M$ that are not contained in the domain of heuristic h, i.e., not assigned a measure. Both are grey, as it is possible that the set is empty; i.e., the heuristic always has non-zero measure, or the heuristic measures to all possible pairings.

Definition 10. We say x is **covered by** y (or y **covers** x), if $x < y$ and $x \leqslant z < y$ implies $z = x$ [22].

We will now use our observation, along with Definitions 9 and 10, to formally characterise Figure 6.5 as a covering relation.

We make the assumption there is at least one possible pairing for which a given heuristic does not apply. For surely there is some analogue of the No Free Lunch Theorem [112] for heuristics that finds no heuristic optimally measures all pairings of claim and methodology.

Theorem 2. Assuming that there is at least one pairing $(c, m) \in \mathcal{X}$, in reasonable pairings of claim and methodology, for which the heuristic of mathematistry does not apply, the order of mathematistry induced by heuristic $h \in \mathbb{H}$ is not dense everywhere.

Proof. To show that the order of mathematistry is not dense everywhere, we will demonstrate that $[[\bot]]_h$ covers $[[\varnothing]]_h$, as described by Definition 10.

Assume there is at least one pairing (c_1, m_1), from reasonable pairings $\mathcal{X}$, of claim and methodology, for which the heuristic of mathematistry does not apply, under some heuristic h

from the heuristics $\mathbb{H}$ of mathematistry. That is, the pairing (c_1, m_1) cannot be found in the domain of h. Thus we have $[[\varnothing]]_h$ exists and $[[c_1, m_1]]_h \equiv [[\varnothing]]_h$ in $\langle \mathfrak{X}; \rightarrow \rangle_h$.

Now take any other pairing (c_2, m_2) from minima measures of mathematistry under h. That is, $(c_2, m_2) \in \min(h)$.

Any other pairing must be in one of three states. We will consider each to show there is no distinct element between $[[c_1, m_1]]_h$ and $[[c_2, m_2]]_h$, as required by Definition 9 to satisfy density, between these equivalence classes.

If the pairing has no measure of mathematistry under h, then it is a member of $[[\varnothing]]_h$. Since $[[\varnothing]]_h \equiv [[c_1, m_1]]_h$, we have not found a distinct element with measure of mathematistry between $[[c_1, m_1]]_h$ and $[[c_2, m_2]]_h$.

If the pairing has a measure of mathematistry, then either its measure is the same as c_2, m_2 or greater, as c_2, m_2 is a member of the minima of h. If its measure is the same then it is in $[[c_1, m_1]]_h$, and not distinct.

And, finally, if the measure is greater than the minima, then its equivalence class is distinct but *above* the minima of h, and thus not between the equivalence class of null measure under h and those of the minima, as required. Hence, we may conclude, by Definition 10, that $[[\bot]]_h$ covers $[[\varnothing]]_h$ in $\langle \mathfrak{X}; \rightarrow \rangle_h$. □

By finding one example of a covering relation, we have shown that the order of mathematistry is not dense everywhere. But we have considered a specific case. Any number of further questions of density could be asked of the Order on any subset. One might be interested in delineating between classes of models, for example. But, as our intention is to provide examples of metascience questions posed in an order theoretic language, we now consider a different aspect of the order of mathematistry.

6.6 The utility of heuristics

This manuscript has, thus far, questioned the utility afforded in using an order-theoretic framework to formalise questions about the practice of science. It is, however, certainly worth posing the counter question, that is, consider where there is little utility afforded. In mathematical parlance, we might say we are questioning under what conditions this order-theoretic construction is trivial. As demonstrated by Theorem 2, for any given h, if we agree that for any heuristic there must be claims and methodological pairings for which h does not apply, then the construction is non entirely trivial, that is it does not comprise a single equivalence class. We

now question to what extent the structure is uninformative or trivial, as well as the potential complexity of the construction.

At first blush, to consider this might seem lacking in purpose. However, a cautionary tale emerges in the application of heuristics, and the limitations inevitably afforded by the domain for which the heuristic was initially conceived. For example, we might question the extent to which preregistration translates to scientific methodology outside of social science. Similarly, reproducibility may be unfeasible in medical or political settings, for example, the research on the COVID-19 pandemic is necessarily irreproducible, for we cannot, and would not wish to, recreate the exact circumstances of this global event. Computational reproducibility may, too, be unattainable where datasets contain sensitive information.

6.6.1 The limitations of a heuristic

Perhaps any given heuristic is necessarily biased by the the conceptual framework from which it began existence. When considering the collection $\mathcal{X}$ of reasonably paired claims C and methodologies M for assessing these claim, we might question for which pairings does the heuristic apply. We might ask if, for *any* heuristic, its utility of measuring how well a methodology furthers the process of scientific development is limited to a small, albeit arguably interdisciplinary, domain.

Claim 2. Let $h \in \mathbb{H}$. Then

$$[[0]]_h \cup [[\varnothing]]_h \;\gg\; \langle \mathfrak{X}; \to \rangle_h \backslash ([[0]]_h \cup [[\varnothing]]_h).$$

Informally, by Claim 2, we are asking, for any heurisitic, are there many more pairings heuristic h does not provide a measure of mathematistry for than there are pairings for which a heuristic provides a measure of mathematistry? Despite possible limitations or inherent bias of any formalism applied to metascience, this manuscript suggests there is yet utility.

What is the purpose of formalism? To define things. In metascience, where conventions are liable to be deeply ingrained in any given discipline, it is worth asking if we may serve ourselves better by first agreeing on *what it is* we are discussing. Formalism via mathematics provides us with a vehicle to fix components, vary other components, and speak in the abstract about what does and does not serve the pursuit of good enough practices in science. However, we know from Theorem 2, that this nomenclature describes non-trivial spaces. We now consider two such non-trivial constructions, one that is at least five levels, and one that demonstrates

the combinatorial complexity that can occur, providing the heuristic differentiates thus.

6.6.2 Non-trivial applications of the order of mathematistry

In particular, the order of pairings that involve methodologies with a statistical intervention, that is, a mathematical calculation, is non-trivial. We consider statistical interventions in the first, and, in the second, we turn to a conceptualisation of building ecological models [36].

The order of mathematistry on statistical models

For our first example of a non-trivial order of mathematistry, we consider a heuristic, h that allows differentiation between completely informative, an approximation, somewhat informative, and not at all informative application of a method m, involving a statistical calculation, to a scientific claim c. That is, we consider cases where there is a mathematical calculation embedded in the methodology. Then this heuristic will have at least five levels, providing a simple example of a non-trivial application of the order of mathematistry.

Theorem 3. Let h be a heuristic of mathematistry, chosen from possible heuristics $\mathbb{H}$ of mathematistry and $\langle \mathfrak{S}; \rightarrow \rangle_h$ be the order of mathematistry induced by h on the subset S of methodologies M where the procedure involves a statistical calculation, where h allows for at least five levels. Then $\langle \mathfrak{S}; \rightarrow \rangle_h$ has five levels.

Proof. Let us assume claim 2 holds, and that under heuristic h we have pairings with both 0 and $\varnothing$ measures in $C \times S$.

From Lemma 2, we know $\langle \mathfrak{S}; \rightarrow \rangle_h$ has a distinct bottom $[[\varnothing]]_h$, covered by $[[\bot]]_h$. Where the correct statistical calculation is applied to a population's data for question of interest, say, a proportion n/N of a group, n, from a larger population, N, we have the precise result for the population of interest, $\top$, truth. Thus, $\langle \mathfrak{S}; \rightarrow \rangle_h$, has a top, $[[\top]]_h$. It remains to show there are two levels between the bottom, $[[\bot]]_h$, and the top, $[[\top]]_h$.

Below a population stastistic, we have a sample statistic, say, $\hat{n}/\hat{N}$, from which we might draw inference about the true proportion n/N, *where the calculation has been performed correctly.* Below this we have an *incorrect* calculation, say, N/n, which, although incorrect, may still be more informative than an application of the statistical method that produces a $\bot$ measure of mathematistry under h. □

We represent the structure described in Lemma 3 as a Hasse diagram in Figure 6.6. When viewed this way, we see why representations such as Figure 6.1, arise, to combat the lamented

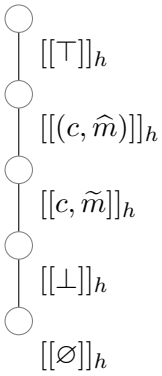


Figure 6.6: If we take a heursitic h that allows for at least five levels of distinction for statistical calculations, a subset of possible methodologies, then the suborder of mathematistry on statistical calculations $\langle \mathfrak{S}; \to \rangle_h$ has at least five levels. Those levels are the equivalence classes generated by: the top, $\top$, the population summary statistic, $\hat{m}$; the statistical calculations that estimate population parameters, $\hat{m}$; an incorrect computation of a population estimator, $\tilde{m}$; pairings that have a measure of 0 under heurstic h; and pairings for which the heuristic does not apply, generated by $\varnothing$.

rise of **dichotomania**, an overemphasis on dichotomising scientific questions to true or false [1]. As we shall further explore in the next section, a mathematical calculation that is a statistical model is a Borgesian[3] garden of forking data in which there are many opportunities, research degrees of freedom, in which the model may be compromised or robustified [33].

The mathematistry of QRP in ecological models

For our final application of the order of mathematistry, we turn to Gould's conceptual framework for questionable research practices in ecological models [36]. This is an extension of Gelman and Loken's conceptualisation of 'researcher degrees of freedom' in null hypothesis significance tests, which recognises that researchers must make several decisions when performing a statistical analysis [33]. Within these degrees of freedom, lie many pitfalls where the researcher may inadvertently engage in QRP [31]. With Gould's adaptation of Gelman and Loken's conceptual framework for statistical inference, that is, inference that does not fall into the family of null hypothesis significance testing, the order of mathematistry is non-trivial.

Gould identifies a six-phase model development process:

[3]The Garden of Forking Paths is 1941 short story by Jorge Borges in which the multitudinous possibilities that arise from each decision made are conceptualised as a labrynthine garden of bifurcating paths [11].

1. Determine model.
2. Collect data.
3. Algorithm selection.
4. Model predictions.
5. Model assessment.
6. Submit for publication or not.

At each stage, Gould identifies questionable research practices, cherry picking, p-hacking, etc., that adversely affect the methodology's ability to further the process of scientific discovery.

To conceptualise this, somewhat crudely, let us restrict ourselves to whether the step in the model development process sufficiently furthers the process of scientific discovery, where we denote $m(s_1)$ as the presence of QRP at step 1., $m(s_1, s_2)$ as the presence of QRP at steps 1. and 2., and so forth.

Given that QRP can occur at any of the six model development steps, the order of mathematistry, even crudely partitioned by meeting the heuristic or not, is a complex structure. Perhaps one QRP affects the mathematistry of the model less than another, so that $h(c, m(s_1)) < h(c, m(s_3))$, for example. Furthermore, multiple QRPs will likely affect a model's ability to inform on the ecological problem in question, so that $h(c, m(s_1, s_2)) > h(c, m(s_1))$.

In Figure 6.7, we construct a crude ordering on the set of ecological models and questions they might answer, which is detailed in the caption. Despite fairly crude simplifications made within the heuristic, the resulting construction is complex. And now that we have defined and constructed the order of mathematistry, as well explored its limitations, and non-trivial cases, we pause to note what is not discussed in this manuscript, before concluding with a note on communication across disciplines.

6.7 Other queerings

We have described how our definition of a heuristic h, from possible heuristics, $\mathbb{H}$, may be drawn from not only mathematics, but metascience literature on questionable research practices, that indicate issues, e.g., experimental design [31] or the literature on the application and interpretation of regression models [69], a statistical modelling tool common to many disciplines,

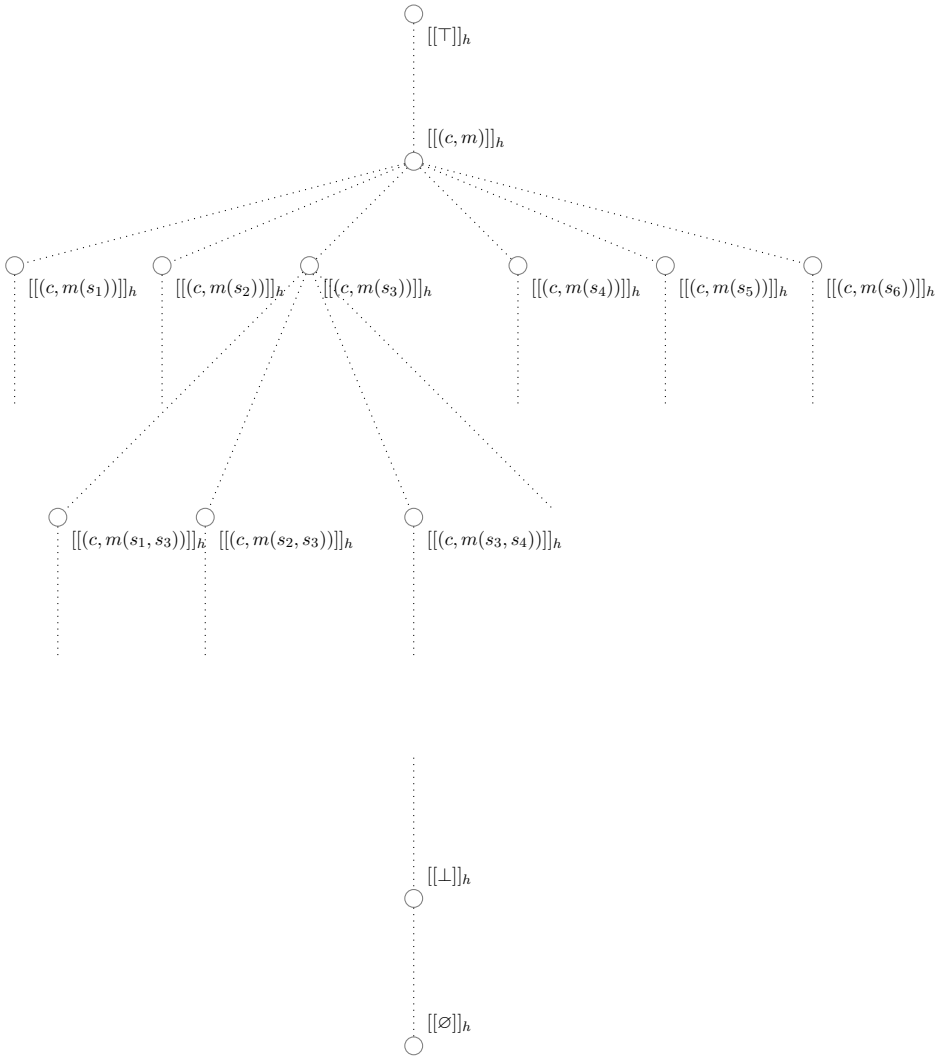


Figure 6.7: Let us, somewhat crudely, suppose the heuristic h assigns a different but equal value to the presence of a QRP at one step in the algorithm, and for two steps, and so forth. Let $[[c, m(s_n)]]$ denote the equivalence class of ecological models and things to model, such there exists a QRP at the nth step. This is crude on two counts: firstly that there is an equivalence class of one or more QRPs at Step 1. in Gould's six phases of building an ecological model; and furthermore, the ordering is parallel for QRP at each step. But, even with crudely simplifying the problem thus, the order of mathematistry on ecological models is complex.

or the literature on misapplication, such as the use of exploratory data analysis statistical techniques in confirmatory research [96]. As our focus was to construct an order-theoretic conceptual framework, this literature is given only shallow measure.

Indeed, there is much not discussed in this manuscript, as well, from within mathematics. Is mathematistry a category? What benefits might be leveraged from measure theory, the study of measures? Perhaps measure theory is even better adapted to this problem than order theory. Another mathematical approach might be a directed graph where we allow all edge-relations $\rightarrow_h$ to exist for all $h \in \mathbb{H}$. Which is to say, consider the space where all heuristics of good and bad science are allowed.

Statistical meta-researchers present a different perspective of conceptualising metascience, and in each case, we might queer the conceptualisation through order theory. For example, it would be interesting to explore an the order of mathematistry on Devezer *et al.*'s mathematical model of scientific discovery, which provides a framework to study reproducibility [26].

If nothing else, this manuscript observes there is much work that can be done in constructing a discipline-agnostic metascience formalism that facilitates friendly, positive collaboration between scientific fields.

6.8 Scientific ways to discuss how to science

It's tempting to dismiss good enough scientific practice of strongly linking a methodology to a claim, in order to elucidate a meaningful scientific result that furthers the process of scientific discovery, as not *that* hard. Especially from the perspective of a well-established discipline with widely agreed-upon conventions.

But what is important, or a pitfall, in one discipline may be irrelevant or trivial to avoid in another. What comprises the process of preregistration in the context of statistical simulations? To learn from each other, we require standardised ways of talking about what types of scientific claims particular methodologies are appropriate for.

The cardinality question presented in Figure 6.4, along with the other questions posed in Sections 6.5 and 6.6, provide a framework for how we can avoid an overemphasis on specific scientific methodologies in our canon of good enough [111] (i.e., realistic) research practices, such as preregistration [97] or reproducible computing [39], to the detriment of scientific claims for which these specific methodologies are, at best, redundant [97].

To learn from and improve each others' science, we need to 'queer,' as Simpson argues, 'rather than invert the existing structures and build a more equitable version of the world' [90].

By posing questions about metascience through the language of mathematics, we queer our understanding of methodologies of the sciences of uncertainties, where inference is practiced but mathematics frequently avoided, or applied with an overenthusiastic and unnuanced fervor. Rather than seek to reinvent new mathematics, this manuscript aims to queer the questions of psychological science through the language of mathematics. If some progress towards a informative discipline-agnostic nomenclature for metascience has been achieved, then this manuscript has achieved its aim.

Chapter 7

Foibles & Limitations

The nature of interdisciplinary work

7.1 A dissertation is never completed

A dissertation of mathematical computation is an apprenticeship in science; a dissertation is never completed, it is ended. As with apprenticeship in anything, there are many things I'd do differently now, given a do over. And, as a work of apprenticeship, it is unlikely to define any field; at best, this intrepid graduate student endeavours to faithfully record my thoughts on these scientific questions at this time[1].

I'll ruminate briefly on but a few examples of the many flaws and things I would now do in another way. The time to end a doctorate is when sufficient work has been completed that the scientist is ready to move onto a new project. Despite the following foibles and limitations, this dissertation is demonstrative of an apprenticeship in interdisciplinary computational metascience.

[1] I am already at the point where a part of me wants to submerge this dissertation in the bathtub, "*Shh shh* it'll all be over soon", until it stops kicking. It's been a daily struggle not to attach a stickynote to the title page, *I pity the poor fool that must needs read this here screed*. But I will only learn to write good essays by writing bad essays, then better essays; at some point rewriting the same essays does not serve to render me a better scientist. I am confident I am at this point where my development as a scientist will be best served by moving on to my next project, a Bayesian network meta-analysis for a Cochrane study. Do get in touch if you wish to contribute to `nmareporting::`, for reporting network meta-analyses according to open science protocols.

7.2 Meta-analysis of medians

One limitation of this chapter is there is more work that would be required to bring this chapter to publication. There have likely been advances in this problem during the disrupted progress on this manuscript.

7.3 Testing and code

One of the major limitations of this thesis is a lack of depth of inquiry into the practice of automated testing. This dissertation does not develop the ideas beyond practices described in the Testing chapter in *R Packages* [104]. There are many such coding limitations in this work, in addition to testing. Functional programming is only engaged with at a cursory level, too.

Whilst this is a limitation, this is not the focus of the mansuscript. instead, this dissertation is an exploration of the minimal tools required for reproducible computing for scientific research. Rather than a comprehensive dive of computational tools and techniques, the manuscripts in this thesis are largely driven by the following question.

> How to practice good enough computational science?

Here we invoke the idea of *good enough* scientific computing [111], as opposed to aiming for an unattainable best practice. After all, researchers are, by and large, not experts in computer science, but those from other domains who wish to practice computing to futher, to borrow from Devezer *et al.*, the process of scientific discovery [26].

There is a seemingly infinite world of programming resources, but a research software engineer is not a software developer in the conventional sense. Some aspects, such as data handling, need to be especially robust, but other aspects of quality assurance, for example, the extent of unit tests, are arguably lower for a research context.

7.4 Mathematistry

There are a couple of examples in this manuscript that I believe need to be reworked before publication. Also, I believe there's a fundamental conflation of the ideas of order and heuristic, and that both would require further clarification.

7.5 Reproducibility and the nature of interdisciplinary work

Limitations in this dissertation abound. Less consideration is given to mathematical statistics than to a single-disciplinary manuscript of mathematical statistics; this is an inevitable consequence of interdisciplinary work. With this in mind, this dissertation has endeavoured to avoid engaging with the more complex offerings of any one discipline, but to explore an achievable, for any mathematical scientist, workflow for answering questions. There are many things to learn, in each of the disciplines engaged with. However, this dissertation is also representative of the first generation of researchers where software engineering has increasingly dominated research practice across a multitude of disciplines, such as archaeology [67], or, for a specific example of interdisciplinary computational work, `faux::` [25], for simulating data with particular structures, such as factorial designs, which are common in experimental psychology.

Here is what I learnt.

Open scientific practice facilitates a shift of emphasis from solution to framing. It is arguably more useful to provide a protocol by which scientists may understand the problem (with, say, a package website with vignette), such as provided by `parameterpal::`, and contribute to the solution, via a GitHub online code repository[2]. Given a goal and a specific set of conditions and experience, it is easy to discern a problem. Via open science practice, one can crowd source a solution efficiently. This framing has scientific utility, but comes at a cost of time and education, just as with clarity of mathematical argument. If this dissertation convinces the reader that the cost, that is to say, investing in education and valuing time spent on research software engineering is worthwhile for the practitioner, and for furthering the process of scientific discovery, then this manuscript has achieved its aim.

In Chapter 5, a somewhat incorrect, but functionally adequate, solution is provided for simulating the proportion allocated to intervention group in a study. As noted in Section 5.4.2, after sharing `parameterpal::` openly, an incorrect assumption was picked up by Daniel Oberski, an academic in the Netherlands. Via gist, he provided an improved mathematical solution, which was incorporated into the code. Yanina Saibene, an agricultural academic in Argentina, further developed the package by converting the vignette an interactive tutorial. Framing the question clearly, how it was solved, and facilitating others' solutions is arguably an essential skill for contemporary computational research.

From different perspectives, this dissertation advocates for a reorientation of mathematical science such that we recognise the challenge and time consuming nature of reproducible

[2] https://github.com/softloud/parameterpal

computation, as well as the value of research software engineering for good enough scientific practice.

Bibliography

[1] Valentin Amrhein, Sander Greenland and Blake McShane. 'Scientists rise up against statistical significance'. In: *Nature* 567.7748 (Mar. 2019), p. 305.

[2] David Auburn. *Proof: A Play*. Farrar, Straus and Giroux, 5th Mar. 2001. 99 pp. ISBN: 978-0-571-19997-6.

[3] Frederik Aust. *Citr: 'RStudio' add-in to insert markdown citations*. 2018.

[4] Lee J. Bain and Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. Brooks/Cole Cengage Learning, 1992. 644 pp. ISBN: 978-0-534-38020-5.

[5] Andrew Belmonte. 'The tangled web of self-tying knots'. In: *Proceedings of the National Academy of Sciences* 104.44 (30th Oct. 2007), pp. 17243–17244. ISSN: 0027-8424, 1091-6490.

[6] Yves Bertot. 'A short presentation of Coq'. In: *International Conference on Theorem Proving in Higher Order Logics*. 2008, pp. 12–16.

[7] Martin Bland. 'Estimating Mean and Standard Deviation from the Sample Size, Three Quartiles, Minimum, and Maximum'. In: *International Journal of Statistics in Medical Research* 4.1 (27th Jan. 2014), pp. 57–64–64. ISSN: 1929-6029.

[8] Mark Blokpoel et al. 'Deep Analogical Inference as the Origin of Hypotheses'. In: 11 (2018), p. 24.

[9] Michael Borenstein. *Introduction to metaanalysis*. JSTOR, 2008.

[10] Michael Borenstein et al. *Introduction to Meta-Analysis*. John Wiley & Sons, 24th Aug. 2011. 434 pp. ISBN: 978-1-119-96437-7.

[11] Jorge Luis Borges and Anthony Boucher. 'The garden of forking paths'. In: *Ellery Queen's mystery magazine.* 12.57 (1948).

[12] George E. P. Box. 'Science and Statistics'. In: *Journal of the American Statistical Association* 71.356 (Dec. 1976), pp. 791–799. ISSN: 0162-1459, 1537-274X.

[13] Charlotte Brontë. *jane eyre*. OUP Oxford, 2000.

[14] Stacy Brown. 'Partial unpacking and indirect proofs: A study of students' productive use of the symbolic proof scheme'. In: *Proceedings of the 16th Annual Conference on Research in Undergraduate Mathematics Education*. Vol. 2. 2013, pp. 47–54.

[15] Jennifer Bryan. 'Excuse Me, Do You Have a Moment to Talk About Version Control?' In: *The American Statistician* 72.1 (2nd Jan. 2018), pp. 20–27. ISSN: 0003-1305.

[16] Colin F. Camerer et al. 'Evaluating replicability of laboratory experiments in economics'. In: *Science* 351.6280 (Mar. 2016), pp. 1433–1436. ISSN: 0036-8075, 1095-9203.

[17] Lewis Carroll. *The Annotated Alice: The Definitive Edition*. Ed. by Martin Gardner. Updated, Subsequent edition. New York: W. W. Norton & Company, 17th Nov. 1999. 352 pp. ISBN: 978-0-393-04847-6.

[18] Sheldon J. Chow. 'Many Meanings of 'Heuristic''. In: *The British Journal for the Philosophy of Science* 66.4 (1st Dec. 2015), pp. 977–1016. ISSN: 0007-0882.

[19] Open Science Collaboration. 'Estimating the reproducibility of psychological science'. In: *Science* 349.6251 (28th Aug. 2015). ISSN: 0036-8075, 1095-9203.

[20] Mia Consalvo. 'Zelda 64 and Video Game Fans: A Walkthrough of Games, Intertextuality, and Narrative'. In: *Television & New Media* 4.3 (1st Aug. 2003), pp. 321–334. ISSN: 1527-4764.

[21] Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. New York: Springer-Verlag, 2008. ISBN: 978-0-387-75970-8.

[22] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 18th Apr. 2002. 316 pp. ISBN: 978-0-521-78451-1.

[23] Brian A. Davey. *When is a Proof?* 2nd ed. La Trobe University, 2009.

[24] Brian A. Davey, Charles T. Gray and Jane G. Pitkethly. 'The Homomorphism Lattice Induced by a Finite Algebra'. In: *Order* 35.2 (1st July 2018), pp. 193–214. ISSN: 1572-9273.

[25] Lisa DeBruine. *faux: Simulation for factorial designs*. manual. Zenodo, Sept. 2020.

[26] Berna Devezer et al. 'Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity'. In: *PLOS ONE* 14.5 (15th May 2019), e0216125. ISSN: 1932-6203.

[27] David L. Donoho. 'An invitation to reproducible computational research'. In: *Biostatistics* 11.3 (July 2010), pp. 385–388. ISSN: 1465-4644.

[28] Fiona Fidler and John Wilcox. 'Reproducibility of Scientific Results'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2018. Metaphysics Research Lab, Stanford University, 2018.

[29] *Firefly*. In collab. with Joss Whedon et al. 20th Sept. 2002.

[30] Denae Ford et al. 'Paradise Unplugged: Identifying Barriers for Female Participation on Stack Overflow'. In: *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. FSE 2016. New York, NY, USA: ACM, 2016, pp. 846–857. ISBN: 978-1-4503-4218-6.

[31] Hannah Fraser et al. 'Questionable research practices in ecology and evolution'. In: *PLOS ONE* 13.7 (16th July 2018), e0200303. ISSN: 1932-6203.

[32] Hannah Fraser et al. 'The role of replication studies in ecology'. In: *Ecology and Evolution* n/a (n/a 2020). ISSN: 2045-7758.

[33] Andrew Gelman and Eric Loken. 'The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time'. In: (2013), p. 17.

[34] Andrew Gelman et al. 'Bayesian Workflow'. In: *arXiv:2011.01808 [stat]* (3rd Nov. 2020). arXiv: 2011.01808.

[35] Sandra M Gilbert and Susan Gubar. *The madwoman in the attic: The woman writer and the nineteenth-century literary imagination*. Yale University Press, 1980.

[36] Elise Gould. *Elise Gould: Questionable Research Practices in non-hypothesis testing research: ecological models for conservation decision-making*. 2019. URL: https://www.metascience2019.org/poster-session/elise-gould/ (visited on 05/05/2020).

[37] Matthew Grainger et al. 'Maximising the leverage of existing knowledge could reduce research waste in applied ecology and conservation'. 20th Sept. 2019.

[38] Charles T. Gray. ‘code::proof: Prepare for Most Weather Conditions’. In: *Statistics and Data Science*. Communications in Computer and Information Science (2019). Ed. by Hien Nguyen, pp. 22–41.

[39] Charles T. Gray and Ben Marwick. ‘Truth, Proof, and Reproducibility: There’s No Counter-Attack for the Codeless’. In: *Statistics and Data Science*. Ed. by Hien Nguyen. Communications in Computer and Information Science. Singapore: Springer, 2019, pp. 111–129. ISBN: 9789811519604.

[40] Garrett Grolemund and Hadley Wickham. *R for Data Science*. 2017.

[41] Susan Haack. *Defending Science - within Reason: Between Scientism And Cynicism*. Prometheus Books, Mar. 2011. ISBN: 978-1-61592-168-3.

[42] Hadley Wickham. ‘Tidy Data’. In: *Journal of Statistical Software* 59.1 (12th Sept. 2014), pp. 1–23. ISSN: 1548-7660.

[43] Alex Hayes. *testing statistical software - aleatoric*. 6th July 2019. URL: https://www.alexpghayes.com/blog/testing-statistical-software/ (visited on 08/06/2019).

[44] Megan L. Head et al. ‘The Extent and Consequences of P-Hacking in Science’. In: *PLOS Biology* 13.3 (Mar. 2015), e1002106. ISSN: 1545-7885.

[45] Pavol Hell and Jaroslav Nesetril. *Graphs and Homomorphisms*. Oxford Lecture Series in Mathematics and Its Applications. Oxford, New York: Oxford University Press, 22nd July 2004. 260 pp. ISBN: 978-0-19-852817-3.

[46] Jim Hester. ‘covr: Bringing test coverage to R’. 9th Jan. 2016.

[47] Jim Hester. *covr: Test Coverage for Packages*. 2018.

[48] Michael Hopkin. ‘Palaeontology journal will ’fuel black market’’. In: *Nature* 445 (17th Jan. 2007), pp. 234–235. ISSN: 1476-4687.

[49] Stela Pudar Hozo, Benjamin Djulbegovic and Iztok Hozo. ‘Estimating the mean and variance from the median, range, and the size of a sample’. In: *BMC Medical Research Methodology* 5.1 (20th Apr. 2005), p. 13. ISSN: 1471-2288.

[50] Michael Hüttermann. *DevOps for Developers*. Apress, 24th Oct. 2012. 183 pp. ISBN: 978-1-4302-4570-4.

[51] Leslie John, George Loewenstein and Drazen Prelec. ‘Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling - Leslie K. John, George Loewenstein, Drazen Prelec, 2012’. 2012.

[52] Franz Kafka. *The Trial*. Trans. by David Wyllie. 1st Apr. 2005.

[53] Daniel S. Katz and Kenton McHenry. *Super RSEs: Combining research and service in three dimensions of Research Software Engineering*. Daniel S. Katz's blog. 12th July 2019. URL: https://danielskatzblog.wordpress.com/2019/07/12/ (visited on 16/07/2019).

[54] Os Keyes, Josephine Hoy and Margaret Drouhard. 'Human-Computer Insurrection: Notes on an Anarchist HCI'. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. the 2019 CHI Conference. Glasgow, Scotland Uk: ACM Press, 2019, pp. 1–13. ISBN: 978-1-4503-5970-2.

[55] George Klees et al. 'Evaluating Fuzz Testing'. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. CCS '18. New York, NY, USA: ACM, 2018, pp. 2123–2138. ISBN: 978-1-4503-5693-0.

[56] E. Kulinskaya, S. Morgenthaler and R. G. Staudte. *Meta analysis: a guide to calibrating and combining statistical evidence*. Vol. 756. John Wiley & Sons, 2008.

[57] Johan Kwisthout, Todd Wareham and Iris van Rooij. 'Bayesian Intractability Is Not an Ailment That Approximation Can Cure'. In: *Cognitive Science* 35.5 (1st July 2011), pp. 779–784. ISSN: 0364-0213.

[58] Imre Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery*. Reissue edition. Cambridge: Cambridge University Press, 8th Oct. 2015. 196 pp. ISBN: 978-1-107-53405-6.

[59] R. J. LeVeque, I. M. Mitchell and V. Stodden. 'Reproducible research for scientific computing: Tools and strategies for changing the culture'. In: *Comput Sci Eng* 14 (2012).

[60] James (JD) Long and Paul Teetor. *R Cookbook, 2nd Edition*. 2019.

[61] Matthew C. Makel et al. 'Questionable and Open Research Practices in Education Research'. preprint. preprint. 31st Oct. 2019.

[62] Review Manager. *RevMan 5*. 2019.

[63] Review Manager. *RevMan Web*. 2019.

[64] Per Martin-Löf. 'Constructive Mathematics and Computer Programming'. In: *Studies in Logic and the Foundations of Mathematics*. Ed. by L. Jonathan Cohen et al. Vol. 104. Logic, Methodology and Philosophy of Science VI. Elsevier, 1st Jan. 1982, pp. 153–175.

[65] Ben Marwick. *rrtools: Creates a reproducible research compendium*. 2018.

[66] Ben Marwick, Carl Boettiger and Lincoln Mullen. *Packaging data analytical work reproducibly using R (and friends)*. e3192v2. PeerJ Inc., 20th Mar. 2018.

[67] Ben Marwick and Sophie Schmidt. 'Tool-driven Revolutions in Archaeological Science'. In: *preprint*. preprint (3rd Jan. 2019).

[68] Miles McBain and Jonathan Carroll. *Datapasta: R tools for data copy-pasta*. 2018.

[69] Richard McElreath. *Statistical rethinking: A bayesian course with examples in R and stan*. CRC Press, 2016.

[70] Robert K. Merton. *On Social Structure and Science*. University of Chicago Press, Sept. 1996. ISBN: 978-0-226-52071-1.

[71] Charles Murray. 'How to Accuse the Other Guy of Lying with Statistics'. In: *Statistical Science* 20.3 (2005), pp. 239–241. ISSN: 0883-4237.

[72] Danielle Navarro. *Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection*. preprint. PsyArXiv, 26th Oct. 2018.

[73] Danielle Navarro. *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6.1)*. 2019.

[74] Danielle Navarro. *Paths in strange spaces*. 2019.

[75] Brian A. Nosek et al. 'Preregistration Is Hard, And Worthwhile'. In: *Trends in Cognitive Sciences* 23.10 (1st Oct. 2019), pp. 815–818. ISSN: 1364-6613.

[76] Anna Nowogrodzki. 'How to support open-source software and stay sane'. In: *Nature* 571 (July 2019), p. 133.

[77] Gary W. Oehlert. 'A Note on the Delta Method'. In: *The American Statistician* 46.1 (1992), pp. 27–29. ISSN: 0003-1305.

[78] Centre for Open Science. *Preregistration*. 2020. URL: https://cos.io/prereg/ (visited on 28/12/2019).

[79] Hilary Parker. 'Opinionated analysis development'. In: *preprint* (2017).

[80] R. D. Peng. 'Reproducible Research in Computational Science'. In: *Science* 334.6060 (Dec. 2011), pp. 1226–1227. ISSN: 0036-8075, 1095-9203.

[81] Andrew Pickering. *The mangle of practice: Time, agency, and science*. University of Chicago Press, 2010.

[82] Melina de Barros Pinheiro et al. 'D-dimer in preeclampsia: Systematic review and meta-analysis'. In: *Clinica Chimica Acta* 414 (24th Dec. 2012), pp. 166–170. ISSN: 0009-8981.

[83] C. Ragkhitwetsagul et al. 'Toxic Code Snippets on Stack Overflow'. In: *IEEE Transactions on Software Engineering* (2019), pp. 1–1. ISSN: 0098-5589.

[84] Jean Rhys. *Wide sargasso sea*. WW Norton & Company, 1992.

[85] David Robinson and Alex Hayes. *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. 2019.

[86] Francisco Rodriguez-Sanchez et al. 'Ciencia reproducible: qué, por qué, cómo'. In: *Revista Ecosistemas* 25.2 (16th July 2016), pp. 83–92–92. ISSN: 1697-2473.

[87] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 25th Sept. 2009. 399 pp. ISBN: 978-0-470-31719-8.

[88] Steven Shapin and Simon Schaffer. *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life (New in paper)*. Vol. 32. Princeton University Press, 2011.

[89] Jiandong Shi et al. 'How to estimate the sample mean and standard deviation from the five number summary?' In: *arXiv preprint arXiv:1801.01267* (2018).

[90] Dan P. Simpson. *What if it's never decorative gourd season? « Statistical Modeling, Causal Inference, and Social Science*. 2019. URL: https://statmodeling.stat.columbia.edu/2019/11/13/what-if-its-never-decorative-gourd-season/ (visited on 14/11/2019).

[91] skyisup. *Deadly Boss Mods Addon Guide*. Wowhead. 2019. URL: https://www.wowhead.com/deadly-boss-mods-addon-guide (visited on 08/10/2019).

[92] Morten Heine Sørensen and Pawel Urzyczyn. *Lectures on the Curry-Howard isomorphism*. Vol. 149. Elsevier, 2006.

[93] Victoria Stodden. *What scientific idea is ready for retirement?* 2014. URL: https://www.edge.org/response-detail/25340.%202014..

[94] Victoria Stodden, Jonathan Borwein and David H. Bailey. '"Setting the Default to Reproducible" in Computational Science Research'. In: *SIAM News* 46.5 (2013).

[95] Victoria Stodden and Sheila Miguez. 'Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research'. In: (6th Sept. 2013).

[96] Aba Szollosi and Chris Donkin. *Arrested theory development: The misguided distinction between exploratory and confirmatory research*. preprint. PsyArXiv, 21st Sept. 2019.

[97] Aba Szollosi et al. *Preregistration is redundant, at best*. preprint. PsyArXiv, 31st Oct. 2019.

[98] UHS. *Universal Hint System: Not your ordinary walkthrough. Just the hints you need.* 2019. URL: http://www.uhs-hints.com/ (visited on 13/08/2019).

[99] Wolfgang Viechtbauer. 'Conducting meta-analyses in R with the metafor package'. In: *Journal of Statistical Software* 36.3 (2010), pp. 1–48.

[100] Eric-Jan Wagenmakers. *A Breakdown of "Preregistration is Redundant, at Best"*. Bayesian Spectacles. 5th Nov. 2019. URL: https://www.bayesianspectacles.org/a-breakdown-of-preregistration-is-redundant-at-best/ (visited on 09/11/2019).

[101] Joshua D. Wallach, Kevin W. Boyack and John P. A. Ioannidis. 'Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017'. In: *PLOS Biology* 16.11 (Nov. 2018), e2006930. ISSN: 1545-7885.

[102] Xiang Wan et al. 'Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range'. In: *BMC Medical Research Methodology* 14.1 (19th Dec. 2014), p. 135. ISSN: 1471-2288.

[103] Martin Westgate et al. *metaverse: Workflows for evidence synthesis projects*. 2019.

[104] H. Wickham. *R Packages: Organize, Test, Document, and Share Your Code*. O'Reilly Media, 2015. ISBN: 978-1-4919-1056-6.

[105] Hadley Wickham. *Advanced R*. 1 edition. Boca Raton, FL: Routledge, 27th Sept. 2014. 478 pp. ISBN: 978-1-4665-8696-3.

[106] Hadley Wickham. *testthat: Get Started with Testing*. 2011.

[107] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*. 2017.

[108] Hadley Wickham and Jennifer Bryan. *usethis: Automate Package and Project Setup*. 2019.

[109] Hadley Wickham, Peter Danenberg and Manuel Eugster. *Roxygen2: in-line documentation for R*. 2019.

[110] Greg Wilson et al. 'Best Practices for Scientific Computing'. In: *PLoS Biology* 12.1 (7th Jan. 2014). Ed. by Jonathan A. Eisen, e1001745. ISSN: 1545-7885.

[111] Greg Wilson et al. 'Good enough practices in scientific computing'. In: *PLOS Computational Biology* 13.6 (22nd June 2017). Ed. by Francis Ouellette, e1005510. ISSN: 1553-7358.

[112] D.H. Wolpert and W.G. Macready. 'No free lunch theorems for optimization'. In: *IEEE Transactions on Evolutionary Computation* 1.1 (Apr. 1997), pp. 67–82. ISSN: 1089778X.

[113] Claire Wyatt. *Research Software Engineers Association*. s. 2019. URL: https://rse.ac.uk/ (visited on 16/07/2019).

[114] Achim Zeileis. 'CRAN Task Views'. In: *R News* 5.1 (2005), pp. 39–40.