Stratified sequential nonparametrics: inferential validity by design, any way you slice it

By
Jacob Spertus

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Statistics
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Philip Stark, Chair
Professor Avi Feller
Professor Sam Pimentel

Spring 2024

Stratified sequential nonparametrics: inferential validity by design, any way you slice it

Abstract

Stratified sequential nonparametrics: inferential validity by design, any way you slice it

by

Jacob Spertus

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Philip Stark, Chair

Modern statistical practice has taken an Icarian flight in its embrace of model-based inference. Models are underwritten by sophisticated assumptions about the origins of data, involving hypothetical populations that conveniently follow parametric distributions. These assumptions are abstract and often demonstrably false, providing ample grounds for skepticism of model-based findings. Models are also obscure—requiring a high degree of mathematical sophistication to understand and interpret—which serves to preclude deliberation between researchers, prevent scrutiny by diverse stakeholders, and obfuscate underlying normative values and possible weaknesses. They support the gathering of professional statisticians and scientists into a priestly class, empowered to steer a technocratic state through the appearance of knowing. They do not support a healthy science—one that knows its limits, ascertains the truths it can, and earns public trust.

The design-based philosophy of statistics might do better. In design-based theory and practice, emphasis is placed on the physics of how the data were collected. Hypotheses are posited in terms of sharply defined, real-world quantities and all assumptions necessary to link the data to those hypotheses flow from the design. The assumptions are generally simple and justified so that the output is rigorous and transparent. Those qualities help ensure conclusions are usually true. They also support inter-subjective belief (i.e., trust) that a given conclusion is true. Moreover, the design-based view clearly circumscribes the kinds of problems that are amenable to rigorous statistics: those with a known or sharply hypothetical (as-if) design. If widely adopted, design-based statistical thinking may engender the circumspection and humility currently lacking under the influence of model-based data science.

This dissertation develops methods for design-based statistics. The chapters are particularly focused on design-based inference from surveys and experiments, motivated by applications in risk-limiting post-election audits (Chapters V, VI, and VII) and soil carbon sequestration (Chapters II, III, and IV). Risk-limiting audits

are fundamentally survey problems. They map a sharp question of interest—who won this contest?—to a collection of null hypotheses about the means of lists of bounded numbers, which represent populations derived from cast ballots. Providing rigorous and transparent evidence that reported election results are accurate is critical to supporting trustworthy elections. Adherence to the design-based paradigm when developing and implementing risk-limiting audits ensures such evidence can be furnished.

The science of soil carbon sequestration involves a large array of statistical problems that can be classified as either surveys (e.g., carbon stock measurement) or experiments (e.g., management experiments) and handled by a design-based approach. Often, studies involve a survey (sampling soil cores from plots) embedded in an experiment (randomly assigning plots to treatment), or vice versa. Soil carbon sequestration is a trending topic for its hypothesized potential to offset emissions and mitigate climate change. Failure to rigorously measure sequestration and provide transparent evidence of its efficacy could squander resources, cause shortfalls in emissions reductions, and shake public confidence in coordinated efforts to fight climate change. The regular practice of design-based statistics in soil science could support effective action and accurate carbon budgets.

The technical emphasis of this dissertation falls on valid inference in the presence of two major design elements: sequential sampling (Chapters IV, V, VI, VII) and stratification (Chapters III, VI, VII). Sequential sampling is a natural, necessary, or expedient feature of many real-world data collection procedures. While most traditional inference procedures compute a single inferential statistic (e.g. a P-value) on a batch of n data points, a sequential procedure returns a valid statistic at any time during an iterative process of sampling (e.g., one-at-a-time as each data point comes, or periodically as rounds of data are collected). Sequential analysis thus allows data collection to expand as needed until there is sufficient evidence to draw a conclusion about a hypothesis of interest. We leverage various old and new ideas from probability, game theory, and statistics in developing and implementing efficient methods for sequential analysis. Chapter IV suggests some uses in soil carbon measurement, especially for adaptive experimental designs. In Chapter V we use the theory of Kelly optimality to develop efficient sequential tests for risk-limiting comparison audits. By minimizing the expected number of ballots needed to confirm the winner(s) of a contest, the tests reduce the cost of implementing risk-limiting audits. In Chapter VI we compare sequential tests constructed from betting test supermartingales to tests constructed from exponential test supermartingales. We find the former to be more efficient for risk-limiting comparison audits. Chapter VII builds on and generalizes Chapter VI, proposing new definitions of optimality and constructing sequential tests for population means when the population is also stratified.

Stratification is widely used in design-based statistics to accommodate logistical constraints, increase statistical efficiency, and lower the costs of sampling. Stratification entails partitioning a population into disjoint strata and drawing some number of samples from each stratum uniformly, with or without replacement, and independently across strata. Traditionally, a batch sample of fixed-size n_k is drawn from stratum k and inference on the population mean proceeds using Gaussian theory and finite-population asymptotics. Chapter III explores this strategy for measuring soil carbon stocks at a single time or verifying stock change over time. Chapter VI constructs finite-sample nonparametric tests for risk-limiting audits that are both stratified and sequential (see above). Chapter VII builds a general framework for sequential stratified testing and develops optimal and efficient tests for the mean of a stratified bounded population. The tests are valid (i) sequentially, (ii) in finite samples, (iii) without parametric assumptions, (iv) under any stratification, (v) and with all probabilities flowing from the design.

To Mim and PopPop.

Contents

List of Figures

List of Tables

Acknowledgements

Thank you to Philip for 6(!) great years of advising. Philip and I worked closely on numerous papers, software, consulting with scientists, and teaching. He showed up for countless hours of collaborative writing and coding, which improved my writing and thinking tremendously. He has also been very supportive of me exploring and pursuing my interests across statistics and science. This PhD was a true apprenticeship—a very special chance to learn through deep collaboration and careful work—for which I am extremely grateful.

Along the way, we had some great collaborations with researchers in soil science including Whendee Silver, Tim Bowles, Paige Stanley, Jessica Chiartas, and Eric Slessarev. Paige, Tim, and Jessica wrote Chapter III with me, while Eric and Whendee helped with Chapter IV. I discovered so much about field and laboratory science from those folks, and am very proud of the work we did together. Mayuri Sridhar and I worked together closely on the math and code for stratified sequential inference (Chapter VI). I also want to shout out Justin Berardino from Orange County elections for working with me, Philip, and Amanda to develop software for simultaneous risk-limiting audits, important work which did not appear in this dissertation.

I'm grateful, too, for generous mentorship in statistics from Sam Pimentel. I really appreciate Sam's thoughtfulness and rare attention not only to careful statistical work, but also to good teaching and mentorship, from which I was fortunate to benefit. The causal inference reading group, led by Sam, Avi, and Peng Ding, was formative in my learning during the first few years of my PhD. I'm also thankful for classes taught by Philip, Sam, Aaditya Guntuboyina, and Jon McAuliffe.

My dear friend Amanda Glazer helped me get through this PhD, especially during the first 4 years. I had a blast collaborating with Amanda on something like 6 papers and 2 teaching jobs. I'll also never forget thrashing out all the first year angst at punk shows, co-charing diversity committee, and our adventures in Davis and Guerneville.

Countless friends supported me through this journey with their light and presence: Walter, Andrew, Andy, Peyten, Jess, Jeanne, Adrian, Ella, Evan, Leigh, Joty, Lana, Liana, Caroline, Damir, Danny, Dan, Tom, Karl, Gina, Imani, Felix, among many others. Thank you for the abundance.

I'm deeply indebted to my ancestors for setting me on my way. All those who came before and crossed, I sense. I carry the water and the flame. Thank you especially to my grandparents, my cousins, my parents, and my siblings, who offer boundless support for my life and pursuits. The love is enormous.

Chapter 1

Introduction

Statistics plays a large and expanding role in knowledge formation and decision making across science, industry, and government. Data are proliferating, as are new techniques for accessing, processing, and interpreting them. In policy making, quantitative data are seen as inherently rigorous or objective and given priority over qualitative evidence, deliberation, or direct experience [Saltelli and Giampietro, 2017]. In research, there is pressure to publish at a high rate and to use pre-existing or easily-attainable data (e.g., from convenience samples) to answer complex (though not always empirically sharp) scientific questions.

These factors have bent statistical theory and practice towards a *model-based* paradigm, wherein distributional assumptions proliferate. For a simple and classical example, a modeler might assume a population of interest was drawn from a Gaussian superpopulation with unknown mean and variance. For a more complex example, the modeler might assume the data came from a population drawn from a hierarchy of various parametric probability distributions, the parameters drawn in turn from known prior distributions or mapped from known functions of independent variables. It is often left ambiguous what exactly the population represents, what the superpopulation represents, why the population can be modeled as a random draw from a superpopulation, or why that superpopulation would have a known parametric form.

The traction of model-based assumptions and the authority of model-based results rests on the inter-subjective accord of researchers within a particular discipline, not on how closely the assumptions track with the physical world. Furthermore, randomness in a hypothetical sampling mechanism is *epistemic*—in the sense that it exists only in the mind, as an expression of belief about the population and where the data came from [Sterba, 2009]. Model-based frequentist statistics shares this use of probability and randomness with Bayesian statistics. Personal probabilities need not be tied to any real sampling mechanism, disconnecting the data and its origins from the theory being tested [Berk and Freedman, 2003]. This can blunt the capacity for deliberation or critique: models can be arbitrarily complex mathematical objects involving various assumptions about physics and randomness that are difficult to reason about or relate to scientific subject-matter.

As a result, the model-based paradigm is conducive to hierarchies of authority. A modeler is not constrained in their ability to weigh in on scientific questions given a set of data, sufficient willingness to make heroic assumptions, and the assent of others in their field, which may be a function of reputation and interpersonal dynamics more than reasoned agreement with the method. When target quantities are vague and modeling assumptions arcane, the interpretations, advice, and predictions of experts who rely on them need not track with reality. In this paradigm, experts may be vested with substantial political power, while lacking the insight necessary to make good decisions. Thus, the model-based paradigm supports technocracy without justifying it: power concentrates while truth does not. The resulting system is both unjust and unstable.

The *design-based* paradigm offers a way out. In design-based statistics, the population is a well-defined (often finite) list of units corresponding to real-world quantities, and the target parameter is a deterministic function of that list (e.g., its mean). Inference and estimation proceed with minimal assumptions on the population; for example, that it is supported on an interval $[a, b]$, with a and b known. Probability assumptions center on the eponymous *design*—the (as-if) random process by which data were drawn from the population—and are *aleatory*, arising from inherent properties of a physical system like the roll of a die. Design-based statistics uses aleatory randomness to ensure that the data are representative of the target population in a structured, probabilistic way,[1] one amenable to meaningful probability statements and long-run error control. The design is either exactly known

[1] There are two other broad uses of aleatory randomness that precede statistics. (i) Mantic experiments (tarot, I Ching, bone divination, etc) use aleatory randomness to create meaning, often as prophecy and in a religious context. While scientific experiments are intended to establish causality, mantic experiments generate meaning, possibly through the interaction between coincidence and the human mind [Diaconis and Mosteller, 1989, Jung, 2010]. Divinatory practices emphasizing expert (i.e., priestly) interpretation over aleatory randomness (haruspicy, augury, etc), may have been replaced by more transparent oracles (e.g., dice) precisely due to suspicion of the mediating human element [Hacking, 2006]. (ii) Games of chance use aleatory randomness to generate excitement. Gambling is thought to be among the first inventions of human society [David, 1955], and provided the basis for the development of probability theory in the 17th century [Hacking, 2006]. The concept of Luck allows aleatory probability to be as personal as epistemic probability. Girolamo Cardano's committed belief in Luck may have delayed an earlier formalization of probability in his 16th century gambling manual *Liber de Ludo Alea* [Gigerenzer et al., 1989, Cardano, 1966 (originally 1525].

to the researcher (as in a probability survey or randomized experiment) or unknown (as in observational causal inference) but requiring relatively simple, transparent, and falsifiable assumptions on the unknown design compared to a more inscrutable model on nature [Rosenbaum, 2002]. This dissertation is about applications in and methods for design-based inference, with an emphasis on problems where the data come from a known design.

Some History

The gulf between model-based and design-based statistics stems from controversies at the foundation of statistics. The earliest statistical analyses—of mortality data in the 17th century—spurred debate over reasonable assumptions on the relationship between age and death. The stakes were high, with substantial financial ramifications for states hinging on the proper price of annuities [Gigerenzer et al., 1989, Hacking, 2006]. The majority of mathematicians at the time—including Graunt, Halley, and De Moivre, all of whom assumed mortality followed a simple arithmetic progression—tended to believe in nature's proclivity for simple and universal laws: their thinking was model-based. The Dutch mathematician Nicholas Struyck raised a rare dissent, warning "nature doesn't listen to our suppositions" [Gigerenzer et al., 1989].

The degree of researcher control over the data generating process plays a key role in determining where the emphasis is placed—on the design or on the model of nature—when making probabilistic assumptions. Without a known design, a hypothetical model is required to make any general claims from the data. Early mortality data was purely observational—coming from convenience samples or censuses—and could not lend itself to the development of a design-based statistics. By the mid-19th century experiments were expanding out of the laboratory and into real-world situations attended by a much higher degree of variability.[2] This was particularly true in agriculture, where aggregate farm yields were of great economic significance to the nation and experiments promised to discover new ways of maximizing production [Gigerenzer et al., 1989]. Johnston [1849] laid out many of the open problems in agricultural science, and in experimental design more generally. He noted the need for comparison, but failed to realize the importance of replication and randomization.

Starting in the late 19th century, British professionals established the groundwork for the expansion of statistical theory and practice throughout the 20th century. Many of the founders were committed eugenicists, interested in furnishing scientific

[2] In the laboratory sciences, it had been possible to experiment on pure substances, theoretically constrain possible causes, and exercise nearly complete control over all potentially relevant conditions [Gigerenzer et al., 1989].

justification for colonial hierarchy during the twilight of the British empire and for a domestic social order empowering a small professional intellectual class over both the traditional working classes and the aristocracy [MacKenzie, 1981]. Common statistical models were developed during this period, with Galton theorizing and employing the bivariate Gaussian distribution and ordinary least squares to study human populations and provide an allegedly empirical basis for eugenicist theories. Pearson later built on multiple regression and the theory of the correlation coefficient to formalize biological arguments for socialist eugenics [Pearson, 1895].

After Pearson, Fisher laid critical groundwork for both the mathematics and applied practice of classical statistics. Working in an agricultural context at Rothamsted, he developed a theory of statistical design that recognized the centrality of replication, randomization, and control [Fisher, 1925, 1935]. He also formulated finite-sample valid methods for nonparametric hypothesis tests, including in the famous "lady tasting tea" experiment presented in Fisher [1935]. Despite his pathbreaking work on experimental design and agnostic analysis, Fisher was also instrumental in cementing the centrality of the model-based view [Lehmann, 2011]. Indeed, Fisher [1922] posits:

> This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters which are sufficient to describe it exhaustively in respect of all qualities under discussion.

Fisher's setup is a succinct summary of the model-based paradigm: a scientific theory is expressible purely in terms of hypothetical populations, and the data to hand furnish evidence for or against the theory via a hypothetical sampling mechanism.

The design-based paradigm was developed primarily by Jerzy Neyman. Neyman [1923] introduced a design-based theory of causal inference, showing that scientific theories can be sharply represented by urns containing a finite amount of fixed numbers (potential outcomes). Randomized experiments draw potential outcomes from the urns at random, and probabilistic inference on well-defined causal parameters (e.g., the average treatment effect) is justified solely by the random design. In the context of surveys, Neyman [1934] presents a strong argument for random sampling and design-based inference over purposive sampling and modeling, then common at the time. Neyman encourages a random sampling design to guarantee that probabilistic assumptions are met and the sample is representative, potentially in conjunction with tools like stratification to accommodate logistical challenges or increase estimation and inferential efficiency.

Contributions of this dissertation

This dissertation is interested in applications and methods in a design-based framework. The applications have a classical flavor: they are surveys and experiments where a finite population of units is sampled or randomly assigned to treatments according to a known design. As a further historical connection, two chapters are applications to agricultural problems, in which the theory of randomized controlled trials was pioneered by Johnston, Fisher, and Neyman. Breaking with their primary concern of maximizing farm yields, our problems pertain to studies of soil carbon sequestration as a climate change mitigation strategy. The other applications are in risk-limiting audits (RLAs) to verify that reported winners of elections really won. RLAs are, fundamentally, instances of survey sampling from non-standard finite populations. The methods we develop solve a very fundamental problem: conducting inference on a population mean with guaranteed validity. In our development, we borrow from both historical work (especially Wald [1945]) and recent advances (especially Waudby-Smith and Ramdas [2023]) in finite-sample, nonparametric, and sequential inference.

Chapter II reviews practical aspects of measuring soil organic carbon stock, including data collection via random field sampling and laboratory assay, both of which are prone to error. It derives a variance-optimal ratio of sampling to assay (a trade-off allowed by compositing and replication), assuming a fixed budget and linear costs. The theory is applied using errors and costs either taken from the literature or computed on data from the Marin Carbon Project. Compositing samples and assaying them by dry combustion is found to minimize estimation variance at a fixed budget.

Chapter III concerns statistical methods to measure soil organic carbon stock and stock change. It is concerned with sources of variability in estimation under simple random and stratified sampling, as well as with finite-sample validity in the weak[3] two-sample problem: wherein the null posits only that two populations share a common mean. The power of tests using simple random or stratified samples is evaluated using soil data gathered from rangeland (Paicines Ranch) and multiple croplands around California.

Chapter IV builds on this work, presenting methods for design-based causal inference about soil carbon sequestration. The chapter reviews practical uncertainties including core-level assay variability, plot-level spatial heterogeneity, study-level variation in potential outcomes, and uncertainties involved in generalizing from a

[3] The strong two-sample problem posits that the two populations are equal in distribution. It is readily solved with permutation tests.

study to a population. It is especially focused on estimation and inference on the population average treatment effect, on identifying moderators that can be used to predict how a plot will respond to treatment, and on making optimal policy decisions for a population of farms.

Chapter V is about optimal sequential inference for risk-limiting comparison audits using betting test supermartingales (TSMs). The chapter maximizes the efficiency of sequential tests using the theory of Kelly optimality in order to reduce workloads for ballot-level comparison audits, which are the most efficient form of risk-limiting audit. It derives the theoretically optimal "oracle" tests, which are not empirically identified, and provides predictable approximations of the oracle tests that can be used in practice, analogous to Stark [2023] on efficient ballot-polling audits.

Chapter VI develops methods for stratified RLAs, wherein the sample of ballots is drawn by a stratified and sequential design and the inference must be anytime valid. The methods leverage the betting TSMs of Waudby-Smith and Ramdas [2023] and the union-of-intersections tests idea proposed in Ottoboni et al. [2018] for sequential stratified inference. They are substantially more efficient than the previous state-of-the-art for stratified RLAs [Ottoboni et al., 2018]. The chapter also presents a computationally scalable risk measure that uses exponential supermartingales [Howard et al., 2021], reducing the union-of-intersections test to a linear program that can accommodate audits with any number of strata.

Chapter VII generalizes and improves on the ideas in Chapter VI, developing a broad theory for stratified sequential testing and providing practical and efficient strategies based on betting supermartingales and union-of-intersections tests. The theory generalizes Kelly-optimality from the sequential to the stratified sequential setting, wherein sequences of bets and a sequence of stratum selections determine the efficiency of a test. We provide an implicit construction of the Kelly-optimal bets under any simple alternative, and an explicit construction for a few parametric simple alternatives (e.g. Bernoulli distributions within strata). Computing the P-value can be challenging. We provide a few ways to implement these strategies in real-world problems and demonstrate their efficiency compared to existing methods.

Chapter 2

Optimal compositing for soil carbon estimation

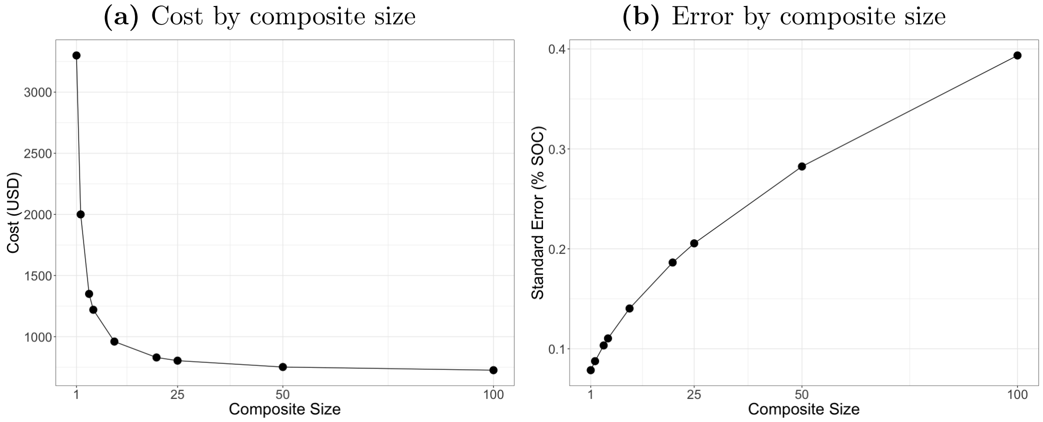
2.1 Introduction

Climate change is likely to put enormous strain on nature and human societies in the coming decades. It is largely driven by the release of atmospheric carbon dioxide (CO_2) that was once sequestered in the earth, either in fossil fuels or as soil carbon. Since the cultivation of soils began, soils have lost about 50-70% of their carbon to the atmosphere. Soil still accounts for the 2nd largest store of carbon on Earth after the ocean, containing about 7.5 times that of the atmosphere [Lal and Stewart, 2018]. However, agriculture is now one of the largest contributors to global carbon emissions.

In pursuit of solutions, a growing movement of farmers and other advocates are highlighting "regenerative agriculture" as a way to make agriculture a net sink, rather than a source, of carbon – drawing CO_2 from the atmosphere and sequestering it in the land as soil organic carbon (SOC). Regenerative agriculture provides a variety of ecosystem services including water use efficiency, biodiversity, and overall soil health. These may be sufficient to support its use, but to pay for regenerative agriculture on the basis of SOC sequestration, decision makers need to know *how much* carbon is sequestered by different strategies.

In order to measure SOC sequestration, at a minimum scientists must be able to measure how much SOC is in a given plot of land at a given point in time. This task is referred to as "SOC stock estimation." Soil scientists accomplish SOC stock estimation by collecting multiple cores of soil from a given plot, preparing/processing the samples, and analyzing (assaying) their SOC concentration by a number of different techniques. SOC is either presented on its own, as a concentration, or it is converted to stock using soil bulk density measured on nearby intact cores. Both sampling and assay of SOC concentration are subject to uncertainties and both become expensive at the volumes necessary to overcome these uncertainties. All else equal, increasing the number of samples and the number of assays will reduce uncertainty while driving up costs. A process called compositing allows investigators to reduce cost by mixing together sampled cores and assaying the mixture(s), but compositing incurs additional error when there is uncertainty in the assay.

Figure 2.1 sketches this trade-off in an example where 100 cores have been collected, and the investigator must now choose how much to composite before assaying the composited samples. Parameters and costs are taken from a survey of California rangelands, detailed later in this paper (Section 2.9). Fig 2.1a shows that, across the range of possible composite sizes, the cost increases by a factor of 7. Correspondingly, Fig 2.1b shows a 5-fold decrease in standard error. Clearly, compositing has substantial implications for both uncertainty and cost.


(a) Cost by composite size

(b) Error by composite size

Figure 2.1: Costs (a) and error (b) associated with estimating SOC concentration across a range of possible composite sizes. Decreasing the size of composites (taking more assays) yields a tradeoff: estimation error will decrease, but costs will increase. These assay costs and error reflect assay of California rangeland topsoils with loss-on-ignition. For details see Sections 2.7 and 2.9, especially Table 2.2. SOC = soil organic carbon; USD = United States Dollars.

In this paper we resolve this trade-off by presenting sampling and assay as an optimization problem. Given a fixed budget, we derive the sampling and assay sizes that minimize estimation uncertainty. Conversely, given a fixed estimation precision we'd like to achieve, we derive the optimal sizes to minimize the budget. The solutions depend on the heterogeneity and mean SOC concentration of the plot(s) under study, the assay error, and the costs associated with sampling and assay.

Our paper is organized as follows. In Section 2.2 we situate our work in the soil science and statistics literature. In Section 2.3 we formalize the objectives of SOC estimation. We then turn to the logistics and statistics of estimation, covering sampling in Section 2.4, compositing in Section 2.5, sample preparation in Section 2.6, and assay in Section 2.7. Section 2.8 contains our main results: optimal sample and assay sizes to maximize precision under budget constraints. Scientists, farmers, and policy-makers can use these results to design their own efficient sampling and compositing strategies. To facilitate practical use of our methods, we demonstrate their use by applying them to data from a soil survey in California in Section 2.9. We derive optimal assay strategies and composite sizes in this setting. Section 2.10 discusses additional nuances, challenges, and extensions of stock estimation, and provides recommendation for practice. All of our work is supported by R software, available at `https://github.com/spertus/soil-carbon-simulations`.

2.2 Other Relevant Literature

As part of this paper we review the components of stock estimation and the processes of sampling, compositing, and assay. We focus on estimating the average concentration of SOC in a plot. In order to make minimal assumptions about the plot under study and for our results to be as general as possible, we take the design-based perspective on estimation. Thus, the model of SOC concentration in the plot is minimal. Specifically, we do not make any assumptions about the spatial distribution of SOC concentration. Inference proceeds from random sampling, while SOC concentration is unknown but fixed. Webster and Lark [2012] and de Gruijter et al. [2006] provide accessible reviews of soil sampling, inference, and optimization from the design-based perspective.

The design-based perspective contrasts with the model-based or “geostatistical” perspective, originally developed to map gold mines [Krige, 1951]. The geostatistical approach to SOC stock estimation conceptualizes SOC content as random or at least well approximated by a random process. Geostatistics is especially useful for estimating an entire function of a soil property, i.e. for mapping. We do not examine the model-based approach in detail here. Diggle and Ribeiro [2007] and de Gruijter et al. [2006] are good references on geostatistics and its applications to natural resource monitoring.

Patil et al. [2011] provides a detailed accounting of the statistics of compositing, and includes an analysis of compositing with additive assay error. The benefits of compositing depend on the relative size of the plot heterogeneity to the assay error. Lark [2012] analyzes properties of various compositing schemes alongside a geostatistical model for spatial variation. The author shows that compositing nearby cores improves the precision of an SOC map, compared to taking a single core at each location. Kosmelj et al. [2001] analyzes compositing alongside a cost model in the context of soil sampling for zinc or calcium, solving an optimization problem for compositing over subplots without considering assay error. In a case study, they found that optimal compositing could reduce costs by around 50% while maintaining estimation precision.

We analyze three laboratory assay methods used to measure SOC concentration in soil samples: loss-on-ignition (LOI), dry combustion in an elemental analyzer (DC-EA), and mid-infrared spectroscopy (MIRS). LOI involves measuring the difference in mass before and after heating samples in a furnace. The heating cooks off the organic matter in the soil – along with an unpredictable amount of "mineral" or structural water. The amount of mass lost can be mapped to the SOC concentration in the sample using ordinary least squares regression [Nayak et al., 2019, De Vos et al., 2005]. DC-EA combusts small aliquots of soil at high temperatures in an elemental analyzer that measures the amount of CO_2 released during the burn. DC-EA machines vary in their specifics, but are generally considered the gold-standard for precise determination of SOC concentration [Nayak et al., 2019, FAO, 2020, Smith et al., 2020]. MIRS assays carbon by shining infrared light on samples and recording the wavelengths absorbed. These wavelengths ("spectra") can then be closely mapped to SOC concentration (determined by DC-EA) using machine learning methods. MIRS requires a considerable upfront investment both in the machinery and in developing a large spectral library that links wavelength signatures to SOC concentrations within a region of interest (e.g. a country or state). MIRS and vis-NIRS could become highly cost-effective assay strategies as prices come down and spectral libraries expand [England and Viscarra Rossel, 2018, Nayak et al., 2019, Wijewardane et al., 2018]. LOI and MIRS are "high-throughput" methods, as many samples can be analyzed quickly and cheaply. However these methods offer less precision than DC-EA, and may be prone to biases.

The core contribution of this paper is similar in spirit to a classical power analysis, which determines how many samples are needed to estimate quantities to within a desired precision or to run a hypothesis test at a desired power. Kravchenko and Robertson [2011] presents basic methods and an application of power analysis to detecting SOC change in tillage experiments. Pringle et al. [2011] derived sample sizes necessary to detect changes in SOC stocks on Australian rangelands. A 2019 report by the Food and Agriculture Organization of the United Nations also includes a section on conducting power analysis [FAO, 2020]. These power analyses do not consider the effects of compositing or assay error, nor do they consider the costs of sampling and assay. In our work we provide a framework to derive optimal composite sizes given a cost model. In the process, we characterize budgets that are needed to achieve reasonable precision when estimating SOC concentration.

There is a precedent for analyzing optimal designs in soil science, but most of this work has been done in the geostatistical literature and generally concerns how to optimally distribute samples given an assumed model. If scientists have access to a reliable variogram describing the spatial distribution of SOC, then the sampling design can be optimized to minimize estimation or prediction variance van Groenigen et al. [1999], Brus et al. [2006]. If SOC exhibits *any* spatial auto-correlation, well-spread random samples can increase efficiency compared to uniform independent random sampling. Traditionally, grid or transect sampling is often used, but these designs may be biased and don't yield accurate standard errors Webster and Lark [2012], Wolter [1984]. Investigators may also use auxiliary variables, like management type, topography, or vegetation, to yield more efficient sampling designs. de Gruijter et al. [2016] presents a recipe to estimate SOC concentration or stock at the farm scale. That paper focuses on reducing costs through an optimally-stratified sampling design, while compositing receives less attention. Other modern design approaches aim to improve spatial coverage or auxiliary variable balance through sophisticated random sampling. Well-spread random samples can be achieved by a kind of nested stratification, as in the generalized random tessellation stratified design [Stevens and Olsen, 2004], or by the cube or local pivotal method, wherein samples repel each other spatially [Tillé and Wilhelm, 2016]. All of these papers seek to optimally distribute sample points and do not account for assay error.

New ways of measuring SOC stocks continue to emerge at a rapid pace, driven by advances in technology and data science. Assay can now be accomplished directly in the field using techniques like mobile infrared spectroscopy, eddy covariance assay, inelastic neutron scattering, and laser-induced breakdown spectroscopy. These techniques tend to involve far more assay error than laboratory analyses [Chatterjee et al., 2009a, Nayak et al., 2019, England and Viscarra Rossel, 2018]. Additionally, an active area of research seeks to combine various assays and remote sensor data using machine learning and geostatistics [Wadoux et al., 2019, Padarian et al., 2019, England and Viscarra Rossel, 2018]. A few of these new technologies do not involve randomly sampling cores, and are thus outside the context of this work. The rest apply readily to framework we present here.

2.3 Estimation Goals

SOC concentration (e.g. percent SOC or grams of SOC per kg of soil) is a (non-random) three-dimensional function in latitude, longitude, and depth. In this paper, we are interested in estimating the average concentration, μ, in a bounded area of land to some fixed depth; or the total stock of SOC $\mathcal{T}$ in the area. Typically, estimation occurs within fixed depth profiles, which can then be aggregated to whole-profile stock or concentration estimates. The equivalent soil mass method provides an important alternative strategy wherein profiles are defined to some predetermined mass, not depth [Wendt and Hauser, 2013].

We follow the convention of estimating concentrations and stocks *within* profiles defined by depth or mass. We thus suppress dependence on depth as we develop our ideas. For concreteness, the reader may imagine we are only discussing top-soil concentration or stock in what follows, though our analysis applies to any profile. We also stress that the maximum depth of the survey is very important. Many physical, chemical, and biological mechanisms can move SOC downward or cause soil loss at depth. Long-term management can impact deep soil SOC, so concentrations and stocks may need to be estimated down to a meter or more to accurately account for the SOC sequestration of different management strategies [Tautges et al., 2019, Luo et al., 2010].

If we are only interested in average concentration, it suffices to estimate μ. If we want to estimate the stock $\mathcal{T}$, we also need the bulk density in grams per cubic centimeter d, the area of the plot in square meters $\mathcal{A}$, and the length of the profile in meters L. Assuming that bulk density is constant within depth, the total amount of carbon within the depth profile is

$$\mathcal{T} \equiv 10^4 \times L \times \mathcal{A} \times \mu \times d.$$

The factor 10^4 includes conversion of %SOC to gram per gram, and bulk density to grams per cubic meter. Different factors may be applied to report SOC in tons per hectare (Mg ha^{-1}).

In reality, SOC is never exactly the same across a study area. The degree of heterogeneity can be expressed as the plot variance, σ_p^2, which is the average squared distance of SOC concentration from the mean μ (for a definition in symbols see Section A.1 in the Appendix). If every point in the plot has the same SOC concentration μ, then $\sigma_p = 0$. On the other hand, if the SOC concentration is highly variable across the plot then σ_p will be large. The maximum value, $\sigma_p = 50$, is attained when half the plot is 0% SOC and the other half is 100% SOC. Along with the assay precision, the plot heterogeneity σ_p allows us to characterize the uncertainty in estimates of μ.

2.4 Sampling

Investigators typically estimate μ by sampling relatively small amounts of soil from the plot under study. Soil samples can be taken using an auger, a corer, or by digging a pit. An augers can mix soil horizons, while with a corer horizons are typically kept distinct. Compaction can occur with either method, which may skew depths or density estimates. Digging a pit and drawing samples from the side may yield the best samples, with clear horizons and no compaction, but is relatively destructive and very labor intensive. In what follows, we typically refer to a distinct (uncomposited) sample as a "core," though in principle it could be drawn by any of the above methods. Taking cores at randomly sampled locations can ensure that estimates of μ are unbiased. In this section, we describe three random sampling approaches that are regularly used in practice: uniform independent random sampling, stratified sampling, and cluster sampling.

Uniform independent random samples (UIRSs) are generated by sampling n points uniformly — no particular locations are favored — and independently — the location of a particular core does not affect the location of any other. In the soil science literature, UIRSs are sometimes equated with "simple random samples" Webster and Lark [2012]. However, in statistics simple random sampling denotes uniform sampling without replacement from a discrete, finite population. We use the more cumbersome UIRS to avoid confusion. Sometimes, plots are conceptually "discretized" by mapping the continuous surface to a fine grid, which then becomes the finite sampling frame so that simple random sampling is equivalent to uniform independent random sampling (UIRSing). UIRSs can provide unbiased estimates of μ no matter how SOC is distributed in the plot. UIRSs also yield unbiased estimates of the heterogeneity σ_p. This allows researchers to characterize the precision of the estimate and thus to conduct hypothesis tests or construct confidence intervals based on a UIRS.

Stratified sampling can be used to take advantage of auxiliary information about the distribution of SOC, which can yield more precise estimates. For example, in rangeland the distribution of SOC may be driven by topography, vegetation type, mineralogy, microclimates, or land-use history [Pringle et al., 2011, Webster and Lark, 2012]. Strata and sample sizes per strata can be selected using algorithms that predict SOC concentrations in order to maximize the expected precision given a fixed overall sample size [de Gruijter et al., 2016]. Like UIRSs, stratified samples can yield unbiased estimates of μ, σ_p, and the variance of estimators.

Finally, cluster random samples are drawn by first choosing a point at random and then deterministically sampling along a regular transect or grid extending from the original point. Cluster random samples with a single random starting point are sometimes called "systematic random samples" in the soil science literature [de Gruijter et al., 2006]. Cluster random samples have the advantage of automatically distributing samples evenly across part of a plot. Logistically, this makes samples relatively easy to collect, since cores can be efficiently taken by moving regular distances along the transect or grid. Statistically, this reduces the variance of sample means from cluster random samples when SOC is positively correlated in space, a standard geostatistical assumption. However, sample means from cluster random samples are not inherently unbiased and do not have a simple variance. Both of these properties depend on further assumptions about how SOC is distributed within the plot [de Gruijter et al., 2006, Webster and Lark, 2012, Wolter, 1984]. If these assumptions are not met, cluster random samples may yield biased or imprecise estimates. Periodicity of the property under study (due to row cropping, for example) can lead to poor inferences.

In this paper, we assume that cores are gathered by UIRSing. This makes our results quite general, and covers the wide range of applied cases where UIRSing is used. Furthermore, the variance of sample means from a stratified or cluster sample is typically *lower* than that of a UIRS — lower variance is the main reason why more sophisticated designs are used. Thus our results can be interpreted as a providing an upper bound on the uncertainty of these other sampling designs. Finally, if we assume that SOC is distributed completely randomly in a given plot (i.e. with no spatial correlation), then the properties of estimates based on a UIRS are equivalent to those based on stratified or cluster random sampling.

There are, however, certain land types or surveys where UIRSing can be logistically infeasible. For example, in row crop studies, only treated rows can be sampled, which is typically much easier to achieve using cluster sampling. Furthermore, note that there is a logistically optimal way to collect n cores by UIRSing. First, sample all n points from the plot, find the shortest path through all n points, and move along that path collecting cores at the sampled points. This is called the "traveling salesman problem" in computer science. The length of the shortest path through a UIRS of size n generated in a plot of area $\mathcal{A}$ tends to be about $0.72\sqrt{n\mathcal{A}}$ [Arlotto and Steele, 2016]. Even compared to this shortest path, cluster random samples can have much shorter paths: a transect sample for a rectangular $a \times b$ plot is no longer than $\sqrt{a^2 + b^2}$ for any n. For example, in the experiments conducted by [Tautges et al., 2019] the plots are 64×64 meters and 10 cores were collected per plot. $\mathcal{A} = 4096$ square meters and the shortest path through $n = 10$ randomly generated points is expected to be about 146 meters. On the other hand, a transect through such a plot is about 91 meters. This makes the transect path length only 60% of the expected length of the best UIRS path.

2.5 Compositing

Compositing is the practice of combining cores together from a particular profile in order to capture variability in the plot while reducing assay costs. Where we call n the number of cores, sampled from the field, k is the number of samples left after compositing. Edge cases are $n = k$, when we do no compositing, and $k = 1$, when we composite down to one sample. We assume here that each composited sample is comprised of equal proportions of the constituent cores. We also assume that n is divisible by k and that each composited sample is comprised of exactly n/k cores. For example, we might take a UIRS of $n = 30$ cores from a plot and composite down to $k = 6$ composited samples of size $n/k = 5$ constituent cores. We also assume that samples are perfectly homogenized after compositing, so that equal parts of constituent samples are present in any given aliquot of the composited sample. Perfect homogenization may be difficult to achieve in some types of soils, like soils with high clay content that tend to clod, which can compromise the validity of compositing. Our final assumption is compositing additivity, which implies that the SOC concentration in a composited sample is equal to the mean SOC concentration of its constituent cores. Compositing additivity is met for SOC, but not for other properties like pH, which needs to be considered if investigators plan to measure such properties using the same samples.

There are two reasons why more compositing is not always better. First, assay error leads to (hopefully) unbiased but still variable assays, which needs to be reduced by assaying multiple cores or else by assaying a single core multiple times. Second, compositing is itself an error prone process. It can be very difficult to achieve exactly equal proportions and perfect homogenization, especially in heavy clay soils. These challenges can be alleviated and the errors are hedged by assaying more, smaller composite samples. Finally, in order to do inference we typically need to estimate the plot heterogeneity σ_p, which can only be estimated when $k \geq 2$, a topic we return to in Section 2.8.3.

Logistically, compositing is almost always done in the field to reduce the labor of transporting all n cores to the laboratory. A drawback is that it may be more difficult to achieve good homogenization in field using crude tools on field-moist soil. Furthermore, it is generally important to composite at random. If nearby cores are composited together, which can arise naturally if compositing is done sequentially along a transect or shortest UIRS path, the properties of the sample variance of composited samples may be different. For example, suppose that nearby points tend to have similar SOC concentrations and that nearby points are systematically composited together. In this case the sample variance of composited samples of nearby samples will underestimate σ_p^2, which will lead to over-optimistic conclusions about the precision of an estimate of μ [Patil et al., 2011].

2.6 Preparing Samples for Assay

Sample preparation affects both the cost and precision of estimates of μ, and generally depends on the assay method (see Table 2.1). For dry combustion in an elemental analyzer (DC-EA), samples must be air dried at room temperature. For loss-on-ignition (LOI), samples should be dried in an oven at 105 degrees Celsius, as they must completely dry. The composition of the soil can also determine the proper drying temperature. Salts present in some soils will hold onto water at temperatures higher than 105 degrees, so Chatterjee et al. [2009a].

After drying, samples are passed through a 2mm sieve, which helps remove large bits of organic material (e.g. large roots) and rock. Nevertheless, it can be challenging to differentiate between aggregates and rocks, and to make sure that all $>$ 2mm aggregate material makes it through the sieve. In particular, some soils are too hard once they dry and must be broken up with a mortar and pestle before they can be sieved. Roots may also be picked out by hand. Some studies aim to isolate and separately quantify root fractions. Furthermore, when comparing plots (e.g. in an experiment), carbon in roots can overshadow differences in SOC content [Ryals et al., 2014, FAO, 2020].

After drying, samples are ground to a fine powder (e.g. in a ball mill), which helps ensure homogenization and accurate assay. MIRS can be very sensitive to the size and uniformity of the grind [England and Viscarra Rossel, 2018]. On the other hand, LOI does not require soils to be ground.

Finally, many elemental analyzers (EAs) used for DC-EA cannot distinguish between SOC and soil inorganic carbon (e.g. carbonates). For such machines, assays give the concentration of *total* carbon, not just organic carbon. Soils must be checked in advance for inorganic carbon before assay. If the pH is greater than 7.4, ground samples may be treated with hydrochloric acid to remove carbonates [Nayak et al., 2019]. Methods like LOI don't get hot enough to combust carbonates, while MIRS can usually distinguish between organic and inorganic carbon in spectra.

2.7 Assay

In this section, we review the three major methods for assaying SOC concentration before introducing the concept of assay error. For more details on these assay methods, as well as newer *in situ* methods see the recent reviews by Nayak et al. [2019] and England and Viscarra Rossel [2018].

DC-EA is the gold standard for SOC assay. EAs are expensive to purchase, maintain, and run, but they measure carbon directly and at a fairly high throughput. EAs combust aliquots of soil at high temperatures (around 1000° C) in a pure oxygen environment, and assay the CO_2 released using gas chromatography. DC-EA is generally the most precise and expensive assay method for SOC, but the precision and cost per analysis will vary by EA model.

To assay a sample by LOI, investigators measure the mass of dried soil samples, bake them at around 550° celsius in a muffle furnace, and then measure how much mass was lost during baking [Chatterjee et al., 2009a, FAO, 2020]. This process (ideally) cooks off all the organic matter in the soil, some fraction of which is SOC. The fraction of organic matter that is SOC is determined by calibrating the LOI assays to DC-EA assays using linear regression, or by using a fixed conversion factor of 0.58 [Chatterjee et al., 2009a]. However, the nature of the relationship between LOI and DC-EA is often site specific, depending in particular on the vegetation, texture, and residual water content in the soil [De Vos et al., 2005, Nayak et al., 2019]. The site level differences make LOI especially tricky for comparing different plots, as opposed to the same plot at different times, because water content and mineralogy may differ substantially. This makes 0.58 suspect as a universally valid fraction. It is well-known that LOI is relatively imprecise, even in the ideal scenario where it is calibrated to soils using DC-EA. However, LOI is considerably cheaper than DC-EA both in terms of upfront costs and costs per sample, and allows investigators to assay many more samples per assay rep than DC-EA [De Vos et al., 2005].

MIRS works by shining light in the mid-infrared range (4,000-400 cm^{-1} or 2500-25,000 nm) on dried samples and measuring the wavelengths that are absorbed [Nayak et al., 2019, Reeves, 2010, Wijewardane et al., 2018, Bellon-Maurel and McBratney, 2011]. MIRS is a high-throughput technology that requires even less resources than LOI. It has the further logistical advantage of simultaneously assaying SOC and soil inorganic carbon (SIC), alongside many other soil properties like pH, texture, and cation exchange capacity [Wijewardane et al., 2018]. MIRS is thus a promising new assay method despite the considerable upfront costs of units. Similar to LOI, MIRS must be initially calibrated to DC-EA assays. A database of samples that contains both spectra and DC-EA SOC assays is called a spectral library. Spectra are unique to soils, so spectral libraries must be constructed within a region of interest and do not transfer well to new regions [Wijewardane et al., 2018]. Furthermore, unlike LOI, the relationship between spectra and SOC content is not simple, necessitating the use of more complex prediction methods that need to be rigorously validated [Bellon-Maurel and McBratney, 2011, Wijewardane et al., 2018]. Calibrations are also highly sensitive to sample prep procedures: samples must be well dried and ground to a consistent size for precise assay [Wijewardane et al., 2018]. Labs can expect to pay a significant upfront cost for purchasing a MIRS unit and establishing a spectral library, but after the initial investment MIRS is cost effective to run, and can be quite precise with proper user training and sample preparation, making it an appealing alternative to DC-EA.

Table 2.1: A table of sample preparation procedures, their costs per sample, and whether they are needed for assay with LOI, DC-EA, or MIRS. Asterisk denotes that sample preparation may vary depending on specific details of the assay technology or soils. IC = inorganic carbon; LOI = loss-on-ignition; DC-EA = dry combustion in an elemental analyzer; MIRS = mid-infrared spectroscopy.

Procedure	LOI	DC-EA	MIRS
Transportation	✓	✓	✓
Oven Drying	✗	✗	✗
Air Drying	✗	✓	✓
Sieving	✓	✓	✓
Grinding	✗	✓	✓
Check for IC	✗	✓*	✗

From a statistical perspective, the assay process is important because additional random error is introduced into the data. Unbiased assays are centered on the true SOC concentration of the (composited) sample. Biased assays systematically overestimate or underestimate the SOC concentration. It is not guaranteed that assays are unbiased (see Bellon-Maurel and McBratney [2011]), though we will assume that they are here. Even when assays are unbiased, they add error to SOC estimation as measurements will not be exactly the same for two or more assays run on the same sample. This variability can be due to errors in weighing, slight differences in aliquots taken from the same sample (especially if homogenization is poor), instrumental drift, or error in predictions or calibrations (especially for LOI and MIRS). We conceptualize assay error on a multiplicative scale so that the amount of error is proportional to the true SOC concentration. Unbiased multiplicative errors are centered at 1, but realizations vary around 1 depending on a variance σ_δ^2, which is roughly the expected *percent* error in assay. We detail how to estimate σ_δ^2 in Sections A.3.1 and A.3.2 of the appendix.

As an example, a realized assay error of 1.1 will cause a true SOC concentration of 1% to appear as 1.1% and a true SOC concentration of 5% to appear as 5.5%. A precise assay method has a small σ_δ^2 so realizations tend to be close to 1, and the measured SOC concentration is close to the true SOC concentration. Note that we will sometimes use an additional subscript to refer to a specific method, e.g. $\sigma_{\delta,\text{DC-EA}}$ is the assay error variance of DC-EA.

2.8 Optimal Sampling and Assay

In this section we highlight our main results. We provide a formula for the precision of estimates of μ given a sample size n and a number of assays k. We derive the optimal n and k that will maximize precision while maintaining a given budget.

2.8.1 Estimation Error

Suppose we have a UIRS of size n and that composites are formed randomly from n/k samples in equal proportions and with perfect homogenization, so that k assays are taken. Suppose S_i^* is the assayed SOC concentration of the ith composited sample. Our estimator is the mean of these assayed composite samples:

$$\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} S_i^*.$$

This is an unbiased estimator of μ, so that $\mathbb{E}[\hat{\mu}] = \mu$. Its variance is

$$\mathbb{V}(\hat{\mu}) = \frac{\sigma_p^2(1+\sigma_\delta^2)}{n} + \frac{\mu^2\sigma_\delta^2}{k}. \tag{2.1}$$

If there is no assay error, this reduces to the usual formula for the variance of a UIRS mean: σ_p/n. Because the estimator is unbiased, it's expected error (mean-squared error) is also equal to (2.1). In order to reduce the error, we can either gather more samples n or make more assays k. The optimal allocation of samples and assays will depend on the plot parameters σ_p and μ, the assay error variance σ_δ^2, and a cost model for sampling and assay.

2.8.2 Optima

We now introduce such a cost model. Call cost_c the cost of sampling a single core, cost_P the cost of sample preparation, and cost_A the cost of assaying a single composited sample. Note that these costs depend on the sampling and assay methods employed. For example, cost_A under LOI is considerably lower than cost_A under DC-EA. We assume that the cost of compositing itself is negligible, but it could easily be included in cost_c. Finally, we assume a fixed cost of the study cost_0, which doesn't vary over n and k. The total cost is:

$$\text{cost}_0 + n \cdot \text{cost}_c + k \cdot (\text{cost}_P + \text{cost}_A). \tag{2.2}$$

We ultimately want to choose both an optimal n and k, which we call n_{opt} and k_{opt} respectively, as well as a sample prep and assay method. We first consider the sample prep and assay methods to be fixed, optimizing only for n and k, and then discuss how to choose among strategies.

Given the cost model along with the plot and assay parameters, the composite size that minimizes the error in Equation (2.1) is:

$$\frac{n_{\text{opt}}}{k_{\text{opt}}} = \frac{\sigma_p \sqrt{1 + \sigma_\delta^2}}{\mu \sigma_\delta} \times \sqrt{\frac{\text{cost}_P + \text{cost}_A}{\text{cost}_c}}. \tag{2.3}$$

The optimal composite size thus depends on the ratio of plot heterogeneity σ_p and the degree of assay error σ_δ. It also depends on the ratio of assay and sampling costs, though it is less sensitive to small changes in cost due to the square root applied to this ratio. Note that there are two boundary conditions that are not reflected in Equation (2.3). Namely, if we initially find $k_{\text{opt}} < 1$ then we take $k_{\text{opt}} = 1$ with the implication that all cores should be fully composited to 1 composite sample. On the other hand, if we find $k_{\text{opt}} > n_{\text{opt}}$, then set $k_{\text{opt}} = n_{\text{opt}}$ with the implication that all sampled cores should be assayed without compositing. Ultimately, there are only gains to compositing if

$$\sigma_p^2(1 + \sigma_\delta^2)(\text{cost}_P + \text{cost}_A) > \mu^2 \sigma_\delta^2 \text{cost}_c.$$

Otherwise no compositing should be done.

Given a fixed budget B, we can compute the optimal variance $\mathbb{V}(\hat{\mu})_{\text{opt}}$. The optimal variance can be difficult to interpret. Taking the square root yields the optimal *standard error* $\text{SE}(\hat{\mu})_{\text{opt}}$ we can achieve at budget B:

$$\text{SE}(\hat{\mu})_{\text{opt}} = \frac{\sigma_p \sqrt{(1 + \sigma_\delta^2)\text{cost}_c} + \mu \sigma_\delta \sqrt{\text{cost}_P + \text{cost}_A}}{\sqrt{B - \text{cost}_0}} \tag{2.4}$$

The optimal standard error is on the same scale as the estimate (i.e. percent SOC).

Finally, different sample prep and assay methods involve trade-offs between the costs and the assay error. Clearly, if a method is both cheaper and less erroneous, it is preferred. But how much error should we tolerate for a cheaper assay? The *relative efficiency* of different methods is the ratio of the minimum errors they are able to achieve, per Equation (2.4). The relative efficiency of method 1 over method 2 is:

$$\frac{\mathrm{SE}(\hat{\mu})_{\text{opt},1}}{\mathrm{SE}(\hat{\mu})_{\text{opt},2}} = \frac{\sigma_p \sqrt{(1+\sigma_{\delta_1}^2)\mathrm{cost}_c} + \mu \sigma_{\delta_1} \sqrt{\mathrm{cost}_{P_1} + \mathrm{cost}_{M_1}}}{\sigma_p \sqrt{(1+\sigma_{\delta_2}^2)\mathrm{cost}_c} + \mu \sigma_{\delta_2} \sqrt{\mathrm{cost}_{P_2} + \mathrm{cost}_{M_2}}} \tag{2.5}$$

A relative efficiency close to 1 suggests a near toss-up between different sample prep and assay strategies. On the other hand, a large relative efficiency suggests that method 2 is more efficient than method 1, and vice versa for a small relative efficiency. The upshot is that for any budget, we can achieve substantially more precise estimates when the relative efficiency is far from 1.

Alternatively, given a maximum variance V that we can tolerate, we might ask for a minimum budget over all ways of allocating the budget to samples and assays. This is the inverse of the previous problem. The expressions for the optimum n and k are fairly complicated. We provide details in Section A.2.2 in our appendix.

2.8.3 Variance estimation

So far we have assumed that we know the parameters σ_δ and σ_p. In practice, these quantities must be estimated with gathered data or, when planning a survey, based on physical reasoning and past studies.

An unbiased estimator of the plot variance σ_p^2 is the usual sample variance with an adjustment factor for the size of composites:

$$\hat{\sigma}_p^2 = \frac{n}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(S_i^* - \hat{\mu})^2\right],$$

where as above, $\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} S_i^*$ As previously noted, this formula will underestimate the sample variance if composite samples are systematically more homogeneous than the plot itself. This can happen, for example, when composites are grouped together by distance instead of randomly.

We can estimate σ_δ using replicated assays, detailed in Section A.3.1 of our appendix. For methods like LOI or MIRS that involve calibration, the additional error due to calibration must be taken into account. See Section A.3.2.

Putting these pieces together, we can estimate the overall standard error of $\hat{\mu}$ by:

$$\widehat{\mathrm{SE}}(\hat{\mu}) = \sqrt{\frac{\hat{\sigma}_p^2(1+\hat{\sigma}_\delta^2)}{n} + \frac{\hat{\mu}^2\hat{\sigma}_\delta^2}{k}}$$

2.8.4 A Confidence Interval

If the sample size n is not too small, then an asymptotic confidence interval based on the the t-distribution with $n-1$ degrees of freedom will be approximately correct. Specifically, denote $t_{(1-\alpha/2)}$ as the $(1-\alpha/2)$ quantile of the t-distribution with $n-1$ degrees of freedom. The interval

$$\left[\hat{\mu} - t_{(1-\alpha/2)} \times \widehat{\mathrm{SE}}(\hat{\mu}),\ \hat{\mu} + t_{(1-\alpha/2)} \times \widehat{\mathrm{SE}}(\hat{\mu})\right]$$

bounds the true mean μ with about 95% probability. Checking if a particular value of μ (say μ_0) is in this interval is equivalent to a level α t-test of the null hypothesis $H_0 : \mu = \mu_0$. Often, a researcher will set $\alpha = .05$ to yield a 95% confidence interval: $[\hat{\mu} - 1.96 \times \widehat{\mathrm{SE}}(\hat{\mu}),\ \hat{\mu} + 1.96 \times \widehat{\mathrm{SE}}(\hat{\mu})]$. In instances where the confidence interval includes values less than 0, in particular if $1.96 \times \widehat{\mathrm{SE}}(\hat{\mu}) > \hat{\mu}$, it is valid to set the lower confidence limit equal to 0.

2.8.5 Estimating a difference

Often, investigators aim to estimate the difference between average SOC concentrations, either between two plots at the same time or within the same plot at different times. Let μ_1 and μ_2 be the mean SOC concentrations in plot 1 and plot 2. Then the parameter of interest is $\mu_1 - \mu_2$. Let $\hat{\mu}_1$ and $\hat{\mu}_2$ be estimators of μ_1 and μ_2, as above. Then the difference in means, $\hat{\Delta}_{1,2} = \hat{\mu}_1 - \hat{\mu}_2$, is an unbiased estimator of $\mu_1 - \mu_2$. Furthermore, assuming independent UIRSing in each plot, the standard error is:

$$\mathrm{SE}(\hat{\Delta}_{1,2}) = \sqrt{\mathbb{V}(\hat{\mu}_1) + \mathbb{V}(\hat{\mu}_2)}.$$

The optimum SE of the difference can be attained by separately optimizing $\mathbb{V}(\hat{\mu}_1)$ and $\mathbb{V}(\hat{\mu}_1)$, as above, yielding sampling and assay sizes of n_1, k_1 for plot 1 and n_2, k_2 for plot 2. A reasonable estimate of the SE is $\widehat{\mathrm{SE}}(\hat{\Delta}_{1,2}) \equiv \sqrt{\widehat{\mathbb{V}}(\hat{\mu}_1) + \widehat{\mathbb{V}}(\hat{\mu}_2)}$. An approximate $(1-\alpha)$ confidence interval on the difference is:

$$\left[\hat{\Delta}_{1,2} - t_{(1-\alpha/2)} \times \widehat{\mathrm{SE}}(\hat{\Delta}_{1,2}),\ \hat{\Delta}_{1,2} + t_{(1-\alpha/2)} \times \widehat{\mathrm{SE}}(\hat{\Delta}_{1,2})\right]$$

where $t_{(1-\alpha/2)}$ is now the $(1-\alpha/2)$ quantile of the t-distribution with $\min(n_1, n_2)$ degrees of freedom.

If sample sizes are fairly small, say $n_1, n_2 < 30$, the difference-in-means will generally not have a normal distribution. In this case, a permutation test should be used to test for a difference between μ_1 and μ_2. Permutation tests provide an exact level α test at any sample size, without assumptions about the distributions of the samples. Permutation confidence intervals can be derived by testing a range of hypotheses over a grid of effect sizes. The corresponding $1-\alpha$ confidence interval contains all effect sizes that are *not* rejected at level α. Pesarin and Salmaso [2010b] and Good [2005] are good references for the theory and implementation of permutation tests.

2.9 Application

In this section we demonstrate a practical application of our analysis. We draw on a variety of sources to estimate parameters and costs. We stress that the results are not intended to provide universal guidance on sampling, sample prep, and assay—they are highly sensitive to the inputs. The open-source software and web tool we provide are intended to enable investigators to draw their own conclusions from their own inputs.

2.9.1 Data

We combine data from multiple sources to estimate σ_δ for DC-EA, LOI, and MIRS: $\sigma_{\delta,\text{DC-EA}}$, $\sigma_{\delta,\text{LOI}}$, and $\sigma_{\delta,\text{MIRS}}$, respectively. $\sigma_{\delta,\text{DC-EA}}$ is estimated from assays on samples taken from rangeland soils in Marin County, California by the Silver Lab at UC Berkeley, referred to here as the Marin data. The samples were run in duplicate on a Carlo Elantech Elemental analyzer at UC Berkeley. We use the method presented in Section 2.8.3 to compute $\hat{\sigma}_{\delta,\text{DC-EA},i}$ on each sample and took the median across samples to get $\hat{\sigma}_{\delta,\text{DC-EA}}$. We applied the methods presented in detail in Section A.3.2 of our appendix to estimate the additional assay error in LOI and MIRS callibrated to DC-EA assays. Briefly, we derived the validation root mean squared error (RMSE_v) for LOI by regressing LOI assays on DC-EA assays taken at the Agricultural Diagnostic Laboratory at the University of Arkansas (ADL). These assays were taken on samples from several sites in Colorado collected by the Wainwright Lab at Lawrence Berkeley National Laboratory. We estimated $\sigma_{\delta,\text{MIRS}}$ using the RMSE_v provided in table 4 of England and Viscarra Rossel [2018]. They computed this estimate from a median of MIRS RMSE_v values reported in a range of studies. These errors were then divided by our estimates of $\hat{\mu}$ and added to the DC-EA error variance estimate to approximate their overall assay error variance on a multiplicative scale. We also computed the SE assuming no assay error and no cost to assay, which represents a typical power analysis and provides a lower bound on the SE across assay methods.

We used the Marin data to get estimates of σ_p and μ in the topsoil (0-10 cm) and in deep soil (50-100 cm). Within depth profiles, we computed the sample mean and standard deviation at each site and then took the median over sites as our estimates $\hat{\sigma}_p$ and $\hat{\mu}$. The samples were collected using transect sampling, not UIRSing, but should provide reasonable estimates.

2.9.2 Results

Inputs

All inputs are summarized in table 2.2. Using the Marin data, we estimated the topsoil plot heterogeneity as $\hat{\sigma}_p = 0.54$ and the mean as $\hat{\mu} = 3.61$. We estimated the deep soil heterogeneity as $\hat{\sigma}_p = 0.12$ and the deep soil mean as $\hat{\mu} = 0.48$. Based on the duplicated DC-EA assays, we obtained the estimate $\hat{\sigma}_{\delta,\text{DC-EA}} = 0.02$. The RMSE_v for LOI was 0.31 in the range of the Marin data assays. Dividing by $\hat{\mu}$ and combining this with the DC-EA error, we estimated an error variance for LOI of $\hat{\sigma}_{\delta,\text{LOI}} = 0.11$ in the top soil and $\hat{\sigma}_{\delta,\text{LOI}} = 0.67$ in deep soil. England and Viscarra Rossel [2018] reported a median RMSE_v of 0.11 for MIRS, yielding an estimate of $\hat{\sigma}_{\delta,\text{MIRS}} = 0.05$ in the top soil and $\hat{\sigma}_{\delta,\text{MIRS}} = 0.25$ in deep soil.

The cost of sampling and the fixed costs of the survey are not well constrained. We set the fixed cost at $\text{cost}_0 \equiv 200$ and cost_c at 5, 20, or 40 USD to reflect cheap, medium, and expensive sampling. We assumed a transport cost of 2.00, a cost of 4.00 for oven drying, 1.00 for air drying, 2.00 for sieving, 4.00 for grinding, and 2.00 for acid testing for inorganic carbon. Without root picking, this puts the cost of sample prep at 8.00 for LOI, 11.00 for DC-EA, and 9.00 for MIRS. The assay costs for DC-EA and MIRS were estimated in [O'Rourke and Holden, 2011]. That paper reported a cost of about 15 USD per sample for DC-EA and 1.30 USD per sample for MIRS. The 12:1 price ratio of DC-EA to LOI reported in De Vos et al. [2005] yields an assay cost of 1.25 USD per sample for LOI. Adding cost_P and cost_A for each method yields $\text{cost}_{\text{DC-EA}} = 26.00$ USD, $\text{cost}_{\text{LOI}} = 9.25$ USD, and $\text{cost}_{\text{MIRS}} = 10.30$ USD.

Table 2.2: Inputs to optimization problem as estimated from Marin and LBL data. $\hat{\sigma}_{\delta,\text{DC-EA}}$ is the assay error of DC-EA; $\hat{\sigma}_{\delta,\text{LOI}}$ is the assay error of LOI; $\hat{\sigma}_p$ is the plot heterogeneity (standard deviation); $\hat{\mu}$ is the plot mean concentration; $\text{cost}_{\text{DC-EA}}$ is the cost of DC-EA assay plus the costs of associated sample prep; cost_{LOI} is the cost of LOI plus the costs of associated sample prep; $\text{cost}_{\text{MIRS}}$ is the cost of MIRS plus the costs of associated sample prep. cost_c is the cost of sampling.

Description	**Notation**	**Value(s)**
DC-EA assay variance	$\hat{\sigma}_{\delta,\text{DC-EA}}$	0.02
LOI assay variance	$\hat{\sigma}_{\delta,\text{LOI}}$	0.11, 0.67
MIRS assay variance	$\hat{\sigma}_{\delta,\text{MIRS}}$	0.05, 0.25
Plot heterogeneity	$\hat{\sigma}_p$	0.68, 0.12
Plot mean	$\hat{\mu}$	3.57, 0.48
Fixed cost	cost_0	200
Cost of DC-EA assay	$\text{cost}_{\text{DC-EA}}$	26.00
Cost of LOI assay	cost_{LOI}	9.25
Cost of MIRS assay	$\text{cost}_{\text{MIRS}}$	10.30
Cost of taking a core	cost_c	5, 20, 40

Outputs

Figure 2.2 plots the optimal SE of estimation attainable for each assay method across a range of budgets. Figure 2.3 plots the same results but rescaling SEs to coefficients of variation. The output indicates that DC-EA is the best assay method in both topsoil and deep soil, yielding the most precise estimate at any given budget. In terms of relative performance, the assay method is more important in the deep soil than in the topsoil: DC-EA represents a major improvement over the other methods in deep soil, while the precision is essentially a toss-up in top soil. Under our inputs, DC-EA gets close to achieving the lower bound implied by no assay error.

Optimal composite sizes are provided in Table 2.3 across the range of sampling costs and depths. Compositing is more valuable as the assay method becomes more precise and expensive, with large gains to compositing under DC-EA and essentially no gain under LOI. Compositing is also more valuable if samples are cheap to gather and the plot is heterogeneous, in which case it becomes beneficial to focus budgets on sampling rather than assay.

Relative efficiencies are given in Table 2.4. Relative efficiencies for topsoil assays are fairly close to 1. For deep soil, DC-EA is at least twice as efficient as LOI — all relative efficiencies are less than 0.5 — and at least 30% more efficient than MIRS for any sampling cost.

Table 2.3: Optimal composite sizes for the three assay methods. Sampling costs are set at 5, 20, or 40 USD (top row). Soil parameters are determined from the topsoil (first three columns) or deep soil (last three columns). An optimal composite size of 1 suggests that no compositing should be done, i.e. that all cores should be measured.

	Topsoil (0-10 cm)			Deep Soil (50-100 cm)		
	5 USD	20 USD	40 USD	5 USD	20 USD	40 USD
DC-EA	20	10	7	26	13	9
MIRS	5	3	2	2	1	1
LOI	2	1	1	1	1	1

Table 2.4: Relative efficiencies of different assay methods compared to DC-EA at different profiles (first column) and sampling costs (second column). A relative efficiency significantly less than 1 suggests DC-EA is more efficient than the alternative method at any given budget, and vice versa for a relative efficiency greater than 1. A relative efficiency near 1 suggests little difference between methods.

Soil Profile	Sampling Cost	$\mathbf{SE_{DC\text{-}EA}}$ / $\mathbf{SE_{LOI}}$	$\mathbf{SE_{DC\text{-}EA}}$ / $\mathbf{SE_{MIRS}}$
Topsoil (0-10 cm)	5 USD	0.70	0.90
Topsoil (0-10 cm)	20 USD	0.79	0.93
Topsoil (0-10 cm)	40 USD	0.83	0.95
Deep Soil (50-100 cm)	5 USD	0.25	0.48
Deep Soil (50-100 cm)	20 USD	0.36	0.61
Deep Soil (50-100 cm)	40 USD	0.42	0.67

2.10 Discussion

In this paper, we statistically formalized the sampling and assay processes to characterize the precision of SOC concentration estimation while making minimal assumptions. We derived optimal composite sizes to maximize precision under a fixed budget. Although we did not discuss it extensively, we also solved the inverse problem of minimizing costs given a fixed precision (see section A.2.2 of our appendix).

We applied our method to data from a California rangeland, bringing in costs and errors of measurement from other studies [De Vos et al., 2005, England and

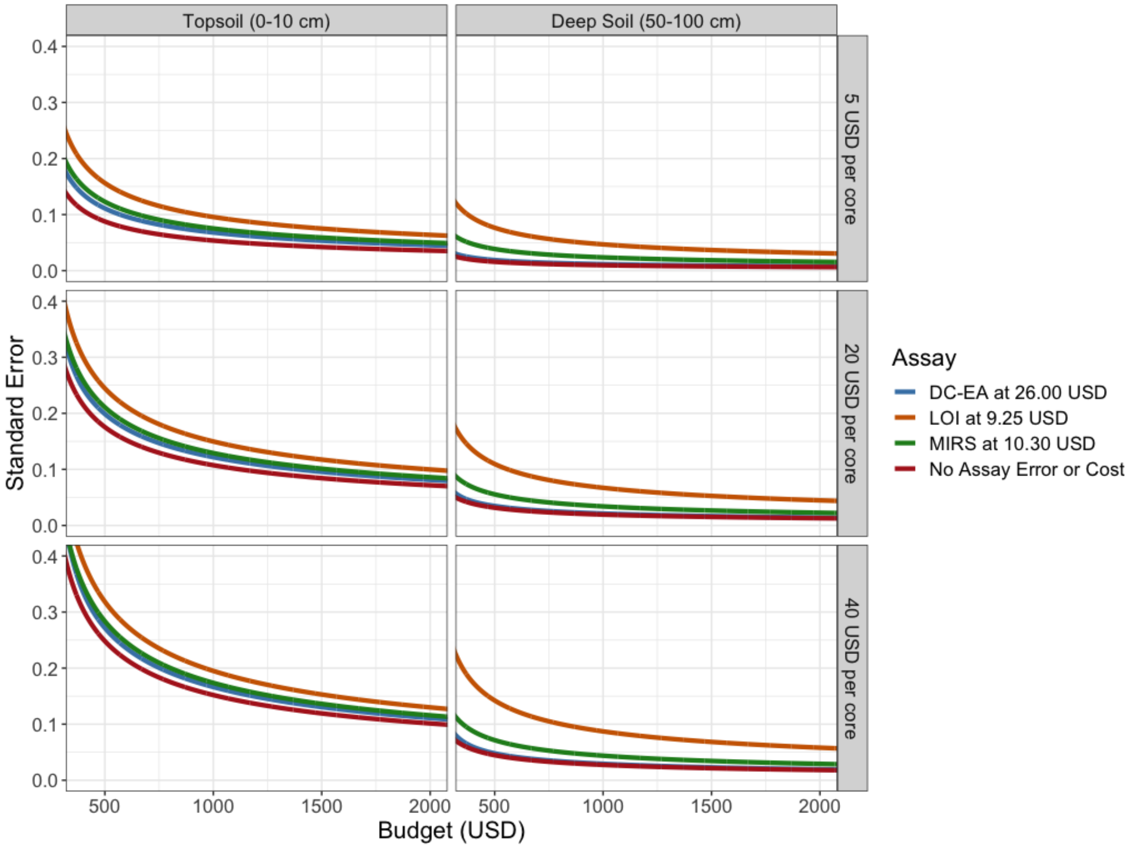


Figure 2.2: Optimal standard errors for estimating μ given parameters in Table 2.2. The x-axis is the budget in US dollars, the y-axis is the standard error in %SOC attained at the given budget. Different colored lines correspond to different assay methods. The cost of sampling varies across rows, and the depth varies across columns. The line labels indicate the combined costs of sample prep and assay for each method. DC-EA = dry combustion in an elemental analyzer; LOI = loss-on-ignition; MIRS = mid-infrared spectroscopy; USD = United States dollars.

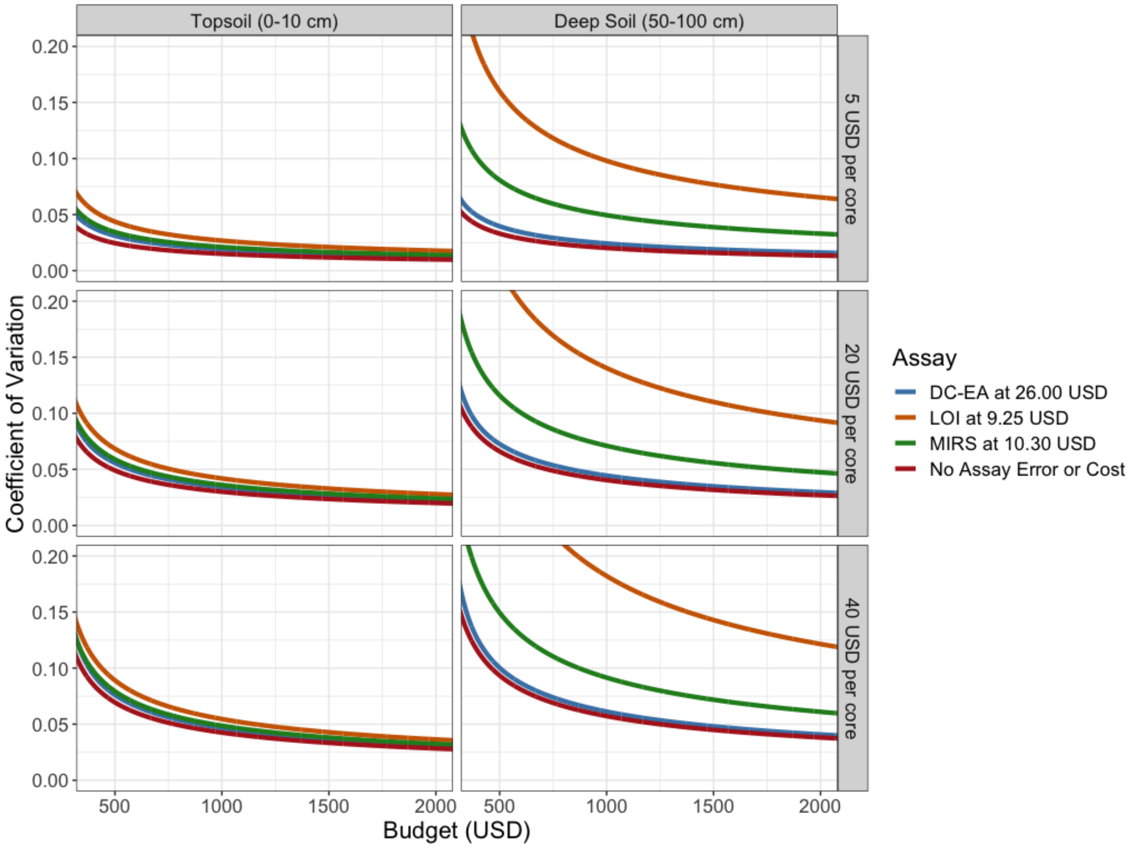


Figure 2.3: Optimal coefficients of variation for estimating μ given parameters in Table 2.2. The x-axis is the budget in US dollars, the y-axis is the coefficient of variation: $\mathrm{SE}(\hat{\mu})_{\mathrm{opt}}/\mu$. Different colored lines correspond to different assay methods. The cost of sampling varies across rows, and the depth varies across columns. The line labels indicate the combined costs of sample prep and assay for each method. DC-EA = dry combustion in an elemental analyzer; LOI = loss-on-ignition; MIRS = mid-infrared spectroscopy; USD = United States dollars.

Viscarra Rossel, 2018, O'Rourke and Holden, 2011]. There are a number of interesting implications from our results.

First, we found assay error to be a significant source of uncertainty in SOC estimation that is usually not taken into account. Indeed, many analyses in the SOC literature compute uncertainty estimates accounting only for plot heterogeneity (or in some cases, only inter-plot heterogeneity). We found that incorporating assay error and costs can double the uncertainty (in terms of standard error) compared to the conventional approach of not incorporating assay error.

Furthermore, the depth of soil under study was an especially important consideration for the assay method employed. We found that efficiencies varied much more across the assay methods when attempting to quantify deep soil concentrations rather than top soil. In terms of the coefficient of variation, top soil can be quantified by any assay method to within about 5% of the mean at a budget of 1000 USD. On the other hand, DC-EA seems far better at accurately quantifying deep soil concentrations than other methods despite its high cost. Equation 2.4 reveals that when the plot heterogeneity σ_p is high the estimation error will be driven largely by the cost of sampling while the cost and precision of assay have little effect. Intuitively, we need many samples to characterize the heterogeneity within the plot, and cheaper, less precise assay methods generally allow many samples to be collected and assayed. Conversely, if the plot heterogeneity is low, it is better to collect a few samples that accurately represent the average plot concentration and focus the budget on assaying them as precisely as possible.

We also found that with our inputs, the benefits of compositing were quite variable. Compositing many cores together is beneficial when the assay method is fairly expensive and precise (e.g. DC-EA), while sampling is fairly cheap (e.g. 5.00 USD per core). Equation 2.3 reveals that compositing may also be resourceful when the plot heterogeneity is large compared to assay error.

We did not incorporate bulk density into our analysis. Estimating bulk density is critical to converting from concentrations to stocks. Estimating SOC stock is especially necessary in studies of carbon sequestration and climate change mitigation, while SOC concentration is typically the parameter of interest from a soil health and functioning perspective. As with SOC concentrations, bulk density tends to vary substantially across a landscape. However, investigators frequently take only one bulk density sample and bulk density is also prone to assay error [Walter et al., 2016b]. Thus converting from concentration to stock will incur substantial additional error, which should ideally be reflected in confidence intervals on stock estimates. Our results on the error in concentration estimation can be seen as a lower bound on the error in stock estimation, i.e. assuming no error in bulk density estimation. In addition to including error in bulk density, future work should incorporate more general sampling schemes that can improve efficiency, like stratified sampling or well-spread sampling, and optimize over the sampling design as well as the assay method. Considering the sampling design alongside the compositing and assay strategy will allow investigators to design economical soil surveys that achieve their desired precision.

2.11 Conclusion

When assays introduce error into an estimation process, compositing samples may have major ramifications for both the precision and cost of estimates. In this paper we detailed the processes involved in soil organic carbon estimation and derived optimal composite sizes to maximize estimation precision when assays are subject to multiplicative errors. An analysis of data from California rangeland indicated that DC-EA would yield more precise estimates than LOI or MIRS, and that, for any given budget, compositing samples before assay would yield more precise estimates than assaying individual samples. Optimal composite sizes and assay methods will depend on parameters of the plot under study and on the costs of sampling, sample preparation, and assay. Thus, our results are not meant to provide universal guidance. We hope that the framework we presented here will be useful to investigators aiming to design efficient soil surveys for soil organic carbon concentration and stock.

Chapter 3

Valid soil organic carbon measurement

3.1 Introduction

Interest in measuring soil organic carbon (SOC) is expanding dramatically because agricultural interventions that sequester C in soil may help to mitigate climate change. Recent policy initiatives and emerging soil C markets designed to accelerate management transitions require practical methods to measure SOC with low uncertainty or they may often reward false positives and fail to reward genuine sequestration. Indeed, the high uncertainty of SOC measurements likely contributed to the 2011 collapse of the Chicago Climate Exchange, the only prior U.S. voluntary C market [Gosnell et al., 2011].

In practice, the accuracy of SOC measurements is limited by spatial heterogeneity, sampling design, variability in bulk density, and variation in soil processing methods and laboratory assays. Reliably detecting and accurately quantifying changes in SOC stocks is challenging because, compared to these sources of variation and uncertainty, the annual changes produced by agricultural management interventions are often small [Bai et al., 2019, Minasny et al., 2017], for instance ranging from < 0.1 % absolute change for conversion to no-till [Franzluebbers, 2005] to approximately 0.5 % with biochar application [Jones et al., 2012, Majumder et al., 2019]. Methods for estimating SOC must be precise and powerful enough to detect such small changes in a heterogeneous medium [Ellert et al., 2002, Homann et al., 1998, Lehmann et al., 2007, Robertson et al., 1997]. Minimizing the errors that arise in each of the many steps in SOC measurement (see SI 1 and SI Table 1 for a full description) is especially important in the context of C offset markets. Only accurate estimates of SOC sequestration with transparent levels of uncertainty should be used for generating credits and allowing governments and industries to offset greenhouse gas (GHG) emissions.

Yet protocols currently being used by C markets for measurement, reporting and verification (MRV) of SOC sequestration may be inadequate [Necpalova et al., 2014, Oldfield et al., 2021]. Importantly, many MRV protocols recommend but do not require—or else make no mention of—powering measurement campaigns using representative spatial heterogeneity information. Agricultural soils used to generate C credits have varying degrees of spatial heterogeneity and require different sample sizes to detect a given absolute or relative change in SOC. For example, spatial heterogeneity is typically higher on rangelands than croplands due to diverse topography, rocky soil horizons, low and patchy soil fertility, and patchiness of grazing and manure deposition. Rather than tailoring sample size requirements to expected levels of heterogeneity, many protocols (including the Climate Action Reserve Soil Enrichment Protocol, Australian Carbon Methodology, and Verra VM0021) simply set a minimum sample size within designated areas (e. g., 3 samples per stratum). If the MRV protocol does not require determining the number of samples necessary to detect a reasonable level of SOC sequestration, it could fail to reward legitimate sequestration or have a large chance of erroneously rewarding nonexistent sequestration.

Addressing knowledge gaps associated with sampling design—including sample placement, stratification, and compositing—could further reduce the measurement uncertainties of SOC offsets. For example, C market protocols often encourage the use of systematic sampling, but samples collected by simple or stratified random sampling are less likely to bias SOC estimates and allow more rigorous statistical analysis. While stratifying soil sampling into more homogeneous land subunits can increase the power to detect SOC sequestration, many protocols lack quantitative guidance for defining strata and some do not require field sampling at all, relying instead on model output [Oldfield et al., 2021], with notable exceptions (e.g. Australian Carbon Methodology). Compositing—combining samples to reduce analysis costs—is a common practice allowed in MRV protocols (e.g. Climate Action Reserve Soil Enrichment Protocol, Australian Carbon Methodology), though the impact on measurement error is often unknown [de Gruijter et al., 2016]. At one extreme, all samples collected within an experimental unit can be combined into a single sample for analysis (i.e., full compositing) [Carey et al., 2020, Tautges et al., 2019], making it impossible to estimate spatial heterogeneity and substantially increasing measurement error [Spertus, 2021].

The impact of compositing on measurement error depends in part on the error of laboratory analyses. The extent to which dry combustion assays contribute to overall error in measuring SOC from either intra-lab (replicated measurements on the same instrument) or inter-lab (measurements on different instruments) analytical variability is not well known [Chatterjee et al., 2009b, O'Rourke and Holden, 2011], limiting the ability to optimize sampling campaigns and the reliability of estimates and inferences. Compositing and subsequent laboratory analyses can be optimized to minimize contribution to error within a given budget, given estimates of spatial heterogeneity, analytical error, and laboratory costs [Spertus, 2021]. To our knowledge, such an analysis has never been done to inform soil-sampling campaigns.

Lastly, the choice of statistical methods for data analysis also influences the likelihood of false positives (Type I errors)—generating C offsets when SOC wasn't sequestered—and false negatives (Type II errors)—failing to generate C offsets when SOC was sequestered. In a C market, Type I error can lead to allocation of payments without actual SOC sequestration, and possibly even increase net C emissions; while a Type II error can fail to generate C offsets when SOC is sequestered [Sanderman and Baldock, 2010]. Both types of error undermine the utility of C markets, leading to missed opportunities for climate change mitigation. When the assumptions required of common statistical methods are not met (e.g., SOC is not normally distributed), standard hypothesis tests can have Type I error rates that greatly exceed their nominal significance level (e.g., 5 %), and confidence intervals can have coverage probabilities far lower than nominal (e.g., 95 %) [Lehmann and Romano, 2005]. For example, the two-sample Student t-test is often used to assess changes in SOC stocks [Brus and de Gruijter, 2011, de Gruijter et al., 2016, Kravchenko and Robertson, 2011]. Student's ttest assumes that SOC at both measurement times is normally distributed with the same variance. Since SOC generally does not have a normal distribution [Yan et al., 2011] and because agricultural management interventions can redistribute SOC without changing the total [Chappell et al., 2012], Student's t confidence intervals can have true coverage probabilities far lower than the nominal confidence level (e.g., 95 %), and Student's t-tests can have true Type I error rates that greatly exceed the nominal significance level (e.g., 5 %) [Lehmann and Romano, 2005]. This undermines the validity of many standard methods for inference—including ANOVA, mixed effects models, geostatistical models, bootstrapping, Wilcoxon rank-sum tests, permutation tests, and Bayesian models. Quantifying the chance of false conclusions about whether and how much SOC has been sequestered is crucial for SOC offsets.

Lesser-known statistical methods can strictly limit the Type 1 error rate and increase reliability. For example, there are nonparametric tests and confidence intervals that are valid for any SOC distribution [Anderson, 1969, Learned-Miller and Thomas, 2019, Romano and Wolf, 2000, Stark, 2009e, 2023, Waudby-Smith and Ramdas, 2023]. These methods are conservative or exact: the probability of Type I errors is not larger than the nominal significance level (e.g., 5 %), and the chance that confidence intervals include the true amount of SOC sequestered is not less than the nominal confidence level (e.g., 95 %). Suitable nonparametric tests and confidence intervals can produce reliable inferences about SOC stocks and changes, though their widespread adoption has been hindered by their relatively low power.

Below, we investigate these uncertainties and knowledge gaps and how they affect the cost and reliability of SOC sequestration measurements. Using new, on-farm data from California crop and rangelands, we 1) evaluate the relative impact of spatial heterogeneity, analytical variability, and compositing on measurement precision and power; 2) use simulations to examine the validity and power of common statistical tests to detect SOC sequestration using different sampling designs on high and low heterogeneity agricultural landscapes; and 3) compare the validity and power of the t-test to those of a new nonparametric method across a range of sample sizes and SOC changes. Based on our findings, we make straightforward recommendations, targeted toward SOC markets, to improve the accuracy and reliability of SOC sequestration measurements, yield more trustworthy C credits, and support progress towards climate change mitigation goals.

3.2 Methods

3.2.1 Collecting SOC data: Rangeland and cropland sampling and laboratory analysis

We leverage new data collected from two intensive field sampling campaigns on California crop and rangelands. While these samples were originally collected for other purposes, we use them to study field-level spatial heterogeneity and to provide an empirical basis for simulations. We outline our sampling methods briefly below, with more details SI 2.1.

Rangeland samples were collected in December 2019 from a ranch in Paicines, California. The data were collected to quantify spatial heterogeneity of SOC in a constrained, field-scale setting, controlling for soil type, catenal position, slope aspect, and vegetation—not to quantify SOC stock for the whole ranch. We used soil survey information within the ranch boundaries to identify Auberry Fine Sandy Loam soils. Samples were collected using a stratified transect design with five 100 m transects on two adjacent hillslopes stratified by slope position: summit/shoulder (1 transect), backslope (2 transects), and footslope (2 transects). Soils were sampled down to 100 cm, or the point of refusal, and divided into 5 depth ranges (0–10 cm, 10–30 cm, 30–50 cm, 50–75 cm and 75–100 cm). We attempted to collect 33 samples along each transect, but time constraints limited us to 25 samples at one transect. In all, we attempted to collect 785 samples, but bedrock or rock obstructions limited the depth of sampling at some locations (mostly along the summit position), resulting in 662 total samples. Each sample was air-dried and sieved to 2 mm. Visible plant materials were removed, and soils were ground using a ball mill (SI 2.1; Retsch, Newtown, PA).

Cropland soil samples were collected in September and October 2019, from seven farms across Southern California (SI Fig. 1) representing various soil types and cropping systems, including two orchards, a vineyard, two intensive cropping systems, and two diversified farms (full soil taxonomy in SI Table 2). Samples were collected along 50 m transects. At each site, transect locations were selected based on the dominant soil type (Soil Survey Geographic (SSURGO) Database, United States Department of Agriculture, Natural Resource Conservation Service), consistent historic and current management, and cropping system. The number of transects ranged from two at the small, diversified farms to six at one of the larger cropland sites. Depth ranges were defined slightly differently at different sites based on tillage depths (0–10, 10–20 cm vs 0–15, 15–30 cm) and genetic horizon in the subsurface. In all, 455 samples were collected from the seven farms. Samples were airdried and sieved to 2 mm; visible plant materials were removed; and then soils were oven-dried at 60 C and ground using a ball mill.

Bulk density (BD) samples for croplands and rangelands were collected using the pit method [Walter et al., 2016a]. Cores were collected from the center of each depth increment used for bulk soil samples. For sampling depths greater than 10 cm, multiple cores were collected to ensure samples were representative. At the rangeland site, three 1.5 m deep soil pits were dug along each transect (one at each end, and one in the center at 50 m) using an excavator, a total of 15 pits. At the cropland sites, one soil pit was dug at the central location of each transect to 1.5 m or the point of refusal. Bulk density samples were oven-dried at 105 degrees C until their weight no longer decreased. Visible rock fragments were removed before weighing the samples and submerged in water to measure their volume. Rock volume was subtracted from core volume in estimating soil density. We used bulk density and TC% to calculate SOC stocks for each depth increment.

Two different dry combustion analyzers were used to measure C concentrations (TC%) of prepared samples. All cropland soils were analyzed with a Costech ECS 4010 elemental analyzer (Costech, Valencia, CA)—a widely used instrument for dry combustion analysis. Rangeland samples were analyzed on an Elementar soliTOC cube (Elementar, Ronkonkoma, NY; see [Natali et al., 2020]), a relatively new instrument designed to improve precision by combusting higher sample masses (up to 3 g of soil vs 50 mg) while separating total organic C (TOC), residual organic C (ROC), and total inorganic C (TIC) via a temperature ramping method, DIN19539. The

ECS 4010 measures only the mass of total C (TC), and thus we compare only TC% between the two instruments.

To quantify the precision and bias of each instrument, we reanalyzed 15 rangeland and 22 cropland samples that had the minimum, median, and maximum TC% for each depth and land-use category (SI Fig. 2). Five analytical replicates of each sample were measured on each instrument. Samples with high TIC (greater than0.1 measured by the soliTOC, were treated with HCl to remove TIC and reassayed on the ECS 4010 (SI Figs. 2 and 3). Finally, we ran 25 additional replicates of two soil standards with known TC% on each instrument (SI Fig. 4).

3.2.2 Assessing spatial heterogeneity of SOC and bulk density

To visualize the relative heterogeneity of TC% by land use, depth, and transect, we used histograms, sample means, and coefficients of variation (CV). While TC% of the non-rocky component of the soil is the focus of this study, we also examined the variability of BD measurements by comparing the CV across depths for rangeland and cropland site 7, which had substantially more BD samples than other cropland sites. To assess differences in spatial heterogeneity across land uses, depth, strata (transects), and sites, we tested the hypotheses that population TC% distributions were equal across depths and transects on rangeland soils, or depths and sites on cropland soils using a nonparametric test called permutation ANOVA, a way of calibrating the ANOVA test statistic to control the rate of false rejections without any assumption about the distribution of SOC [Pesarin and Salmaso, 2010a]. Details of how the permutation ANOVA was performed are in SI 2.2. Code is available at: `github.com/spertus/soil-carbon-statistics`.

3.2.3 Evaluating analytical variability

We repeated analyses of the same samples to estimate the variability of laboratory assays. For each sample and instrument, analytical error was quantified by the estimated relative error (see SI 2.3 for the formula), which is approximately the CV. We report the median estimated relative error for each instrument. (The estimated relative error measures variability but not bias; we estimated bias using measurements of known standards.) To evaluate whether there were systematic differences in measurements between the two instruments (SoliTOC and ECS 4010), we used permutation tests for the two-sample problem, which asks whether the difference between two samples would be unlikely if the samples were created by randomly partitioning their pooled values into the two groups. We used the difference in means as the test statistic and simulated 10,000 draws from the permutation distribution using the R package permuter (see SI 2.3 for more details).

3.2.4 Quantifying relative uncertainty from spatial and analytical heterogeneity

We quantified the contributions of analytical variability and spatial heterogeneity to uncertainty in estimates of the population mean TCusing the delta method [Goidts et al., 2009], which decomposes the total uncertainty into a sum of the contributions from analytical variability and spatial heterogeneity (SI 2.4). If the ratio of the contribution from spatial heterogeneity to total uncertainty is close to 1, spatial heterogeneity contributes more than analytical variability to overall uncertainty, vice versa if the ratio is closer to 0. If the ratio is 0.5, analytical and spatial heterogeneity contribute equally to total uncertainty. To assess how compositing affects the relative contributions to uncertainty, we computed the proportion of total uncertainty due to spatial heterogeneity without compositing, and the corresponding proportion when 90 cores are composited to one analytic sample (an extreme degree of compositing). We computed these ratios within depths for both land use types and both instruments.

3.2.5 Comparing how sources of error affect statistical power

We studied how spatial heterogeneity, assay variability, and compositing affect the ability to detect changes in average TC%. Specifically, we approximated the power of the unpaired two-sample t-test when samples are drawn by simple random sampling and there is no compositing, optimal compositing (derived in Spertus [2021], or full compositing. We only examined the power for relatively large sample sizes (n $\geq$ 90 cores) because Student's t-test is especially unreliable for small sample sizes (see below). Comparing compositing strategies requires a budget; if money were no object, assaying every sample separately (i.e., not compositing) minimizes error. Compositing involves a tradeoff between various costs and errors. To explore the tradeoff, we took the marginal cost of collecting a single soil core in the field to be $20 USD and the cost of laboratory analysis (including sample preparation) in an elemental analyzer to be $13.60 USD per sample, the average price charged by five commercial labs for TC% analysis. Given these unit costs, the cost to collect, prepare, and analyze 90 cores (without compositing) is $3,024 USD. Using the same total budget, we explored what the uncertainty would have been had the money been used to take more cores and composite some of them (optimal compositing, which maximizes power within the budget) or all of them (full compositing) instead of assaying them individually.

The power calculations use the estimates of land-use specific TC% average and spatial heterogeneity (averaged across sites for cropland) and instrument-specific median relative error to approximate the power to detect a change of a given magnitude.

3.2.6 Power and validity of tests for detecting TC% change

We performed two simulations to estimate the true significance level and power of different two-sample hypothesis tests. The "validity simulation" estimated the significance level (i.e., Type 1 error rate) of two tests—the chance a test erroneously rejects the null hypothesis when there is no change in total SOC—in four scenarios. The two tests were the usual two-sample Student t-test and a nonparametric test that uses a pre-specified upper bound on TC concentration (See SI 1.6 for details). We set this bound at 10% or 20% TC, established TC ranges in mineral soils. In each of the four scenarios, SOC means were set exactly equal, but the shapes of the SOC distributions could differ in ways that might plausibly result from agricultural interventions, inferred from our empirical crop and rangeland TC data. In the "unchanged normal distribution" scenario, both distributions were normal with SDs of 0.5%; in the "tilled cropland" scenario, the distribution at the first time was the actual topsoil TC% samples from cropland site 5 (right-skewed, Figure 3.1) and the distribution at the second time was normal with SD 0.5 "change in skew" scenario, the distribution at time 1 was rangeland topsoil samples (right-skewed) and the distribution at time 2 was the same but multiplied by -1 (left-skewed); in the "extreme hotspot" scenario, the distribution at time 1 had 99% of its mass in a normal distribution centered at 2.8 % TC (SD: 0.05 %) and 1 % as a point-mass centered at about 20 % TC [Beem-Miller et al., 2016, Miller et al., 2016]. At the second time, the distribution was normal (SD: 0.05 %). We ran both tests at a nominal 5 % level 5000 times with sample sizes ranging from 5 to 150 at each epoch and recorded the rate of (false) rejections. We compared these simulated significance levels to the nominal 5 % significance level (Fig. 6). The "power simulation" estimated the chance of detecting increases in SOC of various magnitudes with sample sizes 10, 30, 90, and 200 using the Student's t-test with unstratified sampling, Student's t-test with stratified sampling, and the nonparametric test with unstratified sampling. (Stratified nonparametric tests are in development.) The reference population distributions (at time 1) were taken to be the empirical distributions of samples from the rangeland site or from cropland site 5, which had median spatial heterogeneity and the most samples among the cropland sites. The hypothetical change in TC% was an additive shift of the reference distribution, with shifts ranging from 0 % (no change) to 60 % of baseline. For example, the baseline average TC % across our cropland sites was 2.7 % TC, so the simulated TC% at time 2 ranged from 2.7 % to 4.32 %. The stratified Student t-test was used only on rangeland samples because the cropland transects were not stratified and there were few samples per transect. For the purpose of this simulation, we treated each of the 5 transects as if it were a random sample from a distinct stratum. Under this assumption, samples from a transect are representative of the distribution of %TC within the corresponding stratum. This assumption is probably false in a way that favors the stratified Student t-test—within-transect heterogeneity is likely lower and between-transect variation higher than the corresponding quantities in an actual stratified random sampling design. The simulations sampled independently with replacement from each distribution (either pooled or stratified by transect), conducted the tests at nominal significance level 5 %, and recorded whether the null was rejected. The nonparametric test requires the user to specify an upper bound on the concentration: smaller bounds leads to more powerful tests, but misspecification can make the test invalid. We ran nonparametric tests with upper bounds of 10 %, which exceeds the maximum in any of our data (7.8 % TC), and 20 %, the established bound on TC in mineral soils. We also ran the nonparametric tests at a significance level of 10 examine how raising the significance level increases power. We ran each simulation 500 times, with 10, 30, 90, or 200 samples drawn from the population at each epoch. For stratified sampling, sample sizes were allocated proportional to "size," measured by the number of samples in the original transect. All statistical analyses were conducted in R (version 3.6.1). Code is available at: `https://github.com/spertus/soil-carbon-statistics`.

3.3 Results

3.3.1 Spatial heterogeneity of SOC and bulk density

In both rangeland and cropland soils, TC% generally decreased with depth (Figure 3.1). In rangeland soils, mean TC% varied from 3.77 % in topsoils (0–10 cm) of the summit/shoulder transect to 0.47 % at 75–100 cm of the footslope transect. Mean TC% in cropland soils varied from 4.31 % at 0–15 cm (at CROP3) to 0.10 % at 60–100 cm (at CROP7). Permutation ANOVA found that variations in mean TC% were statistically significant across transects ($p < 10^{-4}$) and depth ($p < 10^{-4}$) in rangeland soils and across sites ($p < 10^{-4}$) and depth ($p < 10^{-4}$) in cropland soils. Mean TOC% at the rangeland site was similar to TC% (SI 10): most samples had low TIC%. The spatial heterogeneity of TC% varied with land use, depth, and geographic location (transect and site; Table 3.1). Heterogeneity of TC%, as measured by the coefficient of variation (CV), was higher in the rangeland site than in the cropland

CROPLAND

Depth (cm)	CROP1	CROP2	CROP3	CROP4	CROP5	CROP6	CROP7	Total
DFS	Vineyard	Orchard	Crop	Crop	DFS	Orchard		
0–15	2.45 (0.26)	0.82 (0.26)	4.31 (0.26)	2.37 (0.21)	2.74 (0.25)	0.94 (0.20)	0.64 (0.21)	2.04 (0.24)
15–30	1.21 (0.19)	0.56 (0.41)	3.05 (0.31)	1.14 (0.15)	2.03 (0.21)	0.73 (0.27)	0.17 (0.36)	1.27 (0.25)
30–60	0.73 (0.39)	0.36 (0.58)	2.36 (0.34)	0.88 (0.13)	1.60 (0.22)	0.50 (0.42)	0.10 (0.44)	0.93 (0.31)
60–100	0.42 (0.77)	0.25 (0.62)	–	0.56 (0.21)	1.22 (0.21)	–	0.12 (1.14)	0.52 (0.38)

RANGELAND

Depth (cm)	Bx	By	Mx	My	T	Total
0–10	1.55 (0.27)	1.63 (0.32)	2.02 (0.36)	1.63 (0.32)	3.77 (0.36)	2.16 (0.54)
10–30	0.67 (0.48)	0.90 (0.27)	0.99 (0.32)	0.86 (0.20)	1.99 (0.36)	1.11 (0.56)
30–50	0.60 (0.66)	0.64 (0.28)	0.82 (0.33)	0.71 (0.39)	1.32 (0.48)	0.78 (0.53)
50–75	0.59 (0.75)	0.53 (0.41)	0.75 (0.51)	0.59 (0.36)	1.41 (0.63)	0.71 (0.70)
75–100	0.65 (1.17)	0.47 (0.27)	1.01 (0.71)	0.94 (0.61)	0.96 (0.34)	0.84 (0.75)

Table 3.1: Estimates of TC% means and coefficients of variation (CV; in parentheses) for cropland and rangeland. Mean and CV for rangeland sites are listed by transect. Mean and CV for croplands are listed by site. Cropland depths were not always consistent by site. For example, the second sampling depth ranged from 15–30, 15–35, and 15–40 in some cases. We used the most common depth increments here.

sites. The CV increased with depth in every rangeland transect and in cropland sites with diversified or perennial farming systems (vineyards and orchards), but not in conventionally managed croplands. Bulk density was highly variable with land use and across sites but generally not with depth. Heterogeneity was particularly high in the rangeland soils, where CV ranged from 0.08 at 30–50 cm and 75–100 cm to 0.16 at 0–10 cm and 50–75 cm. Heterogeneity within the cropland sites was lower with CV ranging from 0.04 at 0–15 cm and 15–30 cm to 0.07 at 60–100 cm. Within a given depth, BD varied substantially across rangelands (15 soil pits) and the CROP7 site (16 soil pits), but no consistent patterns emerged (Figure 3.2). BD for the six other cropland sites combined is plotted in SI Fig. 12. Rangeland SOC stocks were 30.3, 31.6, 22.5, 25.9, and 30.4 Mg C/ha at 0–10, 10–30, 30–50, 50–75, and 75–100 cm, respectively. Whole profile stocks (0–100 cm) were 141.7 Mg C/ha, SE: 6.7 (SI Table 3). Like most SOC stock estimates, the estimated SE does not reflect uncertainty and variability of bulk density (although we argue below that those uncertainties should be taken into account). In croplands, whole profile SOC stocks varied by site from 32.6 Mg C/ha (0–100 cm for CROP7) to 230.0 Mg/ha (0–70 cm for CROP3) (SI Table 4).

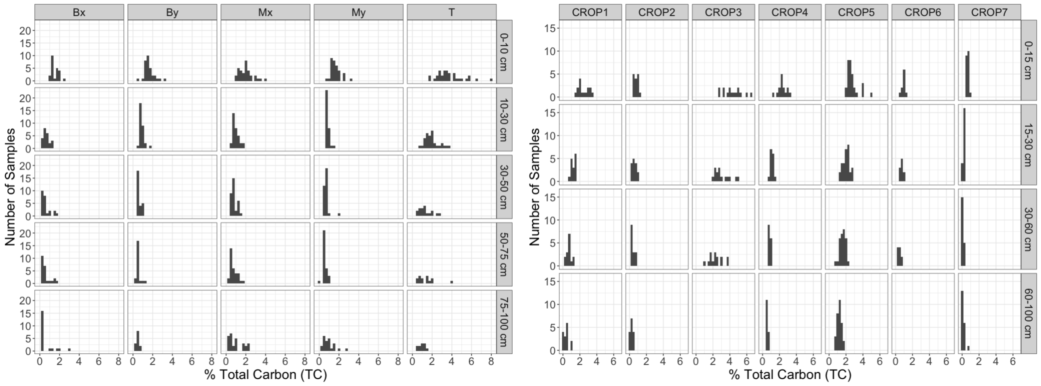


Figure 3.1: Histograms of TC% by transect and depth in rangeland soils (left panel) and site and depth in cropland soils (right panel). Transect labels in (a) refer to catental positions and replicates: Bx (footslope, replicate X), By (footslope, replicate Y), Mx (backslope, replicate X), My (backslope, replicate Y), and T (summit/shoulder, no replication). Site labels in (a) refer to cropland sites: CROP1 is a diversified farming system, CROP2 is a vineyard, CROP3 is an orchard, CROP4 and CROP 5 use conventional cropping, CROP6 is a diversified farming system, and CROP7 is an orchard. Depth increments differed between rangeland and cropland sampling schemes. Plotted values are TC%, which is equal to TOC% in samples with zero TIC.

3.3.2 Analytical variability

We compared measurements of 25 analytical replicates of two standard soils on the soliTOC and ECS 4010 dry combustion analyzers. Both instruments showed low variance and a small but consistent positive bias (SI Fig. 4). Based on analytical replicates of 36 samples measured on both instruments, the estimated median relative errors of the measurements were 0.024 for the soliTOC and 0.061 for the ECS 4010 (Figure 3.3). Permutation tests generally found little evidence of systematic differences between the instruments in replicated TC% measurements, except for samples with TIC% greater than 10 % of TC%. In the most extreme case, average replicated TC% measured on the soliTOC was nearly triple that of ECS 4010 for a rangeland sample with 90 % of TC% as TIC%. Removing inorganic C with HCl improved the agreement of measured TOC% between the two instruments (SI Fig. 2).

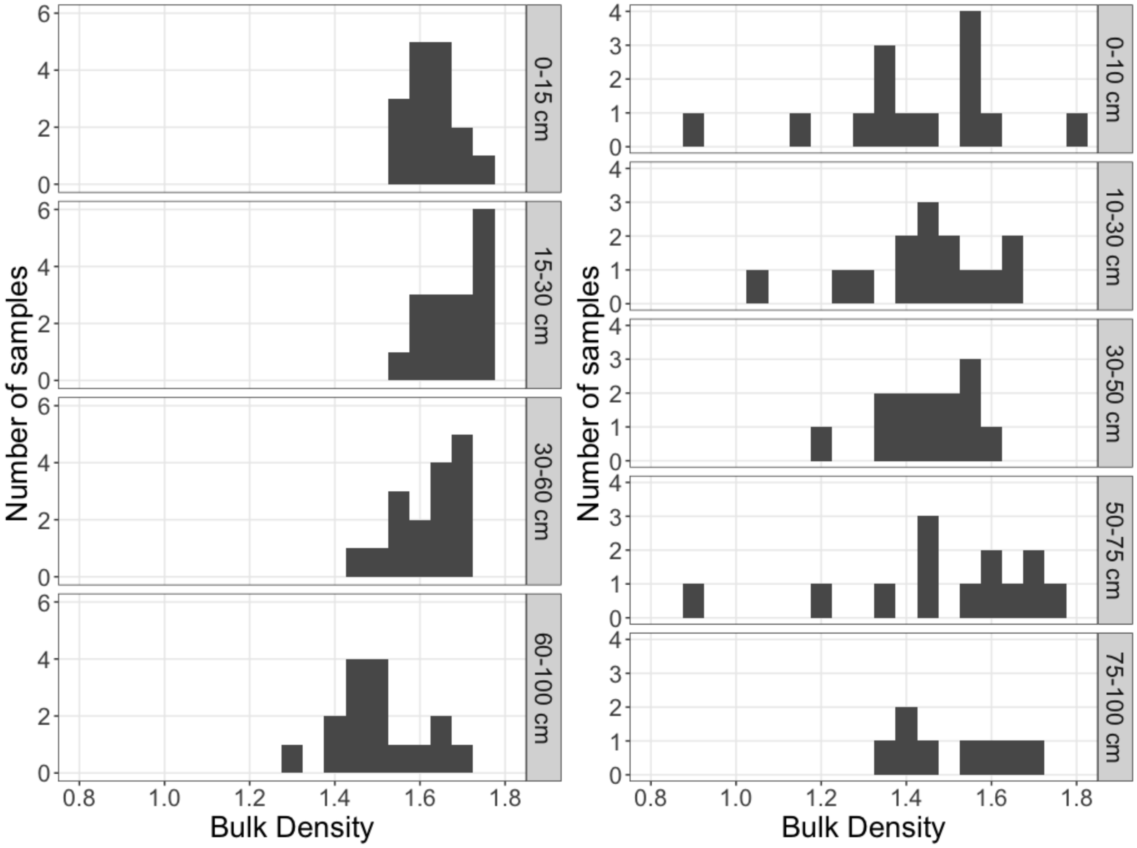


Figure 3.2: Empirical distributions of bulk density (BD) samples across 16 soil pits on CROP7 (left column) and 15 rangeland soil pits (right panel) by depth (in rows).

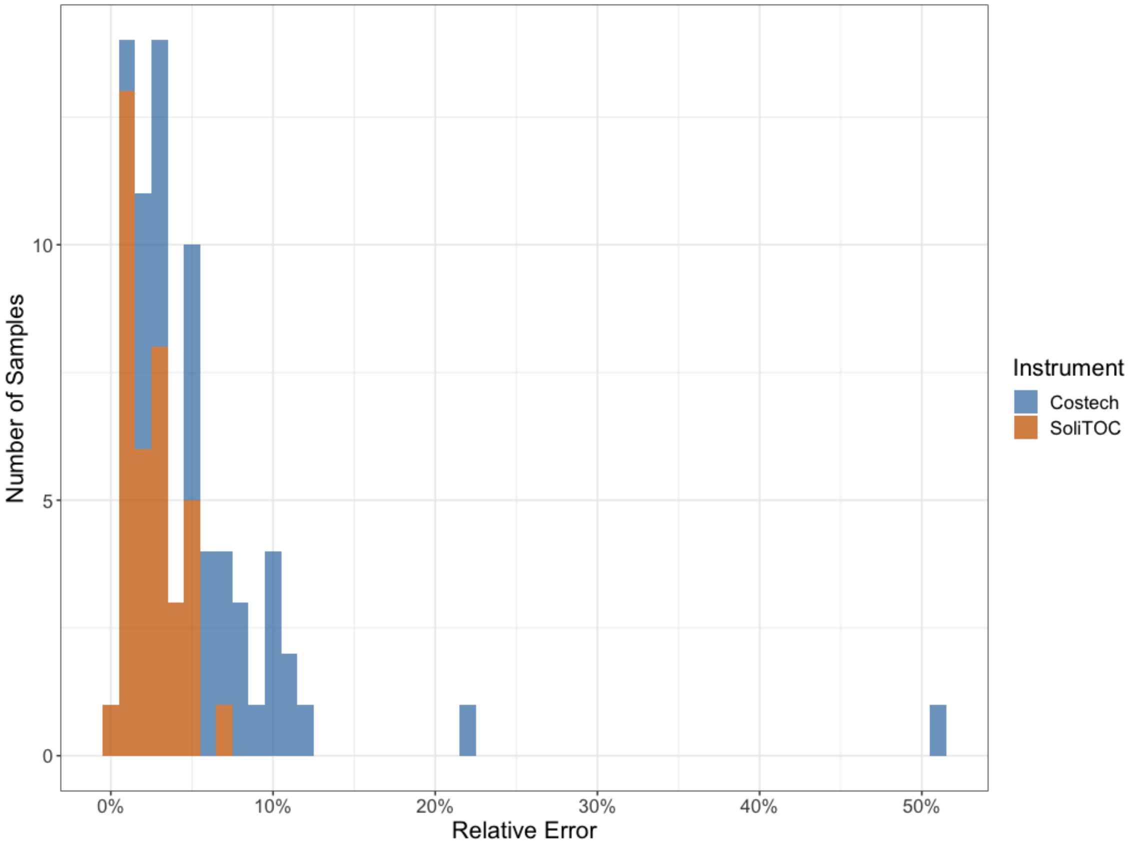


Figure 3.3: Histogram of relative error of replicated assays computed for each sample run on soliTOC (blue) and ECS 4010 (orange). Histograms bins are 1% relative error wide and stacked. The samples with relative error above 20% on ECS 4010 had high proportions of inorganic C.

3.3.3 Sources of uncertainty and their effects on statistical power

In general, spatial heterogeneity contributes much more uncertainty than analytical variability does, both for rangelands and croplands (Figure 3.4). However, compositing can mitigate or exacerbate the relative contributions to uncertainty from spatial heterogeneity and analytical variability. With no compositing, analytical variability contributes little to the overall uncertainty (Figure 3.4). If all n = 90 cores are composited to k = 1 analytic sample ("full compositing"), analytical error becomes a major component of the uncertainty in estimates of TC% for cropland soils, especially for the less precise ECS 4010 analyzer (Figure 3.4). The theoretical power of Student's t-test under various compositing schemes reflects this tradeoff (Figure 3.5). The power of Student's t-test to detect TC% change generally depends more on spatial heterogeneity than analytical variability for both instruments, except for full compositing with the ECS 4010, which had much less power (Figure 3.5).

There was relatively little difference in power between optimal compositing and no compositing for every land use and analytical instrument. When spatial heterogeneity is high (e.g., in rangeland) and lab analysis is precise (e.g., with soliTOC), power is maximized by allocating more of the budget to sampling and using some compositing to reduce the number of assays. On the other hand, when spatial heterogeneity is low (e.g., in cropland) and lab analysis is imprecise (e.g., ECS 4010), accuracy is maximized by allocating more of the budget to assays and reducing or avoiding compositing.

3.3.4 Power and validity of tests for detecting TC% change

The nominal significance level of Student's t-test can greatly understate its actual chance of making a Type I error, i.e., of erroneously rejecting the null hypothesis when the hypothesis is true (Figure 3.6). In the validity simulations, the true significance level of Student's t-test was always larger than its nominal level, except when the distributions were both normal. In the "tilled cropland" and "change in skew" scenarios, the level was close to 10 % at very small sample sizes, but approached its nominal 5 % at larger sizes. In the "extreme hotspot" scenario, the true significance level was always many times higher than the nominal significance level, and remained above 20 % for a sample size of 150. In contrast, the nonparametric test never erroneously rejected the null hypothesis.

Our "power" simulation compared the power of the unstratified Student t-test, stratified Student t-test, and a nonparametric test for detecting SOC shifts, for

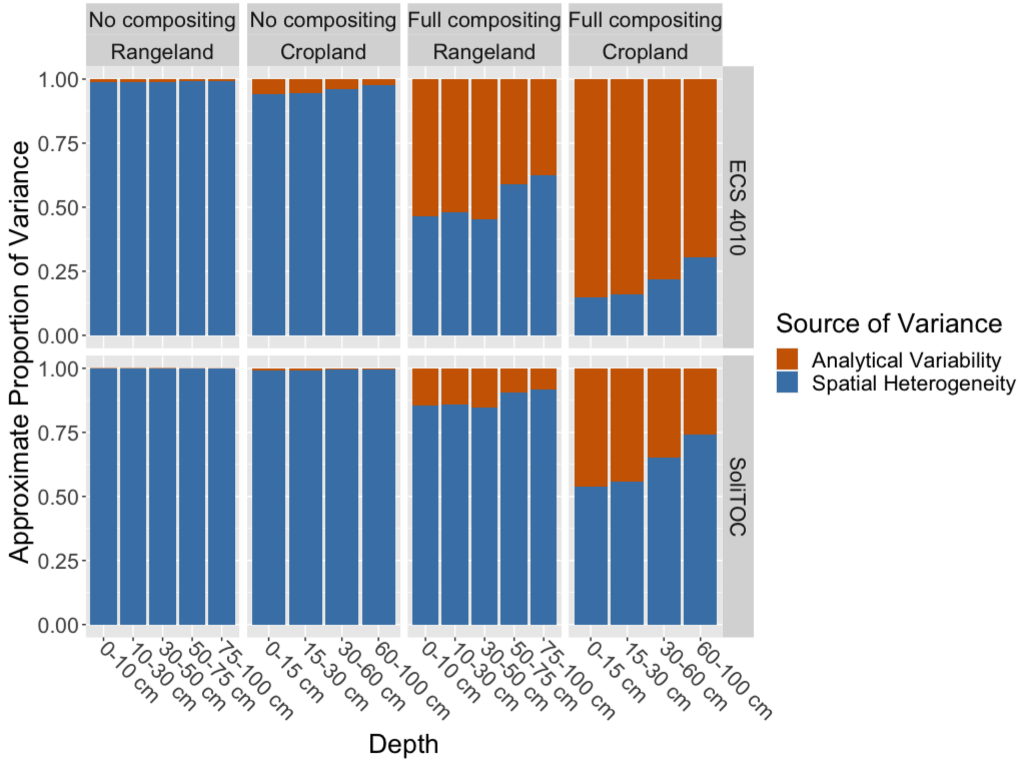


Figure 3.4: Contributions to the variance of the sample mean from assay uncertainty and compositing. Proportion of variance (y-axis) reflects assaying either all field samples individually ("No Compositing" panels) or 90 field samples together ("Full Compositing"). Different panels correspond to different land uses (in columns) and instruments (in rows). Field heterogeneity is estimated using data from the rangeland site or averages across cropland sites, at various depths (x-axis). Assay variability is estimated either on ECS 4010 (top panels) or soliTOC (bottom panels) elemental analyzers. Cropland depths vary slightly by site (see Figure 3.1).

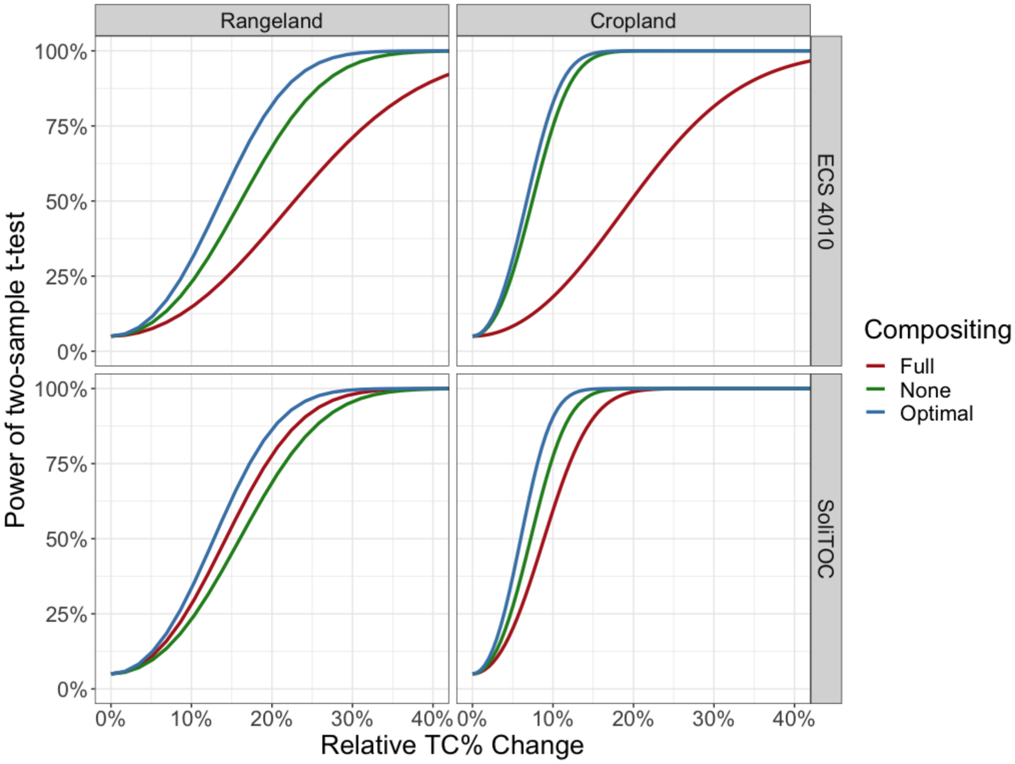


Figure 3.5: Theoretical power of Student's t-test to detect changes in topsoil TC% (if TC% is normally distributed) for two levels of compositing, for a budget that covers the cost of 90 cores and 90 laboratory analyses without compositing, or 150 cores and one laboratory analysis for full compositing. The X-axis shows the relative change in average TC% from baseline (2.16 TC% for rangeland or 2.04 TC% for cropland); the Y-axis is power. Different panels correspond to different land types (in columns) and instruments (in rows). Colors correspond to the compositing scheme. Optimal compositing for the same budget uses 140 cores composited to 19 analytic samples (SI 1.4).

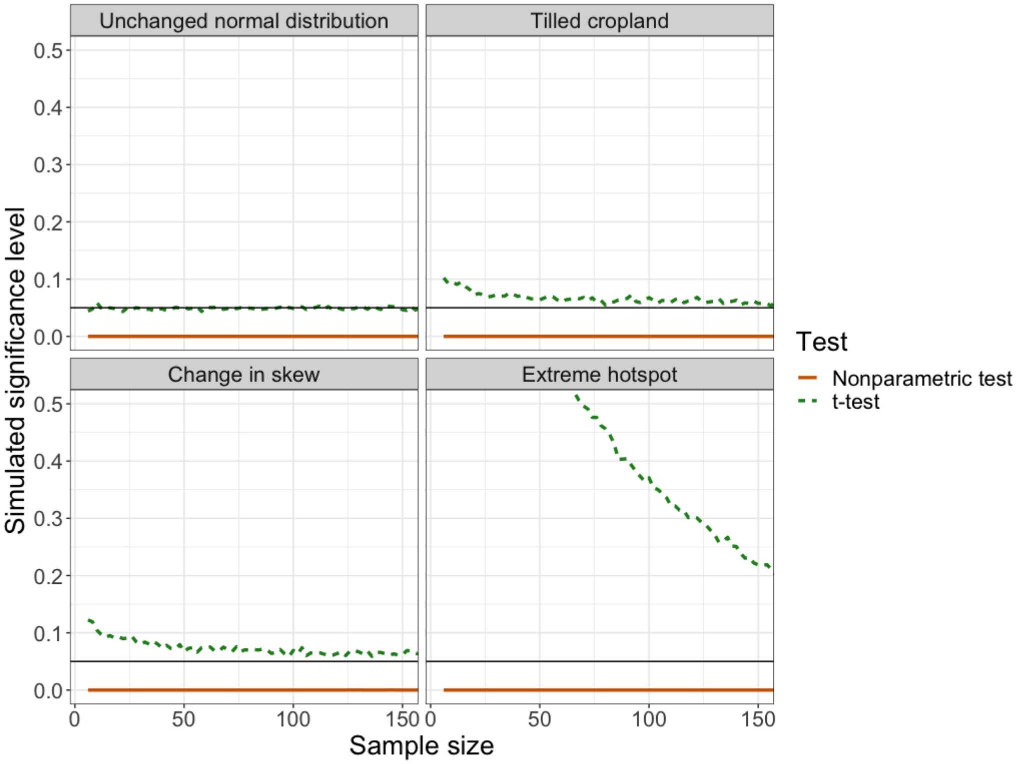


Figure 3.6: Simulated significance levels of two nominal 5% level tests: the two-sample Student t-test and a nonparametric test. Each panel reflects 5000 simulations at each sample size (x-axis); random samples were drawn independently from each of two distributions that had identical means. The y-axis plots the rate of false rejections of the null hypothesis (the Type I error rate). The solid black line is the nominal 0.05 significance level, which both curves should be at or below.

sample sizes of 10, 30, 90, and 200 from each time (Figure 3.7). Both Student t-tests appear to have more power than the nonparametric test to detect shifts in TC% at all sample sizes; the stratified Student t-test was more powerful than the unstratified test at the same level. (Stratified nonparametric tests would presumably have higher power than the unstratified nonparametric test; they are the subject of ongoing research.) However, comparing Student's t-test and the nonparametric test can be misleasing: Student's t-test rejects more often than the nonparametric test when the null hypothesis is false, but it also rejects more often than it should when the null hypothesis is true. Student's t-test at (nominal) significance level 5 % does not limit the true Type I error rate to 5 % unless the population has a normal distribution. In general, when the population distribution is not normal, the true Type I error rate of Student's t-test rate cannot be determined unless the population distribution is known. The power of the nonparametric test improved as the population bounds were tightened and as the level was relaxed: the nonparametric test with 10 % max TC and significance level 0.10 was the most powerful among the nonparametric tests. For example, to have 80 % power to detect a relative change of 20 % from baseline average TC% on rangeland soils requires about 30 samples with the stratified Student t-test, 90 with the unstratified Student t-test, and 200 with the nonparametric test with 10 % max TC and/or a relaxed significance level of 0.1. All tests had more power to detect small changes in cropland soils than in rangeland soils due to lower spatial heterogeneity in croplands. For example, the power of the unstratified Student t-test to detect a 10 % change with 90 samples was 80 % for cropland soils but only about 30 % for rangelands.

3.4 Discussion

3.4.1 Crop and rangelands are spatially heterogeneous

Given the rapid development of C markets, accurate detection and quantification of the impact of management interventions on SOC changes are more important than ever. Our study demonstrates how tailoring sampling and analytical decisions to the high spatial heterogeneity often found in managed lands could improve the reliability and efficiency of SOC sequestration estimates and their associated C credits. As expected, SOC at the rangeland site was more heterogeneous than at the seven cropland sites, with roughly twice as large a CV at every depth (Table 3.1). This is consistent with other surveys of California rangelands, though differences in depths, spatial scales, and measures of variability limit quantitative comparisons ([Carey et al., 2020, Devine et al., 2020, Silver et al., 2010]). This is also consistent with the

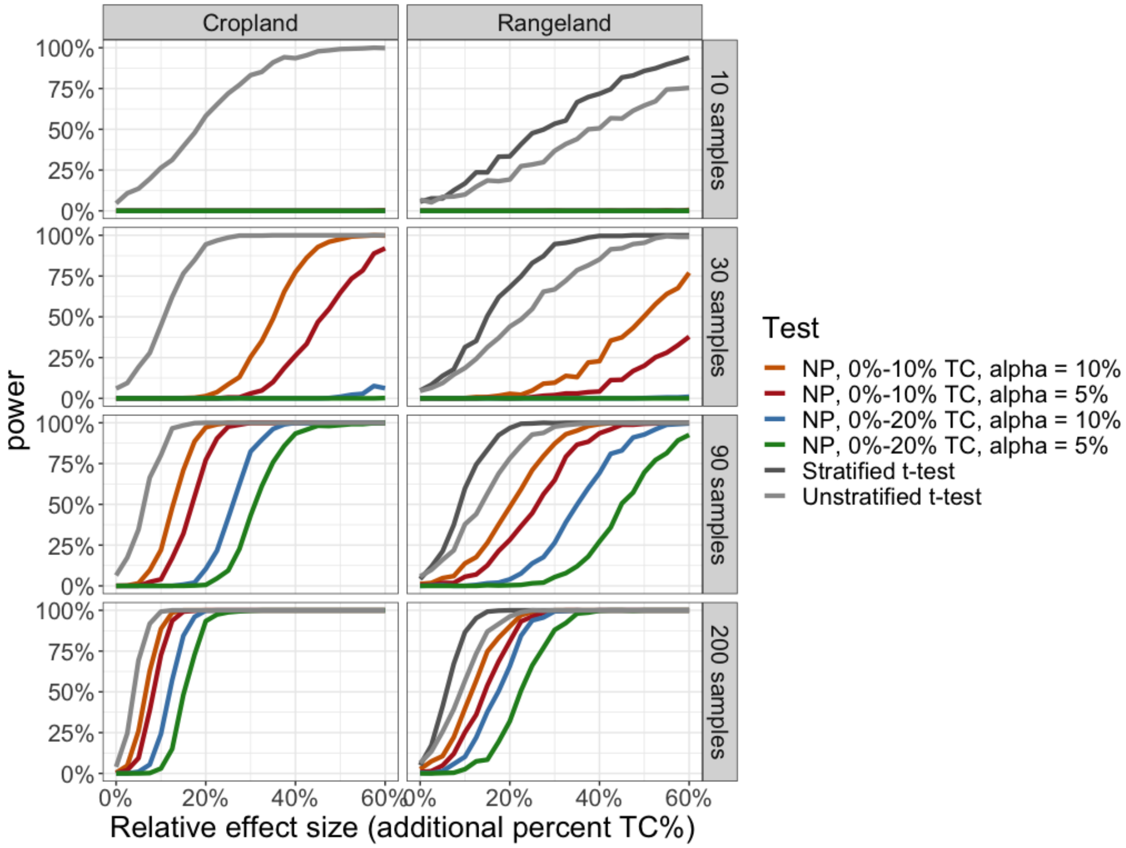


Figure 3.7: Simulated power of three two-sample hypothesis tests (Student's t-test for unstratified and stratified samples and a nonparametric test for unstratified samples based on Learned-Miller and Thomas [2019] to detect relative changes in TC% with sample sizes 10, 30, 90, or 200 from each of two populations. The first population distribution is the empirical distribution of TC% measurements for CROP5 (left column) or the rangeland site (right column) topsoil. The second population distribution is the same as the first, but each value was shifted by 0 % to 60 % of the mean of the first population, 2.7 % TC for the cropland samples and 2.2 % TC for the rangeland samples. Unstratified samples were simple random samples with replacement from the populations. Stratified samples from rangeland were simple random samples with replacement within transects, independent across transects, with sample sizes proportional to the original number of data in each transect. Stratified samples from cropland were not explored because there were no natural strata in the original data. Four curves are presented for the nonparametric test by varying pre-specified upper bounds on the population (10 % TC or 20 % TC) and the nominal significance level (5 % or 10 %); both the Student t-tests used a nominal significance of 5 %, which may understate the chance of false positives. NP = nonparametric.

general notion that rangeland SOC is typically more heterogeneous than croplands because of variation in topography, presence of rock fragments, and patchiness of grazing and manure deposition. This makes accurately estimating SOC change on rangelands more challenging.

Deep sampling is critical for making reliable conclusions about C sequestration and greenhouse gas mitigation [Kravchenko and Robertson, 2011, Kuzyakov and Blagodatskaya, 2015], especially since SOC gains near the surface may be offset by losses at depth [Poffenbarger et al., 2020, Slessarev et al., 2021, Syswerda et al., 2011, Tautges et al., 2019]. The CV of SOC in our study tended to increase with depth, while standard deviations decreased. Hence, a given relative change (e.g., 10 % gain from baseline TC%) is harder to detect in subsurface soils than in topsoil, but a given absolute change (e.g., 0.5 TC% gain) may be easier to detect. Since detecting an equivalent absolute change in the subsurface requires fewer samples, topsoil heterogeneity should generally guide decisions around sample size.

Though we have emphasized TC% measurements, high variability of BD within sites (Fig. 2), especially in rangelands, contributes additional uncertainty to SOC stock estimates [Walter et al., 2016a]. Even where TC% is relatively homogeneous, variability in BD could lead to large uncertainties in estimates of SOC stocks and of SOC stock changes, and ultimately prevent the reliable detection of these changes [Slessarev et al., 2021]. Failing to account for BD variability (e.g., treating BD estimates as fixed) underestimates uncertainty and may lead to erroneous conclusions about SOC stock change. See SI 4.1. for further discussion on combining BD and TOC% uncertainties for SOC stock estimates.

3.4.2 Analytical variability contributes little to measurement error

Variability in assay measurements contributed far less to measurement error than spatial heterogeneity of SOC, except when samples were highly composited (Figure 3.4). Replicated measurements show that both the soliTOC and ECS 4010 analyzers have estimated median relative error below 0.07.

TC%, however, differed substantially between instruments for samples with high TIC%, (see also SI Fig. 3). re-analysis on the ECS 4010 after TIC removal improved agreement between TC% measurements on the ECS 4010 and TOC% on the soliTOC, indicating that TIC should be removed when using elemental analyzers like the ECS 4010. Larger sample masses (10–20x the mass of traditional dry combustion instruments) may explain the higher precision of the soliTOC, which had about one-third the median relative error of the ECS 4010 (Fig. 3). Larger analytical subsamples should better represent the entire sample and reduce variability inherent to small subsamples. In the case of the ECS 4010, increased analytical replication may be necessary. SOC monitoring schemes could mitigate analytical error by using the same instrument, ideally in the same lab, for repeated analysis, and by including standards with comparable amounts of TIC, when analyzing samples known to contain TIC.

3.4.3 Measurement protocol recommendations to reduce uncertainty

Spatial heterogeneity is likely to dominate measurement uncertainty in many scenarios. We recommend three ways to for measurement protocols to reduce uncertainty in SOC estimates and increase the reliability of C credits: provide stratified sampling guidance, minimize compositing, and, most importantly, require larger sample sizes.

Stratification on variables such as catenal position, soil type, topography and historical management can increase the power of detecting SOC sequestration and generally reduces uncertainty for a given total sample size on heterogeneous landscapes [Devine et al., 2020, de Gruijter et al., 2016]. Our simulations provide further evidence that stratification can be a useful sampling strategy: stratified sampling had higher power to detect increases in TC% at the rangeland site than simple random sampling (Figure 3.7). Without stratification, far larger sample sizes are required to reliably detect and quantify SOC changes. While current protocols such as Climate Action Reserve's (CAR) Soil Enrichment Protocol and Verra's VM0021 allow and encourage stratification, they do not provide straightforward and quantitative stratification guidance. Preliminary field surveys, geospatial information regarding soil and landscape features, and expert pedological knowledge are useful for defining strata in research settings [Post et al., 2001]. C market protocols should look to incorporate algorithmic stratification [Devine et al., 2020, de Gruijter et al., 2016], digital soil mapping, and user-friendly software tools (e.g. Stratifi; `https://www.quickcarbon.org/tools`), to help standardize and ease barriers to stratification for SOC measurement.

Compositing can be optimized to minimize uncertainty within a cost budget, given estimates of the analytical precision, spatial heterogeneity, and the (marginal) unit cost of collecting, preparing, and analyzing a sample [Spertus, 2021]. Without such estimates, it is best to avoid compositing (especially when collecting baseline samples), because it reduces information on spatial heterogeneity, complicates sampling designs and analyses, and increases the contribution of analytical error (Figure 3.5). Compositing also tends to reduce power by decreasing the effective sample size. Compositing is most helpful when SOC is highly heterogeneous, the cost of each laboratory assay is high, and the budget is small. In such cases, investigators should consider optimal, rather than full compositing [Spertus, 2021]. We've developed a web app for investigators (including for use in soil C measurement protocols) to help determine optimal compositing schemes, which is accessible at: `https://scf.berkeley.edu/shiny/bosf/soil-carbon-statistics/`.

Finally, many current sampling designs for the sale of C credits use sample sizes that are too small to allow any statistical test to have a reasonable chance of detecting moderate changes in SOC [Necpalova et al., 2014] or quantifying SOC changes on heterogeneous landscapes on relevant timescales. To illustrate, assume that compost application on rangelands increases relative TC% by 20 % (as per Ryals et al. [2014] after 3 years of application). Based on the spatial heterogeneity we observed in rangeland soils, in order to have 80 % power to detect such an increase (using stratified sampling and Student's t-test) would require collecting and analyzing nearly 100 soil samples at baseline and another 100 samples after the compost was applied, with no compositing (Figure 3.7).

Most rangeland management interventions, however, such as improved grazing practices, are expected to produce much smaller C gains. For instance, Conant et al. [2017] found a relative increase of 10 % from grazing improvements. The smaller the anticipated change in SOC, the larger the sample size must be to reliably detect and quantify the change. Similarly, using nonparametric tests—which may be needed to properly control the false positive rate—require larger sample sizes. For instance, it would require more than 200 samples to have an 80 chance of detecting a 10 % relative increase in SOC using either the unstratified Student t-test or the nonparametric test (Figure 3.7). No matter the sampling design or statistical test, the sample sizes typical in current campaigns and protocols (e.g., 8 samples for USDA GRACEnet; 9 samples composited to 1 for CDFA Healthy Soils Program; minimum of 3 samples for the Australian Carbon Credits Methodology; [Davis et al., 2017]) are far too small to have sufficient power to detect and quantify changes in rangeland SOC (Figure 3.7).

Our simulations suggest that detecting SOC changes in croplands may be easier than rangelands, but common sample sizes are still inadequate. Nonparametric tests have little chance of detecting reasonable changes with only 10 samples, and Student's t-test is likely to be misleading for such small samples and to lack sufficient power to detect realistic changes. With only 10 cropland samples, a relative increase of 30%—a very large change—would be needed for Student's t-test to have 80% power [Saby et al., 2008]. At the CROP5 site, Student's t-test required about 90 samples to have an 80% chance of detecting a 10 % relative change in TC%.

Given that sampling campaigns are routinely underpowered, we suggest a priori power analyses to determine site-specific minimum sample sizes [Kravchenko and Robertson, 2011]. This could include conducting a power analysis either by collecting and analyzing reconnaissance samples, or with regional and relevant spatial heterogeneity information (e.g., from prior studies or soil survey information). We also suggest routinely conducting post hoc power analyses to determine whether studies that find no effect of management on SOC had sufficient power to detect expected differences.

3.4.4 Tests must be valid to provide credible evidence of carbon change

Even when sampling is well-designed and executed, statistical analysis matters. Student's t-test and its relatives may erroneously conclude C was sequestered when it was not, at a much higher rate than the nominal significance level. As shown in Fig. 6, this occurs when even one of the TC distributions is skewed. The false positive rate for Student's t-test is particularly high when there are SOC hotspots or the distribution of TC (but not its mean) changes over time. Some management interventions redistribute SOC and create or destroy C hotspots [Baker et al., 2007, Kuzyakov and Blagodatskaya, 2015, Marin-Spiotta et al., 2014]. For example, establishing perennial intercrops or hedgerows and spreading high-C inputs such as biochar and compost can create SOC hotspots. Valid inference is crucial to measure SOC sequestration credibly; Student's t-test and related tests and confidence intervals likely often understate the chance of false positives and have an inordinately large chance of false negatives.

How can monitoring and verification campaigns ensure that estimates and inferences are reliable? An important consideration is whether the soil population of interest might have skewed SOC, including from SOC hotspots. If so, it might be possible to stratify the sample so that SOC distributions within strata are not severely skewed. Skewness in the population distribution makes Student's t-test behave particularly poorly. While transformations (e.g. logarithmic) are possible, skewness in the population that can undermine parametric statistical inferences may not be evident for realistic sample sizes. Larger sample sizes improve the approximations Student's t-test relies on, but in general, it is not possible to determine how large the sample must be for the approximation to have a particular level of accuracy [Cochran, 1977].

If hotspots might exist but their locations are unknown prior to sampling, Student's t-test should not be used. In our simulations, nonparametric tests were less powerful than Student's t-test, but they control the false positive rate for every SOC distribution (Figures 3.6 and 3.7), while Student's t-test can fail for some SOC distributions. Thus, Student's t-test may appear more powerful, but it is wrong more often. Our simulations show that using prior geochemical knowledge to bound TC more tightly (e.g. 10 % instead of 20 %) can increase the power of nonparametric tests, as can testing at a higher significance level (e.g. 10 % instead of 5 %). Deriving more powerful nonparametric tests is an active research area in Statistics [Romano and Wolf, 2000, Waudby-Smith and Ramdas, 2023], which we hope to extend to stratified soil samples (e.g., Wendell and Schmee [1996a]). We have written an R package to facilitate wider use of nonparametric tests, which can be installed from the R console by running `devtools::install_github(\spertus/nptests")`.

3.4.5 Study limitations and future research

Our analyses relied on soil samples that were collected using common approaches, rather than the sampling protocols we recommend here (systematic rather than random samples). This could understate overall spatial heterogeneity, making our findings conservative, if SOC is spatially autocorrelated. However, we found little evidence of spatial autocorrelation in our rangeland samples (SI Figs. 5–9). These simulations are a starting point; other changes to SOC distributions and deeper soil depths should be examined. The geographic extent of the soil sampling was also limited and thus does not fully represent the heterogeneity of croplands and rangelands worldwide, but we expect the qualitative differences in heterogeneity between them will be more broadly applicable.

3.4.6 Broader implications for research, C markets, and policy

There have been numerous calls to standardize protocols for measuring SOC [Bispo et al., 2017, Davis et al., 2017, Jandl et al., 2014], but complete standardization may not be practical given differences among project needs and budgets, landscape heterogeneity, and lab constraints. In particular, given the large contribution of spatial heterogeneity to the uncertainty of SOC estimates, protocols that require fixed sample sizes or generate C credits on the basis of a fixed, small, minimum number of samples are not appropriate. For instance, sampling designs optimized to detect SOC changes for croplands may have little chance of detecting similar changes on rangelands, which typically require larger samples because they are more heterogeneous. Instead, sampling design processes should be standardized, such as the use of algorithmic stratification and a priori power analyses to select sample sizes adequate to detect plausible changes. In the case of C markets, verifiers should ensure that the sample size was adequate to detect and quantify SOC sequestration prior to generating and selling C offsets.

The consequences of inaccurate estimates of SOC for C markets are large. Current verification sampling protocols used to quantify and generate C credits for C markets cannot reliably estimate SOC sequestration, especially on heterogeneous agricultural lands. This could result in SOC offsets having little connection to the true extent of sequestration [Jackson Hammond et al., 2021]. Verification protocols for croplands include Climate Action Reserve's (CAR) Soil Enrichment Protocol, Verra's VM0021, Australia's Carbon Credits Methodology (ACCM), and the Food and Agriculture Organization's (FAO) Global Soil Organic Carbon (GSOC). All four protocols require a minimum of only three or more samples per stratum, far fewer than required to estimate the impact of management changes on a timescale of years. Some—though not all—protocols also lack details on how to stratify and analyze the resulting data to estimate SOC stocks and stock changes. Especially because they sanction such small sample sizes, these protocols may often reward "false positives" and fail to reward genuine sequestration. We recommend revising each of these protocols to require substantially more samples tailored to land-specific spatial heterogeneity, and, following ACCM as an example, provide much more detailed and useful guidelines for participants on when, where, and how to sample to minimize uncertainties. While governments, companies, and society must decide what level of confidence suffices to demonstrate SOC sequestration (e.g. the ACCM accepts SOC sequestration with only 60 % confidence, to encourage participation), protocols must actually be able to deliver that level of confidence.

For some purposes, instead of estimating SOC stock changes on each participating farm or ranch, it might suffice to estimate the aggregate change across many farms/ranches, collecting few samples from each, to minimize costs. Alternatively, one might conduct intensive sampling on a random sample of sites or a network of regional research monitoring sites (which could be supported by the development of funding programs like AgARDA or through increases in funding to LTERs or Climate Hub networks). Limiting sampling efforts to a smaller number of dedicated sites representing a range of climates, soil types, and cropping systems could allow for more intensive sampling—with higher power to detect SOC stock changes. This intensive sampling could then be used to calibrate, validate, and improve models such as MEMS 2.0 (Microbial Efficiency-Matrix Stabilization) [Zhang et al., 2021] that can estimate SOC change across broader landscapes and generate SOC credits for similar farms and ranches. This may be a more efficient use of resources and could drive more accurate verification in the long-term. However, both strategies represent a shift from paying for results to paying for practices that are expected—but not guaranteed—to produce results.

3.5 Conclusions

Spatial heterogeneity of SOC is a primary obstacle to accurately measuring changes in SOC stocks, even with careful sampling design and execution, accurate assays, and rigorous statistical analysis. Attempting to measure or verify SOC sequestration using too few samples, poor sampling design, imprecise laboratory instruments, or inappropriate statistical analysis can undermine climate change mitigation goals. We highlighted errors, quantified uncertainties, and demonstrated potential improvements in design and analysis, using data from California croplands and rangelands. There are several straightforward ways that sampling schemes can be improved, especially for C markets. Collecting information on the degree and pattern of heterogeneity before a comprehensive sampling campaign can make it possible to use stratified sampling to advantage. Such information also makes it possible to perform power calculations and identify optimal compositing approaches, ensuring that the campaign has sufficient statistical power to detect anticipated changes in. In general, reliable inferences about the short-term effect of management interventions on soil C require larger sample sizes and less compositing than is commonly used. We demonstrate that Student's t-test has highly inflated false-positive rates in scenarios that may be common in the field and suggest caution when using Student's t-tests and its relatives for verifying changes, especially when sample sizes are small. Nonparametric statistical methods can control false positives for any sample size, without assumptions about SOC distributions; providing more reliable, trustworthy results. The power of nonparametric tests can be increased using transparent, verifiable assumptions (e.g. geochemical constraints on the maximum SOC). Careful planning and continued collaboration between soil scientists and statisticians will help improve accuracy and precision of SOC measurements.

Supplementary Information

Data and code are available at `https://github.com/spertus/soil-carbon-statistics`. Supplementary tables and figures are available at `https://www.sciencedirect.com/science/article/pii/S0016706122006309#m0005`.

Chapter 4

Soil organic carbon sequestration potential and policy optimization

4.1 Introduction

Global soils contain approximately double the amount of carbon stored in the atmosphere [Le Quéré et al., 2018], despite significant declines since the expansion of industrial agriculture [Sanderman et al., 2017]. If some soil organic carbon (SOC) could be restored, it would reduce the amount of CO2 in the atmosphere. Because negative emissions are necessary to curtail the effects of climate change, governments and scientists have stepped up research into sequestering SOC as a "natural climate solution" [Bossio et al., 2020]. This in turn has fueled enthusiasm in a stewardship philosophy and cluster of agricultural management techniques collectively called "regenerative agriculture." Regenerative agriculture aims to improve ecosystem and soil health in a holistic sense, with SOC sequestration as a (potential) co-benefit [Lal, 2020]. A large and growing body of empirical work aims to evaluate these claims, inquiring into the effects of land management changes (e.g., no-till agriculture, cover cropping, management-intensive grazing, etc) on SOC levels. Such research informs billions of dollars in policy investment [University, 2021, Minasny et al., 2017]. To incentivize regenerative agriculture in service of sequestration[1] policy-makers must attribute SOC increases to management interventions and, ideally, tailor these interventions for maximum impact.

However, several knowledge gaps currently jeopardize the effectiveness of policies aimed at SOC sequestration. In particular, the amount of SOC that could be

[1] Other benefits should not be overlooked, and may justify considerable public investment on their own.

sequestered[2] in a given soil after a given intervention is unknown. This sequestration potential is critical to understand because it determines what a proposed policy can accomplish, including the amount of C sequestered, the rate of drawdown, and the longevity of the sequestered C. Sequestration potential is an inherently causal concept, involving a comparison of the effects of two or more courses of action on the same soil.

To illustrate, suppose a hypothetical policy-maker is tasked with deciding whether to pay for a policy implementing no-till agriculture across cropland in the US corn belt (the target population). Ideally, they would know the total amount of SOC in the top 1 m of soil in each farm in (say) 5 years, **both** if that farm adopted no-till agriculture **and** if that farm were tilled annually. These quantities, only one of which can ever be observed on a given plot, are called potential outcomes[3]. As a function of time, potential outcomes are called potential trajectories (Figure 4.1). If the policymaker knew all the potential trajectories within a population of farms, they could take nuanced actions to optimize sequestration over time given farm-specific costs and benefits of each intervention. Our primary goal is to approximate optimal policy decisions defined in terms of potential trajectories.

4.2 Interventions

Interventions intended to sequester SOC can take many different forms, and a brief taxonomy seems useful. Interventions may be conceptualized as policy changes or as physical changes: incentivizing a farmer to make a change is not the same as making the change, since the farmer may not comply. The former concept is more important to real-world policy decisions and sequestration potential, while the latter is likely interesting to scientists (who seek to understand mechanism) and to individual farmers (who are directly in control of their operation).

A study may examine a point-in-time treatment (e.g., a bolus of an input) or a pattern of treatment continuing over time (e.g., yearly application of inputs). Some interventions (e.g., land-use change or adoption of conservation tillage) involve a

[2] Sequestration should be distinguished from storage: while sequestration refers to net removal of CO2 from the atmosphere, storage refers to the gross increase in SOC in the soil without accounting for inputs (e.g. fossil fuel use or C-rich amendments) [Chenu et al., 2019]. They can be distinguished empirically by accounting for inputs.

[3] Potential outcomes were first described by Jerzy Neyman in his seminal 1923 paper on agricultural experiments [Neyman, 1923]. Hurlbert [1984] provides a non-technical overview of terminology and issues in experimental design. [Imbens and Rubin, 2015] is a good reference on causal inference with potential outcomes.

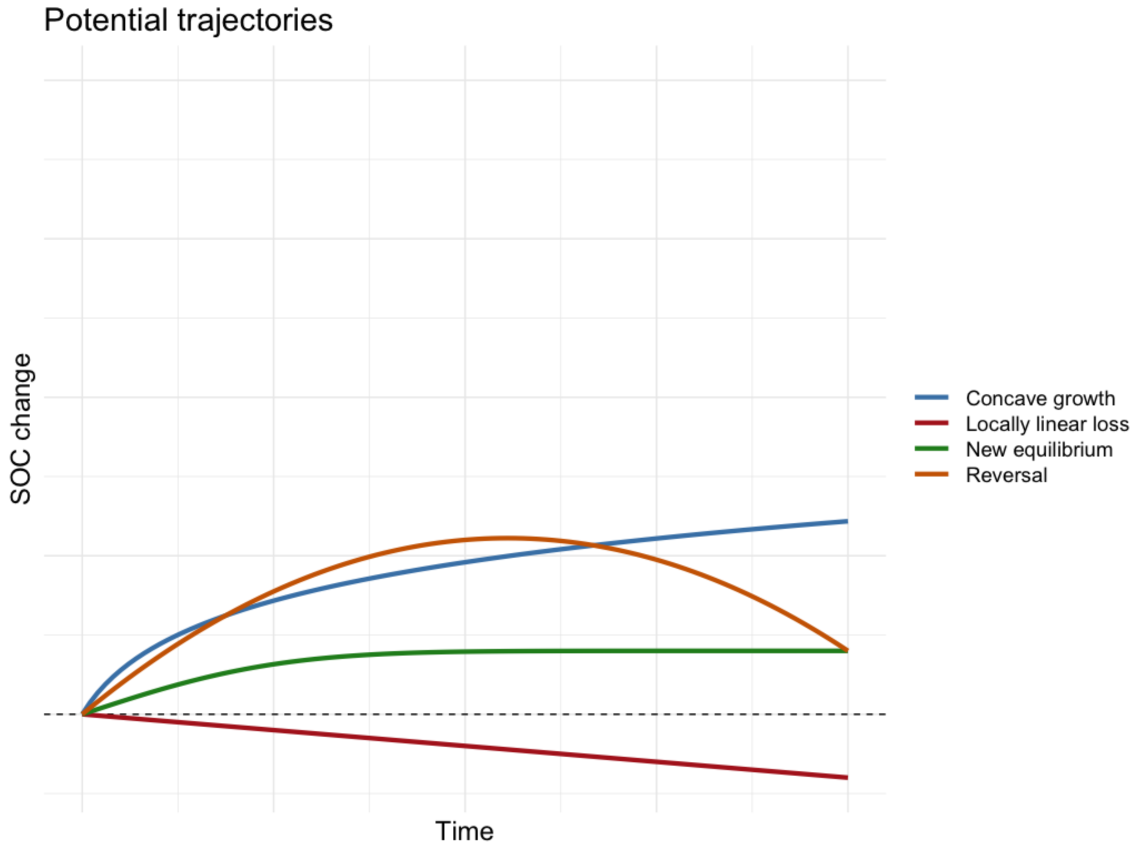


Figure 4.1: An illustration of four idealized population-level potential trajectories (colored lines) in terms of additional SOC sequestered (y-axis) over time (x-axis). From the point a decision is implemented, each course of action leads to a different trajectory of total SOC sequestered across the population. The trajectories are smoothed and do not reflect short-term variation like seasonality. A policy-maker with access to all potential trajectories and a target sequestration timeline could make optimal decisions under budgetary constraints. By definition, potential trajectories are equal at baseline (left most point on x-axis). The baseline SOC level is denoted by the dashed line. Under the green trajectory, SOC approaches a new equilibrium after the change has occurred, as theorized in Stewart et al. [2007] Under the blue and red trajectories, equilibrium is not achieved by the end of the time span plotted. The red trajectory displays a linear loss of SOC, as found, for example, in the study of Sanford et al. [2012]. Under the orange trajectory there is a reversal, a major concern of policies aiming to sequester SOC as a negative emissions strategy [Smith, 2005, Thamo and Pannell, 2016]. Researchers will often need to assume that trajectories are monotone, equilibrium is reached and maintained, and/or there are no future sequestration reversals when defining outcomes, designing studies, and interpreting results to make policy decisions.

phase transition that may or may not continue over time. To avoid ambiguity, is best to explicitly define the timing of the intervention when designing and interpreting studies (e.g., does "conservation tillage" entail merely a switch at baseline or indefinite maintenance of the practice?).[4] In any event, in this paper we will assume that the intervention is fully defined at baseline either as a point-in-time or a pattern over time.

Interventions can be conceptualized as quantities, discrete categories, or both. Quantitative interventions include amendments like compost, manure, or biochar application, which are readily summarized in terms of the amount applied (in total, per hectare, and/or per unit time). Categorical interventions include conversion tillage, adaptive multi-paddock (AMP) grazing, and land-use change. Sometimes interventions under study can include both, for example when comparing manure to compost at different levels or compost application to AMP grazing.

Interventions are often framed in terms of their levels of C input, which may be more or less abstract. In particular, the level of C input may be empirically constrained for some quantitative interventions (e.g. biochar) but rarely for categorical interventions. It is nevertheless useful to conceptualize interventions simply in terms of their C input, especially when comparing across treatment types and theorizing relations between treatment intensity (understood in terms of C input) and response. Such relations are an important element of sequestration potential.

4.3 Sequestration potential and saturation

The effectiveness of interventions is determined by the mechanisms that govern SOC sequestration. Interventions change SOC levels by changing the amount of C input to the soil and by changing the biological, chemical, and physical conditions that fix C. Broadly speaking, sequestration occurs when C input exceeds loss from decomposition, which varies as a function of environmental factors and soil properties. When C inputs and losses are balanced, SOC is at equilibrium [Chenu et al., 2019]. After an external change, SOC tends toward a new equilibrium over time (Figure 4.1). Sequestration potential thus depends on both the nature of the management intervention (influencing the C input rate and potentially the decomposition rate) and on the capacity of a given soil volume to retain additional SOC under that intervention. Critically, the starting concentration of SOC in a given soil may influence the decomposition rate, and hence the soil's capacity to store new C after an intervention. In the most

[4] Such questions are related to the concept of policy or intervention reversal, which could cause a reversal of sequestration. We discuss this further in Section 4.8.

extreme manifestation of this pattern, soils would have no additional capacity to store additional SOC under any intervention after the SOC level has reached a maximum amount: the SOC saturation hypothesis [Hassink, 1997]. In its original form, this hypothesis states that soils have a finite capacity to store SOC because SOC is protected from decomposition by complexation with silt and clay sized minerals; once the available mineral surface area is exhausted, no additional C can be stored [Hassink, 1997]. The precise dynamics depend on a complex interplay between the minerality and microbial biology of the soil, which can themselves change as a function of time and inputs. Empirical evidence for the SOC saturation hypothesis remains debated [Slessarev et al., 2023, Begill et al., 2023].

Regardless of the exact mechanism, C saturation is expected to result in diminishing storage of additional SOC as C inputs increase, with an eventual plateau once the soil is fully saturated [Stewart et al., 2007]. Conversely, if saturation does not occur, increasing C inputs might yield a linear increase in SOC storage. Different soil types may exhibit behavior in response to different types of treatment and/or the intensity of quantitative treatment. For example, drier mineral soils could exhibit saturation while organic wetland soils could exhibit linear returns to inputs [Gorham, 1991]. Alternatively, the reality for a given soil may be intermediate between these two extreme cases, with gains in SOC diminishing but not plateauing as C inputs increase. Four possibilities are presented in Figure 4.2.

However, not all interventions can be quantified in terms of C inputs, nor does total input sufficiently characterize the impact of any given intervention on SOC dynamics. In particular, additional complexity is necessary to theorize saturation when interventions are discrete (e.g., tillage vs no-till or conventional vs regenerative grazing) or when comparing quantitative C inputs applied in different forms (e.g., compost vs manure vs biochar). In canonical formulations of the saturation hypothesis, a discrete intervention may create a threshold–an effective stabilization capacity–below the absolute saturation limit by altering the decomposition rate of the soil [Stewart et al., 2007]. For instance, conventional tillage and no-till strategies may produce different effective stabilization capacities, both of which lie below the saturation threshold. The actual saturation threshold is then defined by a counterfactual scenario in which C inputs are maximized and disturbance is minimized (as in a native or wild state).

If the saturation hypothesis is true, management interventions are expected to have less impact on fields near their saturation threshold. Thus, all else being equal, baseline SOC levels may moderate the effect of interventions on SOC sequestered, proxy sequestration potential, and predict treatment effects. Fractionating mineral soils into mineral associated organic carbon (MAOC) and particulate organic carbon (POC) could yield better empirical proxies to long-term sequestration potential: the

MAOC pool is hypothesized to be both more stable and more susceptible to saturation than the POC pool [Cotrufo et al., 2019, Begill et al., 2023].

Proxies for sequestration potential–total baseline SOC, baseline MAOC, or other features–could play two key roles in policy decisions. First, they could improve effect estimates for policies applied to whole populations, which remain poorly constrained [Georgiou et al., 2022, Viscarra Rossel et al., 2024]. Such estimates are critical to regional, national, and international C accounting and mitigation projections within larger portfolios of positive and negative emissions [e.g., International Energy Agency [2023]]. However, real-world policy decisions involve more refined actions than treating or not treating an entire population: resources are constrained and usually only a limited number of plots in a population can be treated. In this context, proxies of sequestration potential would help determine which plots to treat and how in order to maximize the overall impact of a policy.

In the remainder of this paper, we address challenges that currently hinder empirical measurement of SOC and present a strategy to optimize sequestration given empirical proxies for sequestration potential. The next section lays out the hierarchy of uncertainties inherent to quantifying the effects of interventions on SOC stock. Then we formalize the overall policy objective of maximizing SOC sequestration across a population. We then provide a strategy for estimating sequestration potential and the optimal policy using empirical data. In a simulation study based on data from a study of compost application on California rangelands, we compare the effectiveness of our proposed methods against various alternatives. We conclude with a discussion of additional uncertainties, limitations of our methods, policy implications, and directions for future research in this area.

4.4 Measuring the effects of management interventions

Sequestration studies aim to evaluate whether and how much a proposed treatment will increase SOC stock. They are subject to a hierarchy of uncertainties–at the core, plot, study, and population scale–all of which must be controlled and, when possible, quantified in uncertainty estimates like confidence intervals. Additional uncertainties arise when translating study interventions to policy interventions. We begin our exposition at the most granular of SOC uncertainties–the core-level–and zoom out from there.

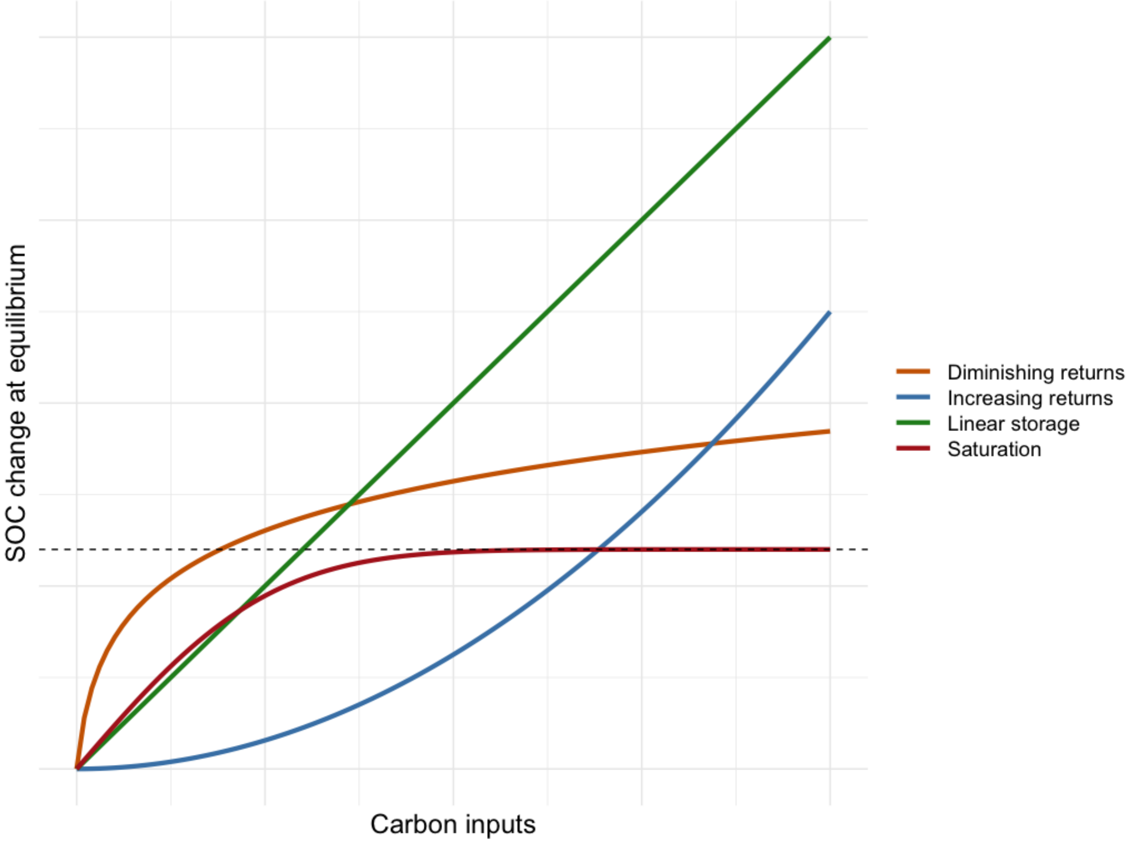


Figure 4.2: SOC at equilibrium (y-axis) as a function of C inputs from the intervention (x-axis) for various response possibilities (colors). C inputs may be thought of as a single pulse for a point treatment, or a total or average over time for a continuing treatment. The plot assumes all underlying potential trajectories reach equilibrium (see Figure 4.1). The red curve demonstrates complete saturation, wherein increasing inputs cannot increase equilibrium SOC above a certain point, the saturation threshold (dotted line) [Stewart et al., 2007]. The orange curve represents partial saturation, wherein there are diminishing returns to inputs but returns do not encounter a threshold. This possibility could arise under partial saturation, whereby a soil can be partitioned into two pools, one of which saturates entirely while the other does not [Hassink, 1997, Stewart et al., 2007, Cotrufo et al., 2019] The green curve entails no saturation whatsoever: returns are linear at any level of input. Peat soils are an instance of linear storage, as anaerobic conditions prevent decomposition [Gorham, 1991]. Note that when all inputs come from the intervention, this scenario implies perfect storage but no sequestration. Finally, the blue curve represents increasing returns to inputs, which could occur, for example, if a dead soil is progressively restored as inputs are increased.

4.4.1 Core-level uncertainties

To measure SOC stock, investigators take soil cores and measure them for SOC concentration (%SOC) and dry bulk density (BD) or equivalent soil mass (ESM; see next section). At the core level, uncertainties in %SOC arise from laboratory sample preparation protocols, human error, subsampling variability, and instrument error, which we collectively call assay variability. Assay variability can be addressed through careful laboratory work, including the use of precise instruments (e.g., elemental analyzers) and analytical replicates to control and estimate subsampling and instrument error. BD measurements are also subject to core-level uncertainties, including compression of soils during sampling, the presence of coarse fragments (e.g., gravel) in BD cores, and residual moisture. All of these factors need to be controlled through clear protocols, careful field work, and accounting. Finally, when MAOC is used to proxy saturation, the soil fractionation process requires a specific range of shaking or ultrasonic intensities to release the fine fraction without breaking up the coarse fraction and thereby contaminating the MAOC measurement with POC Amelung and Zech [1999], Six et al. [2024].

4.4.2 Plot-level uncertainties

Plot-level uncertainties are often more substantial than core-level [Stanley et al., 2023]. The main source of plot-level uncertainty is the variability of %SOC and BD across space–spatial heterogeneity–which can be estimated by random sampling. When heterogeneity is high, no single core is representative of the total stock in a given plot. Larger samples are needed to detect and quantify realistic stock changes when spatial heterogeneity is high, and typical sample sizes may be far too small [Kravchenko and Robertson, 2011, Stanley et al., 2023]. In some cases, stratification or balanced sampling can help to control spatial heterogeneity while keeping sample sizes and costs relatively low [de Gruijter et al., 2016, Potash et al., 2023]. Compositing sampled cores before assay may also reduce costs or increase precision, but the potential benefits are sensitive to the costs and errors associated with sampling and assay; compositing also introduces more opportunities for user-error [Spertus, 2021, Stanley et al., 2023]. BD change contributes uncertainty to stock change measurements depending on how sampling depth is determined. In particular, if samples are taken to a fixed depth, changes in BD (e.g., from soil compaction) can impact the measured SOC stock, even when the true SOC stock has not changed. The equivalent soil mass (ESM) procedure [Wendt and Hauser, 2013], attempts to navigate this issue, but changes the subject somewhat: stock change is expressed on a mass-mass basis, with BD measurements used only to determine the depth window over which the mass-mass average is taken.

ESM also introduces a modeling uncertainty, as the cumulative soil mass to SOC mass relationship must be estimated using a cubic spline.

Other sources of plot-level uncertainty are often overlooked but may be substantial. One such source is the presence of large rocks (i.e. boulders) in plots (especially on rangelands) that affect the volume of actual SOC-bearing soil. When rocks are not taken into account, stock estimates may be biased upwards and estimates of change may be attenuated. Unfortunately, precise accounting for large rocks–those larger than a soil corer or bulk density ring–is difficult and is not done in most studies, but at the very least their presence should be noted. Another source of uncertainty in volumetric[5] SOC measurement is negative correlation between %SOC and BD, which necessitates joint core-wise estimates of the two quantities across a plot. The common practice of estimating plot-level BD at a single pit using the ring method does not produce the data necessary to avoid this bias. Accurate, unbiased plot-level stock estimates require %SOC and BD to be measured on each core, volumetric concentrations to be computed at the core level, and the average of these concentrations to be taken as the average stock estimate (the estimate of the total being scaled according to the volume of soil).

4.4.3 Study-level uncertainties

Study-level uncertainties arise when multiple plots are involved in an observational study[6] or in a randomized controlled trial (RCT), which assigns different interventions to different plots at random and (ideally) measures SOC stock before and after the interventions are applied. For now, assume the study is an RCT with n equal-area plots enrolled and only two interventions: n1 plots are assigned to treatment and $n_0 = n - n_1$ are assigned to control. Every plot in the study has two potential outcomes (as above, a slice in time of a potential trajectory) recording what its stock would be after some years if it received treatment and if it received control. In symbols, let $Y_i(1)$ denote the SOC stock of plot i at the end of the study if it received treatment, and $Y_i(0)$ denote its stock if it received control. Ideally, we would know both potential outcomes for every plot, and could compute any causal summary of interest, including the individual plot treatment effect

$$\tau_i = Y_i(1) - Y_i(0),$$

[5] The ESM procedure of Wendt and Hauser [2013] does not have this issue since BD is not explicitly measured and ESM depths are determined separately for each core.

[6] Including cross-sectional "space-for-time" studies, as well as longitudinal studies without randomization.

which captures the amount of SOC sequestered in plot i by the end of the study attributable to treatment. In reality, we can only ever observe one potential outcome for each plot. If a plot is assigned to treatment, its potential outcome on treatment is observed while its outcome on control is counterfactual; vice versa if the plot is assigned to control. This means we cannot estimate any individual plot treatment effect i without entirely assuming its counterfactual (e.g., that $Y_i(0)$ is equal to baseline SOC), which is rarely justifiable. While we cannot estimate τ_i, we can draw conclusions about study-level causal parameters. In particular, randomized treatment assignment allows us to make unbiased estimates and valid inferences for the study average treatment effect (SATE):

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)].$$

The SATE records the utility of treatment in the study. For example, a SATE of 5 Mg SOC ha^{-1} indicates that 5 Mg of additional SOC per hectare would have been sequestered during the duration of the study had all plots been treated.

Even if there were no plot-level uncertainty, the SATE and other study-level causal parameters would be unknown (because only one potential outcome is observed for each plot), and estimates of them would be subject to random noise. In particular, inter-plot variability contributes uncertainty to estimates of the SATE. Inter-plot variability arises from spatial heterogeneity in %SOC and BD across plots, and from potentially variable responses of plots to treatment. It is thus critical for studies to enroll enough plots to estimate and control inter-plot variability. This can balloon the cost of the study, especially when attempting to detect relatively small treatment effects on a short time horizon. Control measures like blocking or pairing plots based on underlying features (e.g., location, soil type, historical land use, topography, etc) can help constrain inter-plot variability and improve precision without expanding the size of a study. Such design choices must be reflected in the method of data analysis.

There are important sources of study-level uncertainty beyond inter-plot variability. One such source is interference of treatment assignment between plots, wherein one plot's treatment assignment affects another plot's outcome. For example, interference may arise if two plots are adjacent to each other on a topographic gradient, so that C inputs on a treated plot (e.g. compost amendments) run-off onto the adjacent control plot. Another study-level uncertainty is adherence to protocol and its antithesis, noncompliance, wherein treatment received is not identical to treatment assigned. For example, noncompliance occurs when a plot assigned by the design to receive treatment (e.g. management intensive grazing) actually receives control (e.g. conventional grazing) for any reason (e.g. economic contingencies that incentivize a land manager to deviate). Noncompliance can bias estimates of treatment effects, and is difficult to account for at the analysis stage. It should be kept to a minimum as much as possible, and any deviations from protocol recorded.

4.4.4 Population-level (generalization) uncertainty

Population-level uncertainties arise when study results are generalized across space or time, which is a necessary step in interpreting findings, designing new studies, and informing policy decisions. To generalize across time typically requires assuming that potential trajectories have a particular shape; for example, that sequestration continues along a linear trajectory or reaches a plateau (equilibrium) and does not reverse. The linear trajectory assumption is implicit in the common practice of reporting change in terms of Mg C ha^{-1} y^{-1} and multiplying by years to extrapolate total sequestration over time, but is rather dubious. Reaching and maintaining a new SOC equilibrium is more likely on theoretical grounds, and reversals are possible. Furthermore, the trajectories cannot depend on the absolute time at which a management change occurs: they are assumed to be stationary. For example, a stationary treatment effect implies that switching the population from control to treatment now would create the same trajectory as making the same switch in 20 years. This assumption is reasonable in the absence of major changes in a population over time, but such changes could be caused (for example) by long-term climate change, which usually can't be entirely ruled out. Unfortunately, there is no way to account for non-stationarity in the design if it does exist, but results can be checked for their sensitivity to the stationarity assumption. On the other hand, generalizing across space–also called upscaling–can be made more or less rigorous by design. To ensure external validity, the plots enrolled in a study should be representative of the larger population to which the findings will be applied. The ideal strategy in this regard is to randomly sample plots from the population of interest and enroll them into an RCT. Such random enrollment is rarely feasible in practice, and studies more often enroll plots by systematic or convenience samples. When enrollment is done in this way, generalization must be based on informal reasoning, though auxiliary data may help. Such data record features of the plot and the population of interest that could influence the effect of treatment, such as climate, topography, soil type, or land use history. In some cases, quantitative data may be used to add precision and estimate uncertainty in the generalization [Egami and Hartman, 2023]. Quantitative methods for generalization may use inverse probability weighting [Cole and Stuart, 2010], outcome modeling [Nguyen et al., 2017], or doubly-robust estimation [Dahabreh et al., 2019]. All of these estimators require accounting for treatment effect moderators like baseline SOC or MAOC fraction.

4.5 Causal model of sequestration

There is a population of N volumes or *plots* of soil, each defined by a geographic area and a depth. For example, a plot could be all the soil to 1 meter on a particular ranch or a particular field, and the population could be all fields classified as California rangeland. In our setup, only a subset of $n \leq N$ plots in a given population will be randomly enrolled into the study and randomly assigned to treatment. Data from the study will be used to draw inferences about the population. In particular, we are interested in identifying an optimal or near optimal policy—one that will sequester the most SOC across the population given a fixed budget.

Throughout this section we will generally use lowercase to indicate fixed quantities and calligraphic font for sets (random or fixed). Exceptions are matrices, which will be bold and uppercase without italics (e.g., $\mathbf{A}$) but may be fixed or random, and N—the size of the population. Vectors will be bold and defined using square brackets, e.g. $\boldsymbol{x} = [x_1, ..., x_N]$ for a fixed vector and $\boldsymbol{X} = [X_1, \dots, X_N]$ for a random vector. Scalar multiplication is denoted $a\boldsymbol{x}$, while the dot product between two vectors is $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_i x_i y_i$. If $\mathcal{A}$ is a collection, $|\mathcal{A}|$ is its cardinality. Collections may contain multiple copies of the same element (technically, they are *multisets* or *bags*). Totals over collections or vectors will generally be denoted using bars—e.g., $\bar{x} := \sum_{i=1}^{N} x_i$—in contrast to their typical use to denote averages.

4.5.1 Potential trajectories and outcomes

For each possible treatment z taking a value in the set $\mathcal{Z}$, each plot i in the population has a fixed *potential trajectory* of soil organic carbon stock denoted $y_i(z, t)$, where $t \in \mathbb{R}_+$ indicates time from the application of treatment. Implicitly, by writing a potential trajectory in this way, we have assumed (i) *no interference*—the treatment plot i receives does not affect any potential trajectory of plot j if $j \neq i$—and (ii) *temporal stationarity*—the response may depend on the time since the treatment was applied, but the absolute time is irrelevant: $y_i(z, 10)$ is identical whether the experiment started, for example, in 1980 or in 2000.

We define *baseline time* to be $t = 0$ and *baseline stock* to be $y_i(z, 0)$. By definition $y_i(z, 0) = y_i(z', 0)$ for any two treatments $z, z' \in \mathcal{Z}$. We will express time in years, so $y_i(z, 5)$ is the carbon stock in plot i 5 years from baseline if it received treatment z. A *potential outcome* (PO) $y_i(z)$ is derived from a potential trajectory by fixing time. So, for example, we might define $y_i(z) := y_i(z, 5)$ for all i to be the stock on treatment z after 5 years. To simplify notation we will generally work with potential outcomes rather than trajectories, with the understanding that trajectories are relevant to extrapolating in time. For example, a 10 year study can only assess permanence to 100 years by making assumptions about the potential trajectory, such as $y_i(z) := y_i(z, 10) = y_i(z, 100)$.

The universe of possible treatments $\mathcal{Z}$ may be binary, categorical, or continuous. In all cases, we will assume that $0 \in \mathcal{Z}$ corresponds to a control treatment. For example, if we are interested in comparing between management intensive and conventional grazing, $\mathcal{Z}$ is binary and we use $z = 0$ to denote conventional grazing. Soil amendments like compost or manure may be continuous, expressed in Mg Ha^{-1}, in which case $\mathcal{Z}$ is a positive real number. Different amendments are sometimes put on the same scale by translating them to carbon inputs, though this may hide genuine differences among treatments: a ton of carbon from compost may have an effect on SOC quite different from that of a ton of carbon from biochar.

4.5.2 Agents and Goals

There are countless actors—human and nonhuman—who stand to gain or lose from actions taken on any given plot of land, let alone across a population of N plots. We reduce this network to $N + 1$ unique agents: a land manager for each plot i, and a policy-maker in charge of managing the population.

As an individual agent, the manager of plot i may wish to maximize $y_i(z_i)$ alongside additional outcomes, costs, and constraints (financial, environmental, biological, social, aesthetic, etc.) by choosing among possible actions $z_i \in \mathcal{Z}$. This is a complex goal requiring fine-grained knowledge and experience, perhaps only accessible to managers in close relationship to lands over long periods of time (i.e. a smallholder or indigenous community) [Scott, 1999].

This paper instead focuses on the role of a policy-maker with the narrow goal of finding a course of action that will approximately maximize (over all possible courses of action) the total amount of SOC stored in the population some number of years after the action is taken. This is the goal of SOC sequestration considered purely as a means of sinking atmospheric CO$_2$ to mitigate climate change. We will suspend normative judgements of that goal and, taking it as a given, focus on the practicalities of identifying effective policies. Along the way, we will also discuss how to quantify average treatment effects, which are important target parameters in scientific studies and policy decisions. Empirical proxies of sequestration potential may be useful for estimating both average treatment effects and optimal policies.

4.5.3 Parameters

Let $\boldsymbol{z} := [z_1, \ldots, z_N] \in \mathcal{Z}^N$ be a *treatment regime*, such that z_i is the treatment received by plot i. Let $\boldsymbol{w} := [w_1, \ldots, w_N]$ be the known relative areas of each plot such that $\sum_{i=1}^N w_i = 1$. Under a given treatment regime $\boldsymbol{z}$, the *population average potential outcome* ($\boldsymbol{z}$-PAPO) is

$$\bar{y}(\boldsymbol{z}) := \sum_{i=1}^N w_i y_i(z_i),$$

the average mass of SOC in the population after some years under treatment regime $\boldsymbol{z}$. We note that the total SOC sequestered is just $\sum_{i=1}^N y_i(z_i)$, and that estimation and inference on the average and on the total are equivalent up to the known factor N. We will work with the average to avoid the arbitrary scaling by N. We will also assume plots are of equal size (implying $\bar{y}(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^N y_i(z_i)$) to avoid carrying around the weight vector $\boldsymbol{w}$.

The canonical goal of causal inference is to estimate contrasts between aggregate potential outcomes. Under treatment regime $\boldsymbol{z}$, the *population average treatment effect* ($\boldsymbol{z}$-PATE) is:

$$\tau(\boldsymbol{z}) := \bar{y}(\boldsymbol{z}) - \bar{y}(\boldsymbol{0}) = \frac{1}{N}\sum_{i=1}^N (y_i(z_i) - y_i(0)) = \frac{1}{N}\sum_{i=1}^N \tau_i(z_i)$$

where $\tau_i(z_i) = y_i(z_i) - y_i(0)$ is the *individual treatment effect* (ITE) for plot i on treatment z_i. In this definition, the control treatment level ($z = 0$) always serves as a baseline for the contrast. Often, interest centers on $\boldsymbol{z}$-PATEs with $\boldsymbol{z} := \boldsymbol{1}z$, under which every plot in the population receives the same treatment.

For example, consider the important case when $\mathcal{Z}$ is binary. The ITEs are $\tau_i = y_i(1) - y_i(1)$. If every plot must receive the same treatment the (only) $\boldsymbol{z}$-PATE is written

$$\tau := \bar{y}(\boldsymbol{1}) - \bar{y}(\boldsymbol{0}) = \frac{1}{N}\sum_{i=1}^N \tau_i,$$

and denotes the additional mass of SOC in the population after some years if all plots were treated, compared to if they were all on control. We call the **1**-PATE simply the PATE and evaluate various estimators of it below. A positive PATE is usually taken to mean that treatment should be favored over control as a blanket policy, but if the ITEs $\{\tau_i\}_{i=1}^N$ vary—if there is *treatment effect heterogeneity*—then the PATE may be substantially lower than the best $\boldsymbol{z}$-PATE over all $\boldsymbol{z} \in \{0,1\}^N$. That is, a policy may have higher bang for its buck when treatments are allowed to vary across the population compared to when the treatment must be the same for all units.

4.5.4 Optimal treatment regimes

Let $\boldsymbol{y}(\boldsymbol{z}) := [y_1(z_1), \ldots, y_N(z_N)]$ record the POs for all plots in the population under treatment regime $\boldsymbol{z}$. The objective of the policymaker is to maximize the PTPO over $\boldsymbol{z}$ under resource constraints. Assume each treatment incurs a plot-specific cost $c_i(z_i)$ with $\boldsymbol{c}(\boldsymbol{z}) := [c_1(z_1), \ldots, c_N(z_N)]$, and that costs are additive across plots (e.g., there are no economies of scale in applying the same treatment to nearby plots) so that the total cost of implementing $\boldsymbol{z}$ is $\mathbf{1} \cdot \boldsymbol{c}(\boldsymbol{z})$. The policy has overall budget C_0. Define the *treatment portfolio*:

$$\mathcal{P} := \{\boldsymbol{z} : \boldsymbol{z} \in \mathcal{Z}^N, \mathbf{1} \cdot \boldsymbol{c}(\boldsymbol{z}) \leq C_0\}.$$

It is the set of all treatment regimes that meet the budgetary constraint. The *optimal PTPO* is

$$\bar{y}^* := \max_{\boldsymbol{z} \in \mathcal{P}} \bar{y}(\boldsymbol{z}) = \max_{\boldsymbol{z} \in \mathcal{P}} \mathbf{1} \cdot \boldsymbol{y}(\boldsymbol{z})/N.$$

An *optimal regime* $\boldsymbol{z}^*$ is a treatment regime $\boldsymbol{z}^*$ that achieves the optimal PTPO, satisfying $\bar{y}^* = \bar{y}(\boldsymbol{z}^*)$. If $\boldsymbol{y}(\boldsymbol{z})$ was a known and simple (e.g., linear) function of $\boldsymbol{z}$, finding the optimal regime would be straightforward. However, the set $\{\boldsymbol{y}(\boldsymbol{z})\}_{\boldsymbol{y} \in \mathcal{P}}$ is inherently unknown[7] since only one PO can be observed per unit.

A common way to simplify the problem is to consider only policies within a *restricted treatment portfolio*:

$$\mathcal{R} := \{\boldsymbol{z} : \boldsymbol{z} = \mathbf{1}z \text{ for } z \in \mathcal{Z}, \mathbf{1} \cdot \boldsymbol{c}(\boldsymbol{z}) \leq C_0\} \subset \mathcal{P},$$

and search for the optimum

$$\max_{\boldsymbol{z} \in \mathcal{R}} \bar{y}(\boldsymbol{z}) \leq \bar{y}^*.$$

When $\mathcal{Z}$ is discrete, this only requires estimating the $|\mathcal{Z}|$ values $\{\bar{y}(z\mathbf{1})\}_{z \in \mathcal{Z}}$, and is a well-studied problem in the causal inference, survey sampling, and multi-armed bandit literatures. However, when there is high TEH, the best single treatment may perform much worse than the best vector of (possibly different) treatments: treating all plots the same is far from the optimal regime if plots exhibit substantially different benefits to different treatments.

4.5.5 Predicting stock and predicting treatment effects

To better approximate $\bar{y}^*$, we can employ additional information about individual plots in the form of *covariates*. Previous work on SOC measurement has leveraged

[7] Unlike in a usual survey problem, where the population(s) could be enumerated by census.

covariates to stratify or balance sampling designs [de Gruijter et al., 2016, Potash et al., 2023], to increase the precision of total stock estimates [Viscarra Rossel et al., 2016, Särndal et al., 1992, Brus, 2000], or to make local predictions of stock (i.e., to construct a map) [Padarian et al., 2019, Devine et al., 2020]. In these contexts, covariates are effective when they predict SOC measurements. For example, when using ordinary least squares, a higher R^2 of the regression of measurements on covariates corresponds to a more precise stock estimate or a more accurate SOC map. Predictive and accessible covariates may include geography, topography, soil series, land-use history, wetness, and spectral data captured by satellite or drone. Crucially, the use of covariates in a working model does not necessitate that the model is correct in the sense that it accurately captures a "true data generating process." Rather, in a model-assisted approach [Särndal et al., 1992], estimation of total SOC stock is more precise if covariates are predictive, but inferences about SOC are asymptotically valid only under random sampling, which can be guaranteed by the study design.

Our strategy is similar to model-assisted stock estimation, but instead of a single population of actual mean SOC stock we are targeting:

1. The $|\mathcal{Z}| - 1$ mean differences $\bar{y}(z) - \bar{y}(0)$ representing $z\mathbf{1}$-PATEs under uniform treatment
2. The $|\mathcal{P}|$ population means $\{\bar{y}(\boldsymbol{z})\}_{\boldsymbol{z} \in \mathcal{P}}$ representing $\boldsymbol{z}$-PAPOs under different treatment regimes.

Both tasks can be accomplished by positing a regression model for each restricted regime population $\boldsymbol{y}(z)$ with $\boldsymbol{z} \in \mathcal{R}$ or a combined model with treatment-covariate interactions [Freedman, 2008a,b, Lin, 2013, Ding et al., 2019, Künzel et al., 2019].

4.5.6 Decomposing potential outcome populations

Let $\boldsymbol{x}_i$ be a length-p vector of covariates for plot i with constant term 1 in the first index. For any $z \in \mathcal{Z}$, a unit-level PO can be decomposed into:

$$y_i(z) := \boldsymbol{\beta}(z) \cdot \boldsymbol{x}_i + \varepsilon_i(z)$$

where

$$\boldsymbol{\beta}(z) := \arg \min_{\boldsymbol{a} \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i(z) - \boldsymbol{a} \cdot \boldsymbol{x}_i)^2$$

is the finite-population least squares linear regression coefficient of $\boldsymbol{y}(z\mathbf{1})$ on the matrix of covariates $\boldsymbol{\mathcal{X}} := [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T]^T \in \mathbb{R}^{N \times p}$, and $\varepsilon_i(z) := y_i(z) - \boldsymbol{\beta}(z) \cdot \boldsymbol{x}_i$ is idiosyncratic variation in the POs. By definition $\sum_{i=1}^{N} \varepsilon_i(z) = 0$, so the variance of the idiosyncratic effects is $\sigma_\varepsilon^2(z) := \sum_{i=1}^{N} \varepsilon_i^2(z)$. Since everything so far is at the population level, all these quantities are fixed (not random variables) and no distributional or parametric assumptions are involved in the decomposition above.

Considering a fixed $z \in \mathcal{Z}$, a closed-form solution for $\boldsymbol{\beta}(z)$ is available through standard linear regression theory. Assuming the columns of $\boldsymbol{\mathcal{X}}$ are linearly independent we have

$$\boldsymbol{\beta}(z) := (\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}^T\boldsymbol{y}(z\mathbf{1}).$$

While the covariance matrix $\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}$ is fully known, $\boldsymbol{y}(z\mathbf{1})$ is not, so $\boldsymbol{\beta}(z)$ is unknown. However $\boldsymbol{\beta}(z)$ can be consistently estimated under random sampling and random treatment assignment, as we will show. Given such an estimate $\widehat{\boldsymbol{\beta}}(z)$ of $\boldsymbol{\beta}(z)$, we can estimate each population PO by $\hat{y}_i(z) := \widehat{\boldsymbol{\beta}}(z) \cdot \boldsymbol{x}_i$ of $y_i(z)$. Then, letting

$$\hat{\boldsymbol{y}}(\boldsymbol{z}) := [\hat{y}_1(z_1), \ldots, \hat{y}_N(z_N)],$$

we estimate the optimal regime $\boldsymbol{z}^*$ by

$$\hat{\boldsymbol{z}}^* := \arg\max_{z \in \mathcal{P}} \mathbf{1} \cdot \hat{\boldsymbol{y}}(\boldsymbol{z}). \tag{4.1}$$

While $\hat{\boldsymbol{z}}^*$ is fully identified statistically, the optimization may be computationally challenging depending on the nature of $\mathcal{P}$. A full treatment of that optimization is outside the scope of this paper, but we provide the solution when $\mathcal{P}$ is entirely unrestrictive (i.e., when $\mathcal{P} = \mathcal{Z}^N$) and discuss more general solutions in Section 4.8. We next describe the study design and estimation procedure.

4.6 Data, estimation, and inference

4.6.1 Study design

Suppose plots are enrolled in the study by simple random sampling from the population. Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a collection of n random indices recording the enrolled plots. Then the POs for plots in the study are $\{y_i(z)\}_{i \in \mathcal{S}} = \{Y_i(z)\}_{i=1}^{n}$. We emphasize that $Y_i(z)$ involves a random re-indexing so that $Y_i(z) \neq y_i(z)$, but instead $Y_i(z) = y_{S_i}(z)$. The expected value of a potential outcome in the sample is $\mathbb{E}[Y_i(z)] = \bar{y}(z)$. A length-$p$ random vector of covariates $\boldsymbol{X}_i$ is also observed for each unit.

From here we will assume that treatment is categorical [8] with K levels. Generically,

[8] Naturally continuous treatments can be handled in this framework by discretizing them to an arbitrarily fine grid. The design and analysis of studies while maintaining the continuity of treatments requires additional assumptions beyond the scope of this paper.

let $\mathcal{Z} := \{0, \dots, K-1\}$. Treatment is assigned as in a *completely randomized experiment*, so that n_k plots are assigned to treatment k uniformly across plots with $\sum_{k=0}^{K-1} n_k = n$. Denoting treatment assignments $\boldsymbol{Z} := [Z_1, \dots, Z_n]$, we have

$$\mathbb{P}(\boldsymbol{Z} = [z_1, \dots, z_n]) = \left(\binom{n}{n_1} \binom{n - n_1}{n_2} \dots \binom{n - \sum_{k=1}^{K-2} n_k}{n_{K-1}} \right)^{-1}$$

That is, all ways of partitioning the n plots into the K treatments with fixed group sizes $[n_0, n_1, \dots, n_{K-1}]$ are equally likely.

4.6.2 Observed data

After the study, we observe the outcome $Y_i := Y_i(Z_i)$ for each experimental plot $i \in \{1, \dots, n\}$. Take $\boldsymbol{Y}_k := [Y_i]_{i:Z_i=k}$ and $\mathbf{X}_k := [\boldsymbol{X}_i]_{i:Z_i=k}$ to be the observed outcomes and covariates for treatment group k. Under simple random sampling of plots and completely random assignment of treatments, the data $(\boldsymbol{Y}_k, \mathbf{X}_k)$ are a simple random sample of size n_k from the population $(y_i(k), \boldsymbol{x}_i)_{i=1}^N$. The samples are also dependent across k because they are necessarily for mutually exclusive sets of plots. Both dependencies are trivial when $n << N$. Pooling across k and appending the vector of assigned treatments, the observed data are denoted $(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{Z})$. The data suffice to estimate treatment effects and the optimal policy.

4.6.3 Estimation and inference for the PATE

When $\mathcal{Z}$ is binary, we can estimate the PATE τ using the observed outcomes and (potentially) the covariates. Denote a generic estimator of τ by $\hat{\tau}$. We describe three possibilities for $\hat{\tau}$: the difference-in-means, difference-in-differences, and an OLS-adjusted estimator. Each has a consistent or conservative variance estimator $\widehat{\mathbb{V}}[\hat{\tau}]$, which allows for asymptotically valid Wald-style equal-tailed $(1-\alpha)$ confidence intervals:

$$\hat{\tau} \pm \Phi^{-1}(\alpha/2)\sqrt{\widehat{\mathbb{V}}[\hat{\tau}]},$$

where $\Phi^{-1}(\alpha/2)$ is the $\alpha/2$ quantile of the normal distribution, about 1.96 for $\alpha = 0.05$. The intervals may not achieve their nominal $(1-\alpha)$ coverage in small studies. We evaluate their true coverage in some simple settings in our simulations, and discuss the possibility of using nonparametric methods with guaranteed finite-sample validity in Section 4.8.

The unbiased *difference-in-means* (DiM) estimator is:

$$\hat{\tau}^{\text{DiM}} := \bar{Y}_1 - \bar{Y}_0,$$

where $\bar{Y}_k = \frac{1}{n_k} \sum_{i:Z_i=k} Y_i$ is the sample mean of $\boldsymbol{Y}_k$. If σ_0^2 is the variance of the control population POs $\boldsymbol{y}(\boldsymbol{0})$ and σ_1^2 is the variance for the treatment population POs $\boldsymbol{y}(\boldsymbol{1})$, then the variance of $\hat{\tau}^{\text{DiM}}$ is $\mathbb{V}(\hat{\tau}^{\text{DiM}}) = \sigma_0^2/n_0 + \sigma_1^2/n_1$. It can be estimated without bias by plugging in the sample variances:

$$\widehat{\mathbb{V}}(\hat{\tau}^{\text{DiM}}) = \hat{\sigma}_0^2/n_0 + \hat{\sigma}_1^2/n_1,$$

where $\hat{\sigma}_z^2 := (n_z - 1)^{-1} \sum_{i:Z_i=z} (Y_i - \bar{Y}_z)^2$.

Now suppose one of the measured covariates is baseline SOC B_i, and let $D_i = Y_i - B_i$ be the difference between observed followup and baseline SOC. Also let $\bar{B}_z$ and $\bar{D}_z$ be the sample means of B_i and D_i, respectively, in treatment group $z \in \{0, 1\}$. The *difference-in-differences* (DiD) estimator is:

$$\hat{\tau}^{\text{DiD}} := \bar{D}_1 - \bar{D}_0 = (\bar{Y}_1 - \bar{B}_1) - (\bar{Y}_0 - \bar{B}_0).$$

Like the DiM, the DiD is unbiased for τ. It has variance $\mathbb{V}(\hat{\tau}^{\text{DiD}}) = \sigma_{D0}^2/n_0 + \sigma_{D1}^2/n_1$, where σ_{Dz}^2 is the variance of the population of differences $\{d_i\}_{i=1}^N$ where $d_i := y_i(z) - b_i$. When b_i tends to be close to $y_i(z)$, the variance of the differences σ_{Dz}^2 is less than the variance of the raw POs σ_z^2 and the DiD is more efficient than the DiM: $\mathbb{V}(\hat{\tau}^{\text{DiD}}) < \mathbb{V}(\hat{\tau}^{\text{DiM}})$. Letting, $\hat{\sigma}_{Dz}^2 = n_z^{-1} \sum_{i:Z_i=z} (D_i - \bar{D}_z)^2$ be the sample variance of the differences in treatment group z, the variance estimate

$$\widehat{\mathbb{V}}(\hat{\tau}^{\text{DiD}}) = \hat{\sigma}_{D0}^2/n_0 + \hat{\sigma}_{D1}^2/n_1$$

is unbiased.

The final estimator of τ we consider is the OLS-interaction estimator of Lin [2013]. The OLS-interaction estimator $\hat{\tau}^{\text{OLS}}$ is the coefficient on Z_i in the OLS regression of Y_i on Z_i, $\boldsymbol{X}_i$, and the interaction $Z_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$, where $\bar{\boldsymbol{X}}$ are the column means of $\mathbf{X}$. Lin [2013] shows that $\hat{\tau}^{\text{OLS}}$ is consistent and asympotically normal, has lower asymptotic variance than $\hat{\tau}^{\text{DiM}}$, and may be substantially more precise in finite-samples. However, the estimate has a small finite-sample bias of order $1/n$. The "sandwich" variance of the OLS coefficient estimates provides a consistent estimate for $\mathbb{V}[\hat{\tau}^{\text{OLS}}]$. Lin [2013] demonstrates that Wald-style intervals with this variance estimate achieve their nominal coverage in simulations of a relatively small experiment ($n_1 = 58$, $n_0 = 99$). The sandwich covariance estimator is not necessary in balanced binary experiments, where $n_1 = n_0$; the usual covariance estimate suffices. We primarily consider OLS-interaction strategy when $\boldsymbol{X}_i = [1, B_i]$ includes the single covariate B_i, representing baseline SOC (or another one-dimensional proxy of sequestration potential). In this case, the estimator can be written

$$\hat{\tau}^{\text{OLS}} := (\bar{Y}_1 - (\widehat{\beta}_B + \widehat{\beta}_{\text{mod}})\bar{B}_1) - (\bar{Y}_0 - \widehat{\beta}_B \bar{B}_0),$$

where $\widehat{\beta}_B$ estimates the association of B_i with Y_i among the control group (the coefficient on B_i in the interacted OLS), and $(\widehat{\beta}_B + \widehat{\beta}_{\text{mod}})$ estimates the association of B_i with Y_i in the treatment group ($\widehat{\beta}_{\text{mod}}$ is the coefficient on $Z_i(B_i - \bar{B})$). The resemblance to $\hat{\tau}^{\text{DiD}}$ is apparent: $\hat{\tau}^{\text{DiD}}$ fixes these coefficients to 1, analogous to an unbiased fixed-slope estimator in survey sampling [Cochran, 1977]. Thus, compared to $\hat{\tau}^{\text{DiD}}$, the OLS estimator $\hat{\tau}^{\text{OLS}}$ sacrifices a small finite-sample bias for a potential gain in asymptotic precision [Lin, 2013].

4.6.4 Estimation of the optimal policy

Now suppose that, in addition to the observed study data $(\boldsymbol{Y}, \mathbf{X}, \boldsymbol{Z})$, we have the covariate matrix $\boldsymbol{\mathcal{X}}$ for the entire population of interest. We can estimate the coefficients $\boldsymbol{\beta}(k)$ using the study data, and impute POs for the entire population using $\boldsymbol{\mathcal{X}}$ and $\widehat{\boldsymbol{\beta}}(k)$. That is, we propose to estimate each $y_i(k)$ by $\hat{y}_i(k) := \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(k)$ for $i \in \{1, \ldots, N\}$, and plug these fitted values in to (4.1).

When there are no budgetary constraints on the treatment portfolio ($C_0 = \infty$), the estimated optimal regime $\hat{\boldsymbol{z}}^*$ can be found by separately maximizing each $\hat{y}_i(k)$, i.e., by letting $\hat{k}_i^* := \arg\max_{k \in \mathcal{Z}} \hat{y}_i(k)$ and $\hat{\boldsymbol{z}}^* := [\hat{k}_1^*, \ldots, \hat{k}_N^*]$. When there are budgetary constraints and varying costs of treatment, the estimates are inputs to an optimization routine. Under the cost-additivity assumption, the optimal regime could be found by linear programming. We return to the plausibility of that assumption and the linear programming solution in our discussion. We will also suggest some uses for inference on $\boldsymbol{z}^*$ and some possible avenues to construct confidence sets in our discussion, but implementing them is outside the scope of this paper. Hence, in what follows we will evaluate only the quality of point estimates of the optimal policy and compare them to the business-as-usual scenario of using the DiM estimate of the PATE to choose the policy, which is equivalent to optimizing over the restricted treatment portfolio $\mathcal{R}$. The restriced optimal policy estimate is thus $\hat{\boldsymbol{z}}_{\mathcal{R}}^* := (\arg\max_{z \in \mathcal{Z}} \bar{Y}_z)\mathbf{1}$. It sets the treatment for every unit in the population equal to the treatment that gives the largest observed mean in the study.

4.7 Simulations

4.7.1 Simulated populations and design

The populations in our simulations were based on data from an RCT of compost application conducted at multiple sites around California, representing a range of soil types, climates, biologies, and land use histories. The data was collected as part of an initiative for the Natural Resources Conservation Service (NRCS) and described in Silver et al. [2018]. The NRCS study enrolled pairs of 30m × 62.5m plots at 14 different sites. Within each pair one plot was chosen at random to be treated with 0.64cm of compost, while the other plot was an untreated control. Plots were measured for %TC by drawing 5 soil cores along a diagonal transect to a depth of 10cm. Cores were assayed for %TC by dry combustion in an elemental analyzer. BD was not measured, so all our results are expressed in terms of %TC on a mass-to-mass basis. This entire measurement procedure was done pre-treatment in 2016 and repeated every subsequent year after the application of treatment until 2019. We took followup SOC to be the 2019 measurements and baseline SOC to be the 2016 measurements.

Using the NRCS data to inform population parameters (see Table 4.1), we simulated populations with plots as units recording average baseline SOC (b_i), 3-year followup SOC under treatment ($y_i(1)$), and 3-year followup SOC under control ($y_i(0)$). The latter two quantities are potential outcomes, while b_i is the covariate proxying sequestration potential. There were $N = 2480$ plots in each population. For each plot, baseline SOC was generated by drawing (independently of other plots) from a normal distribution with mean equal to the empirical mean %TC in 2016 in the NRCS Data (μ_b; pooled across treatment and control plots) and SD equal to the empirical across-plot SD in %TC in 2016 (σ_b^{ap}). Follow-up SOC on control was equal to baseline SOC plus random noise:

$$y_i(0) = b_i + \varepsilon_{0i},$$

where ε_{0i} was generated by drawing from a normal distribution with mean equal to the empirical average of the change between 2016 and 2019 for control plots in the NRCS data (Δ_0) and SD equal to the empirical across-plot SD of the change between 2016 and 2019 for control plots ($\sigma_{\Delta_0}^{\text{ap}}$). Thus, control potential outcomes randomly differed from baseline according to the variation observed in the NRCS data. Treatment potential outcomes were drawn from the linear model:

$$y_i(1) = y_i(0) + \tau + \beta_{\text{mod}} \times \tilde{b}_i + \varepsilon_{1i},$$

where τ is the ATE, reflecting how effective treatment is on average; ε_{1i} is random noise drawn from a normal distribution with mean 0 and variance $\sigma_{\varepsilon_1}^2$, parameterizing idiosyncratic variation in the individual treatment effect; $\tilde{b}_i := (b_i - \bar{b})/(N^{-1}\sum_{i=1}^{N}(b_i - \bar{b})^2)$, i.e. baseline SOC standardized to have mean 0 and variance 1; β_{mod} is the moderator effect, parameterizing the association between standardized baseline SOC and the individual treatment effect for plot i. A negative value of β_{mod} implies higher baseline SOC is associated with a lower treatment effect.

We simulated populations for a range of parameter settings $\{\tau, \beta_{\text{mod}}, \sigma^2_{\varepsilon_1}\}$ from null to large values of each parameter. Specifically, we let $\tau \in \{0, 0.05, 0.1, 0.3\}/\sigma^2_{\text{ap}}$ where $\sigma^2_{\text{ap}} = 0.66$ is the empirical baseline across-plot standard deviation in the NRCS data; $\beta_{\text{mod}} \in \{0, -0.1, -0.5\}$; and $\sigma^2_{\varepsilon_1} = \{0, 0.1\}$. For each combination of parameter values, a single population was generated by the process described above.

Then for each of 200 simulation replicates, we drew study data by enrolling $n = \{10, 100, 1000\}$ plots into a balanced RCT, randomizing half to treatment and half to control according to a completely randomized design. The observed data (Y_i, B_i) also involved independent measurement error from sampling $\delta_{ij} \sim \mathcal{N}(0, \sigma^2_\delta)$ for $i \in \{1, ..., n\}$, $j \in \{1, 2\}$ with δ_{i1} representing the measurement error added to baseline:

$$B_i = b_{S_i} + \delta_{i1},$$

and δ_{i2} representing the measurement error added to the follow-up observation:

$$Y_i = Y_i(Z_i) + \delta_{i2}.$$

The degree of measurement error was set at $\sigma_\delta \in \left\{\sigma_b^{\text{wp}}/\sqrt{5}, \sigma_b^{\text{wp}}/\sqrt{30}, \sigma_b^{\text{wp}}/\sqrt{100}\right\}$ where $\sigma_b^{\text{wp}} = 1.02$ is the empirical average within plot variance (spatial heterogeneity) in the 2016 NRCS data. The measurement noise was thus computed assuming 5, 30, or 100 samples were taken within each plot. Additional variability due to assay was assumed negligible and not included.

4.7.2 Parameters, estimators, and evaluation criteria

We were interested in estimators of the PATE τ, the moderator effect β_{mod}, and the optimal policy z^*. For the ATE, we evaluated estimates and corresponding Wald confidence intervals of the difference-in-means (DiM), difference-in-differences (DiD), and ordinary least squares adjusted (OLS) estimators. For the moderator effect, we evaluated the estimator $\hat{\beta}_{\text{mod}}$ taken from the estimated coefficient on the interaction $Z_i(B_i - \bar{B})$ of the OLS estimator. For the optimal policy, we assumed there was no budgetary constraint and compared the optimal PTPO $\bar{y}^* = \bar{y}(\boldsymbol{z}^*)$ to the PTPO using the estimated optimal regime $\bar{y}(\hat{\boldsymbol{z}}^*)$ and to the PTPO using the estimated restricted optimal regime $\bar{y}(\hat{z}^*_r \mathbf{1})$ where $\hat{z}^*_r = \text{argmax}_z \bar{Y}_z$ is just the larger of the two observed means.

For each estimator we recorded the estimate and 95% Wald confidence interval for each simulation. Using the replicate simulations, we estimated the bias, root mean squared error (RMSE), confidence interval (CI) coverage, and expected CI width. The bias is the difference between the expected value of the estimator and the parameter. Some of our estimators (DiM, DiD) are known to be exactly unbiased, while some (OLS) can have a small finite-sample bias. The RMSE captures the expected deviation of an estimate from the parameter, taking into account both the bias and variability of an estimator. The CI coverage reflects the validity of inference using that procedure: a valid 95% CI should cover the parameter in around 95% of the simulations or more. Finally, the expected CI width measures the "perceived" error of the estimator. Smaller CI width is preferable (the method has higher power), conditional on the coverage being near 95% (the method is valid).

4.7.3 Simulation results

Summaries of the parameters estimated from the NRCS data and used in the simulations appear in Table 4.1.

Parameter	**Value (TC%)**
Baseline average (μ_b)	2.34
Baseline within-plot SD (σ_b^{wp})	1.02
Baseline across-plot SD (σ_b^{ap})	0.47
Average control change (Δ_0)	0.16
Across-plot SD control change ($\sigma_{\Delta_0}^{\text{ap}}$)	0.14

Table 4.1: Parameters estimated from NRCS data and used to simulate populations. TC% = percent total carbon.

The performance of the PATE estimators is tabulated at a few trial sizes n in Table 4.2. The results are averaged across the population settings described in Section 4.7.1. None of the estimators had any finite-sample bias to two decimal places. However, in small trials ($n \leq 10$) the 95% Wald CIs were not strictly valid for any method, since they only cover the PATE in around 90% of simulations. All CIs achieved their nominal coverage by $n = 100$. The OLS estimator $\hat{\tau}^{\text{OLS}}$ had the best performance in terms of RMSE and CI width. At $n = 100$, $\hat{\tau}^{\text{OLS}}$ could resolve the PATE to 0.4 TC% on average, while $\hat{\tau}^{\text{DiM}}$ could resolve it to 0.5 TC%. The performance of $\hat{\tau}^{\text{DiD}}$ was in between. In Figure 4.3 we plot the CI width as a function of the trial size, the number of samples drawn per plot, and the moderator effect for each estimator. The OLS estimator $\hat{\tau}^{\text{OLS}}$ always had the shortest CIs on average, while the CI width of $\hat{\tau}^{\text{DiD}}$ and $\hat{\tau}^{\text{DiM}}$ depended on the moderator effect and the samples per plot. The DiD estimator $\hat{\tau}^{\text{DiD}}$ was better when the moderator was positively associated with the treatment effect ($\beta_{\text{mod}} = 0.5$) and when there were more samples per plot. This is because (a) $\hat{\tau}^{\text{DiD}}$ needed to estimate both baseline and follow up averages, and incurred more error in those estimates when there are fewer samples per plot; and (b) when the moderator effect was negative, baseline SOC was correlated positively with control POs but negatively with treatment POs, so the treatment group differences $\boldsymbol{Y}_1 - \boldsymbol{B}_1$ are more variable than the raw values $\boldsymbol{Y}_1$, increasing the standard error of $\hat{\tau}^{\text{DiD}}$.

In terms of the optimal policy, the average oracle value $\bar{y}(\boldsymbol{z}^*)$ across the simulated populations was 2.63 TC%. On average the return under the estimated optimal policy $\bar{y}(\hat{\boldsymbol{z}}^*)$ was 2.61 TC%, with a loss of 0.02 TC% compared to the oracle value. The return under the restricted optimal policy estimate $\bar{y}(\hat{\boldsymbol{z}}^*_{\mathcal{R}})$ averaged to 2.50 TC%, with a loss of 0.13 TC%. As expected, the differences were driven primarily by the moderator effect. We observed the relation $\bar{y}(\hat{\boldsymbol{z}}^*) \geq \bar{y}(\hat{\boldsymbol{z}}^*_{\mathcal{R}})$, with the inequality holding strictly when and only when $\beta_{\text{mod}} \neq 0$.

n	Estimator	Bias	CI Width	CI Coverage	RMSE
10	Difference-in-differences	0.00	1.42	0.91	0.37
	Difference-in-means	0.00	1.49	0.90	0.39
	OLS-adjusted	0.00	1.36	0.89	0.37
100	Difference-in-differences	0.00	0.46	0.95	0.12
	Difference-in-means	0.00	0.49	0.94	0.13
	OLS-adjusted	0.00	0.41	0.94	0.11
1000	Difference-in-differences	0.00	0.15	0.95	0.04
	Difference-in-means	0.00	0.16	0.96	0.04
	OLS-adjusted	0.00	0.13	0.95	0.03

Table 4.2: Simulation results for various estimators (described in Section 4.6.3) of the PATE τ. Results are averaged over the range of populations described in Section 4.7.1 and over 200 studies with random sampling and treatment assignment simulated on each population. The width and true coverage of nominal 95% confidence intervals is shown. CI = confidence interval; RMSE = root mean-squared error; OLS = ordinary least squares.

4.8 Discussion

We provided an integrated review of uncertainties in studies targeting SOC sequestration along with a design-based causal model for estimating treatment effects and optimizing sequestration across a population of interest.

We found that the regression adjusted estimator of the population average treatment effect tended to have marginally lower error and confidence interval width than

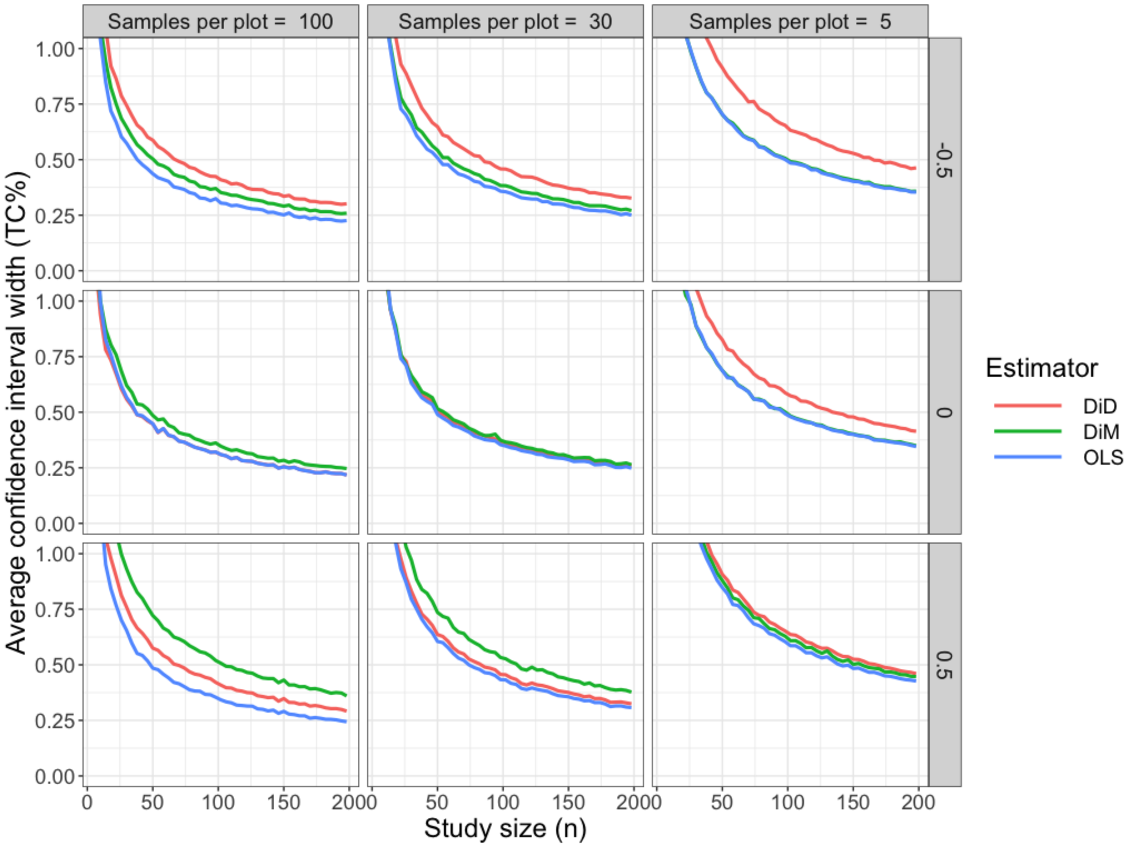


Figure 4.3: Confidence interval widths (y-axis) at a range of trial sizes n (x-axis) for various estimators (line colors). The columns demarcate different numbers of samples per plot (5, 30, or 100); more within-plot sampling reduces the uncertainty due to within-plot spatial heterogeneity. The rows denote different moderator effects β_{mod}. DiD = difference-in-differences; DiM = difference-in-means; OLS = ordinary least squares adjusted; TC% = percent total carbon.

the difference-in-means or difference-in-differences on average across a wide range of simulation settings. Our proposed estimate of the optimal policy also improved on using the standard, restricted estimate based on uniformly applying treatment to the group with the larger estimated effect, especially when there was treatment effect heterogeneity described by an observed moderator (i.e., sequestration potential). In the rest of this section, we will discuss additional considerations for sequestration studies and policies, limitations of our work here, and directions for future research.

4.8.1 Other study designs and potential pitfalls

We assumed a very particular study design: a randomized controlled trial (RCT) with units sampled uniformly at random from a broader population of interest. This setup represents an ideal wherein the internal and external validity of all estimates and inferences are rigorously justified by the design alone. Rarely, if ever, is such a design feasible in real-world soil science experiments. Some of the assumptions can be relaxed.

The assumption of simple random sampling into a completely randomized experiment is not necessary: there exist estimators that retain unbiasedness and inferential validity under a much broader range of designs, including non-uniform sampling and assignment, stratification or blocking, rerandomization, and cluster sampling or treatment assignment [Aronow and Middleton, 2013, Imbens and Rubin, 2015, Egami and Hartman, 2023, Li and Ding, 2020]. Furthermore, the plots in the experiment may constitute a convenience sample from the larger population or may not be embedded in a population at all. In that case all inferences may be confined to the experiment itself. There is a rich literature on such *randomization inference* dating back 100 years ago to early agricultural statistics [Neyman, 1923], which continues to flourish today [Lin, 2013, Ding et al., 2019]. The internal validity of such experiments is guaranteed by the design alone, but the experiment has no immediate external validity: nothing can be said about a larger population without further assumptions. Section 4.4.4 mentioned a few ways to support external validity. We emphasize the importance of careful consideration of the population to which results are being extrapolated, and of replicating studies. Replication is the strongest way to establish that causal effects generalize across contexts.

Purely observational studies are more problematic, requiring hypothetical populations and sampling designs premised on unverifiable assumptions. The design-based view of observational causal inference aims to explicitly approximate a hypothetical RCT and to use simple methods for statistical analysis, which are relatively interpretable and transparent so that weaknesses in the study can be easily identified

[Rosenbaum, 2002]. In contrast, the model-based and Bayesian views tend towards intricate distributional and functional assumptions that are often obscure, strained, and highly sensitive to misspecifications [Berk and Freedman, 2003].

All the designs discussed above assume a longitudinal study. An RCT is inherently longitudinal since the response must be observed some time after the application of treatment. However, some observational studies in soil science are cross-sectional or “space-for-time”, involving measurement at only one time point Walker et al. [2010]. Such studies are vulnerable to vagueries and delicate assumptions. For example, when the exact land-use history is unknown, it may be difficult even to precisely define the intervention under study or to ensure that plots within a group received the same intervention. Longitudinal observational studies and RCTs are subject to similar challenges. For example, the longterm Highfield ley-arable experiment at Rothamsted regularly collected data for nearly 75 years, but the treatment of assigned plots and even the study population evolved over that time [Blyth et al., 2023]. Nevertheless, as a rule of thumb, space-for-time studies are likely to be less rigorous than longitudinal observational studies, which are likely to be less rigorous than RCTs.

Finally, some studies may not rely on empirical data at all or use empirical data as simple inputs to a mechanistic model. The CENTURY and DAYCENT biogeochemical models [Parton, 1996] are commonly used as part of theoretical studies of SOC sequestration and may be used to inform policy decisions as well, especially to extrapolate measurements over space and time [Silver et al., 2018] or as part of sequestration crediting protocols [Mathers et al., 2023]. The reliability of the output of such models needs to be distinguished from that of empirical results. In particular model results are a lesser form of evidence because their validity is premised on accurately capturing all relevant aspects of the complex, multi-causal, physical mechanisms governing SOC sequestration. The models greatly simplify these mechanisms and must, given that the physics of SOC sequestration (including saturation) remain poorly constrained and highly variable under real-world conditions. The models must be calibrated empirically and tested for their ability to predict SOC changes and causal effects in a wide-range of contexts [Necpálová et al., 2015]. Properly conducted randomized trials can rigorously estimate causal effects without the need to understand or control all the physics, which is currently impossible.

4.8.2 Saturation in a causal context

The method we developed can be used to furnish empirical evidence for the saturation hypothesis. A large body of past work has failed to provide such evidence, in large part because saturation has not been couched in a causal model, which has led to confusion when analyzing and interpreting the data [Slessarev et al., 2023]. In short, investigators have used baseline SOC B_i and followup SOC Y_i, computed the difference $D_i := Y_i - B_i$, regressed D_i on B_i, and interpreted a significantly negative regression coefficient as evidence that baseline SOC is negatively associated with change. This is not a meaningful result: the difference between any two independent random variables will be correlated with either one of the original random variables. Slessarev et al. [2023] note this as an instance of regression to the mean, and propose a correction.

Our approach instead suggests formalizing the saturation hypothesis in a causal model, with the phenomenon manifesting either as moderation by baseline SOC or as diminishing returns to inputs. In the binary treatment case, baseline moderation is equivalent to a nonzero value of β_{mod} from Section 4.6.3. More treatment levels and multivariate or non-linear moderating effects can be estimated using the general framework in Section 4.5.6. On the other hand, an empirical study of saturation in terms of diminishing returns might estimate the effects of a continuous treatment at varying levels of intensity, as in a dose-response design or analysis [Holland-Letz and Kopp-Schneider, 2015, Efron and Feldman, 1991]. Combining ideas of moderation and dose-response in the design and analysis of studies targeting sequestration is an interesting area of future research, which could help to shed light on the saturation hypothesis and to improve the efficacy of SOC sequestration policies.

4.8.3 Cost model and optimization

In characterizing the available treatment portfolio $\mathcal{P}$, we assumed that the overall cost of a policy was a linear and additive function of the costs of each individual treatment. This assumption creates a linear constraint set, which allows the optimal policy to be computed readily: the optimization is a linear program, for which fast algorithms exist at any scale [Karmarkar, 1984].

In many cases, a linear cost model is not appropriate. For example, policies are likely to exhibit economies of scale, so that marginal costs diminish as more plots receive a given treatment. Roughly speaking, diminishing marginal costs would suggest a more parsimonious and balanced treatment regime, where only a few intervention levels are prescribed in relatively equal proportions. In the extreme, it suggests treating all plots the same, i.e., choosing a treatment from $\mathcal{R}$. Furthermore, the real costs of interventions will often depend on geography, especially for treatments involving shipping costs: it is cheaper to spread x amount of an amendment on the same plot than to distribute $x/2$ to two plots (depending on the distance between them).

These considerations could make the optimization much more complex and potentially intractable, even if the population potential outcomes $\{\boldsymbol{y}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{Z}^N}$ were fully known. Brute-force solutions are generally impossible, since they involve enumerating $|\mathcal{Z}|^N$ averages $\bar{y}(\boldsymbol{z})$. If geography is a primary concern, the population could be partitioned into G clusters with all plots in a cluster constrained to receive the same treatment, reducing the burden of enumeration to $|\mathcal{Z}|^G$. Defining an accurate cost model and finding tractable routines for optimization is an important area for further research.

We also note that the true costs and benefits of different courses of action are not strictly internal to the project. Externalities are generally very difficult if not impossible to account for. Policy costs are not incurred in isolation, but in relation to other existing and potential policies. For example, converting a corn field to native grasses requires resources, but those resources may be diverted from an existing policy (e.g., a corn subsidy). An existing practice may impose a wide range of costs on nature or society—environmental, social, medical, etc—that are difficult to capture in a cost model and compare to a counterfactual action. Many costs are not even quantifiable for an individual, let alone universally.

4.8.4 Valid inference with non-normal data, complex target parameters, and sequential designs

The inference methods we proposed in this paper are asymptotically valid: the level of a confidence interval or hypothesis is approximately correct in large samples. How large a given sample must be for the approximation to be close is generally unclear. In our simulations of small studies ($n = 10$), Wald-style confidence intervals had true coverage probabilities below their nominal level (90% true vs 95% nominal). Those simulations drew the populations from normal distributions and we expect the coverage would be worse in populations that depart from normality, especially those with strongly varying skew [Stanley et al., 2023]. For skew to vary strongly across potential outcome populations, there must be a high heterogeneity of the treatment effect. If the treatment effect is constant, the PATE can be bounded with a guaranteed level using a randomization test [Ding et al., 2016] and furthermore the restricted optimal policy is the optimal policy (since all units respond the same to a given treatment). That assumption is generally implausible in the context of SOC sequestration, and is incongruous with all versions of the saturation hypothesis. Conservative confidence intervals for the PATE can be constructed by adapting the nonparametric method of Stanley et al. [2023] to the context of an RCT, though that may entail a subsantial loss in power compared to Wald-style intervals. Another option is to target the quantiles of the treatment effects $\{\tau_i\}_{i=1}^N$ and bound them using the permutation-based method of Caughey et al. [2023]. That method may provide additional insight into treatment effect heterogeneity, but changes the subject somewhat from this paper since it does not immediately bound average or total sequestration.

Another open problem is inference about the optimal policy $\boldsymbol{z}^*$. Earlier we suggested a method that bases the policy decision only on the estimate $\hat{\boldsymbol{z}}^*$. However, if the average $\bar{y}(\boldsymbol{z})$ is relatively flat around $\boldsymbol{z}^*$, there are many approximately optimal solutions, some of which may have considerably lower cost than $\hat{\boldsymbol{z}}^*$. Incorporating the uncertainty of $\hat{\boldsymbol{z}}^*$ is thus a valuable goal. To that end, a $(1-\alpha)$ confidence set $\mathcal{C}_{\boldsymbol{z}^*}$ might contain all policies $\boldsymbol{z}$ for which the hypothesis

$$H_0(\boldsymbol{z}) : \max_{\boldsymbol{z}' \in \mathcal{P}} \bar{y}(\boldsymbol{z}) \geq \bar{y}(\boldsymbol{z}') - \epsilon$$

does not reject at level α. In other words, $\mathcal{C}_{\boldsymbol{z}^*}$ is a set of policies that are potentially nearly optimal. The decision maker might then choose the least costly $\boldsymbol{z}$ such that $\boldsymbol{z} \in \mathcal{C}_{\boldsymbol{z}^*}$. The setup is similar to the null hypothesis in a bioequivalence problem [Westlake, 1979], inference on optimal average treatment effects [Kasy and Sautmann, 2021], and problems in multiple testing, especially procedures for multiple comparisons with the sample best [Hsu, 1996]. Constructing $\mathcal{C}_{\boldsymbol{z}^*}$ may be challenging.

Finally, SOC studies could be fruitfully embedded in a sequential design. While myriad details would need to be worked out, in essence the enrollment of plots and their assignment to treatment could take place on a rolling basis, with covariate and outcome data collected regularly over time. Treatments for an individual plot need not be fixed in advance (as a point treatment or pattern over time), but could change in response to measurements or other decisions. The sequential setup is more complex, but tracks more closely with how agriculture is usually practiced, may improve theoretical frameworks for SOC sequestration policy, and may lead to better practical studies if carefully implemented. Furthermore, some long running agricultural RCTs are best understood as sequentially evolving over time, including examples at Rothamsted Blyth et al. [2023]. The relevant literature includes adaptive clinical trials and policy experiments [Pallmann et al., 2018, Kasy and Sautmann, 2021], multi-armed bandits (especially contextual bandits) [Slivkins, 2024], reinforcement learning [Sutton and Barto, 1998], and dynamic treatment regimes [Chakraborty and Murphy, 2014]. In these settings, recent advances in sequential inference could be leveraged to ensure finite-sample and sequential validity without parametric assumptions [Howard et al., 2021, Ramdas et al., 2023, Waudby-Smith et al., 2024].

4.8.5 Uncertainties at the policy level

In our review of measurement in Section 2, we did not address uncertainties that arise specifically around policy design. These include additionality, permanence, and leakage. Additionality is often defined to mean that the policy creates SOC sequestration that would not have occurred otherwise, but this generally assumes the effectiveness of interventions [Indigo Agriculture, 2024]. Separating the implementation of an action from its effectiveness, we define additionality to mean that the policy stimulates an action that would not have occurred otherwise. For example, when a land manager is paid to stop tilling, they do, and they would not stop if they were not paid: the policy causes the management change to occur. In this respect, additionality is closely related to compliance in the experimental setting, and may be evaluated by considering existing regulatory, financial, physical, and social (dis)incentives to adopt an intervention as well as by recording information about compliance in a study. While the scientific effects of an intervention are best estimated if there is strong pressure to comply, additionality and policy effects are best estimated if the study measures the level of compliance when the policy is applied to a wider population. That is, while spatial generalization calls for plots to be representative of the population, determining additionality calls for the experimental intervention to be representative of the policy intervention. Additionality of a policy evokes the well-established tension between internal and external validity [Bates and Glennerster, 2017].

Permanence often refers to the longevity of additional sequestered SOC in a population Indigo Agriculture [2024], Smith [2005], Thamo and Pannell [2016], where a return to baseline SOC before 100 years is called a reversal. We differentiate between (a) the permanence of an intervention and (b) the permanence of sequestered SOC. Concept (a) is an additional population-level policy concern and uncertainty, which, like additionality, can be evaluated by monitoring cross-over in a longitudinal study with an externally-valid intervention. Concept (b) is a more complex question that needs to be addressed in the context of the specific intervention and its permanence (i.e. (a)), in addition to scientific knowledge about the SOC trajectory after the intervention is implemented. Such knowledge needs to be generated at the study-level and generalized across time and space, as above, before being tied to (a).

Finally, leakage refers to the externalities of a particular intervention in terms of its greenhouse effect. An intervention that stores SOC but releases a large amount of methane exhibits leakage (e.g., in flooding fields for conversion to rice cultivation [Minami, 1994, Nitta, 2022], as does an intervention that burns fossil fuel as part of a supply chain (e.g., in transporting compost long-distances [Silver et al., 2018]). Sequestration is storage minus leakage, and need not be positive even when storage is positive. Hence, it is critical to account for leakage through life-cycle assessment, tracking on-farm fossil fuel use, and measuring externalities like methane, nitrous oxide, and nitric oxide production [Minami, 1994, Pilegaard, 2013, Ryals and Silver, 2013]. Again, comparison to appropriate counterfactual scenarios is crucial and can be formalized using potential outcomes.

Chapter 5

Optimal sequential risk-limiting comparison audits

5.1 Introduction

Machines count votes in most American elections, and (reported) election winners are declared on the basis of these machine tallies. Voting machines are vulnerable to bugs and deliberate malfeasance, which may undermine public trust in the accuracy of reported election results. To counter this threat, risk-limiting audits (RLAs) can provide routine, statistically rigorous evidence that reported election outcomes are correct—that reported winners really won—by manually checking a demonstrably secure ballot trail [Lindeman and Stark, 2012]. RLAs have a user-specified maximum chance—the *risk limit*—of certifying a wrong reported outcome, and will never overturn a correct reported outcome. They can also be significantly more *efficient* than full hand counts, requiring fewer manually tabulations to verify a correct reported outcome and reducing costs to jurisdictions.

There are various ways to design RLAs. Data can be sampled as batches of ballots (i.e. precincts or machines) or as individual ballot cards (hereafter, we refer to cards simply as "ballots"). Sampling individual ballots is more statistically efficient than sampling batches. In a *polling audit*, sampled ballots are checked directly without reference to machine interpretations. Ballot-polling audits sample and check individual ballots. In a *comparison audit*, manual interpretations of ballots are compared to their machine interpretations. Ballot-level comparison audits check each sampled ballot against a corresponding *cast vote record* (CVR)—a digital receipt detailing how the machine tallied the ballot. Not all voting machines can produce CVRs, but ballot-level comparison audits are the most efficient type of RLA.

Ideally, RLAs will simultaneously check multiple (potentially all) contests within a jurisdiction—a task made considerably more efficient by targeting samples with card-style data [Glazer et al., 2021]. Card-style data are most feasibly derived from CVRs, in which case each contest can be audited using ballot-level comparison. Because the overall workload of the audit is aggregated across contests, optimizing the efficiency for individual contests can provide substantial workload reductions for the audit as a whole. Thus, constructing sharper ballot-level comparison audits is paramount to the implementation of real-world RLAs auditing multiple contests.

The earliest RLAs were formulated for batch-level comparison audits, which are analogous to historical, statutory audits [Stark, 2008a]. Subsequently, the maximum across contest relative overstatement (MACRO) was used for comparison RLAs [Stark, 2009b,d, 2010, Ottoboni et al., 2018], but its efficiency suffered from conservatively pooling observed errors across candidates and contest. SHANGRLA [Stark, 2020] unified RLAs as hypotheses about means of lists of bounded numbers and provided sharper methods for batch and ballot-level comparisons. Each null hypothesis tested in a SHANGRLA-style RLA posits that the mean of a bounded list of *assorters* is less than $1/2$. If all the nulls are declared false at risk limit α, the audit can stop. Any valid test for the mean of a bounded finite population can be used to test these hypotheses, allowing RLAs to use a wide range of *risk-measuring functions*.

Betting supermartingales (BSMs)—described in Waudby-Smith et al. [2021] and Stark [2023]—provide a particularly useful class of risk-measuring functions. BSMs are *sequentially valid*, allowing auditors to update and check the measured risk after each sampled ballot while maintaining the risk limit. They can be seen as generalizations of risk-measuring functions used in earlier RLAs, including Kaplan-Markov, Kaplan-Kolmogorov, and related methods [Stark, 2009b, 2020]. They have tuning parameters λ_i called *bets*, which play an important role in determining the efficiency of the RLA. Previous papers using BSMs for RLAs have focused on setting λ_i for efficient ballot-polling audits; betting for comparison audits has been treated as essentially analogous [Waudby-Smith et al., 2021, Stark, 2023]. However, as we will show, comparison audits are efficient with much larger bets than are optimal for ballot-polling.

This paper details how to set BSM bets λ_i for efficient ballot-level comparison audits, focusing on audits of plurality contests. Section 5.2 reviews SHANGRLA notation and the use of BSMs as risk-measuring functions. Section 5.3 derives optimal "oracle" bets under the Kelly criterion [Kelly Jr., 1956], which assumes knowledge of true error rates in the CVRs. In reality, these error rates are unknown, but the oracle bets are useful in constructing practical betting strategies, which plug in estimates of the true rates. Section 5.4 presents three such strategies: guessing the error rates *a priori*, using past data to estimate the rates adaptively, or positing a distribution of likely rates and diversifying bets over that distribution. Section 5.5 presents two simulation studies: one comparing the oracle strategy derived in Waudby-Smith et al. [2021] for ballot-polling against our comparison-optimal strategy, and one comparing practical strategies against one another. Section 5.6 sketches some extensions to betting while sampling without replacement and to social choice functions beyond plurality. Section 5.7 concludes with a brief discussion and recommendations for practice.

5.2 Notation

5.2.1 Population and parameters

Following SHANGRLA [Stark, 2020] notation, let $\{c_i\}_{i=1}^N$ denote the CVRs, $\{b_i\}_{i=1}^N$ denote the true ballots, and $A()$ be an *assorter* mapping CVRs or ballots into $[0, u]$. We will assume we are auditing a plurality contest, in which case $u := 1$, $A(b_i) := 1$ if the ballot shows a vote for the reported winner, $A(b_i) := 1/2$ if it shows an undervote or vote for a candidate not currently under audit, and $A(b_i) := 0$ if it shows a vote for the reported loser. The *overstatement* for ballot i is $\omega_i := A(c_i) - A(b_i)$. $\bar{A}^c := N^{-1} \sum_{i=1}^N A(c_i)$ is the average of the assorters computed on the CVRs. Finally, the comparison audit population is comprised of *overstatement assorters* $x_i := (1 - \omega_i)/(2 - v)$, where $v := 2\bar{A}^c - 1$ is the *diluted margin*: the difference in votes for the reported winner and reported loser, divided by the total number of ballots cast.

Let $\bar{x} := N^{-1} \sum_{i=1}^N x_i$ be the average of the comparison audit population and $\bar{A}^b := N^{-1} \sum_{i=1}^N A(b_i)$ be the average of the assorters applied to ballots. Section 3.2 of Stark [2020] establishes the relations

$$\text{reported outcome is correct} \iff \bar{A}^b > 1/2 \iff \bar{x} > 1/2.$$

As a result, rejecting the *complementary null*

$$H_0 : \bar{x} \leq 1/2 \tag{5.1}$$

at risk limit α provides strong evidence that the reported outcome is correct.

Throughout this paper, we ignore *understatement* errors—those in favor of the reported winner with $\omega_i < 0$. Understatements help the audit end sooner, but will generally have little effect on the optimal bets. We comment on this choice further in Section 5.7. With this simplification, overstatement assorters comprise a list of numbers $\{x_i\}_{i=1}^N \in \{0, a/2, a\}^N$ where $a := (2-v)^{-1} > 1/2$ corresponds to the value on correct CVRs, $a/2$ corresponds to 1-vote overstatements, and 0 corresponds to 2-vote overstatements. This population is parameterized by 3 fractions:

- $p_0 := \#\{x_i = a\}/N$ is the rate of correct CVRs.
- $p_1 := \#\{x_i = a/2\}/N$ is the rate of 1-vote overstatements.
- $p_2 := \#\{x_i = 0\}/N$ is the rate of 2-vote overstatements.

The population mean can be written $\bar{x} = ap_0 + (a/2)p_1$.

5.2.2 Audit data

Ballots may be drawn by sequential simple random sampling with or without replacement, but we first focus on the with replacement case for simplicity. Implications for sampling without replacement are discussed in Section 5.6. We have a sequence of samples $X_1, X_2, \ldots \overset{\text{iid}}{\sim} F$, where F is a three-point distribution with mass p_0 at a, p_1 at $a/2$, and p_2 at 0.

5.2.3 Risk measurement via betting supermartingales

Let $T_i := 1 + \lambda_i(X_i - 1/2)$ where $\lambda_i \in [0, 2]$ is a freely-chosen tuning parameter that may depend on past samples $X_1, \ldots, X_{i-1}$. Define $M_0 := 1$ and

$$M_t := \prod_{i=1}^{t} T_i = \prod_{i=1}^{t} [1 + \lambda_i(X_i - 1/2)].$$

M_t is a *betting supermartingale* (BSM) for any *bets* $\lambda_i \in [0, 2]$ whenever (5.1) holds because

$$\bar{x} \leq 1/2 \implies \mathbb{E}[X_i \mid X_{i-1}, ..., X_1] \leq 1/2 \implies \mathbb{E}[M_t \mid X_{t-1}, \ldots, X_1] \leq M_{t-1}$$

where the first implication comes from simple random sampling with replacement.

Intuitively, M_t can be thought of as the wealth accumulated by a gambler who starts with 1 unit of capital at time $t = 0$ and at time $t = i$ stakes proportion λ_i of their current capital on observing $X_i > 1/2$. If $\lambda_i = 0$, they stake nothing and can neither gain nor lose capital on round i. If $\lambda_i = 2$, they stake everything and can lose all their capital if $X_i = 0$. For any bets that depend only on past data, the gambler cannot expect to accumulate wealth by betting that $X_i > 1/2$ when (5.1) is true.

Ville's inequality [Ville, 1939] then guarantees that it is unlikely that the gambler's wealth ever becomes large:

$$\mathbb{P}(\exists\ t \in \mathbb{N} : M_t \geq 1/\alpha) \leq \alpha.$$

For example, when (5.1) holds, the probability that the gambler ever accumulates more than 20 units of wealth is no more than 0.05.

As a matter of risk measurement, Ville's inequality implies that the truncated reciprocal $P_t := \min\{1, 1/M_t\}$ is a sequentially-valid P-value for the complementary null in the sense that $\mathbb{P}(\exists\ t \in \mathbb{N} : P_t \leq \alpha) \leq \alpha$ when $\bar{x} \leq 1/2$ for any risk limit $\alpha \in (0, 1)$. More details on BSMs are given in Waudby-Smith and Ramdas [2023], Waudby-Smith et al. [2021] and Stark [2023]. To obtain an efficient RLA, we would like to make M_t as large as possible (P_t as small as possible) when $\bar{x} > 1/2$.

5.3 Oracle betting

We begin by deriving "oracle" bets by assuming we can access the true error rates p_0, p_1, and p_2 and optimizing the expected growth of the logarithm of the martingale under these rates. We call these oracle bets because they are exactly optimal for this objective, but depend on unknown parameters and hence cannot be implemented in practice. However, oracle bets can be approximated to run efficient comparison audits with the practical betting strategies discussed in Section 5.4.

5.3.1 Error-free CVRs

In the simple case where there is no error at all in the CVRs, $p_0 = 1$ and $x_i = \bar{x} = a$ for all i. When computing the BSM, it doesn't matter which ballot is drawn:

$$T_i = 1 + \lambda_i(a - 1/2) \text{ and } M_t = [1 + \lambda_i(a - 1/2)]^t.$$

Because $(a - 1/2) > 0$, the best strategy is to bet as aggressively as possible, setting $\lambda_i := 2$. Under such a bet, $M_t = (2a)^t$. Setting this equal to $1/\alpha$ yields the stopping time:

$$t_{\text{stop}} = \frac{\log(1/\alpha)}{\log(2a)} = \frac{-\log(\alpha)}{\log(2) - \log(2 - v)} \tag{5.2}$$

where v is the diluted margin. Ignoring understatement errors, (5.2) is a deterministic lower bound on the sample size of a comparison audit when risk is measured by a BSM. Figure 5.1 plots this as a function of the diluted margin and risk limit.

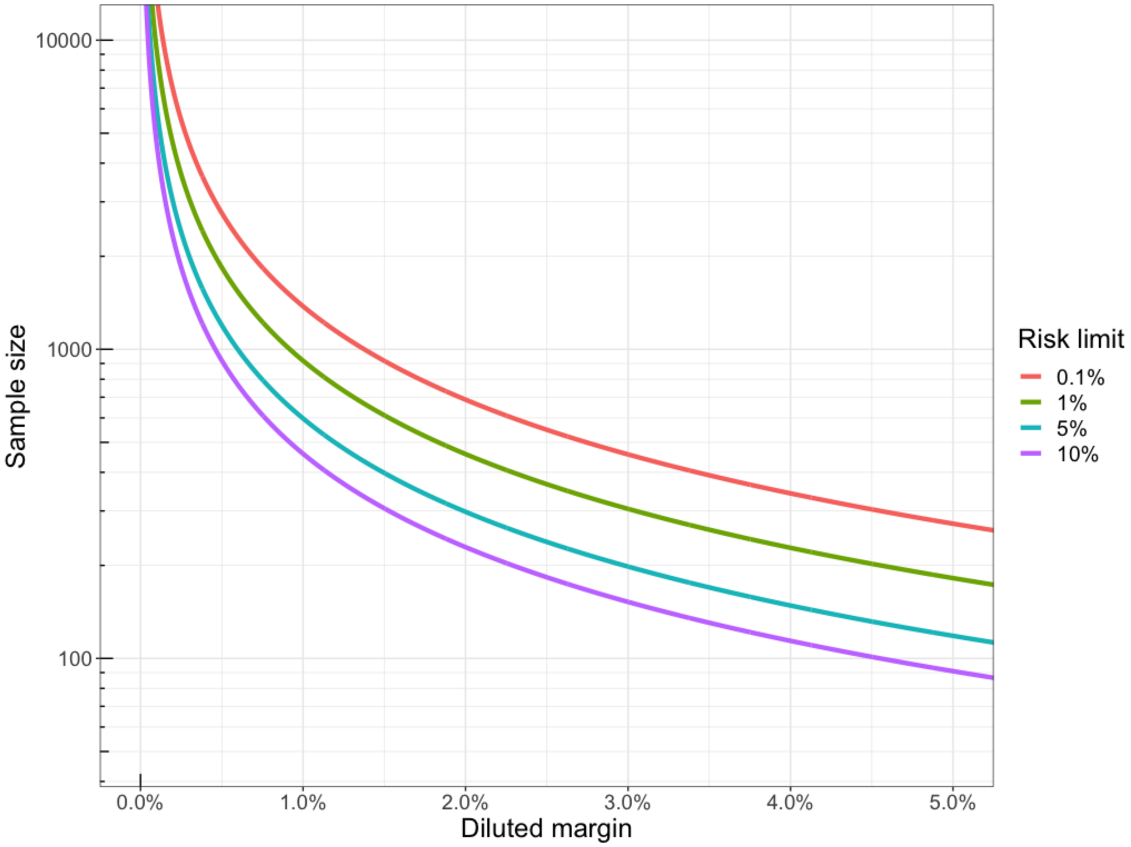


Figure 5.1: Deterministic sample sizes (y-axis; $\log_{10}$ scale) for a comparison audit of a plurality contest with various diluted margins (x-axis) and risk limits (colors), with no error in CVRs and a maximal bet of $\lambda = 2$ on every draw.

5.3.2 Betting with CVR Error

Usually CVRs will have at least some errors, and maximal bets are far from ideal when they do. We now show why this is true before deriving an alternative oracle strategy. In general,

$$T_i = \begin{cases} 1 + \lambda_i(a - 1/2) & \text{with probability } p_0 \\ 1 + \lambda_i(a/2 - 1/2) & \text{with probability } p_1 \\ 1 - \lambda_i/2 & \text{with probability } p_2. \end{cases}$$

If we fix $\lambda_i := \lambda$ and try to maximize M_n by maximizing the expected value of each T_i, we find $\mathbb{E}_F[T_i] = p_0[1 + \lambda(a - 1/2)] + p_1[1 - \lambda(1 - a)/2] + p_2[1 - \lambda/2] = 1 + (ap_0 +$

$\frac{a}{2}p_1 - 1/2)\lambda$. This is linear with a positive coefficient on λ, since $ap_0 + \frac{a}{2}p_1 = \bar{x} > 1/2$ under any alternative. Therefore, the best strategy seems to be to set $\lambda := 2$ as before. However, unless $p_2 = 0$, M_t will eventually "go broke" with probability 1: $T_i = 0$ if a 0 is drawn while the bet is maximal. Then $M_t = 0$ for all future times and we cannot reject at any risk limit α. In this case, we say the audit *stalls*: it must proceed to a full hand count to confirm the reported winner really won.

To avoid stalls we follow the approach of Kelly Jr. [1956], instead maximizing the expected value of $\log T_i$. The derivative is

$$\frac{d}{d\lambda}\mathbb{E}_F[\log T_i] = \frac{(a-1/2)p_0}{1+\lambda(a-1/2)} + \frac{(a-1)p_1}{2-\lambda(1-a)} + \frac{p_2}{2-\lambda}. \tag{5.3}$$

The oracle bet λ^* can be found by setting this equal to 0 and solving for λ using a root-finding algorithm.

Alternatively, we can find a simple analytical solution by assuming no 1-vote overstatements and setting $p_1 = 0$. In this case, solving for λ yields:

$$\lambda^* = \frac{2-4ap_0}{1-2a} \tag{5.4}$$

Note that $\lambda^* > 0$ since $ap_0 > 1/2$ under the alternative, and $\lambda^* < 2$ since $a > 1/2$.

5.3.3 Relation to ALPHA

There is a one-to-one correspondence between oracle bets for the BSM M_t and oracle bets for the ALPHA supermartingale, which reparameterizes M_t. Note that the list of overstatement assorters $\{x_i\}_{i=1}^N$ is upper bounded by the value of a 2-vote understatement, $u := 2/(2-v) = 2a$. Section 2.3 of Stark [2023] shows that the equivalently optimal η for use with ALPHA is:

$$\eta^* := 1/2(1+\lambda^*(u-1/2)) = \frac{1-2ap_0}{2-4a} + 2ap_0 - 1/2.$$

Naturally, when $p_0 = 1$, $\eta^* = 2a = u$, which is the maximum value allowed for η^* while maintaining ALPHA as a non-negative supermartingale.

5.4 Betting in Practice

In practice, we have to estimate the unknown overstatement rates to set bets. We propose and evaluate three strategies: fixed, adaptive, and diversified betting. Throughout this section, we use $\tilde{p}_k$ to denote a generic estimate of p_k for $k \in \{1, 2\}$. When the estimate adapts in time, we use the double subscript $\tilde{p}_{ki}$. In all cases, the estimated overstatement rates are ultimately plugged into (5.3) to estimate the optimal bets.

5.4.1 Fixed betting

The simplest approach is to make a fixed, *a priori* guess at p_k using historic data, machine specifications, or other information. For example, $\tilde{p}_1 := 0.1\%$ and $\tilde{p}_2 := 0.01\%$ will prevent stalls and may perform reasonably well when there are few overstatement error. This strategy is analagous to apKelly for ballot-polling, which fixes λ_i based on an *a priori* estimate of the population assorter mean (typically derived from reported tallies). However, Waudby-Smith et al. [2021] and Stark [2023] show that apKelly can become quite poor when the estimate is far from correct. This frailty motivates more sophisticated strategies.

5.4.2 Adaptive betting

In a BSM, the bets need not be fixed and λ_i can be a *predictable* function of the data $X_1, \ldots, X_{i-1}$, since we condition on these data when establishing M_t as a martingale. Intuitively, the gambler can adapt their bets based on outcomes of previous rounds and, if the null is true, still cannot expect to gain capital in the next round. This fact allows us to estimate error rates based on past samples in addition to *a priori* considerations when setting λ_i. We adapt the "truncated-shrinkage" estimator introduced in Section 2.5.2 of Stark [2023] to rate estimation. For $k \in \{1, 2\}$ we set a value $d_k \geq 0$, capturing the degree of shrinkage to the *a priori* estimate $\tilde{p}_k$, and a truncation factor $\epsilon_k \geq 0$, enforcing a lower bound on the estimated rate. Let $\hat{p}_{ki}$ be the sample rates at time i, e.g., $\hat{p}_{2i} = i^{-1} \sum_{j=1}^{i} 1\{X_j = 0\}$. Then the truncated-shrinkage estimate is:

$$\tilde{p}_{ki} := \frac{d_k \tilde{p}_k + i \hat{p}_{k(i-1)}}{d_k + i - 1} \vee \epsilon_k \tag{5.5}$$

The rates are allowed to learn from past data in the current audit through $\hat{p}_{k(i-1)}$, while being anchored to the *a priori* estimate $\tilde{p}_k$. The tuning parameter d_k reflects the degree of confidence in the *a priori* rate, with large d_k anchoring more strongly to $\tilde{p}_k$. Finally, ϵ_k should generally be set above 0 to prevent stalls.

At each time i, the truncated-shrinkage estimated rate $\tilde{p}_{ki}$ can be plugged into (5.3) and set equal to 0 to obtain the bet λ_i. Fixing $\tilde{p}_{1i} := 0$ allows us to use (5.4), in which case $\lambda_i = (2 - 4a(1 - \tilde{p}_{2i}))/(1 - 2a)$.

5.4.3 Diversified betting

A weighted average of BSMs:

$$\sum_{b=1}^{B} \theta_b \prod_{i=1}^{t} [1 + \lambda_b(X_i - 1/2)],$$

where $\theta_b \geq 0$ and $\sum_{b=1}^{B} \theta_b = 1$, is itself a BSM. The intuition is that our initial capital is split up into B pots, each with θ_b units of wealth. We then bet λ_b on each pot at each time, and take the sum of the winnings across all pots as our total wealth at time t. Waudby-Smith and Ramdas [2023] construct the "grid Kelly" martingale by defining λ_b along an equally spaced grid on $[0, 2]$ and giving each the weight $\theta_b = 1/B$. Waudby-Smith et al. [2021] refine this approach into "square Kelly" for ballot-polling RLAs by placing more weight at close margins.

We adapt these ideas to the comparison audit context by parameterizing a discrete grid of weights for p_1 and p_2. We first note that (p_1, p_2) are jointly constrained by the hyperplane $ap_2 + (a/2)p_1 \leq a - 1/2$ under the alternative, since otherwise there is enough error to overturn the reported result. A joint grid for (p_1, p_2) can be set up by separately constructing two equally-spaced grids from 0 to v/k, computing the Cartesian product of the grids, and removing points where $ap_2 + (a/2)p_1 \geq a - 1/2$. Once a suitable grid has been constructed, the weights at each point can be flexibly defined to reflect the suspected rates of overstatements. At each point (p_1, p_2), λ_b is computed by passing the rates (p_1, p_2) into (5.3) and solving numerically; the weight for λ_b is θ_b. Thus a distribution of weights on the grid of overstatement rates induces a distribution on the bets.

Figure 5.2 illustrates two possible weighted grids for a diluted margin of $v = 10\%$, and their induced distribution on bets $\{\lambda_b\}_{b=1}^{B}$. In the top row, the weights are uniform with $\theta_b = 1/B$. In the bottom row, the weights follow a bivariate normal density with mean vector and covariance matrix respectively specified to capture a prior guess at (p_1, p_2) along with the uncertainty in that guess. The density is truncated, discretized, and rescaled so that the weights sum to unity.

5.5 Numerical evaluations

We conducted two simulation studies. The first evaluated stopping times for bets using the oracle comparison bets in (5.4) against the oracle value of apKelly from Waudby-Smith et al. [2021]. The second compared stopping times for oracle bets and the 3 practical strategies we proposed in Section 5.4. All simulations were run in R (version 4.1.2).

5.5.1 Oracle simulations

We evaluated stopping times of oracle bets at multiple diluted margins and 2-vote overstatement rates when sampling with replacement from a population of size $N = 10000$. At each combination of diluted margin $v \in \{0.05, 0.10, 0.20\}$ and 2-vote overstatement rates $p_2 \in \{1.5\%, 1\%, 0.5\%, 0.1\%, 0\%\}$ we ran 400 simulated comparison audits. We set $p_1 = 0$: no 1-vote overstatements.

The bets corresponded to oracle bets λ^* in Equation (5.4) or to $\lambda^{\text{apK}} := 4\bar{x} - 2$, the "oracle" value of the apKelly strategy in Section 3.1 of Waudby-Smith et al. [2021] and Section 2.5 of Stark [2023][1], which were originally derived for ballot-polling. λ^{apK} uses the true population mean instead of an estimate based on reported tallies. In each scenario, we estimated the expected and 90th percentile workload from the empirical mean and 0.9 quantile of the stopping times at risk limit $\alpha = 5\%$ over the 400 simulations. To compare the betting strategies, we computed the ratios of the expected stopping time for λ^* over λ^{apK} in each scenario. We then took the geometric mean across scenarios as the average reduction in expected workload.

Table 5.1 presents the mean and 90th percentile (in parentheses) stopping times over the 400 simulations. BSM comparison audits with λ^* typically require counting fewer than 1000 ballots, and fewer than 100 for wide margins without CVR errors. On average, betting by λ^* provides an enormous advantage over λ^{apK}: the geometric mean workload ratio is 0.072, a 93% reduction.

5.5.2 Practical simulations

We evaluated oracle betting, fixed *a priori* betting, adaptive betting, and diversified betting in simulated comparison audits with $N = 20000$ ballots, a diluted margin of 5%, 1-vote overstatement rates $p_1 \in \{0.1\%, 1\%\}$, and 2-vote overstatement rates $p_2 \in \{0.01\%, 0.1\%, 1\%\}$.

Oracle bets were set using the true values of p_1 and p_2 in each scenario. The other methods used prior guesses $\tilde{p}_1 \in \{0.1\%, 1\%\}$ and $\tilde{p}_2 \in \{0.01\%, 0.1\%\}$ as tuning parameters in different ways. The fixed method derived the optimal bet by plugging in $\tilde{p}_k$ as a fixed value. The adaptive method anchored the truncated-shrinkage estimate $\tilde{p}_{ki}$ displayed in equation (5.5) to $\tilde{p}_k$, but updated using past data in the sample. The tuning parameters were $d_1 := 100$, $d_2 := 1000$, $\epsilon_1 = \epsilon_2 := 0.001\%$. The larger value for d_2 reflects the fact that very low rates (expected for 2-vote overstatements) are harder to estimate empirically, so the prior should play a larger role. The diversified method used $\tilde{p}_k$ to set the mode of a mixing distribution, as in the lower panels of

[1] λ^{apK} implies a bet of $\eta_i := \bar{x}$ in the ALPHA parameterization.

		Stopping times	
DM	**2-vote OR**	**apKelly** (λ^{apK})	**Oracle** (λ^*)
5%	1.5%	10000 (10000)	1283 (2398)
	1.0%	10000 (10000)	482 (813)
	0.5%	7154 (7516)	242 (389)
	0.1%	4946 (5072)	146 (257)
	0.0%	4559 (4559)	119 (119)
10%	1.5%	2233 (2464)	177 (323)
	1.0%	1705 (1844)	131 (233)
	0.5%	1346 (1429)	83 (116)
	0.1%	1130 (1167)	65 (60)
	0.0%	1083 (1083)	59 (59)
20%	1.5%	339 (371)	52 (78)
	1.0%	304 (335)	42 (57)
	0.5%	272 (289)	35 (61)
	0.5%	249 (258)	30 (29)
	0.0%	245 (245)	29 (29)

Table 5.1: Mean (90th percentile) stopping times of 400 simulated comparison audits run with oracle bets (λ^*) or apKelly bets (λ^{apK}) under a range of diluted margins and 2-vote overstatement rates. DM = diluted margin; OR = overstatement rate.

Figure 5.2. Specifically, the mixing distribution was a discretized, truncated, bivariate normal with mean vector $(\tilde{p}_1, \tilde{p}_2)$, standard deviation $(\sigma_1, \sigma_2) := (0.5\%, 0.25\%)$, and correlation $\rho := 0.25$. The fact that $\sigma_2 < \sigma_1$ reflects more prior confidence that 2-vote overstatement rates will be concentrated near their prior mean, while $\rho > 0$ encodes a prior suspicion that overstatement rates are correlated: they are more likely to be both high or both low. After setting the weights at each grid point according to this normal density, they were rescaled to sum to unity.

We simulated 400 audits under sampling with replacement for each scenario. The stopping times were capped at 20000, the size of the population, even if the audit hadn't stopped by that point. We estimated the expected value and 90th percentile of the stopping times for each method by the empirical mean and 0.9 quantile over the 400 simulations. We computed the geometric mean ratio of the expected stopping times of each method over that of the oracle strategy as a summary of their performance across scenarios.

Table 5.2 presents results. With few 2-vote overstatements, all strategies performed relatively well and the audits concluded quickly. When the priors substantially underestimated the true overstatement rates, the performance of the audits degraded significantly compared to the oracle bets. This was especially true for the fixed strategy. For example, when $(p_1, p_2) = (0.1\%, 1\%)$ and $\tilde{p}_2 = 0.01\%$, the expected number of ballots for the fixed strategy to stop was more than 20 times that of the oracle method. On the other hand, the adaptive and diversified strategies were much more robust to a poor prior estimate. In particular, the expected stopping time of the diversified method was never more than 3 worse than that of the oracle strategy, and the adaptive method was never more than 4 times worse. The geometric mean workload ratios of each strategy over the oracle strategy were 2.4 for fixed, 1.3 for adaptive, and 1.2 for diversified. The diversified method was the best practical method on average across scenarios.

5.6 Extensions

5.6.1 Betting while sampling without replacement

When sampling without replacement, the distribution of X_i depends on past data $X_1, ..., X_{i-1}$. Naively updating an *a priori* bet to reflect what we know has been sampled may actually harm the efficiency of the audit.

Specifically, recall that, for $k \in \{1, 2\}$, $\hat{p}_{ki}$ denotes the sample proportion of the overstatement rate at time i. If we fix initial rate estimates to $\tilde{p}_k$, then the updated estimate at time i given that we have removed $i\hat{p}_{k(i-1)}$ would be

$$\tilde{p}_{ki} = \frac{N\tilde{p}_k - i\hat{p}_{k(i-1)}}{N - i + 1} \quad \text{for} \quad k \in \{1, 2\}.$$

This can be plugged into (5.3) to estimate the optimal λ_i^* for each draw. Fixing $\tilde{p}_{1i} = 0$ and using equation (5.4) yields the closed form optimum:

$$\lambda_i^* = \frac{2 - 4a\tilde{p}_{2i}}{1 - 2a} \wedge 2,$$

where we have truncated at 2 to guarantee that λ_i^* is even a valid bet. This is necessary because the number of 2-vote overstatements in the sample can exceed the number $N\tilde{p}_2$ hypothesized to be in the entire population. If this occurs, the audit will stall if even one more 2-vote overstatement is discovered. More generally, this strategy has the counterintuitive (and counterproductive) property of betting *more* aggressively as more overstatements are discovered. To avoid this pitfall we suggest using the betting strategies we derived earlier under IID sampling, even when sampling without replacement.

5.6.2 Other social choice functions

SHANGRLA [Stark, 2020] encompasses a broad range of social choice functions beyond plurality, all of which are amenable to comparison audits. Assorters for approval voting and proportional representation are identical to plurality assorters, so no modification to the optimal bets is required. Ranked-choice voting can also be reduced to auditing a collection of plurality assertions, though this reduction may not be the most efficient possible [Blom et al., 2019]. On the other hand, some social choice functions, including weighted additive and supermajority, require different assorters and will have different optimal bets.

In a supermajority contest, the diluted margin v is computed differently depending on the fraction $f \in (1/2, 1]$ required to win, as well as the proportion of votes for the reported winner in the CVRs. In the population of overstatement assorters error-free CVRs still appear as $a = (2 - v)^{-1}$, but 2-vote overstatements are $(1 - 1/(2f))a > 0$ and 1-vote overstatements are $(3/2 - 1/(2f))a$. So that the population attains a lower bound of 0, we can make the shift $x_i - (1 - 1/(2f))a$ and test against the shifted mean $1/2 - (1 - 1/(2f))a$. Because there are only 3 points of support, the derivations in Section 5.3.2 can be repeated, yielding a new solution for λ^* in terms of the rates and the shifted mean.

Weighted additive schemes apply an affine transformation to ballot scores to construct assorters. Because scores may be arbitrary non-negative numbers, there can be more than 3 points of support for the overstatement assorters and the derivations in Section 5.3.2 cannot be immediately adapted. If most CVRs are correct then most values in the population will be above 1/2, suggesting that an aggressive betting strategy with $\lambda := 2 - \epsilon$ will be relatively efficient. Alternatively, a diversified strategy weighted towards large values of $\lambda \in (0, 2]$ can retain efficiency when there are in fact high rates of error. It should also be possible to attain a more refined solution by generalizing the optimization strategy in Section 5.3.2 to populations with more than 3 points of support.

5.6.3 Batch-level comparison audits

Batch-level comparison audits check for error in totals across batches of ballots, and are applicable in different situations than ballot-level comparisons, since they do not require CVRs. SHANGRLA-style overstatement assorters for batch-level comparison audits are derived in Stark [2023]. These assorters generally take a wide range of values within $[0, u]$. Because they are not limited to a few points of support, there is not a simple optimal betting strategy. However, assuming there is relatively little error in the reported batch-level counts, will again place the majority of the assorter distribution above 1/2. This suggests using a relatively aggressive betting strategy, placing more weight on bets near 2 (or near the assorter upper bound in the ALPHA parameterization).

Stark [2023] evaluated various BSMs in simulations approximating batch-level comparison audits, though the majority of mass was either at 1 or spread uniformly on $[0, 1]$, not at a value $a \in (1/2, 1]$. Nevertheless, in situations where most of the mass was at 1, aggressive betting ($\eta \geq 0.9$) was most efficient. Investigating efficient betting strategies for batch-level comparison audits remains an important area for future work.

5.7 Conclusions

We derived optimal bets for ballot-level comparison audits of plurality contests and sketched some extensions to broader classes of comparison RLAs. The high-level upshot is that comparison should use considerably more aggressive betting strategies than polling in practice, a point made abundantly clear in our oracle simulations. Our practical strategies approached the efficiency of oracle bets, except in cases where $p_2 = 1\%$. Such a high rate of 2-vote overstatements is unlikely in practice, and would generally imply something has gone terribly wrong: votes for the loser should not be flipped to votes for the winner.

Future work should continue to flesh out efficient strategies for batch-level comparison, and explore the effects of understatement errors. We suspect that understatements will have little effect on the optimal strategy. If anything, they imply bets should be even *more* aggressive, but we already suggest placing most weight near the maximal value of $\lambda_i = 2$, diversifying or thresholding to prevent stalls if 2-vote overstatements are discovered. We hope our results will guide efficient real-world comparison RLAs, and demonstrate the practicality of their routine implementation for trustworthy, evidence-based elections.

Code

Code implementing our simulations and generating our figures and tables is available on Github at `https://github.com/spertus/comparison-RLA-betting`.

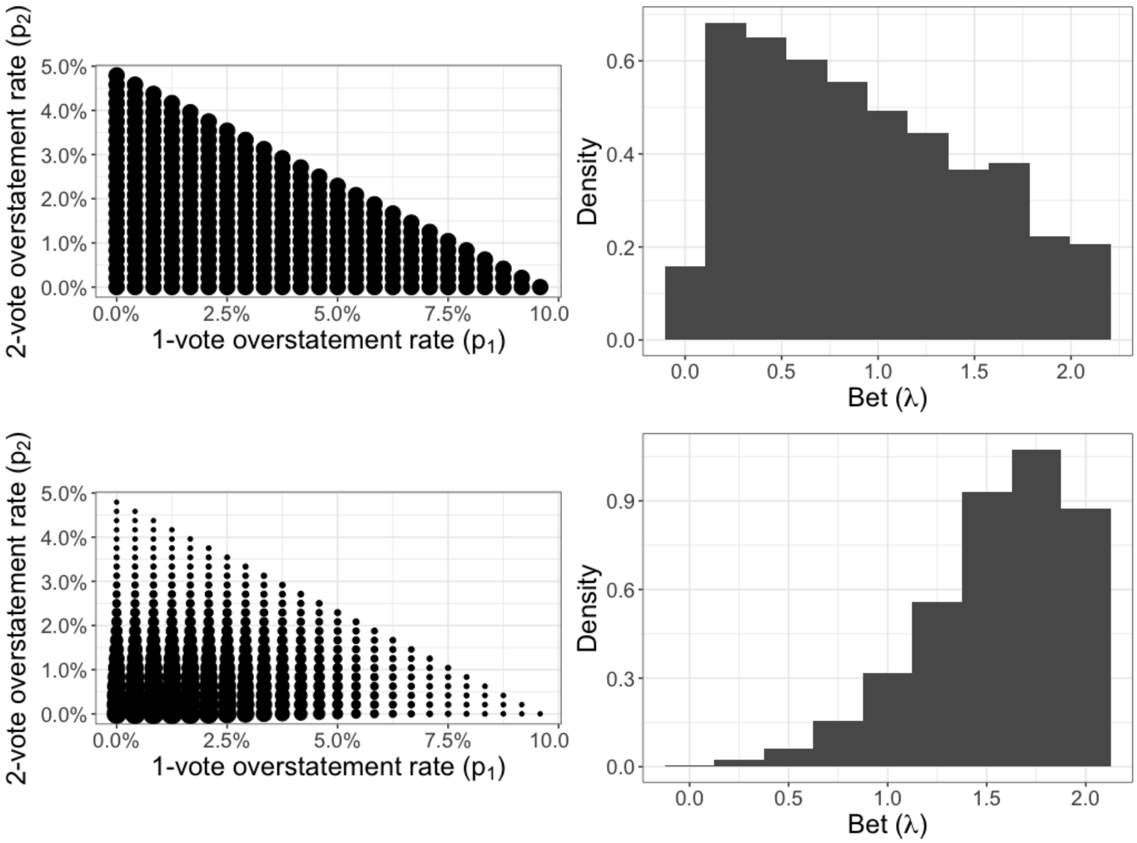


Figure 5.2: Plots showing two mixture distributions over overstatement rates (left column; y-axis = 2-vote overstatment rate, x-axis = 1-vote overstatement rate; point size = mixture weight) and their corresponding induced distributions over the bets (right column; x-axis = bet, y-axis = density). The diluted margin of 10% constrains possible overstatement rates. The upper row shows a uniform grid of weights over all overstatement rates (left column) and its induced distribution on λ (right column). The bottom row plots discretized, truncated, and rescaled bivariate normal weights with parameters $(\mu_1, \mu_2) = (.01, .001)$, $(\sigma_1, \sigma_2) = (.02, .01)$, and $\rho = 0.25$ (left column) and its induced distribution on λ (right column).

True ORs		**Prior ORs**		**Stopping Times**			
p_2	p_1	$\tilde{p}_2$	$\tilde{p}_1$	Oracle	Fixed	Adaptive	Diversified
0.01%	0.1%	0.01%	0.1%	124 (147)	125 (119)	124 (147)	131 (152)
			1%	124 (147)	125 (147)	125 (147)	131 (154)
		0.1%	0.1%	125 (147)	129 (151)	131 (151)	133 (155)
			1%	127 (147)	132 (153)	130 (152)	135 (157)
	1%	0.01%	0.1%	174 (229)	167 (229)	166 (229)	177 (236)
			1%	168 (229)	172 (229)	167 (229)	180 (235)
		0.1%	0.1%	176 (229)	169 (232)	175 (262)	181 (262)
			1%	159 (205)	174 (233)	180 (265)	184 (264)
0.1%	0.1%	0.01%	0.1%	146 (256)	153 (338)	159 (350)	149 (271)
			1%	151 (256)	154 (174)	150 (147)	145 (154)
		0.1%	0.1%	147 (256)	152 (256)	146 (182)	153 (259)
			1%	149 (256)	151 (244)	147 (256)	152 (265)
	1%	0.01%	0.1%	209 (351)	227 (420)	225 (460)	214 (400)
			1%	200 (324)	240 (457)	232 (500)	211 (378)
		0.1%	0.1%	204 (351)	208 (364)	210 (358)	208 (344)
			1%	208 (324)	205 (324)	205 (341)	219 (371)
1%	0.1%	0.01%	0.1%	526 (996)	13654 (20000)	1581 (3517)	888 (2090)
			1%	525 (984)	12685 (20000)	1585 (3731)	739 (1708)
		0.1%	0.1%	528 (1032)	9589 (20000)	1112 (2710)	812 (1982)
			1%	534 (985)	7247 (20000)	915 (2294)	686 (1586)
	1%	0.01%	0.1%	999 (1908)	15205 (20000)	3855 (7811)	2637 (5873)
			1%	1110 (2002)	15641 (20000)	3477 (7529)	1803 (4331)
		0.1%	0.1%	1030 (1868)	13113 (20000)	2795 (5996)	2064 (4884)
			1%	1127 (2256)	13094 (20000)	2437 (5452)	1604 (3758)

Table 5.2: Mean (90th percentile) stopping times over 400 simulated comparison audits with diluted margin of $v = 5\%$ and varying overstatement rates at risk limit $\alpha = 5\%$. The true overstatement rates are in the first two columns. The second two columns contain the prior guesses of the true overstatement rates, used to set bets differently in each strategy as described in Section 5.5.2. The oracle strategy uses the true rates to set the bets, so all variation over ($\tilde{p}_1$, $\tilde{p}_2$) in the results for that strategy is Monte Carlo variation. Monte Carlo variation also accounts for any differences in the fixed and oracle strategies when $(\tilde{p}_1, \tilde{p}_2) = (p_1, p_2)$, since the bets are identical. Note that some stopping time distributions are highly skewed, e.g. the 90th percentile is lower than the mean for fixed bets with $(\tilde{p}_1, \tilde{p}_2) = (p_1, p_2) = (0.1\%, 0.01\%)$. OR = overstatement rate.

Chapter 6

Stratified risk-limiting audits

6.1 Introduction

Most U.S. jurisdictions use computers to tabulate votes. Like all computers, vote tabulators are vulnerable to bugs, human error, and deliberate malfeasance—a fact that has been exploited (rhetorically, if not in reality) to undermine trust in U.S. elections [Levine, 2020, Chaitlin, 2020, Kahn, 2020, Baker and Haberman, 2020].

To deserve public trust, elections must be trustworthy, despite relying on untrustworthy software, hardware, and people: they should provide convincing affirmative evidence that the reported winners really won [Stark and Wagner, 2012, Appel et al., 2020, Appel and Stark, 2020]. Risk-limiting audits (RLAs) are a useful tool for conducting such *evidence-based elections*. RLAs have a specified maximum chance—the *risk limit* α—of not correcting the reported outcome if it is wrong, and never change the reported outcome if it is correct. Below we present methods to reduce the number of ballots that must be manually inspected in an RLA when the reported outcomes are correct, for stratified audit samples.

In a ballot-level *comparison* RLA, manual interpretations of the votes on randomly sampled ballot cards are compared to their corresponding *cast vote records* (CVRs), the system's interpretation of the votes on those cards. In a *ballot-polling* RLA, votes are read manually from randomly selected cards, but those votes are not compared to the system's interpretation of the cards. All else equal, ballot-level comparison RLAs are more efficient than ballot-polling RLAs, but they require the voting system to export CVRs in a way that the corresponding card can be uniquely identified. Not all voting systems can.

Stratified random sampling can be mandatory or expedient in RLAs. Some states' laws require audit samples to be drawn independently across jurisdictions (e.g., California Election Code § 336.5 and § 15360), in which case the audit sample for any contest that crosses jurisdictional boundaries is stratified. Stratifying on the technology used to tabulate votes can increase efficiency by allowing *hybrid audits* [Ottoboni et al., 2018, Howard et al., 2019], which use ballot-level comparison in strata where the voting technology supports it and ballot-polling elsewhere. Another reason to use stratification is to allow RLAs to start before all ballots have been tabulated [Stark, 2019].

The next section briefly reviews prior work on stratified audits. Section 6.3 introduces notation and stratified risk measurement, then presents our improvements: (i) sharper P-values from new risk-measuring functions; (ii) sequential stratified sampling that adapts to the observed data in each stratum to increase efficiency; and (iii) a computationally efficient method for an arbitrary number of strata. Section 6.4 evaluates the innovations using case studies and simulations. Section 6.5 discusses the results and gives recommendations for practice.

6.2 Past Work

The first RLAs involved stratified batch comparison, using the maximum error across strata and contests as the test statistic [Stark, 2008a,b, 2009a, Hall et al., 2009], a rigorous but inefficient approach. Higgins et al. [2011] computed sharper P-values for the same test statistic using dynamic programming. SUITE [Ottoboni et al., 2018, Howard et al., 2019] uses *union-intersection tests* to represent the null hypothesis that one or more reported winners actually lost as a union of intersections of hypotheses about individual strata; it involves optimization problems that are hard to solve when there are more than two strata.

More recently, SHANGRLA [Stark, 2020] has reduced RLAs to a canonical form: testing whether the means of finite, bounded lists of numbers (representing ballot cards) are all less than 1/2, which allows advances in statistical inference about bounded populations to be applied directly to RLAs. Stark [2020] showed that union-intersection tests can be used with SHANGRLA to allow *any* risk-measuring function to be used in any stratum in stratified audits.

Stark [2023] provided a new approach to union-intersection tests using nonnegative supermartingales (NNSMs): *intersection supermartingales*, which open the possibility of reducing sample sizes by adaptive *stratum selection* (using the first t sampled cards to select the stratum from which to draw the $(t + 1)$th card). Stark [2023] does not provide an algorithm for stratum selection or evaluate the performance of the approach; this paper does both.

6.3 Stratified audits

We shall formalize stratified audits using the SHANGRLA framework [Stark, 2020], which unifies comparison and polling audits. We then show how to construct a stratified comparison audit using SHANGRLA, how to measure the risk based on a stratified sample, and how adaptive sequential stratified sampling can improve efficiency.

6.3.1 Assorters and assertions

Ballot cards are denoted $\{b_i\}_{i=1}^N$. An assorter A assigns a number $A(b_i) \equiv x_i \in [0, u]$ to ballot card b_i [Stark, 2020] and the value $A(c_i)$ to CVR i. The value an assorter assigns to a card depends on the votes on the card, the social choice function, and possibly on the machine interpretation of that card and others (for comparison audits). Stark [2020] describes how to define a set of assorters for many social choice functions (including majority, multiwinner majority, supermajority, Borda count, approval voting, all scoring rules, D'Hondt, STAR-Voting, and IRV) such that the reported winner(s) really won if the mean of every assorter in the set is greater than 1/2. The claim that an assorter mean is $> 1/2$ is called an *assertion*. An RLA with risk limit α confirms the outcome of a contest if it rejects the *complementary null* that the assorter mean is $\leq 1/2$ at significance level α for every assorter relevant to that contest.

In a stratified audit, the population of ballot cards is partitioned into K disjoint *strata*. Stratum k contains N_k ballot cards, so $N = \sum_k N_k$. The *weight* of stratum k is $w_k := N_k/N$; the weight vector is $\boldsymbol{w} := [w_1, ..., w_K]^T$. For each assorter A there is a set of assorter values $\{x_i\}_{i=1}^N$. Each assorter may have its own upper bound u_k in stratum k.[1] The true mean of the assorter values in stratum k is μ_k; $\boldsymbol{\mu} := [\mu_1, ..., \mu_K]^T$. The overall assorter mean is

$$\mu := \frac{1}{N}\sum_{i=1}^N x_i = \sum_{k=1}^K \frac{N_k}{N}\mu_k = \boldsymbol{w}^T\boldsymbol{\mu}.$$

Let $\boldsymbol{\theta} = [\theta_1, ..., \theta_K]^T$ with $0 \leq \theta_k \leq u_k$. A single *intersection null* is of the form $\boldsymbol{\mu} \leq \boldsymbol{\theta}$, i.e., $\cap_{k=1}^K \{\mu_k \leq \theta_k\}$. The *union-intersection form* of the *complementary null* that the outcome is incorrect is:

$$H_0 : \bigcup_{\boldsymbol{\theta}:\boldsymbol{w}^T\boldsymbol{\theta}\leq\frac{1}{2}} \bigcap_{k=1}^K \{\mu_k \leq \theta_k\}. \tag{6.1}$$

[1] The notation we use does not allow u to vary by draw, but the theory in Stark [2023] permits it, and it is useful for batch-comparison audits.

From stratum k we have n_k samples $X_k^{n_k} := \{X_{1k}, ..., X_{n_k k}\}$ drawn by simple random sampling, with or without replacement, independently across strata. Section 6.3.3 shows how to use single-stratum hypothesis tests (of the the null $\mu_k \le \theta_k$) to test (6.1). First, we show how to write stratified comparison audits in this form.

6.3.2 Stratified comparison audits

In SHANGRLA, comparison audits involve translating the original assertions about the true votes into assertions about the reported results and discrepancies between the true votes and the machine's record of the votes [Stark, 2020, Section 3.2]. For each assertion, the corresponding *overstatement assorter* assigns ballot card b_i a bounded, nonnegative number that depends on the votes on that card, that card's CVR, and the reported results. The original assertion is true if the average of the overstatement assorter values is greater than 1/2.

We now show that for stratified audits, the math is simpler if, as before, we assign a nonnegative number to each card that depends on the votes and reported votes, but instead of comparing the average of the resulting list to 1/2, we compare it to a threshold that depends on the hypothesized stratum mean θ_k.

Let u_k^A be the upper bound on the original assorter for stratum k and $\omega_{ik} := A(c_{ik}) - A(b_{ik}) \in [-u_k^A, u_k^A]$ be the *overstatement* for the ith card in stratum k, where $A(c_{ik})$ is the value of the assorter applied to the CVR and $A(b_{ik})$ is the value of the assorter for the true votes on that card. Let $\bar{A}_k^b$, $\bar{A}_k^c$, and $\bar{w}_k = \bar{A}_k^c - \bar{A}_k^b$ be the true assorter mean, reported assorter mean, and average overstatement, all for stratum k.

For a particular $\boldsymbol{\theta}$, the intersection null claims that in stratum k, $\bar{A}_k^b \le \theta_k$. Adding $u_k^A - \bar{A}_k^c$ to both sides of the inequality yields

$$u_k^A - \bar{\omega}_k \le \theta_k + u_k^A - \bar{A}_k^c.$$

Letting $u_k := 2u_k^A$, take $B_{ik} := u_k^A - \omega_{ik} \in [0, u_k]$ and $\bar{B}_k := \frac{1}{N_k} \sum_{i=1}^{N_k} B_{ik}$. Then $\{B_{ik}\}$ is a bounded list of nonnegative numbers, and the assertion in stratum k is true if $\bar{B}_k > \beta_k := \theta_k + u_k^A - \bar{A}_k^c$, where all terms on the right are known. Testing whether $\bar{B} \le \beta_k$ is the canonical problem solved by ALPHA [Stark, 2023]. The intersection null can be written

$$\bar{B}_k \le \beta_k \ \text{ for all } \ k \in \{1, \dots, K\}.$$

Define $\boldsymbol{u} := [u_1, \dots, u_K]^T$. As before, we can reject the complementary null if we can reject *all* intersection nulls $\boldsymbol{\theta}$ for which $\mathbf{0} \le \boldsymbol{\theta} \le \boldsymbol{u}$ and $\boldsymbol{w}^T\boldsymbol{\theta} \le 1/2$.

6.3.3 Union-intersection tests

A union-intersection test for (6.1) combines evidence across strata to see whether any intersection null in the union is plausible given the data, that is, to check whether the P-value of any intersection null in the union is greater than the risk limit.

Consider a fixed vector $\boldsymbol{\theta}$ of within-stratum nulls. Let $P(\boldsymbol{\theta})$ be a valid P-value for the intersection null $\boldsymbol{\mu} \leq \boldsymbol{\theta}$. Many functions can be used to construct $P(\boldsymbol{\theta})$ from tests in individual strata; two are presented below. We can reject the union-intersection null (6.1) if we can reject the intersection null for all feasible $\boldsymbol{\theta}$ in the half-space $\boldsymbol{w}^T\boldsymbol{\theta} \leq 1/2$. Equivalently, $P(\boldsymbol{\theta})$ maximized over feasible $\boldsymbol{\theta}$ is a P-value for (6.1):

$$P^* := \max_{\boldsymbol{\theta}} \ \{P(\boldsymbol{\theta}) : \mathbf{0} \leq \boldsymbol{\theta} \leq \boldsymbol{u} \ \text{ and } \ \boldsymbol{w}^T\boldsymbol{\theta} \leq 1/2\}.$$

This method is fully general in that it can construct a valid P-value for (6.1) from stratified samples and any mix of risk-measuring functions that are individually valid under simple random sampling. However, the tractability of the optimization problem depends on the within-stratum risk-measuring functions and the form of P used to pool risk. So does the efficiency of the audit.

We next give two valid combining rules $P(\boldsymbol{\theta})$. Section 6.3.6presents some choices for within-stratum risk measurement to construct $P(\boldsymbol{\theta})$.

6.3.4 Combining Functions

Ottoboni et al. [2018] and Stark [2020] calculate P for the intersection null using Fisher's combining function. Let $p_k(\theta_k)$ be a P-value for the single-stratum null $H_{0k} : \mu_k \leq \theta_k$. Define the pooling function

$$P_F(\boldsymbol{\theta}) := 1 - \chi^2_{2K}\left(-2\sum_{k=1}^{K} \log p_k(\theta_k)\right),$$

where χ^2_{2K} is the CDF of the chi-squared distribution with 2K degrees of freedom. The term inside the CDF, $-2\sum_{k=1}^{K} \log p_k(\theta_k)$, is Fisher's combining function[2]. Because samples are independent across strata, $\{p_k(\theta_k)\}_{k=1}^{K}$ are independent random variables, so Fisher's combining function is dominated by the chi-squared distribution with $2K$ degrees of freedom [Ottoboni et al., 2018]. The maximum over $\boldsymbol{\theta}$, P_F^*, is a valid P-value for (6.1).

[2] Other combining functions could be used, including Liptak's or Tippett's. See Chapter 4 of Pesarin and Salmaso [2010a]

6.3.5 Intersection supermartingales

Stark [2023] derives a simple form for the P-value for an intersection null when supermartingales are used as test statistics within strata. Let $M^k_{n_k}(\theta_k)$ be a supermartingale constructed from n_k samples drawn from stratum k when the null $\mu_k \leq \theta_k$ is true. Then the product of these supermartingales is also a supermartingale under the intersection null, so its reciprocal (truncated above at 1) is a valid P-value [Stark, 2023, Waudby-Smith et al., 2021]:

$$P_M(\boldsymbol{\theta}) := 1 \wedge \prod_{k=1}^{K} M^k_{n_k}(\theta_k)^{-1}.$$

Maximizing $P_M(\boldsymbol{\theta})$ (equivalently, minimizing the intersection supermartingale) yields P^*_M, a valid P-value for (6.1).

6.3.6 Within-stratum P-values

The class of within-stratum P-values that can be used to construct P_F is very large, but P_M is limited to functions that are supermartingales under the null. Possibilities include:

- **SUITE**, which computes P^*_F for two-stratum hybrid audits. The P-value in the CVR stratum uses the MACRO test statistic [Stark, 2009c]; the P-value in the no-CVR stratum takes a maximum over many values of Wald's SPRT indexed by a nuisance parameter representing the number of non-votes in the stratum. The maximations in MACRO and over a nuisance parameter in the SPRT make SUITE less efficient than newer methods based on SHANGRLA [Stark, 2020].

- **ALPHA**, which constructs a betting supermartingale as in Waudby-Smith and Ramdas [2023], but with an alternate parameterization [Stark, 2023]. Such methods are among the most efficient for RLAs [Waudby-Smith et al., 2021, Stark, 2023], but the efficiency depends on how the tuning parameter τ_{ik} is chosen. Stark [2023] offers a sensible strategy based on setting τ_{ik} to a stabilized estimate of the true mean μ_k. We implement that approach and a modification that is more efficient for comparison audits. Both P^*_M and P^*_F can be computed from stratum-wise ALPHA supermartingales. However, finding the maximum P-value over the union is prohibitively slow when $K > 2$.

- **Empirical Bernstein** (EB), which is a supermartingale presented in Howard et al. [2021] and Waudby-Smith and Ramdas [2023]. Although they are generally not as efficient as ALPHA and other betting supermartingales [Waudby-Smith and Ramdas, 2023], EB supermartingales have an exponential analytical form that makes $\log P_M(\boldsymbol{\theta})$ or $\log P_F(\boldsymbol{\theta})$ linear or piecewise linear in $\boldsymbol{\theta}$. Hence, P_M^* and P_F^* can be computed quickly for large K by solving a linear program.

We compare the efficiency of these risk-measuring functions in Sections 6.4.1 and 6.4.2.

6.3.7 Sequential stratum selection

The use of sequential sampling in combination with stratification presents a new possibility for reducing workload: sample more from strata that are providing evidence against the intersection null and less from strata that are not helping. To set the stage, suppose we are conducting a ballot-polling audit with two strata of equal size and testing the intersection null $\boldsymbol{\theta} = [0.25, 0.75]^T$. We have drawn 50 ballot cards from each stratum and found sample assorter means of $[0.5, 0.6]^T$. Given the data, it seems plausible that drawing more samples from the first stratum will strengthen the evidence that $\mu_1 > 0.25$, but additional sampling from the second stratum might not provide evidence that $\mu_2 > 0.75$: to reject the intersection null, it might help to draw disproportionately from the first stratum. Perhaps suprisingly, such adaptive sampling yields valid inferences when the P-value is constructed from supermartingales and the stratum selection function depends only on past data. We now sketch why this is true.

For $t \in \mathbb{N}$ and a particular vector of hypothesized stratum means $\boldsymbol{\theta}$, let

$$\kappa_t(\boldsymbol{\theta}) \in \{1, ..., K\}$$

denote the stratum from which the t-th sample was drawn for testing the hypothesis $\boldsymbol{\mu} \leq \boldsymbol{\theta}$. We call $\kappa(\boldsymbol{\theta}) := (\kappa_t(\boldsymbol{\theta}))_{t \in \mathbb{N}}$ the *stratum selector* for null $\boldsymbol{\theta}$. Crucially, $\kappa(\boldsymbol{\theta})$ is a *predictable sequence* with respect to $(X_t)_{t \in \mathbb{N}}$ in the sense that $\kappa_t(\boldsymbol{\theta})$ can depend on $X^{t-1} := \{X_1, \ldots, X_{t-1}\}$ but not on X_i for $i \geq t$; it could be deterministic given X^{t-1} or may also depend on auxiliary randomness.

For example, a stratum selector could ignore past data and select strata in a deterministic round-robin sequence or at random with probability proportional to stratum size. Alternatively, a rule might select strata adaptively, for instance picking a stratum at random with probability proportional to the current value of each within-stratum supermartingale, so that strata with larger $M_{t_k}^k(\theta_k)$ are more likely to be chosen—an "exploration–exploitation" strategy. In what follows we suppress the dependence on $\boldsymbol{\theta}$ except when it is explicitly required for clarity.

Now, let $M_t^\kappa(\boldsymbol{\theta}) := \prod_{i=0}^t Z_i$ be the test statistic for testing the null hypothesis that the vector of stratumwise means is less than or equal to $\boldsymbol{\theta}$. This is a supermartingale if the individual terms Z_i satisfy a simple condition. Let $Z_0 = 1$ and $Z_i \geq 0$ for all i. If

$$\mathbb{E}_{\boldsymbol{\theta}}[Z_t | X^{t-1}] \leq 1, \tag{6.2}$$

then $(M_t^\kappa(\boldsymbol{\theta}))_{t \in \mathbb{N}_0}$ is a nonnegative supermartingale starting at 1 under the null. By Ville's inequality [Ville, 1939], the thresholded inverse $(1 \wedge M_t^\kappa(\boldsymbol{\theta})^{-1})_{t \in \mathbb{N}_0}$ is an anytime P-value sequence when $\boldsymbol{\mu} \leq \boldsymbol{\theta}$.

Condition (6.2) holds if the Z_i are terms extracted from a set of within-stratum supermartingales using a predictable stratum selector: Let

$$\nu_t^\kappa := \#\{i \leq t : \kappa_i = \kappa_t\} \tag{6.3}$$

be the number of draws from stratum k as of time t. Suppose that for $k \in \{1, \ldots, K\}$, $M_t^k(\theta_k) := \prod_{i=1}^t Y_i^k(\theta_k)$ is a nonnegative supermartingale starting at 1 when X_{ik} is the ith draw from stratum k and the kth stratum mean is $\mu_k \leq \theta_k$. Then if

$$Z_i := Y_{\nu_i^\kappa}^{\kappa_i}(\theta_{\kappa_i}), \tag{6.4}$$

condition (6.2) holds and the interleaved test statistic $M_t^\kappa(\boldsymbol{\theta})$ is an intersection supermartingale under the null. We compare two stratum selection rules in Section 6.4.1.

6.4 Evaluations

6.4.1 Combination and allocation rules

We simulated a variety of two-stratum ballot-level comparison audits at risk limit $\alpha = 5\%$, with assorters defined as in Section 6.3.2. The strata each contained $N_k = 1000$ ballot cards, all with valid votes. Cards were sampled without replacement. The stratum-wise true margins were $[0\%, 20\%]$, $[0\%, 10\%]$ or $[0\%, 2\%]$, corresponding to global margins of 10%, 5%, and 1%, respectively. Stratum-wise reported margins were also $[0\%, 20\%]$, $[0\%, 10\%]$ or $[0\%, 2\%]$, so error was always confined to the second stratum. Each reported margin was audited against each true margin in 300 simulations. Risk was measured by ALPHA or EB combined either as intersection supermartingales (P_M^*) or with Fisher's combining function (P_F^*), with one of two stratum selectors: proportional allocation or lower-sided testing.

In proportional allocation, the number of samples from each stratum is in proportion to the number of cards in the stratum. Allocation by lower-sided testing involves testing the null $\mu_k \geq \theta_k$ sequentially at level 5% using the same supermartingale (ALPHA or EB) used to test the main (upper-sided) hypothesis of interest. This allocation rule ignores samples from a given stratum once the lower-sided hypothesis test rejects, since there is strong evidence that the null is true in that stratum. This "hard stop" algorithm is unlikely to be optimal, but it leads to a computationally efficient implementation and illustrates the potential improvement in workload from adaptive stratum selection.

Tuning parameters were chosen as follows. ALPHA supermartingales were specified either with τ_{ik} as described in Stark [2023, Section 2.5.2] (ALPHA-ST, "shrink-truncate") or with a strategy that biases τ_{ik} towards u_k: (ALPHA-UB, "upward bias"). The ALPHA-UB strategy helps in comparison audits because the distribution of assorter values consists of a point mass at $u_A^k = u_k/2$ and typically small masses (with weight equal to the overstatement rates) at 0 and another small value. This concentration of mass makes it advantageous to bet more aggressively that the next draw will be above the null mean; that amounts to biasing τ_{ik} towards the upper bound u_k. Before running EB, the population and null were transformed to [0,1] by dividing by u_k. The EB supermartingale parameters λ_{ik} were then specified following the "predictable mixture" strategy [Waudby-Smith and Ramdas, 2023, Section 3.2], truncated to be below 0.75. Appendix B.1 gives more details of the ALPHA-ST and ALPHA-UB strategies and the computations.

Sample size distributions for some combinations of reported and true margins are plotted in Figure 6.1 as (simulated) probabilities of stopping at or before a given sample size. Table 6.1 gives estimated expected and 90th percentile sample sizes for each scenario and method. Table 6.2 lists aggregate scores, computed by finding the ratio of the workload for each method over the smallest workload in each scenario, then averaging over scenarios by taking the geometric mean of these ratios.

Intersection supermartingales tend to dominate Fisher pooling unless the stratum selector is chosen poorly (e.g., the bottom-right panel of Figure 6.1 and the last row of Table 6.2). Stratum selection with the lower-sided testing procedure is about as efficient as proportional allocation for the ALPHA supermartingales, but far more efficient than proportional allocation for EB. The biggest impact of the allocation rule occurred for EB combined by intersection supermartingales when the reported margin was 0.01 and the true margin was 0.1: proportional allocation produced an expected workload of 752 cards, while lower-sided testing produced an expected workload of 271 cards—a 74% reduction. Table 6.2 shows that ALPHA-UB with intersection supermartingale combining and lower-sided testing is the best method overall; ALPHA-UB with intersection combining and proportional allocation is a close second; EB with intersection combining and lower-sided testing is also relatively sharp; ALPHA-ST with Fisher combining is least efficient.

We also ran simulations at risk limits 1% and 10%, which did not change the relative performance of the methods. However, compared to a 5% risk limit, a 10% risk limit requires counting about 17% fewer cards and a 1% risk limit requires about 38% more, on average across scenarios and methods.

6.4.2 Comparison to SUITE

SUITE was used in a pilot RLA of the 2018 gubernatorial election in Michigan [Howard et al., 2019]. Three jurisdictions—Kalamazoo, Rochester Hills, and Lansing—were audited, but only Kalamazoo successfully ran a hybrid audit. We recalculated the risk on audit data from the closest race in Kalamazoo (Whitmer vs Schuette) using ALPHA with the optimized intersection supermartingale P-value P_M^*, ALPHA with the optimized Fisher P-value P_F^*, EB with P_F^*, and EB with P_M^*, and compared these with the SUITE P-value. Because we could not access the original order of sampled ballots in the ballot-polling stratum, we simulated P-values for 10,000 random ballot orders with the marginal totals in the sample. We computed the mean, standard deviation, and 90th percentile of these P-values for each method.

To get the ALPHA P-values, we used ALPHA-UB in the CVR stratum and ALPHA-ST in the no-CVR stratum. For EB P-values, we used the predictable mixture parameters of Waudby-Smith and Ramdas [2023] to choose λ_{ik}, truncating at 0.75 in both strata. Sample allocation was dictated by the original pilot audit: 8 cards from the CVR stratum (5,294 votes cast; diluted margin 0.55) and 32 from the no CVR stratum (22,732 votes cast; diluted margin 0.57).

Table 6.3 presents P-values for each method. For ALPHA, the mean P_F^* is about half the SUITE P-value; for P_M^*, the mean is more than an order of magnitude smaller than the SUITE P-value. The P-value distributions for ALPHA are concentrated near the mean. On the other hand, the EB P_M^* and P_F^* P-values are both an order of magnitude larger than the SUITE P-value and their distributions are substantially more dispersed than the distributions of ALPHA P-values.

6.4.3 A highly stratified audit

As mentioned in Section 6.3.6, many within-stratum risk-measuring functions do not yield tractable expressions for $P_F(\boldsymbol{\theta})$ or $P_M(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, making it hard to find the maximum P-value over the union unless K is small. Indeed, previous implementations of SUITE only work for $K = 2$. However, the combined log-P-value for EB is linear in $\boldsymbol{\theta}$ for P_M^* and piecewise linear for P_F^*. Maximizing the combined log-P-value over the union of intersections is then a linear program that can be solved efficiently even when K is large.

To demonstrate, we simulated a stratified ballot-polling audit of the 2020 presidential election in California, in which $N = 17,500,881$ ballots were cast across $K = 58$ counties (the strata), using a risk limit of 5%. The simulations assumed that the reported results were correct, and checked whether reported winner Joseph R. Biden really beat reported loser Donald J. Trump. The audit assumed that every ballot consisted of one card; workloads would be proportionately higher if the sample were drawn from a collection of cards that includes some cards that do not contain the contest. Sample sizes were set to be proportional to turnout, plus 10 cards, ensuring that at least 10 cards were sampled from every county. Risk was measured within strata by EB with predictable mixture λ_{ik} thresholded at 0.9 [Waudby-Smith and Ramdas, 2023]. Within-stratum P-values were combined using P_F^* (P_M^* did not work well for EB with proportional allocation in simulations). To approximate the distribution of sample sizes needed to stop, we simulated 30 audits at each increment of 5,000 cards from 5,580 to 100,580 cards. We then simulated 300 audits at 70,580 cards, roughly the 90th percentile according to the smaller simulations.

In 91% of the 300 runs, the audit stopped by the time 70,580 cards had been drawn statewide. Drawing 70,580 ballots by our modified proportional allocation rule produces within-county sample sizes ranging from 13 (Alpine County, with the fewest voters) to 17,067 (Los Angeles County, with the most). A comparison or hybrid audit using sampling without replacement would presumably require inspecting substantially fewer ballots. It took about 3.5 seconds to compute each P-value in R (4.1.2) using a linear program solver from the `lpSolve` package (5.6.15) on a mid-range laptop (2021 Apple Macbook Pro).

6.5 Discussion

ALPHA intersection supermartingales were most efficient compared to the SUITE pilot audit in Michigan and in simulations. Lower-sided testing allocation was better than proportional allocation, especially for EB. Fisher pooling limits the damage that a poor allocation rule can do, but is less efficient than intersection supermartingales with a good stratum selection rule. For comparison audits, it helps to bet more aggressively than ALPHA-ST by using ALPHA-UB or EB. However, EB was not efficient compared to SUITE when replicating the Michigan hybrid audit due to poor performance in the ballot-polling stratum.

Our general recommendation for hybrid audits is: (i) use an intersection supermartingale test with (ii) adaptive stratum selection and (iii) ALPHA-UB (or another method that can exploit low sample variance to bet more aggressively) as the risk-measuring function in the comparison stratum and (iv) ALPHA-ST (or a method that "learns" the population mean) as the risk-measuring function in the ballot-polling stratum. When the number of strata is large, audits can leverage the log-linear form of the EB supermartingale to quickly find the maximum P-value, as illustrated by our simulated audit spread across California's 58 counties.

In future work, we hope to construct better stratum allocation rules and characterize (if not construct) optimal rules. The log-linear structure of the EB supermartingale may make it simpler to derive optimal allocation rules.

While stratum selection is not an instance of a traditional multi-armed bandit (MAB) problem, there are connections, and successful strategies for MAB might help. For instance, stratum selection could be probabilistic and involve continuous exploration and exploitation, in contrast to the "hard stop" rules we used in our simulations here.

Data and code

All code used in this paper is available at `https://github.com/spertus/sweeter-than-SUITE`. SUITE was applied to the Michigan RLA data in a Jupyter notebook available at `https://github.com/kellieotto/mirla18`. Reported results from California's 2020 presidential election are available at `https://elections.cdn.sos.ca.gov/sov/2020-general/sov/csv-candidates.xlsx`.

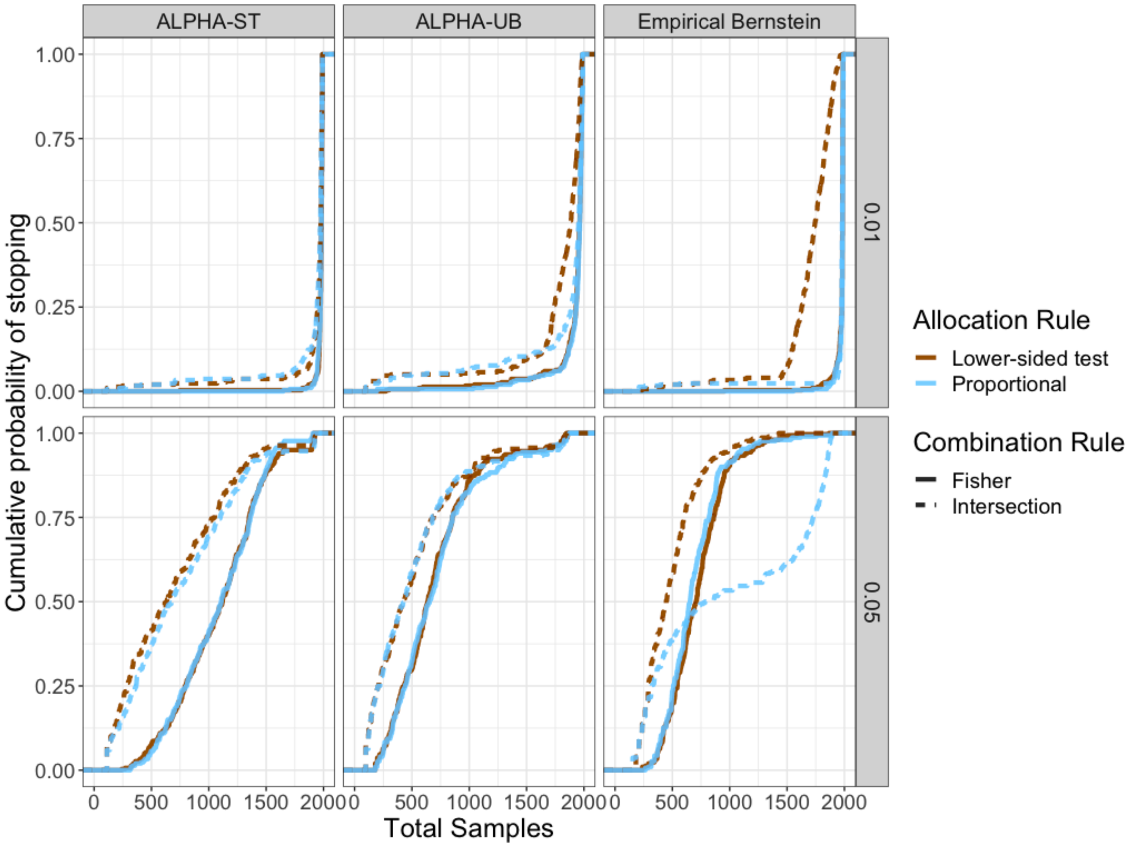


Figure 6.1: Probability that the audit will stop (y-axis) at or before different given sample sizes (x-axis) under different allocation rules (indicated by line color: orange for lower-sided testing and blue for proportional allocation) for different combining functions (indicated by line type: solid for Fisher's combining function and dashed for the intersection supermartingale) at risk limit $\alpha = 5\%$. The true margins are in the rows (1% or 5%) while the reported margin is always 10%. Overstatement errors are confined to one stratum. ALPHA-ST = ALPHA with shrink-truncate τ_{ik}; ALPHA-UB = ALPHA with τ_{ik} biased towards u_k.

Reported margin	supermartingale	Combination	Allocation rule	True margin					
				0.01		0.05		0.1	
				Mean	90th	Mean	90th	Mean	90th
0.01	ALPHA-ST	Fisher	Lower-sided test	1970	1970	1011	1274	338	506
			Proportional	1970	1970	1009	1274	338	540
		Intersection	Lower-sided test	1940	1940	558	848	181	284
			Proportional	1940	1940	554	835	182	298
	ALPHA-UB	Fisher	Lower-sided test	1402	1402	544	754	252	360
			Proportional	1402	1402	548	748	248	354
		Intersection	Lower-sided test	1106	1106	344	504	149	238
			Proportional	1106	1106	342	510	148	232
	Empirical Bernstein	Fisher	Lower-sided test	1438	1438	649	768	384	498
			Proportional	1438	1438	647	782	376	464
		Intersection	Lower-sided test	1102	1102	478	652	271	378
			Proportional	1102	1102	982	1856	752	1728
0.05	ALPHA-ST	Fisher	Lower-sided test	1973	1986	908	908	305	426
			Proportional	1972	1984	908	908	298	412
		Intersection	Lower-sided test	1930	1980	428	428	145	212
			Proportional	1933	1982	428	428	151	228
	ALPHA-UB	Fisher	Lower-sided test	1769	1970	428	428	217	292
			Proportional	1769	1972	428	428	217	288
		Intersection	Lower-sided test	1611	1884	256	256	122	176
			Proportional	1651	1962	256	256	122	180
	Empirical Bernstein	Fisher	Lower-sided test	1882	1986	448	448	306	356
			Proportional	1870	1986	448	448	304	354
		Intersection	Lower-sided test	1610	1858	296	296	199	234
			Proportional	1924	1982	296	296	302	376
0.10	ALPHA-ST	Fisher	Lower-sided test	1971	1990	1088	1536	240	240
			Proportional	1974	1990	1080	1509	240	240
		Intersection	Lower-sided test	1910	1991	694	1312	112	112
			Proportional	1894	1988	755	1347	112	112
	ALPHA-UB	Fisher	Lower-sided test	1904	1984	696	1107	180	180
			Proportional	1914	1984	715	1263	180	180
		Intersection	Lower-sided test	1756	1968	521	1046	98	98
			Proportional	1804	1990	534	1079	98	98
	Empirical Bernstein	Fisher	Lower-sided test	1968	1988	716	987	238	238
			Proportional	1974	1988	686	928	238	238
		Intersection	Lower-sided test	1697	1901	487	799	154	154
			Proportional	1939	1990	1000	1846	154	154

Table 6.1: Expected and 90th percentile sample sizes for various risk-measurement functions, reported margins, and true margins, estimated from 300 simulated audits at risk-limit $\alpha = 5\%$. The best result for each combination of reported margin, true margin, and summary statistic is highlighted. Comparison audit sample sizes are deterministic when there is no error, so the expected value and 90th percentile are equal when the reported and true margins are equal.

supermartingale	**Combination**	**Allocation**	**Score**
ALPHA-ST	Fisher	Lower-sided test	2.11
		Proportional	2.10
	Intersection	Lower-sided test	1.35
		Proportional	1.37
ALPHA-UB	Fisher	Lower-sided test	1.47
		Proportional	1.48
	Intersection	Lower-sided test	1.01
		Proportional	1.02
Empirical Bernstein	Fisher	Lower-sided test	1.73
		Proportional	1.71
	Intersection	Lower-sided test	1.25
		Proportional	1.78

Table 6.2: Score for each method: the geometric mean of the expected workload over the minimum expected workload in each scenario. A lower score is better: a 1.00 would mean that the method always had the minimum expected workload. The best score is highlighted. A score of 2 means that workloads were twice as large as the best method, on average, across simulations and scenarios.

	***P*-value**		
Method	**Mean**	**SD**	**90th**
SUITE	0.037	*	*
ALPHA P_F^*	0.018	0.002	0.019
ALPHA P_M^*	0.003	0.000	0.003
EB P_F^*	0.348	0.042	0.390
EB P_M^*	0.420	0.134	0.561

Table 6.3: Measured risks (P-values) computed from the 2018 Kalamazoo MI audit data. For SUITE, the original P-value is shown. For replications, the mean, standard deviation (SD), and 90th percentile of P-values in 10,000 reshufflings of the sampled ballot-polling data are shown.

Chapter 7

Sequential stratified inference

7.1 Introduction

A ubiquitous problem in applied statistics is to make inferences about the mean μ of a population $\mathcal{X}$, a bag (multiset) of real numbers that could be finite, countable, or uncountable. It is straightforward to construct an unbiased estimate of μ from any probability sample from $\mathcal{X}$, but constructing a valid hypothesis test—one with a Type I error rate guaranteed not to exceed α—is harder. Common methods for inference about the mean involve parametric assumptions about the joint probability distribution of the data or rely on asymptotic arguments. In practice, these methods can have true significance levels much greater than their nominal levels.

For instance, Student's t-test is invalid for general $\mathcal{X}$ at any finite sample size [Lehmann and Romano, 2005]. Absent some restriction on $\mathcal{X}$, there is *no* finite-sample valid test with power exceeding its significance level [Bahadur and Savage, 1956], but it is enough to assume that there are known bounds for each element of $\mathcal{X}$. A one-sided bound (e.g. non-negativity) suffices for a one-sided test. Past work has used such bounds to construct conservative alternatives to Student's t-test when the data are drawn as a uniform independent (with replacement) random sample (UIRS) or a simple (without replacement) random sample (SRS) from $\mathcal{X}$ either of fixed size [Hoeffding, 1963, Anderson, 1967, Bickel, 1993, Fienberg et al., 1977, Romano and Wolf, 2000, Maurer and Pontil, 2009] or sequentially expanding [Kaplan, 1987, Waudby-Smith and Ramdas, 2023, Orabona and Jun, 2022, Stark, 2023].

Many applications use *stratified sampling*, wherein $\mathcal{X}$ is first partitioned into disjoint strata and a UIRS or SRS is taken from each stratum, independently across strata. Stratified sampling is often employed to accommodate logistical constraints or to reduce the variance of unbiased estimates. In particular, variance reduction is

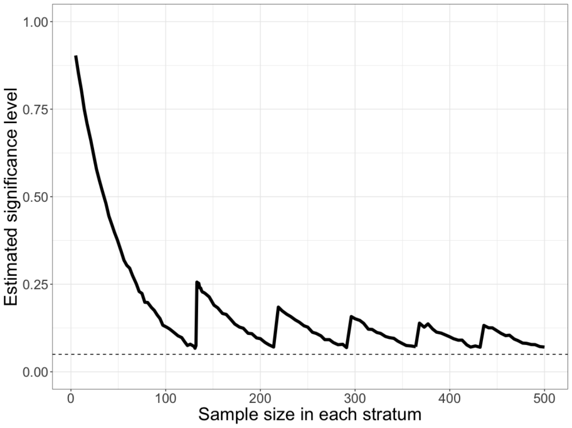


Figure 7.1: Estimated significance level of a stratified, one-sample t-test of the null hypothesis $H_0 : \mu \leq 1/2$ at nominal significance level 5%, for a stratified sample from a population with mean $\mu = 1/2$, as a function of the (fixed) sample size from each stratum. The strata are equal in size and have the same distribution: a mixture of a point mass at 0 (total mass 1%) and a point mass at 0.5050505 (total mass 99%). IID samples of the same size were drawn from the two strata independently. The true significance level was estimated at each sample size using 10,000 simulations. The solid black line is the estimated significance level (y-axis) of the test for a range of sample sizes within each stratum (x-axis). For example, when 50 samples are taken from each stratum (total sample size 100) the nominal 5% significance level test has true level $\sim 35\%$. If the test were valid, the solid line would never be above 5% (dashed line).

possible when strata are more homogeneous than the population as a whole.

Stratification complicates inference. Canonical texts on stratified surveys and experiments suggest using a stratified version of Student's t-test, which approximates the distribution of a weighted sum of stratum-wise sample means by Student's t-distribution [Neyman, 1934, Kish, 1965, Cochran, 1977]. The approximation is good when $\mathcal{X}$ is approximately normal within strata, but the test can be anti-conservative when the within-stratum distributions are skewed. For example, Figure 7.1 plots the true significance level at various sample sizes of a nominal level 5% stratified t-test for a skewed population with two strata. The t-test is anti-conservative even when hundreds of samples are drawn from each stratum. The upshot is that for many real-world problems in stratified inference standard methods are invalid: they do not have a known level over the class of possible populations to which the problem belongs. Bounded populations are one such class.

7.1.1 Nonstandard distributions, gauranteed validity, and auditing

Methods with guaranteed finite-sample validity for bounded populations have been motivated largely by applications in auditing. Audits typically involve (i) high-stakes decisions with a premium on Type I error control; (ii) bounded, but highly skewed populations, for which asymptotic methods are inaccurate; (iii) probability sampling under the control of the auditor, including stratified and/or sequential sampling.

Financial audits aim to determine whether the stated value of a set of assets or invoices *materially*[1] overstates the total true value [Panel on Nonstandard Mixtures of Distributions, 1988], and to draw inferences about the total overstatement. Large sums of money are often involved. For example, the United States Center for Medicare and Medicaid Services relies on random sampling [US CMS, 2023] to estimate and recover billions of dollars in overpayments [Bittinger et al., 2022]. The populations involved are bounded because the amount by which the value of an asset or the amount of an invoice has been overstated is at most the stated value of the asset or invoice. Stratified random sampling is often employed in financial audits [US DHS, 2020] and may dramatically lower the cost of the audit itself [Fienberg et al., 1977].

Election audits are mathematically similar to financial audits, but with a different notion of materiality: the total error is material if it caused any losing candidate to appear to win. If there is a trustworthy, organized record of the votes (see [Appel and Stark, 2020] for steps to establish whether the record is trustworthy), risk-limiting audits (RLAs) can provide evidence that reported winners of an election really won—or ensure a high probability of correcting the reported results if the reported winners did not really win [Stark, 2008a, 2020, 2023]. Elections have high stakes, and RLAs using a trustworthy paper trail provide a sound basis for public confidence in the democratic process [National Academies of Sciences, Engineering, and Medicine, 2018]. For RLAs that compare human reading of individual ballots to the machine interpretation of the same ballots, the population is bounded by the amount that machine error in the interpretation of a ballot could have overstated the margin between a reported winner and a reported loser—in the case of plurality elections, at most two votes when the audit samples individual ballots, or at most the difference between the reported tally in each cluster and a unanimous tally for the loser when the audit samples clusters of ballots [Stark, 2008a,b, 2020]. The population of errors is generally skewed because nonzero errors are typically rare. For audits of plurality contests that use the human reading of votes but do not compare that to

[1] Materiality is often defined indirectly as an amount that would cause a decision maker to decide differently; in practice, it often taken to be a fixed percentage, e.g., 10%.

the machine interpretation, the population is bounded by the maximum number of votes or points that a ballot can award any candidate; other bounds apply to other social choice functions [Stark, 2020]. Stratification may be used to accommodate legal or logistical constraints (e.g., California law requires ballots to be sampled independently by each county[2]) or to increase efficiency (e.g., when heterogeneous voting equipment is involved). Ballots or batches of ballots are drawn one-at-a-time or in larger "rounds." The (sequentially valid or multiplicity-adjusted) P-value of the hypothesis that one or more reported winners actually lost is calculated from the sample. If the P-value is less than the risk limit, the audit stops; otherwise, the auditors may continue sampling or opt to conduct a full hand count. The process proceeds until either the hypothesis has been rejected (i.e., there is strong statistical evidence that the outcome is correct) or there has been a full hand count (which reveals the correct outcome).

7.1.2 Contributions and outline of this paper

This paper introduces methods to make conservative, non-asymptotic inferences about μ from stratified samples, without relying on parametric assumptions. The tests are *sequentially valid*: the analyst may check results as each sample arrives and decide whether to stop sampling or continue gathering data. We call such methods *SFSNP-valid* (sequential, finite-sample, nonparametric). Audits are our lead example, but the methods are also useful for stratified surveys or blocked randomized controlled trials in many regulatory or scientific applications. It is straightforward to construct SFSNP-valid inference by summing confidence sequences constructed within each stratum, but this approach is unnecessarily conservative. Our central goal is to develop a more *efficient* method—one that requires fewer overall samples to meet a given level of evidence.

In broad brush, the new method works as follows: the "global" null hypothesis $H_0 : \mu \leq \eta_0$ is represented as a union of intersection hypotheses. Each intersection hypothesis specifies the mean in every stratum and corresponds to a population mean not greater than η_0. The union is over all such intersections. The global null hypothesis is rejected if every intersection null hypothesis is rejected. For a given intersection null, information about each within-stratum mean is summarized by a test statistic that is a nonnegative supermartingale starting at 1 if the the stratum mean is equal to its hypothesized value—a *test supermartingale* (TSM). The test supermartingales for different strata are combined by multiplication and the combination is converted to a P-value for the intersection null. We explore how the

[2] California Elections Code § 15360.

choices of test supermartingales and the interleaving of samples across strata jointly affect the computational and statistical performance of the test.

The next section reviews the literature on stratified sampling, sequential sampling, finite-sample inference, and inference in the presence of nuisance parameters. Section 7.3 introduces notation and foundational concepts including sequential stratified sampling and TSMs. By Ville's inequality [Ville, 1939], the reciprocal of a TSM is a P-value. We review how to test intersections of hypotheses by combining tests of the individual hypotheses. Section 7.4.1 presents perhaps the simplest strategy for stratified inference, based on summing independently constructed confidence bounds across strata, each with a confidence level adjusted for multiplicity. That method is easy to conceptualize and implement, but is unnecessarily conservative. Section 7.4.2 describes a sharper approach: union-of-intersections tests. Section 7.5 defines notions of consistency and efficiency for sequential-stratified inference. We describe tests that are theoretically optimal but not identified under a composite alternative, and tests that are approximately optimal under the composite alternative. We describe strategies for computation in Section 7.6. Section 7.7 compares the efficiency of different strategies via a few simulation studies. Section 7.8 discusses our results and sketches future directions for research.

7.2 Background and related work

In the early days of statistics Cournot [1843] showed that, for binary populations, stratified sampling with proportional allocation is never less precise than simple random sampling. Tchouproff [1923] was first to derive the optimal stratum-wise sample allocation strategy (for mean squared error) based on stratum sizes and variances, now known as *Neyman allocation*. Neyman independently built on Tchouproff [1923] in a seminal paper on survey sampling, arguing for random sampling over purposive sampling, which was common at the time [Neyman, 1934, Fienberg and Tanur, 1996]. Neyman [1934] addressed both estimation and inference, including confidence intervals for means of stratified populations. Even though Neyman elsewhere codified strict level control as a basic goal in statistical inference [Lehmann, 1993], Neyman [1934] suggests that asymptotic normal-theory confidence intervals are sufficient for applied problems when the sample is not smaller than 15. Foundational textbooks on survey sampling—including Hansen et al. [1953], Kish [1965], and Cochran [1977]—have echoed this idea, promulgating asymptotic tests and confidence intervals for various designs. Those methods can have arbitrarily poor coverage depending on the true shape of the population distribution (Figure 7.1 and Lehmann and Romano [2005]).

On the other hand, finite-sample nonparametric inference procedures for fixed-size, unstratified samples from bounded or nonnegative populations have been around for some time. Early versions of such procedures include Hoeffding's bound [Hoeffding, 1963] for bounded distributions and Anderson's bound [Anderson, 1967] for nonnegative distributions. These methods found particular use in auditing, where heavily-skewed population distributions clearly invalidate normal-theory inference, but can be very conservative [Bickel, 1993, Fienberg et al., 1977]. Romano and Wolf [2000] proposed a confidence interval that is conservative and has asymptotically optimal width. A simpler, "empirical Bernstein" bound was derived in the computer science literature by Maurer and Pontil [2009].

Recently, online applications have rekindled interest in sequential inference, leveraging pioneering results by Ville [1939], Wald [1945], and Robbins [1952] to significantly advance SFSNP-valid inference for unstratified sampling [Howard et al., 2019, Waudby-Smith and Ramdas, 2023, Orabona and Jun, 2022, Stark, 2023]. Essentially every fully sequential method uses TSMs and applies Ville's inequality [Ville, 1939] to yield an SFSNP-valid P-value, from which confidence sets can be derived through the usual duality. While Wald [1945] proved the validity of the sequential probability ratio test without explicitly relying on TSMs or Ville's inequality, it is in fact a special case; see also Kaplan [1987].

Exponential supermartingales constitute a large class of TSMs that includes sequential analogues of the Hoeffding and empirical Bernstein bounds [Howard et al., 2021]. Foundational work on E-values (nonnegative random variables with expected value 1 under the null) by Shafer and Vovk [2019] and recent contributions by Waudby-Smith and Ramdas [2023], Cho et al. [2024], and Orabona and Jun [2022] applying TSMs to construct confidence sequences have simplified, unified, and sharpened SFSNP-valid inference. That body of work has encouraged interpreting E-values and TSMs in the intuitive framework of betting. In particular, betting TSMs correspond to the fortune of a gambler in a sequence of bets. The gambler starts with a bankroll of 1 unit of currency and is allowed to wager a fraction of their current fortune on the outcome of the next round, at odds that are fair or unfavorable under the null, but is not allowed to go into debt. Tests based on betting TSMs can take advantage of the outcomes of previous rounds to adapt bets, increasing efficiency by increasing future wagers on games where the bettor has made money. [3] This paper extends betting TSMs to stratified inference.

Despite their importance, conservative methods for inference from stratified samples have received fairly little attention. A basic method, proposed in Wright [1991] and applicable to small populations with binary elements, sums multiplicity-corrected confidence bounds constructed separately within each stratum. Wendell

[3] See Ek et al. [2023] for an example in election audits.

and Schmee [1996b] improved on Wright's method for small binary populations by testing the composite hypothesis using as the P-value the maximum P-value over the simple hypotheses whose union comprises the composite hypothesis. That method inspired the "stratified union-of-intersections tests of elections" (SUITE) approach of Ottoboni et al. [2018], further generalized by Stark [2020], as well as the EESI method[4] for binary populations with an arbitrary number of strata. Both SUITE and EESI use Fisher combining to pool evidence across strata. Following on that work, Stark [2023] proposed using union-of-intersections tests with product combining of TSMs for SFSNP-valid inference in stratified risk-limiting audits. Vovk and Wang [2021] prove that product combining dominates other methods of combining independent E-values when none of the E-values is less than 1. Spertus and Stark [2022] investigated sample sizes for a range of combining functions, TSMs, and allocation strategies for stratified comparison audits. Cho et al. [2024] use betting TSMs for best arm identification and other hypotheses of interest in multi-armed bandits. Their method is designed to test composite hypotheses about means of multiple data streams under adaptive sampling and could be applied to stratified sampling. We suspect their use of average combining and a particular betting rule may sacrifice statistical efficiency compared to the methods for sequential stratified inference presented here.

Currently, no general theory characterizes desirable properties for stratified tests nor constructs optimal tests. We provide that theory below. Consistent intersection and union-of-intersection tests are those that eventually reject a false null as the sample size increases. We give some examples, noting that many tests are consistent in this sense. In the fixed-size setting, power and relative asymptotic efficiency [Lehmann and Romano, 2005] are common desiderata. Sequential testing has analogous notions, typically defined in terms of minimizing the expected stopping time [Wald, 1945] or the expected P-value at a given time [Kelly Jr., 1956]. These criteria lead to the same strategy: maximize the expected log growth rate of the underlying TSM [Breiman, 1961]. Tests that maximize expected log growth rate have been called *Kelly-optimal* under simple alternatives [Waudby-Smith and Ramdas, 2023] or (more generally) *growth-rate optimal* (GRO) under composite alternatives [Grünwald et al., 2023]. Chapter V above used the theory of Kelly-optimality to derive efficient bets for risk-limiting comparison audits. We generalize these definitions to the sequential stratified setting, in which the Kelly-optimal test involves two coupled multi-dimensional tuning parameters: an optimal betting strategy and an optimal stratum allocation rule. We also note that a more general framing of efficiency in terms of minimal regret and optimal portfolios is possible, and may yield sharper sequential tests in some cases [Cover and Thomas, 2006, Orabona and Jun, 2022].

[4] See `https://github.com/pbstark/Strat`.

Union-of-intersections tests are useful for stratified inference and more broadly to make inferences in the presence of nuisance parameters. In stratified inference, the population mean μ is a scalar parameter of interest, a linear function of the full vector of stratum-wise means $[\mu_1, \ldots, \mu_K]$; the problem thus involves a $(K-1)$-dimensional nuisance parameter. Real-world statistical problems often involve such projections of the parameter space. For instance, Student's t-test was originally devised to conduct inference about a Gaussian mean, treating the standard deviation as a nuisance parameter [Student, 1908]. The t-test uses a pivotal statistic—one whose distribution does not depend on the nuisance parameter—to finesse the issue, but pivotal statistics do not exist for most problems. More generally, nuisance parameters can be dealt with by maximizing a P-value over the possible values of the nuisance parameter or over a confidence set for the nuisance parameter [Tsui and Weerahandi, 1989, Berger and Boos, 1994, Silvapulle, 1996]. In frequentist inference, maximizing the likelihood over nuisance parameters leads to a profile likelihood for the parameter of interest. In Bayesian inference, the marginal likelihood, which integrates over nuisance parameters, plays an analogous role. In some cases, such as the Cox proportional hazards model, partial likelihoods obviate the need to estimate a high dimensional nuisance parameter (e.g., the baseline hazard) [Cox, 1975]. Union-of-intersections tests leverage the simplest of these ideas, maximizing a P-value over the possible values of the nuisance parameter. Our contribution can be viewed as developing a method for rigorous inference in the presence of a multi-dimensional nuisance parameter in a nonparametric problem.

7.3 Preliminaries and notation

7.3.1 Population and parameters

We use calligraphic font for sets and bags, and bold font for vectors (and sometimes for tuples). Tuples are denoted using parentheses, e.g., $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$; finite-dimensional vectors are denoted using square brackets, e.g., $[x_1, \ldots, x_n]$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two vectors with the same dimension K, we write $\boldsymbol{x} \le \boldsymbol{y}$ iff $x_k - y_k \le 0$, $k = 1, \ldots, K$; and we define the dot product $\boldsymbol{x} \cdot \boldsymbol{y} := \sum_k x_k y_k$. The vector of all zeroes is $\mathbf{0}$ and the vector of all 1s is $\mathbf{1}$, with dimension implicit from context. The set of nonnegative integers including 0 is $\mathbb{N}$. If $\mathcal{I}$ is a set or a bag, then $|\mathcal{I}|$ is its cardinality. For two scalars a and b, we denote their minimum as $a \wedge b$ and their maximum as $a \vee b$.

The population of interest is a bag of real numbers $\mathcal{X} := \wr x_i \int_{i \in \mathcal{I}}$. The development below takes $\mathcal{X}$ to be a finite population, with $\mathcal{I} := \{1, \ldots, N\}$, but the results apply with few changes when $\mathcal{I}$ is countable or uncountably infinite. As noted above, only one-sided bounds on the population are needed to get one-sided confidence bounds or one-sided tests; however, for simplicity of notation, we assume that each element of the population is in $[0,1]$.[5] The population mean is $\mu = \mu(\mathcal{X}) := N^{-1} \sum_{i \in \mathcal{I}} x_i$, and we would like to test the *global null* hypothesis:

$$H_0 : \mu \leq \eta_0, \tag{7.1}$$

for *global null mean* η_0. A lower confidence bound can be obtained by inverting tests of H_0 as η_0 varies. If there are upper bounds for each element of the population, an upper one-sided test can be obtained by subtracting each element from its upper bound and then using a lower one-sided test, *mutatis mutandis*.

Let $\boldsymbol{N} := [N_1, \ldots, N_K]$ denote the vector of stratum sizes. The symbol $\mathcal{X}_{\boldsymbol{N}} := (\mathcal{X}_k)_{k=1}^K$ denotes a stratified population, a tuple of K bags with N_k items in the kth bag, $\mathcal{X}_k$, so that $N = \sum_k N_k$. The symbol $\aleph_{\boldsymbol{N}}$ represents all K-tuples of bags of numbers in $[0,1]$ such that the kth bag, $\mathcal{X}_k$, has N_k items; that is, $\aleph_{\boldsymbol{N}}$ denotes all stratified $[0,1]$-valued populations with the requisite number of items in each stratum. We use x_{ki} to denote a generic element of the kth stratum, e.g., $\mathcal{X}_k = \Lbag x_{ki} \Rbag_{i=1}^{N_k}$. The vector of *stratum-wise means* is $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}) := [\mu(\mathcal{X}_1), \ldots, \mu(\mathcal{X}_K)].$, where $\mu(\mathcal{X}_k) = \sum_i x_{ki}/N_k$. The vector of *stratum weights* is $\boldsymbol{w} := [w_1, \ldots, w_k]$, where $w_k := N_k/N$. The mean of a stratified population $\mathcal{X}_{\boldsymbol{N}}$ is

$$\mu = \mu(\mathcal{X}_{\boldsymbol{N}}) := \mu(\cup_k \mathcal{X}_k) = \boldsymbol{w} \cdot \boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}).$$

Let $\aleph_{\boldsymbol{N}}^0 := \{\mathcal{Y} \in \aleph_{\boldsymbol{N}} : \mu(\mathcal{Y}) \leq \eta_0\}$ denote the set of null populations; the global null hypothesis can be written $H_0 : \mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$. Also let $\aleph_{\boldsymbol{N}}^1 := \{\mathcal{Y} \in \aleph_{\boldsymbol{N}} : \mu(\mathcal{Y}) > \eta_0\}$ denote the set of alternative populations. Together, $\aleph_{\boldsymbol{N}}^0$ and $\aleph_{\boldsymbol{N}}^1$ partition $\aleph_{\boldsymbol{N}}$. An *intersection null hypothesis* is the assertion

$$\boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}) \leq \boldsymbol{\eta}$$

for the *intersection null mean* $\boldsymbol{\eta} \in [0,1]^K$. In words, the intersection null hypothesis posits that *each* stratum-wise mean μ_k is below a corresponding *stratum-wise null mean* η_k. The global null hypothesis can be written as a union of intersection null hypotheses:

$$H_0 : \bigcup_{\boldsymbol{\eta} \in \mathcal{E}_0} \{\boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}) \leq \boldsymbol{\eta}\} \tag{7.2}$$

[5] Any known lower bound on elements can be accommodated by shifting: if $x_i \geq a_i$ then $y_i := (x_i - a_i) \geq 0$. If all the elements have the same upper and lower bounds a and b, they can be rescaled to $[0,1]$ with an affine transformation $y_i := (x_i - a)/(b - a) \in [0,1]$; the mean of the resulting population $\Lbag y_i \Rbag$ is the same affine transformation applied to the original mean.

where

$$\mathcal{E}_0 := \{\boldsymbol{\zeta} : \boldsymbol{w} \cdot \boldsymbol{\zeta} \leq \eta_0, \ \mathbf{0} \leq \boldsymbol{\zeta} \leq \mathbf{1}\} \tag{7.3}$$

is the set of all intersection nulls for which the global null is true[6].

7.3.2 Sampling design

We consider samples drawn uniformly at random within strata, with or without replacement, but generalizations to sampling with probability proportional to a measure of "size" are straightforward [Stark, 2023]. Recall that a fixed-size stratified sample consists of K independent (unordered) samples $(\Lbag X_{ki} \Rbag_{i=1}^{n_k})_{k=1}^{K}$, where the sample from stratum k, $\Lbag X_{ki} \Rbag_{i=1}^{n_k}$, is drawn by uniform random sampling with or without replacement. When draws are with replacement, the data are IID uniform draws from $\mathcal{X}_k$. When draws are without replacement, $\Lbag X_{ki} \Rbag_{i=1}^{n_k}$ is uniform over all subsets of size n_k from $\mathcal{X}_k$.

Sequential stratified samples. Unlike the fixed-size case, sequential stratified samples have an order within and across strata, which necessitates a more detailed specification. Let (X_{ki}) be a sequence of random variables representing sampling sequentially from stratum k. For sampling without replacement, i can run from 1 to N_k and $(X_{ki})_{i=1}^{N_k}$ is a random permutation of the stratum values $\Lbag x_{ki} \Rbag_{k=1}^{N_k}$. For sampling with replacement, i runs from 1 to ∞ and the elements of $(X_{ki})_{i=1}^{\infty}$ are IID. Regardless, the variables $\{X_{ki}\}$ are exchangeable for each k and the sequences (X_{ki}) and (X_{ji}) are independent of each other for $k \neq j$.

The hypothesis tests we consider are constrained to use samples from each stratum in the order in which those samples are drawn. But in general, tests of different intersection nulls may interleave the samples across strata differently.

An *interleaving* of samples across strata is a stochastic process (Y_t) indexed by "time" t; $Y^t := (Y_i)_{i=1}^{t}$ is the t-prefix of $(Y_i)_{i \in \mathbb{N}}$. Each interleaving is characterized by a stochastic process, the *stratum selection* S_t: the item in the tth position in the interleaving comes from stratum S_t. Let $S^t := (S_i)_{i=1}^{t}$. The variable S_t is *predictable*—meaning that it can depend on past data Y^{t-1} but not on Y_i for $i \geq t$—and may also involve auxiliary randomness. We emphasize that it does not depend on Y_t, and its value is observed before Y_t is observed: it specifies the stratum from which the tth sample will be drawn. For $k \in \{1, \ldots, K\}$, let

$$p_{kt} := \mathbb{P}\left(S_t = k \ \mid \ Y^{t-1}, S^{t-1}\right), \tag{7.4}$$

[6] If there are constraints on the support of each stratum (e.g., that each element is 0 or 1), that information can be used to sharpen inferences. See Section 7.4.2 below.

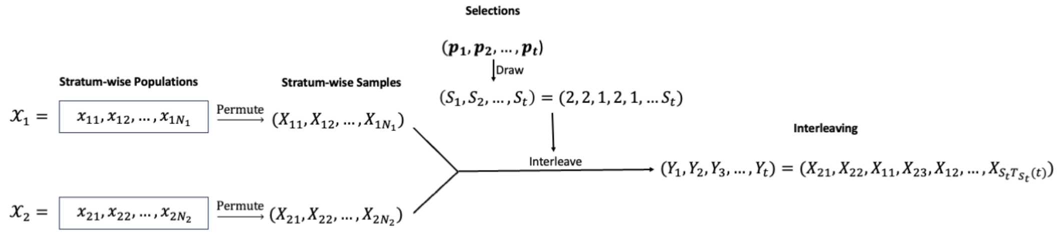

Selections
$(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_t)$
Draw
$(S_1, S_2, \ldots, S_t) = (2, 2, 1, 2, 1, \ldots S_t)$
Stratum-wise Populations
Stratum-wise Samples
$\mathcal{X}_1 =$
$x_{11}, x_{12}, \ldots, x_{1N_1}$
Permute
$(X_{11}, X_{12}, \ldots, X_{1N_1})$
Interleaving
Interleave
$(Y_1, Y_2, Y_3, \ldots, Y_t) = (X_{21}, X_{22}, X_{11}, X_{23}, X_{12}, \ldots, X_{S_t T_{S_t}(t)})$
$\mathcal{X}_2 =$
$x_{21}, x_{22}, \ldots, x_{2N_2}$
Permute
$(X_{21}, X_{22}, \ldots, X_{2N_2})$

Figure 7.2: Depiction of sequential stratified sampling without replacement. The population depicted consists of $K = 2$ strata, each a bag of N_k numbers in $[0, 1]$. Each stratumwise sample is a random permutation of its corresponding bag. The stratum selectors $(\boldsymbol{p}_t)$ then determine the selections (S_t), which interleave the stratumwise samples into a single sequence of data (Y_t).

and define $\boldsymbol{p}_t := [p_{1t}, \ldots, p_{Kt}]$, $t \in \mathbb{N}$. Naturally, for each t, $\mathbf{0} \leq \boldsymbol{p}_t \leq \mathbf{1}$ and $\mathbf{1} \cdot \boldsymbol{p}_t = 1$. If for each t, $\boldsymbol{p}_t$ has one component equal to 1 and the rest equal to zero (i.e., $\boldsymbol{p}_t$ is *one-hot*), the stratum selection is deterministic (conditional on the past). We refer to $\boldsymbol{p}_t$ as the *stratum selector*. To summarize, the stratum *selection* (S_t) is a stochastic process taking values in $\{1, \ldots, K\}$ while the stratum *selector* $(\boldsymbol{p}_t)$ is a vector-valued stochastic process taking values in $\mathbb{R}^K$, specifying the chance that the next draw will be from each of the strata, given the sampling history so far.

As of time t, the number of items in the interleaving Y^t that came from stratum k is $T_k(t)$, and $X_k^{T_k(t)} := (X_{ki})_{i=1}^{T_k(t)}$ are the data from stratum k. Thus, the tth item in the interleaving,

$$Y_t = X_{S_t T_{S_t}(t)},$$

is the $T_{S_t}(t)$th item drawn from stratum S_t, so $Y^t = (X_{S_i T_{S_i}(i)})_{i=1}^t$.

7.3.3 SFSNP-valid hypothesis tests

Our goal is to construct an SFSNP-valid test of the global null $H_0 : \mu(\mathcal{X}_{\boldsymbol{N}}) \leq \eta_0$, i.e., $\mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$.

Definition 1 (SFSNP-valid P-value)**.** *Let $(P_t)_{t \in \mathbb{N}}$ be a $[0, 1]$-valued stochastic process. Then (P_t) is an SFSNP-valid P-value for the global null hypothesis if for all $q \in [0, 1]$,*

$$\mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}(\exists\ t \in \mathbb{N} : P_t \leq q) \leq q \quad \textit{when } \mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0.$$

If (P_t) is an SFSNP-valid P-value, then rejecting the null hypothesis if $P_t \leq \alpha$ for some t is a SFSNP-valid hypothesis test with significance level α. The random variable $\tau := \inf\{t \in \mathbb{N} : P_t \leq \alpha\}$ is the *stopping time* of the test.

Finding SFSNP-valid tests such that τ is small in expectation when the global null is false is the primary goal of this paper. Minimizing stopping times for sequential tests is analogous to maximizing power in the usual fixed-sample, Neyman-Pearson theory of testing. We return to this idea in Section 7.5, where our primary concern is efficient SFSNP-valid tests of H_0.

7.3.4 Test supermartingales

In recent years, *martingales*—fundamental objects in probability theory with close connections to betting—have been used to construct efficient sequential tests and confidence intervals. One can construct an SFSNP-valid P-value from any statistic that is a nonnegative supermartingale starting at 1 when H_0 is true; such statistics are called *test supermartingales* [Waudby-Smith and Ramdas, 2023]:

Definition 2 (Test supermartingale (TSM))**.** *The stochastic process* $(M_t)_{t \in \mathbb{N}}$ *is a* test supermartingale (TSM) *for* H_0 *if, when* H_0 *is true,* $(M_t)_{t \in \mathbb{N}}$ *satisfies*

1. $\mathbb{E}[M_t | M^{t-1}] \leq M_{t-1}$
2. $\mathbb{P}\{M_t \geq 0\} = 1$
3. $M_0 = 1$.

TSMs are linked to SFSNP-valid P-values by the following special case of a result of Ville [1939].

Proposition 1 (Ville, 1939)**.** *Let* M_t *be a TSM for* H_0*. Then if* H_0 *is true, for all* $q \in [0, 1]$

$$\mathbb{P}\{\exists\ t \in \mathbb{N} : M_t \geq 1/q\} \leq q.$$

Ville's inequality is analogous to Markov's inequality, but holds uniformly over $t \in \mathbb{N}$, allowing sequentially valid inference. The truncated reciprocal of a TSM, $1 \wedge (1/M_t) \in [0, 1]$, is an SFSNP-valid P-value. Vovk and Wang [2021] show that $1 \wedge (1/M_t)$ is essentially the only admissible mapping from TSMs to P-values.

Constructing TSMs from random samples

We start by considering a single stratum, generically stratum k. The *within-stratum null mean* is η_k. We construct a process $(M_{kt})_{t \in \mathbb{N}}$ that is a TSM with respect to $(X_{kt})_{t \in \mathbb{N}}$ for the *stratum null* $\mu(\mathcal{X}_k) \leq \eta_k$. Recall that $T_k(t-1)$ is the sample size from stratum k at time $t-1$. The *conditional* stratum-wise null mean η_{kt} is the implied average of the values remaining in $\mathcal{X}_k$ at time t if $\mu_k = \eta_k$, given that we have already observed $(X_{ki})_{i=1}^{T_k(t-1)}$. There are two cases we consider:

- **Sampling with replacement**, in which case $\eta_{kt} := \eta_k$.
- **Sampling without replacement**, in which case
$$\eta_{kt} := \frac{\eta_k - \sum_{i=1}^{T_k(t-1)} X_{ki}}{N_k - T_k(t-1)}.$$

A generic form for a *within-stratum TSM* for the null $\mu_k \leq \eta_k$ is

$$M_{kt_k}(\eta_k) := \prod_{i=1}^{t_k} Z_{ki}(\eta_k),$$

where $Z_{ki}(\eta_k)$ is any term such that $\mathbb{E}[Z_{ki}(\eta_k) \mid X_k^{i-1}] \leq 1$ when $\mu_k \leq \eta_k$. We will use statistics of the form

$$Z_{ki}(\eta_k) := 1 + \lambda_{ki}(X_{ki} - \eta_{ki}),$$

which defines $M_{kt_k}(\eta_k)$ as a *betting TSM*. Waudby-Smith and Ramdas [2023], Orabona and Jun [2022] show that for suitable choices of λ_{ki}, betting martingales provide sharper inferences than many other options for $Z_{ki}(\eta_{ki})$, such as exponential supermartingales [Howard et al., 2021]. The TSM within stratum k at time t can be written:

$$M_{kt}(\eta_k) := \prod_{i=1}^{T_k(t)} [1 + \lambda_{ki}(X_{ki} - \eta_{ki})].$$

We next discuss how to combine the K within-stratum TSMs to form a single P-value for an intersection null.

Intersection test supermartingales (I-TSMs)

Consider a particular intersection null $\boldsymbol{\eta}$. Define the *intersection TSM* (I-TSM):

$$M_t(\boldsymbol{\eta}) := \prod_{k=1}^{K} M_{kT_k(t)}(\eta_k) = \prod_{k=1}^{K} \prod_{i=1}^{T_k(t)} Z_{ki}(\eta_k) = \prod_{i=1}^{t} \tilde{Z}_i,$$

where $\tilde{Z}_i := Z_{S_i T_{S_i}(i)}(\eta_{S_i})$ is the $T_{S_i}(i)$th term of the within-stratum TSM for stratum S_i. By commutativity, the I-TSM at time t can be written as an interleaving of terms $\tilde{Z}_i$ from the within-stratum TSMs, where the (possibly random, but predictable) interleaving is defined by the selections (S_t). Because the selections are predictable, the I-TSM is indeed a TSM under the intersection null $\boldsymbol{\eta}$:

$$\begin{aligned}
\mathbb{E}[M_t(\boldsymbol{\eta}) \mid Y^{t-1}] &= \mathbb{E}\left[\prod_{k=1}^{K}\prod_{i=1}^{T_k(t)} Z_{ki}(\eta_k) \mid Y^{t-1}, S^t\right] \\
&= M_{t-1}(\boldsymbol{\eta})\mathbb{E}\left[\tilde{Z}_t \mid Y^{t-1}, S^t\right] \\
&= M_{t-1}(\boldsymbol{\eta})\mathbb{E}\left[\tilde{Z}_t \mid X_{S_t}^{T_{S_t}(t-1)}\right] \\
&\leq M_{t-1}(\boldsymbol{\eta}).
\end{aligned}$$

The 3rd equality holds because samples from different strata are independent, and the inequality in the 4th line holds by construction of the within-stratum TSMs paired with the fact that $\mu_k \leq \eta_k$ under the intersection null.

The truncated reciprocal of M_t, $P_t(\boldsymbol{\eta}) := 1 \wedge 1/M_t(\boldsymbol{\eta})$ is a sequentially valid P-value for the intersection null $\boldsymbol{\eta}$. Employing betting within-stratum TSMs leads to the *betting I-TSM*:

$$M_t(\boldsymbol{\eta}) := \prod_{k=1}^{K}\prod_{i=1}^{T_k(t)} [1 + \lambda_{ki}(X_{ki} - \eta_{ki})].$$

7.4 Stratified inference with TSMs

In this section, we describe two methods for stratified inference that leverage I-TSMs. The first method is a weighted sum of TSM-based, sequentially valid lower confidence bounds for the K stratum-wise means. This is akin to Wright's method for conservative stratified inference on binary populations [Wright, 1991]. We show that (perhaps surprisingly) using TSMs avoids the need to explicitly adjust for multiplicity. The second method forms an I-TSM for every intersection null $\boldsymbol{\eta} \in \mathcal{E}_0$ and takes the smallest. The second method is trickier to implement, but has less slack than the first method.

7.4.1 Simple stratified inference: combining confidence sequences

We use a linear combination of independent lower confidence sequences (LCBs) for the K stratum-wise means $\{\mu_k\}$ to form a global LCB for μ. The weights in the linear combination are the stratum weights $\boldsymbol{w}$. If the linear combination is above the global null mean, the test rejects. For a level α global test, the individual LCBs must be *simultaneously* valid at level α, controlling the family-wise error rate. If each LCB were generic, we might use Šidák's correction for independent tests, constructing each LCB at level $(1-\alpha)^{1/K}$. This incurs a steep penalty as K increases, nearly equaling the α/K penalty of a Bonferonni correction.

However, we are basing each LCB on a TSM. The following lemma, which is proved in Appendix C.2.1, shows that by using a TSM in each stratum we can avoid the multiplicity penalty.

Lemma 1 (Validity of TSM confidence sequences)**.** *For a within-stratum TSM $M_{kt}(\eta_k)$, define the lower $(1-\alpha)$ lower confidence sequence (LCB) in stratum k by (L_{kt}), where*

$$L_{kt} := \sup_{\eta_k \in [0,1]} \{\eta_k : M_{kt}(\eta_k) > 1/\alpha\}.$$

Then:

1. *$\{(L_{kt})\}_{k=1}^K$ are separately valid, in the sense that when $\mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$:*

$$\sup_{k \in \{1,\dots,K\}} \mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}(\exists\ t \in \mathbb{N} : L_{kt} > \mu_k) \leq \alpha.$$

2. *$\{(L_{kt})\}_{k=1}^K$ are simultaneously valid, in the sense that when $\mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$:*

$$\mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}\left(\exists\ (t_1,\dots,t_K) \in \mathbb{N}^K : \bigcup_{k=1}^K \{L_{kt_k} > \mu_k\}\right) \leq \alpha.$$

The first result simply states that we can construct confidence sequences from TSMs, which follows immediately from the duality of tests and confidence sets [Lehmann and Romano, 2005], as used extensively in [Waudby-Smith and Ramdas, 2023]. The second result is stronger and follows from the closure principle of Marcus et al. [1976], which was employed by Vovk and Wang [2021] to adjust E-values for multiplicity. The surprising result is that LCBs constructed from TSMs are simultaneously valid without the need for an explicit multiplicity correction.

Proposition 2. *Let $L_{kT_k(t)}$ be a $(1-\alpha)$ LCB constructed from a TSM with $T_k(t)$ samples from stratum k at time t, $k \in \{1, \dots, K\}$. Consider the stratum-weighted sum*

$$L_t := \sum_{k=1}^{K} w_k L_{kT_k(t)}.$$

Since $\{L_{kt}\}_{k=1}^{K}$ are simultaneously and sequentially valid LCBs for $\{\mu_k\}_{k=1}^{K}$, whenever $\mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$,

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}(\exists\, t \in \mathbb{N} : L_t > \mu) &= \mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}\left(\exists\, t \in \mathbb{N} : \sum_{k=1}^{K} w_k L_{kT_k(t)} > \sum_{k=1}^{K} w_k \mu_k\right) \\
&\leq \mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}\left(\exists\, t \in \mathbb{N} : \bigcup_{k=1}^{K} L_{kT_k(t)} > \mu_k\right) \\
&\leq \alpha.
\end{aligned}$$

Thus L_t is an SFSNP-valid $(1-\alpha)$ LCB for μ, and a test that rejects H_0 when $L_t > \eta_0$ is an SFSNP-valid level α test with stopping time $\tau(L_t) := \min\{t \in \mathbb{N} : L_t > \mu_0\}$.

The proof is contained in the theorem statement; the final step applies Lemma 1. This approach is easy to implement when computing the LCBs $\{L_{kT_k(t)}\}_{k=1}^{K}$ is straightforward and efficient; Waudby-Smith and Ramdas [2023] and Orabona and Jun [2022] describe a number of possibilities. The selection rule $\boldsymbol{p}_t$ will influence the efficiency of the bound.

There are two sources of slack in using Proposition 2 to test H_0. One is due to the second inequality in Proposition 2: the LCB method bounds each of the K components of $\boldsymbol{\mu}$ separately, but we only need to bound $\boldsymbol{w}^T\boldsymbol{\mu}$. The other source of slack is in controlling for multiplicity using closed testing, which requires each of the K TSMs to reach $1/\alpha$. In contrast, to test the intersection null $\boldsymbol{\eta}$, we only need the *product* of the stratum-wise TSMs to reach $1/\alpha$. In particular, the test could reject $\boldsymbol{\eta}$ with all K TSMs equal to $\alpha^{-1/K}$, a lower hurdle in every stratum. We now develop a test of H_0 that avoids these two sources of slack.

7.4.2 Union-of-intersections test statistic (UI-TS)

Definition 3 (Union-of-Intersections Test Statistic)**.** *A stochastic process $(U_t)_{t\in\mathbb{N}}$ is a* Union-of-Intersections Test Statistic *(UI-TS) for the composite null hypothesis $H_0 : \mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$ if for all $\mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$,*

$$\mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}\big(\exists\, t \in \mathbb{N} : U_t \geq 1/\alpha\big) \leq \alpha.$$

Equivalently, (U_t) *is a UI-TS if its thresholded-reciprocal* $P_t := 1 \wedge 1/U_t$ *is an anytime-valid* P*-value for the null* H_0.

A UI-TS need not be a supermartingale under H_0, but it is an E-process [Shafer, 2021, Vovk and Wang, 2021, Ramdas et al., 2023, Grünwald et al., 2023].

Recall the union-of-intersections form of the global null in Equation (7.2) and the definition of $\mathcal{E}_0$ in Equation (7.3). As shown in Section 7.3.4, we can test a particular $\boldsymbol{\eta} \in \mathcal{E}_0$ using an I-TSM. We can reject H_0 if the P-value for every $\boldsymbol{\eta} \in \mathcal{E}_0$ is less than α, i.e., if the *smallest* I-TSM evaluated over $\mathcal{E}_0$ is at least $1/\alpha$. Therefore,

$$M_t := \min_{\boldsymbol{\eta} \in \mathcal{E}_0} M_t(\boldsymbol{\eta}) \tag{7.5}$$

is a UI-TS for H_0. When every $M_t(\boldsymbol{\eta})$ is a betting I-TSM, M_t is a *betting UI-TS.*

The boundary of $\mathcal{E}_0$

In this paper, we consider only betting UI-TSs based on TSMs $M_{kt}(\eta_k)$ that are monotone decreasing in η_k, so that the I-TSM $M_t(\boldsymbol{\eta})$ is componentwise monotone in $\boldsymbol{\eta}$. As a result, the minimum of $M_t(\boldsymbol{\eta})$ occurs on the boundary of $\mathcal{E}_0$. The boundary depends on the support of each stratum population, which in general will be unknown.

Specifically, let Ω_k be the set of all possible means μ_k in stratum k. For example, if stratum k is binary, then $\Omega_k = \{0, 1/N_k, \dots, (N_k - 1)/N_k, N_k\}$. Let $\Omega = \prod_{k=1}^{K} \Omega_k$ be the Cartesian product of all possible stratumwise means $\boldsymbol{\mu}$. The *boundary* of $\mathcal{E}_0$ is

$$\mathcal{B} := \{\boldsymbol{\eta} \in \Omega : \boldsymbol{w} \cdot \boldsymbol{\eta} \leq \eta_0 \text{ and } \Omega \ni \boldsymbol{\zeta} > \boldsymbol{\eta} \implies \boldsymbol{w} \cdot \boldsymbol{\zeta} > \eta_0\}.$$

For the TSMs we consider,[7] the value of $\boldsymbol{\eta}$ that minimizes $M_t(\boldsymbol{\eta})$ is in $\mathcal{B}$. Futhermore, define

$$\mathcal{C} := \{\boldsymbol{\eta} : \boldsymbol{w} \cdot \boldsymbol{\eta} = \eta_0,\ \mathbf{0} \leq \boldsymbol{\eta} \leq \mathbf{1}\} \subset \mathcal{E}_0. \tag{7.6}$$

Because of the componentwise monotonicity, optimizing over the set $\mathcal{C}$ rather than $\mathcal{B}$ gives a conservative result. In what follows, we will generally define $M_t = \min_{\boldsymbol{\eta} \in \mathcal{C}} M_t(\boldsymbol{\eta})$. The set $\mathcal{C}$ is a polytope, the intersection of the K-cube with the hyperplane $\boldsymbol{w} \cdot \boldsymbol{\eta} = \eta_0$. The geometry of $\mathcal{C}$ is important for the computational tractability of some UI-TSs (see Appendix C.3).

[7] For the TSMs we consider, if $\boldsymbol{\eta} \in \mathcal{E}_0 \setminus \mathcal{B}$, then there is some point $\boldsymbol{\eta}' \in \mathcal{B}$ with $\boldsymbol{\eta} < \boldsymbol{\eta}'$ for which we enforce that $M_t(\boldsymbol{\eta})$ uses the same selections and bets as $M_t(\boldsymbol{\eta}')$. As result, $M_t(\boldsymbol{\eta}') < M_t(\boldsymbol{\eta})$ for all t because of the monotonicity of the stratumwise TSMs in η_k.

Global stopping time and global sample size

Before we discuss efficiency, we draw some distinctions between the *stopping time* of a UI-TS and its *workload* measured by the number of samples it requires to stop. To do so, we embellish the notation to highlight dependence on the intersection null $\boldsymbol{\eta}$. For each $\boldsymbol{\eta} \in \mathcal{C}$, denote the selections by $(S_t(\boldsymbol{\eta}))_{t \in \mathbb{N}}$, the selection rule by $(\boldsymbol{p}_t(\boldsymbol{\eta}))_{t \in \mathbb{N}}$, and the stratum-wise sample sizes by $\{T_k(t, \boldsymbol{\eta})\}_{k=1}^K$. The stopping time of the level α test induced by UI-TS M_t at $\boldsymbol{\eta}$ is

$$\tau(M_t; \boldsymbol{\eta}) := \min\{t \in \mathbb{N} : M_t(\boldsymbol{\eta}) \geq 1/\alpha\}.$$

This is the number of samples needed to reject intersection null $\boldsymbol{\eta}$ using the constituent I-TSM $M_t(\boldsymbol{\eta})$, and the quantity $\tau_k(M_t; \boldsymbol{\eta}) := T_k(\tau(M_t; \boldsymbol{\eta}), \boldsymbol{\eta})$ is the number of samples needed from stratum k.

Now, we define the *global stopping time* as in Section 7.3.3 and note some equivalences:

$$\begin{aligned}
\tau(M_t) :=& \inf\{t \in \mathbb{N} : M_t \geq 1/\alpha\} \\
=& \sup_{\boldsymbol{\eta} \in \mathcal{C}} \tau(M_t; \boldsymbol{\eta}) \\
=& \sup_{\boldsymbol{\eta} \in \mathcal{C}} \sum_{k=1}^K \tau_k(M_t; \boldsymbol{\eta}).
\end{aligned}$$

The global stopping time τ is simply the sample size needed for the "last" I-TSM, considered on its own, to hit or cross $1/\alpha$.

On the other hand, the *global sample size* n_τ is the total number of samples drawn across all strata when H_0 is rejected:

$$n_\tau(M_t) := \sum_{k=1}^K \sup_{\boldsymbol{\eta} \in \mathcal{C}} \tau_k(M_t; \boldsymbol{\eta}).$$

Because n_τ is the sum of stratumwise maxima, $n_\tau(M_t) \geq \tau(M_t)$. However, for a broad class of tests, $n_\tau(M_t) = \tau(M_t)$— for instance, when $S_t(\boldsymbol{\eta}) = S_t$. In what follows, when M_t is clear from context, we will drop it from the notation.

7.5 Desirable properties: consistency and efficiency

This section defines *consistency* and *efficiency* for sequential stratified inference.

7.5.1 Consistency

Loosely speaking, a test is consistent if, with probability 1, it eventually rejects the global null at every fixed level $\alpha \in (0, 1]$ when the global null is false. We first define consistency for intersection nulls, then global consistency. Let $\xrightarrow{\mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}}$ denote convergence in probability, where the randomness arises from sampling from the population $\mathcal{X}_{\boldsymbol{N}}$, including any randomness in the stratum selection.

Definition 4 (Intersection consistency)**.** *Consider a test of the intersection null $\boldsymbol{\eta}$, and let $P_t(\boldsymbol{\eta})$ denote the P-value for that test at time t. The test is* intersection consistent *iff*

$$P_t(\boldsymbol{\eta}) \xrightarrow{\mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}} 0$$

whenever $\boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}) \not\leq \boldsymbol{\eta}$.

An intersection-consistent test has finite stopping time $\tau(\boldsymbol{\eta})$ with probability 1 when $\boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}) \not\leq \boldsymbol{\eta}$, i.e., when at least one element of $\boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}})$ is greater than the corresponding element of $\boldsymbol{\eta}$. An intersection-consistent test can be constructed from an I-TSM that, with probability 1, grows without bound whenever $\boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}) \not\leq \boldsymbol{\eta}$. Constructing such an I-TSM requires protecting against two failure modes. First, we must ensure that in at least one stratum in which $\mu_k > \eta_k$, the stratumwise TSM $M_{kt}(\eta_k)$ grows.[8] Second, we must ensure that the stratumwise TSM for any strata in which $\mu_k \leq \eta_k$ do not drive the I-TSM towards zero. That can be accomplished either by curtailing sampling in strata where there is evidence that $\mu_k \leq \eta_k$ or by reducing the bets in those strata to zero.[9]

Definition 5 (Global consistency)**.** *Consider a test of the global null $\mu \leq \eta_0$, and let P_t denote the P-value for that test at time t. The test is* globally consistent *iff*

$$P_t \xrightarrow{\mathbb{P}_{\mathcal{X}_{\boldsymbol{N}}}} 0$$

whenever $\mu(\mathcal{X}_{\boldsymbol{N}}) > \eta_0$.

[8] Recalling a classic example, suppose that in each stratum the population is binary and we are using the Bernoulli SPRT as the TSM. It is well known that, even when $\mu_k > \eta_k$, if the tuning parameter corresponding to the suspected alternative mean is incorrectly specified, the stratumwise TSM $M_{kt}(\eta_k)$ has a positive probability of never crossing $1/\alpha$. Predictably estimating the true mean or mixing over a prior distribution can remedy this issue (c.f. [Robbins, 1952, Stark, 2023, Waudby-Smith and Ramdas, 2023]).

[9] Instead of combining stratumwise TSMs by multiplication, we could combine them by averaging [Cho et al., 2024]; or instead of combining stratumwise TSMs into an I-TSM, we could combine the (independent) stratumwise P-values derived from the TSMs into a single P-value, for instance, using Fisher's combining function [Ottoboni et al., 2018, Spertus and Stark, 2022].

With probability 1, a globally consistent test has a finite stopping time τ and finite sample size n_τ when the global null is false (i.e., it is a test of power one). For a UI-TS using I-TSMs $M_t(\boldsymbol{\eta})$ that are monotone in $\boldsymbol{\eta}$, we have

$$P_t := \max_{\boldsymbol{\eta} \in \mathcal{E}_0} P_t(\boldsymbol{\eta}) = \max_{\boldsymbol{\eta} \in \mathcal{C}} P_t(\boldsymbol{\eta}).$$

So if every P-value in the set $\{P_t(\boldsymbol{\eta}) : \boldsymbol{\eta} \in \mathcal{C}\}$ is intersection consistent, the test defined by P_t is globally consistent. In theory, we can use any set of tests that are intersection-consistent for all $\boldsymbol{\eta} \in \mathcal{C}$ to construct a globally consistent test this way. In practice, $\mathcal{C}$ is in general uncountably infinite, so maximizing $P_t(\boldsymbol{\eta})$ over $\eta \in \mathcal{C}$ may be intractable unless $P_t(\cdot)$ has special structure. Furthermore, even if a test is consistent, it might require an impractical sample size n_τ.

7.5.2 Efficiency

Our main objective is to test the global null $\mu(\mathcal{X}_{\boldsymbol{N}}) \leq \eta_0$ against the global, composite alternative $\mu(\mathcal{X}_{\boldsymbol{N}}) > \eta_0$. These hypotheses do not completely specify the population, only its mean. In contrast, a *simple alternative* completely specifies a stratified population $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$. Considering simple alternatives allows us to construct theoretically optimal tests that can be approximated in practice.

An *efficient* test keeps n_τ "as small as possible" in some sense. In general, no method is best for all $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$; moreover, n_τ is random. One could consider minimizing a summary of the distribution of sample sizes for different $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$, for instance, the supremum (over $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$) or a weighted average expected sample size. One might also define *admissibility* with respect to some summary of sample size. For example, we might define a test to be inadmissible with respect to expected sample size if there is a second test such that for every $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$ the expected sample size of the second test is not greater than that of the first test, and there is some $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$ for which the expected sample size of the second test is strictly less than that of the first test.

We have not identified an admissible test for any summary of n_τ over all $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$. To construct tests that perform reasonably well—albeit not necessarily optimally—we begin by constructing tests that minimize $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau]$ for a simple alternative $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$. Such tests are called *Kelly optimal*, and we construct Kelly optimal rules for both I-TSMs and UI-TSs. A Kelly-optimal test maximizes the expected log growth of the corresponding test statistic; by Wald's identity [Wald, 1944], this minimizes the expected stopping time $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau]$. Since $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau] \leq \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[n_\tau]$ for any test, the expected stopping time of a Kelly optimal test is a lower bound on the optimal expected sample size. The growth of a particular I-TSM under a simple alternative depends on the selection strategy $(\boldsymbol{p}_t(\boldsymbol{\eta}))$ in addition to the betting rule $(\boldsymbol{\lambda}_t(\boldsymbol{\eta}))$.

Then, we characterize (a) Kelly-optimal strategies for known $\mathcal{X}_{\boldsymbol{N}}$; (b) approximately Kelly-optimal strategies that do not assume $\mathcal{X}_{\boldsymbol{N}}$ is known, instead "learning" how to bet from the sample; and (c) computable approximations to (b). In practice, $\mathcal{X}_{\boldsymbol{N}}$ is not known and neither the composite alternative $\mu(\mathcal{X}_{\boldsymbol{N}}) > \eta_0$ nor the intersection alternative $\boldsymbol{\mu}(\mathcal{X}_{\boldsymbol{N}}) \not\leq \boldsymbol{\eta}$ completely specifies $\mathcal{X}_{\boldsymbol{N}}$. While we can characterize (b) implicitly, constructing examples is challenging absent constraints on the betting and allocation strategies. Instead, Kelly optimality provides a benchmark that we hope to approximate with the predictable, computationally tractable strategies of (c). The computations are dramatically simpler when selections are restricted to be fixed across $\boldsymbol{\eta} \in \mathcal{E}_0$, i.e., when $S_t(\boldsymbol{\eta}) = S_t$. For non-adaptive S_t, $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau] = \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[n_\tau]$, so minimizing the expected stopping time minimizes the expected sample size.

Definitions of efficiency

The simplest definition of efficiency is efficiency at a point: it is the ratio of the expected sample size of the test that minimizes the expected global sample size when $\mathcal{X}_{\boldsymbol{N}} = \mathcal{X}_{\boldsymbol{N}}^1$ to the expected sample size of the test in question when $\mathcal{X}_{\boldsymbol{N}} = \mathcal{X}_{\boldsymbol{N}}^1$.

Definition 6 (Relative efficiency at $\mathcal{X}_{\boldsymbol{N}}^1$)**.** *Consider testing the composite null $H_0 : \mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$ against the simple alternative $\mathcal{X}_{\boldsymbol{N}} = \mathcal{X}_{\boldsymbol{N}}^1 \in \aleph_{\boldsymbol{N}}^1$. Given two UI-TSs for H_0, M_t and M_t', the relative efficiency of M_t to M_t' at $\mathcal{X}_{\boldsymbol{N}}^1$ (for expected sample size) is*

$$0 \leq \frac{\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}^1}[n_\tau(M_t')]}{\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}^1}[n_\tau(M_t)]}.$$

The efficiency of a test M_t at $\mathcal{X}_{\boldsymbol{N}}$ is its relative efficiency to any level-α test M_t^* that minimizes the expected sample size.

Another possible metric of test performance is *regret*, which measures the shortfall from the best method as the gap in expected sample size at the simple alternative.

Definition 7 (Regret at $\mathcal{X}_{\boldsymbol{N}}^1$)**.** *Consider testing the composite null $H_0 : \mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$ against the simple alternative $\mathcal{X}_{\boldsymbol{N}} = \mathcal{X}_{\boldsymbol{N}}^1 \in \aleph_{\boldsymbol{N}}^1$. Let M_t be any UI-TS for H_0. Its regret is:*

$$\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}^1}[n_\tau(M_t^*)] - \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}^1}[n_\tau(M_t)],$$

where M_t^* is a test that minimizes the expected sample size at $\mathcal{X}_{\boldsymbol{N}}^1$ over all tests.

Minimizing regret and maximizing efficiency are equivalent. Efficiency is a common metric in statistics [Wasserman, 2004, Lehmann, 1999, Wald, 1945], while regret is more common in computer science and especially in the multi-armed bandit literature [Lai and Robbins, 1985, Berry and Fristedt, 1985, Auer et al., 2002]. However, regret has roots in statistics. It was first proposed by Savage [1951] as a less pessimistic loss measure than the unnormalized quantity $-\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}^1}[n_\tau(M_t)]$, which was at the forefront of Wald's pioneering work in decision theory [Wald, 1950]. The distinction is equivalent to that between REGROW and GROW in the context of E-values [Grünwald et al., 2023]. We will primarily discuss performance in terms of efficiency.

Sometimes efficiency can be computed or bounded theoretically. For instance, Wald [1945] proves that the SPRT essentially minimizes the expected sample size for testing a point null against a point alternative (for any pre-specified sequential sampling design, e.g., for the stratified case, the stratum selection cannot be adaptive). In other words the SPRT is efficient. He derives a general formula for $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[n_\tau] = \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau]$ for the SPRT and provides explicit formulas for IID Bernoulli and normal data. That theory does not apply to stratified sampling with adaptive stratum selection. In particular, finding an analytical formula for the expected sample size of a UI-TS is intractable even for basic parametric alternatives (e.g. binomials) and simple betting and allocation rules (e.g. round-robin allocation and fixed bets $\boldsymbol{\lambda}_t(\boldsymbol{\eta}) := \boldsymbol{\lambda}$). Thus, we primarily assess the efficiency of various methods through simulation.

Before analyzing efficiency and Kelly optimality for UI-TSs, we note that we can generalize the above definition in two ways. First, while Definition 6 took a simple alternative $\mathcal{X}_{\boldsymbol{N}}^1$, we may summarize efficiency over the composite alternative $\aleph_{\boldsymbol{N}}^1$. There are two standard approaches. (1) Summarize the efficiency by a weighted average over the composite alternative (the weights might correspond to a Bayesian prior on the alternative). This is analogous to the GRO criterion in Grünwald et al. [2023]. (2) Summarize the efficiency by its minimum over the composite alternative (this is like the maximum 'loss,' which a minimax method minimizes). This is analogous to the REGROW criteria of Grünwald et al. [2023]. Efficiency can be also be generalized by considering functionals of the distribution of n_τ other than its expected value. For example, efficiency could be defined with respect to the 80th percentile of n_τ. Such a definition would bring sequential efficiency more in line with the traditional, fixed-size notion of *power*, reflecting the sample size at which the test has a high probability of rejecting the null. While stopping-time percentiles can be evaluated using simulations, it is difficult to characterize them theoretically. Efficiency with respect to expected sample size is more tractable.

Kelly-optimality and efficiency

A Kelly-optimal test maximizes the expected log growth rate, while an efficient test minimizes the expected sample size. The following lemma links the expected log growth to the expected stopping time.

Lemma 2 (Wald [1945])**.** *Consider testing the composite null $H_0 : \mathcal{X}_{\boldsymbol{N}} \in \aleph^0_{\boldsymbol{N}}$ at level α using the UI-TS M_t. Let $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log \Delta M_t]$ be its expected log growth and $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau(M_t)]$ its expected stopping time under the true population $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$. We have:*

$$\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau(M_t)] = \frac{-\log \alpha}{\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log \Delta M_t]}.$$

Because Kelly-optimal tests maximize $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log \Delta M_t]$, they minimize $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau(M_t)]$. Kelly-optimal tests are efficient precisely when $\tau = n_\tau$, which holds for the important subclass of UI-TS with fixed selections $S_t(\boldsymbol{\tau}) = S_t$. Furthermore, the expected stopping time of the Kelly-optimal test always lower bounds the expected sample size of the efficient test.

Kelly optimality for a composite null and simple alternative

Definition 8 (Kelly-optimal betting UI-TS)**.** *A betting UI-TS M_t^* for the composite null $H_0 : \mu(\mathcal{X}_{\boldsymbol{N}}) \leq \eta_0$ is Kelly-optimal for the simple alternative $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$ if its expected log-growth under that alternative is maximal among all betting UI-TSs for H_0. That is:*

$$\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log \Delta M_t^*] = \sup_{M_t} \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log \Delta M_t],$$

where $\Delta M_t := M_t/M_{t-1} = \min_{\boldsymbol{\eta} \in \mathcal{C}} M_t(\boldsymbol{\eta}) / \min_{\boldsymbol{\eta} \in \mathcal{C}} M_{t-1}(\boldsymbol{\eta})$. Identifying a particular betting UI-TS with its collection of bets and selection rules $\{(\boldsymbol{\lambda}_t(\boldsymbol{\eta}), \boldsymbol{p}_t(\boldsymbol{\eta}))\}_{\boldsymbol{\eta} \in \mathcal{C}}$, the supremum is taken over all such collections. The expectation is taken with respect to random sampling from $\mathcal{X}^1_{\boldsymbol{N}}$ and the stratum selections $S_t(\boldsymbol{\eta})$.

We show below how to construct a Kelly-optimal UI-TS; there may be more than one. To do so, we first define Kelly-optimality for an I-TSM for the intersection null $\boldsymbol{\eta} \in \mathcal{C}$ with respect to a simple alternative.

Definition 9 (Kelly-optimal betting I-TSM)**.** *A betting I-TSM $M_t^*(\boldsymbol{\eta})$ for the intersection null $\boldsymbol{\eta}$ is* Kelly-optimal *for the simple alternative $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}}$ if its expected log-growth is maximal among all betting I-TSMs for $\boldsymbol{\eta}$:*

$$\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log \Delta M_t^*(\boldsymbol{\eta})] = \sup_{M_t(\boldsymbol{\eta})} \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log \Delta M_t(\boldsymbol{\eta})],$$

where $\Delta M_t(\boldsymbol{\eta}) := M_t(\boldsymbol{\eta})/M_{t-1}(\boldsymbol{\eta}) = \sum_{k=1}^K 1\{S_t(\boldsymbol{\eta}) = k\} Z_{kT_k(t)}(\eta_k)$ *is the change in the I-TSM at time* t*. The supremum is taken over all choices of bets and selection rules* $(\boldsymbol{\lambda}_t(\boldsymbol{\eta}), \boldsymbol{p}_t(\boldsymbol{\eta}))$*. The expectation is with respect to the random sampling from* $\mathcal{X}_{\boldsymbol{N}}$ *and the stratum selections* $S_t(\boldsymbol{\eta})$*.*

With these definitions in hand, we construct the Kelly-optimal betting I-TSM. We can achieve this by finding the Kelly-optimal bets $(\boldsymbol{\lambda}_t(\boldsymbol{\eta}))_{t\in\mathbb{N}}$ and selection rules $(\boldsymbol{p}_t(\boldsymbol{\eta}))_{t\in\mathbb{N}}$ that uniquely parameterize $M_t^*(\boldsymbol{\eta})$. The optimal bets $\boldsymbol{\lambda}_t^*(\boldsymbol{\eta})$ do not depend on the selection rule $\boldsymbol{p}_t(\boldsymbol{\eta})$. The following Lemma, proved in Appendix C.2.2, yields the Kelly-optimal I-TSM.

Lemma 3 (Construction of the Kelly-optimal I-TSM)**.** *Fix an intersection null* $\boldsymbol{\eta} \in \mathcal{C}$ *and simple alternative* $\mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^1$*.*

1. ***Bets:*** *for every selection rule* $\boldsymbol{p}_t(\boldsymbol{\eta})$*, the Kelly-optimal bets are*

$$\boldsymbol{\lambda}_t^*(\boldsymbol{\eta}) := [\lambda_{1t}^*(\boldsymbol{\eta}), \dots, \lambda_{Kt}^*(\boldsymbol{\eta})]$$

   where

$$\lambda_{kt}^*(\boldsymbol{\eta}) := \underset{\lambda\in[0,1/\eta_{kt}]}{\arg\max}\, \mathbb{E}\left\{\log[1 + \lambda(X_{kt} - \eta_{kt})]\right\}$$

   is the Kelly-optimal bet for the within-stratum TSM $M_{kt}(\eta_k)$*.*

2. ***Selection rules:*** *Let* $\mathcal{A} := \arg\max_{j\in\{1,\dots,K\}} \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\log Z_{jt}(\eta_j)]$ *be the set of strata with maximal expected log-growth under the alternative* $\mathcal{X}_{\boldsymbol{N}}$ *for the bets* $\boldsymbol{\lambda}_t(\boldsymbol{\eta})$*. The Kelly-optimal selection rule is*

$$\boldsymbol{p}_t^*(\boldsymbol{\eta}) := [p_{1t}^*(\boldsymbol{\eta}), \dots, p_{Kt}^*(\boldsymbol{\eta})]$$

   where

$$p_{kt}^*(\boldsymbol{\eta}) := 1\{k \in \mathcal{A}\}/|\mathcal{A}|.$$

   Through $Z_{jt}(\eta_k)$*,* $\boldsymbol{p}_t(\boldsymbol{\eta})$ *depends on the bets* $\boldsymbol{\lambda}_t(\boldsymbol{\eta})$*, and selects a stratum with the highest expected log-growth for that vector of bets.*

The I-TSM $M_t^*(\boldsymbol{\eta})$ *that uses* $\boldsymbol{\lambda}_t^*(\boldsymbol{\eta})$ *and* $\boldsymbol{p}_t^*(\boldsymbol{\eta})$ *is Kelly-optimal.*

If the sampling is with replacement (the draws from each stratum are IID), then the Kelly-optimal bets $\boldsymbol{\lambda}^*(\boldsymbol{\eta})$ and selection rules $\boldsymbol{p}^*(\boldsymbol{\eta})$ are both fixed across time.

Kelly optimality for I-TSMs provides a recipe for constructing a Kelly-optimal UI-TS.

Lemma 4 (A Kelly-optimal UI-TS)**.** *For each $\boldsymbol{\eta} \in \mathcal{C}$, let $M_t^*(\boldsymbol{\eta})$ be Kelly-optimal I-TSM for $\boldsymbol{\eta}$ when the true alternative is $\mathcal{X}_{\boldsymbol{N}}$. The UI-TS:*

$$M_t^* := \min_{\boldsymbol{\eta} \in \mathcal{C}} M_t^*(\boldsymbol{\eta})$$

is globally Kelly-optimal for $\mathcal{X}_{\boldsymbol{N}}$. It is identified with the collection of intersection Kelly-optimal bets and selection rules $\{(\boldsymbol{\lambda}_t^*(\boldsymbol{\eta}), \boldsymbol{p}_t^*(\boldsymbol{\eta}))\}_{\boldsymbol{\eta} \in \mathcal{C}}$.

The proof is essentially self-evident: any TSM under H_0 is of the form $\min_{\boldsymbol{\eta} \in \mathcal{C}} M_t(\boldsymbol{\eta})$; we can always maximize the growth of the minimum by maximizing the growth of every I-TSM in $\mathcal{C}$, which is achieved by using the Kelly-optimal strategy $M_t^*(\boldsymbol{\eta})$ for every $\boldsymbol{\eta}$. Even when $\mathcal{X}_{\boldsymbol{N}}$ is known, it is very difficult to explicitly construct a Kelly-optimal UI-TS, except in very simple cases. We give such an example in Appendix C.1, where the population consists of two point mass strata ($x_{ik} = \mu_k$ for all i). For more general simple alternatives, M_t^* could be constructed numerically, but this will not work when the alternative is composite. We now describe a strategy that achieves efficiency by "learning" aspects of $\mathcal{X}_{\boldsymbol{N}}$ that allow us to approximate the Kelly-optimal solution.

Approximate Kelly optimality under the intersection-composite alternative

Consider again a fixed intersection null $\boldsymbol{\eta} \in \mathcal{C}$ and the intersection-composite alternative $\boldsymbol{\mu} \not\leq \boldsymbol{\eta}$. Within stratum k, the alternative $\mathcal{X}_k$ is not specified, so we cannot directly apply Lemma 3. Instead, we borrow from Waudby-Smith and Ramdas [2023], who provide *predictable* betting methods that approximate the Kelly-optimal strategy using past data. Waudby-Smith and Ramdas [2023] call these methods "growth rate adapted to the particular alternative" (GRAPA) and "approximate GRAPA" (aGRAPA). GRAPA betting directly estimates the Kelly-optimal bet at time t using the data available at time $(t-1)$. Because GRAPA involves solving nonlinear equations, Waudby-Smith and Ramdas [2023] propose aGRAPA, based on a Taylor series approximation. Within stratum k, the aGRAPA bet is

$$\lambda_{kt}^{\text{aG}}(\eta_k) := 0 \vee \frac{\hat{\mu}_{k(t-1)} - \eta_k}{\hat{\sigma}_{k(t-1)}^2 + (\hat{\mu}_{k(t-1)} - \eta_k)^2} \wedge \frac{c}{\eta_k},$$

where $\hat{\mu}_{k(t-1)}$ and $\hat{\sigma}_{k(t-1)}^2$ are the lagged empirical mean and variance, and $c \leq 1$ is a user-specified truncation parameter. When the empirical mean is above the null mean and the variance is relatively small, aGRAPA bets more aggressively: the data give confidence that the next draw will be above η_k and the bet will make money. The truncation at 0 ensures the bet is 0 in strata where the empirical mean is below the null; the truncation at c/η_k prevents bets from becoming too aggressive ($c < 1$ ensures that $M_t > 0$ for all t). Using aGRAPA bets in each stratum $\boldsymbol{\lambda}_t^{\text{aG}}(\boldsymbol{\eta}) := [\lambda_{kt}^{\text{aG}}(\eta_k)]_{k=1}^K$ provides a fairly efficient betting strategy under $\boldsymbol{\eta}$ and the intersection-composite alternative.

The bandit literature offers a range of potential options for $\boldsymbol{p}_t(\boldsymbol{\eta})$. For example, we could pull from the stratum with the largest upper confidence bound on $\mathbb{E}[Z_{kt}(\eta_k)]$ (a UCB algorithm [Lai and Robbins, 1985]) or according to a posterior probability that $\mathbb{E}[Z_{kt}(\eta_k)]$ is largest for an assumed prior (a Thompson sampling algorithm [Thompson, 1933]). All of these strategies approximate the Kelly-optimal solution when $\mathcal{X}_{\boldsymbol{N}}$ is unknown.

Other betting strategies

We propose two additional betting strategies that, when paired with fixed stratum selections $S_t := S_t(\boldsymbol{\eta})$ have some nice computational properties (compared to aGRAPA and other approximately Kelly-optimal bets). We present the strategies now and discuss computation in the next section.

The first strategy uses the same bet for all $\boldsymbol{\eta}$:

$$\boldsymbol{\lambda}_t(\boldsymbol{\eta}) := \boldsymbol{\lambda}_t.$$

The bet may vary over time as a predictable function of past data, but the bets must be identical for every I-TSM $\{M_t(\boldsymbol{\eta})\}_{\boldsymbol{\eta} \in \mathcal{C}}$. Let $\bar{X}_{kT_k(t)}$ denote the sample mean in stratum k at time t. The second strategy is

$$\lambda_{kt}^{\text{nE}}(\boldsymbol{\eta}) := \exp\left(\bar{X}_{kT_k(t-1)} - \eta_k\right).$$

The superscript "nE" stands for negative exponential, the functional dependence of the bet on η_k. The bet $\boldsymbol{\lambda}_t^{\text{nE}}(\boldsymbol{\eta}) = [\lambda_{1t}^{\text{nE}}(\boldsymbol{\eta}), \ldots, \lambda_{Kt}^{\text{nE}}(\boldsymbol{\eta})]$ varies smoothly as a function of $\boldsymbol{\eta}$. It bets more when the running sample mean in stratum k is further above the null mean in stratum k.

7.6 Computational Tractability

The results of Section 7.5.2 in principle allow one to construct an approximately optimal I-TSM for every $\boldsymbol{\eta} \in \mathcal{C}$; Lemma 4 then suggests an efficient UI-TS under the global-composite alternative $\mu > \eta_0$. In practice, this is computationally infeasible.

The computational tractability of a UI-TS depends on how the constituent I-TSMs are constructed and in particular how they depend on $\boldsymbol{\eta}$. We classify I-TSMs accordingly. The term $\boldsymbol{\eta}$-*aware* refers to betting and selection strategies that depend on $\boldsymbol{\eta}$; $\boldsymbol{\eta}$-*oblivious* refers to strategies that do not. Similarly, we call a predictable strategy *adaptive* if it explicitly depends on past data, and *nonadaptive* otherwise. A strategy may be adaptive but $\boldsymbol{\eta}$-oblivious, non-adaptive but $\boldsymbol{\eta}$-aware, etc. The strategies proposed in Section 7.5.2 are adaptive and $\boldsymbol{\eta}$-aware. A UI-TS is called $\boldsymbol{\eta}$-oblivious only if both the stratum selectors and bets are $\boldsymbol{\eta}$-oblivious.

Small K, small N, discrete support: Recall from Section Section 7.4.2 that $\mathcal{B}$ is the (possibly finite) boundary of $\mathcal{E}_0$. If the support of $\aleph_{\boldsymbol{N}}^0$ is small and known (e.g., $\mathcal{X}_{\boldsymbol{N}}$ is known to be binary), if there are few strata (e.g., $K \leq 3$), and if the strata are small (e.g., $\max_k N_k \leq 1000$), it is feasible to enumerate $\mathcal{B}$. That makes it feasible to use an arbitrary $\boldsymbol{\eta}$-aware strategy $\{(\boldsymbol{\lambda}_t(\boldsymbol{\eta}), \boldsymbol{p}_t(\boldsymbol{\eta}))\}_{\boldsymbol{\eta} \in \mathcal{B}}$ by brute-force minimization over $\mathcal{B}$. We discuss the complexity of enumerating $\mathcal{B}$ in Appendix C.3.

Small K: When the strata are large or when the support is unknown, it is not possible to minimize over $\mathcal{B}$ by brute force. In Appendix C.3.2 we show that the I-TSM $M_t(\boldsymbol{\eta})$ is log-concave over $\boldsymbol{\eta} \in \mathcal{C}$ when the bets and selections are $\boldsymbol{\eta}$-oblivious. As a result, when bets and selections are $\boldsymbol{\eta}$-oblivious over a convex subset of $\mathcal{C}$, the I-TSM is log-concave over that subset. In turn, log-concavity implies that the minimum over a convex subset must occur on one of its extreme points. If $\mathcal{C}$ is partitioned into G convex subsets within each of which the bets and selections are $\boldsymbol{\eta}$-oblivious.

This strategy is illustrated in Figure 7.3, which divides the regions into convex *bands* and evaluates the I-TSM at their vertices. It requires specifying G bets and selections $(\boldsymbol{\lambda}_t(\boldsymbol{\eta}_g), S_t(\boldsymbol{\eta}_g))_{g=1}^G$. The number of I-TSM evaluations required depends on the number of bands that share each vertex. The evaluation point $\boldsymbol{\eta}_g$ for each band's betting and selection rule might be set as the centroid of the band intersected with $\mathcal{C}$. We focus on the case $K = 2$, in which case the bands can be defined by a grid of $G + 1$ points along the line $\mathcal{C}$. The resolution of the grid will affect how much the I-TSMs can adapt to $\boldsymbol{\eta}$ and (we conjecture) the efficiency of the UI-TS. We demonstrate this by varying G when $K = 2$. We do not evaluate the band method for $K > 2$, because it is difficult to set up and because we expect it to scale poorly in K: the resolution decreases as K increases with G fixed.

Moderate K: The coarsest banding strategy is to set $G = 1$, which returns the $\boldsymbol{\eta}$-oblivious approach using the same bets and selections for every $\{M_t(\boldsymbol{\eta})\}_{\boldsymbol{\eta} \in \mathcal{C}}$. The resulting I-TSM $M_t(\boldsymbol{\eta})$ is log-concave in $\boldsymbol{\eta}$ over all of $\mathcal{C}$, and its minimum is attained at a vertex of the polytope $\mathcal{C}$: we can compute M_t by evaluating $M_t(\boldsymbol{\eta})$ at the set of vertices $\mathcal{V}$. The number of vertices scales combinatorially in K, as $|\mathcal{V}| = \binom{K}{K/2}$ when K is even and stratum-sizes are equal. Enumerating the vertices is tractable for any $\boldsymbol{N}$ and any support if the number of strata is not too big (e.g. $K \leq 16$).

Arbitrary K: If we use $\boldsymbol{\eta}$-oblivious selects $S_t(\boldsymbol{\eta}) := S_t$, the $\boldsymbol{\eta}$-aware negative exponential betting strategy $\boldsymbol{\lambda}_t^{\text{nE}}(\boldsymbol{\eta})$ leads to an I-TSM $M_t(\boldsymbol{\eta})$ that is smooth and log-convex in $\boldsymbol{\eta}$ (see Appendix C.3.3). The minimum generally occurs on the interior of $\mathcal{C}$ and can be found numerically. In particular, we have a convex objective with a convex constraint and can use projected gradient descent to find the UI-TS for arbitrary $\boldsymbol{N}$ and K. While we do not provide an implementation of that idea here, we evaluate the statistical efficiency of negative exponential bets computed using the banding strategy.

In the next section, we use the banding strategy to evaluate the statistical performance of various methods.

7.7 Simulations

In this section we evaluate the efficiency of our proposed methods in three simulation studies. The first set of simulations examines behavior on distributions that are point-masses within strata. This simulation is an idealization of some common auditing populations (e.g. risk-limiting comparison audits), wherein the vast majority of values are concentrated at a single point reflecting correctly reported values, with a small outlying mass reflecting errors. Point-masses are thus analogous to error-free auditing populations. The second set of simulations models applications to binary populations by drawing samples from Bernoulli distributions with varying μ_k, the probability of observing a 1 in stratum k. The third set draws population values from truncated Gaussian superpopulations within strata. The Gaussians are specified with a range of means and standard deviations, and are truncated to $[0, 1]$. This constitutes a more traditional simulation study, modeling a range of applications with continuous populations.

7.7.1 Point-mass simulations

We evaluated methods on populations with stratum sizes $\boldsymbol{N} = [200, 200]$ and point-masses as within-stratum distributions, so that $x_{ik} := \mu_k$ for $i \in \{1, \ldots, 200\}$ and $k \in \{1, 2\}$. We defined populations with global means $\mu = \frac{1}{2}(\mu_1 + \mu_2)$ ranging between 0.51 and 0.75, and gaps between strata of either 0 (i.e., $\mu_1 = \mu_2 = \mu$) or 0.5 (i.e., $\mu_1 = \mu_2 - 0.5$).

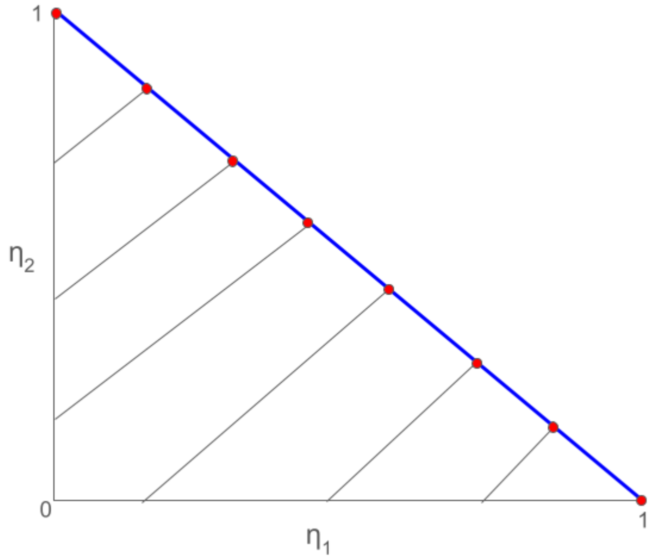


Figure 7.3: A, illustration of the banding strategy for computing the UI-TS when $G = 7$, $K = 2$, $N_1 = N_2$, and $\eta_0 = 1/2$. The space of intersection nulls is plotted with η_1 on the x-axis and η_2 on the y-axis. The feasible null $\mathcal{E}_0$ is represented by the entire right triangle, and split into bands over which bets and selections are fixed. Element-wise monotonicity of $M_t(\boldsymbol{\eta})$ implies the minimum must occur on the boundary $\mathcal{C}$, represented by the blue line. Along that boundary, $M_t(\boldsymbol{\eta})$ is log-concave when bets and selections are $\boldsymbol{\eta}$-oblivious. Since they are fixed over bands, $M_t(\boldsymbol{\eta})$ is log-concave within the intervals demarcated by the red dots, and the minimum over all $\mathcal{E}_0$ (i.e., M_t) is attained at one of the red dots.

The method for inference was either global lower confidence bounds (LCBs) as in Section 7.4.1 or UI-TSs. The banding strategy discussed in Section 7.6 was used with an equally-spaced grid, and we explored settings for $G \in \{3, 10, 100, 500\}$. Bets were specified as aGRAPA described in Section 7.5.2, the fully fixed strategy with $\lambda_{ki} := 0.75$, or the negative exponential approach described in Section 7.5.2. The allocation rule was specified as $\boldsymbol{\eta}$-oblivious, nonadaptive "round robin," alternately drawing from each stratum, an $\boldsymbol{\eta}$-aware and adaptive strategy "predictable Kelly" strategy using a UCB-style algorithm on every $\boldsymbol{\eta} \in \mathcal{C}$, or an $\boldsymbol{\eta}$-oblivious "greedy" strategy using the UCB-style algorithm iteratively on the minimizing $\boldsymbol{\eta}$ at the last time t to produce a single selection for all I-TSMs. In more detail, for each $\boldsymbol{\eta} \in \mathcal{C}$, each stratum k, and each time t, the predictable Kelly strategy computed the average of the past terms $\bar{Z}_{k(t-1)}$ and the standard error of that average $\widehat{\mathrm{SE}}_{k(t-1)}$, estimated as the sample standard deviation of the terms divided by $\sqrt{T_k(t-1)}$. For each k, an upper confidence bound (UCB) on the expected gain was computed as $\bar{Z}_{k(t-1)} + 2\widehat{\mathrm{SE}}_{k(t-1)}$, and the next sample was drawn from the stratum with the largest UCB. This predictable Kelly strategy was only implemented for the UI-TSs, since the LCB approach cannot use $\boldsymbol{\eta}$-aware allocation strategies. The greedy strategy was similar, but instead of varying over each $\boldsymbol{\eta} \in \mathcal{C}$, it implemented the UCB-algorithm for $\boldsymbol{\eta}^*_{t-1}$, the most recent minimizer.

We tested each combination of inference method, betting strategy, and allocation strategy on each population and recorded the stopping time. Because the populations are point masses, there is no randomness whatsoever so that the sample sizes only needed to be computed once. The sample sizes for UI-TSs with aGRAPA bets at various G—averaged across populations and selection rules—appear in Table 7.1. As expected, the efficiency of an adaptive UI-TS using the banding strategy increases as G increases. In this case the efficiency gains level off around $G = 100$. Sample sizes at $G = 100$ appear in Figure 7.4. UI-TSs are always more efficient than the LCB method. The gap can be considerable (the y-axis is on the log scale). However, UI-TSs are highly inefficient when used with both $\boldsymbol{\eta}$-oblivious bets and allocations. It seems most important to make the betting strategy $\boldsymbol{\eta}$-aware, as evidenced by the performance of aGRAPA and negative exponential bets. The greedy allocation rule was slightly more efficient than round robin for aGRAPA bets, but they exhibited identical performance using the negative exponential bets. Finally, despite its theoretically low stopping times, the predictable Kelly rule led to poor sample sizes because different selections were used for each intersection null. This highlights the importance of using $\boldsymbol{\eta}$-oblivious selections when aiming to minimize the global sample size n_τ.

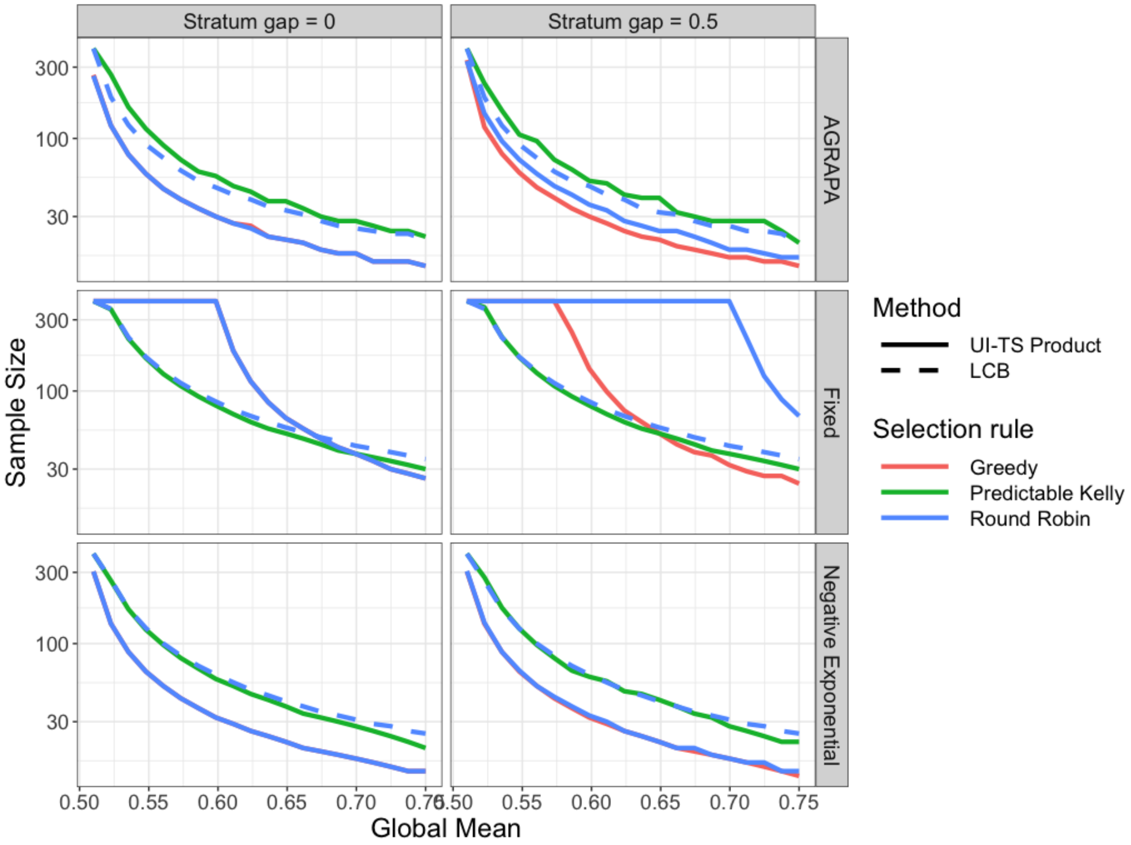


Figure 7.4: Sample sizes (y-axis, log scale) for various sequential stratified tests (line colors and types) of the global null $H_0 : \mu \leq 1/2$ for 2-stratum point-mass alternatives with varying global means (x-axis) and between-stratum spread (columns). For example, when the global mean is 0.6 and the spread is 0.5, the within-stratum means are $\boldsymbol{\mu} = [0.35, 0.85]$. UI-TSs were computed using the banding strategy, with $G = 100$. Note that the round robin (blue) and greedy Kelly (red) lines are overlapping in many panels. All methods assumed sampling was with replacement, but sample sizes were capped at 400: a sample size of 400 should be read as 400 or greater. LCB = lower confidence bound; UI-TS = union-intersection test statistic.

G	Average sample size
3	179.7
10	90.1
100	48.3
500	46.8

Table 7.1: Sample sizes at various G, the number of grid points used in the banding strategy for computing a product UI-TS using aGRAPA bets, averaged across point-mass populations with $K = 2$ strata. The efficiency of the resulting UI-TS increases as G increases, but exhibits diminishing returns: the difference between $G = 100$ and $G = 500$ is trivial.

7.7.2 Bernoulli simulations

We drew 2-stratum populations of size $\boldsymbol{N} = [200, 200]$ from Bernoulli distributions with stratum-wise success probabilities $\boldsymbol{\mu} = [\mu, \mu]$ or $\boldsymbol{\mu} = [\mu - 0.25, \mu + 0.25]$ and population-level success probability μ ranging from 0.51 to 0.70. We tested the global null $H_0 : \mu \leq 0.5$ against these alternatives and recorded the sample sizes for the same methods described in Section 7.7.1. For this simulation we added two new betting strategies based on Kelly optimality for Bernoulli populations. Within each stratum, the "shrink-trunc Bernoulli" strategy took the mean estimate $\hat{\mu}_{kt}$ described in Section 2.5.2 of Stark [2023],[10] and transformed it into a bet $\boldsymbol{\lambda}_t^{\text{ST-Bern}}(\boldsymbol{\eta})$ as described in Section 2.3 of that paper. In detail, the apriori estimate of the alternative was set at $(\eta_k + 1)/2$ for each stratum and null mean; the anchoring factor was $d := 20$ so that the sample mean quickly dominates the estimate; the estimate is truncated to $\eta_k + 1/(2\sqrt{d + T_k(t) - 1})$. This mean estimate $\hat{\mu}_{kt}$ was transformed to a bet by taking $\lambda_{kt}^{\text{ST-Bern}}(\eta_k) = (\hat{\mu}_{kt}/\eta_k - 1)/(1 - \eta_k)$. This is akin to using the Bernoulli SPRT with a plug in estimate for the alternative mean [Wald, 1947]. We used the same method with the *true* alternative mean in each stratum. This is the Kelly-optimal method bet. While not usable in practice since the alternative is unknown, it provides a lower bound on the expected stopping time of a UI-TS using a given selection rule. We replicated each simulation of each population and method 1000 times, and took the empirical mean sample sizes as an estimate of the expected sample size.

Results for round robin allocation and stratum gap equal to 0.5 appear in Figure 7.5. The other allocation rules were comparable in terms of performance to the results in Figure 7.4 and removing them simplifies the plot. The UI-TSs were nearly always

[10] The notation in Stark [2023] is essentially flipped from ours. It uses η_t for our $\hat{\mu}_{kt}$ (the alternative mean estimate) and μ_t for our η_t (the null mean).

sharper than any LCB method, except when fixed bets were used. The Kelly optimal bets corresponding to the Bernoulli SPRT with the true alternative had the lowest expected sample size, followed by the negative exponential bets and the ALPHA-ST bets. Perhaps surprisingly, aGRAPA was not nearly as efficient as the other strategies, though it was still much better than fixed betting.

7.7.3 Gaussian simulations

We drew finite populations of size $\boldsymbol{N} = [200, 200]$ from truncated Gaussian superpopulations within each stratum. The true grand mean μ ranged along a discrete grid from 0.5 to 0.7. Before truncation, the superpopulation Gaussians had standard deviations $\sigma_1 = \sigma_2 = \sigma$ with $\sigma \in \{0.01, 0.05, 0.1\}$. The means were allowed to vary across strata according to a parameter $\delta \in \{0, 0.2\}$ specifying the distance between the strata so that $\boldsymbol{\mu} = [\mu - 0.5\delta, \mu + 0.5\delta]$. We truncated the Gaussians by redrawing samples that landed outside $[0, 1]$, which was rare since for all simulation settings the within-stratum means were bounded away from $\{0, 1\}$ and σ was relatively small.

We implemented the same methods as the point mass and Bernoulli simulations, and used the banding strategy with $G = 100$ to compute the UI-TSs. For this simulation we also implemented a betting TSM drawing *unstratified* samples from the pooled population. That method served as a benchmark that could be used in practice if the statistician had control over the design. In each simulation of each superpopulation scenario, we drew a new finite population from the truncated Gaussian superpopulation, ran each method on the finite population, and recorded the sample size. If the method did not stop by the time the finite population had been consumed, we recorded a stopping time of 400. We replicated the simulations 500 times, and recorded the empirical mean stopping times as a conservative estimate of the true expected stopping times of each method on each population.

The estimated expected stopping times for $\sigma = 0.05$ and round robin allocation are plotted in Figure 7.6. To simplify the plots we do not include other allocation strategies and σ, as variation across these factors is unsurprising. In particular, the allocation strategies had similar performance to the point mass and Bernoulli simulations, with predictable Kelly producing considerably higher stopping times than round robin and greedy Kelly producing slightly lower stopping times when the strata were spread apart ($\delta = 0.2$). The methods' performance was similar for different σ as well, with expected stopping times slightly lower for $\sigma = 0.01$ and higher for $\sigma = 0.1$.

UI-TS with fixed bets and LCB (regardless of the betting strategy) required the most samples to stop. The unstratified TSM generally had the lowest expected sample

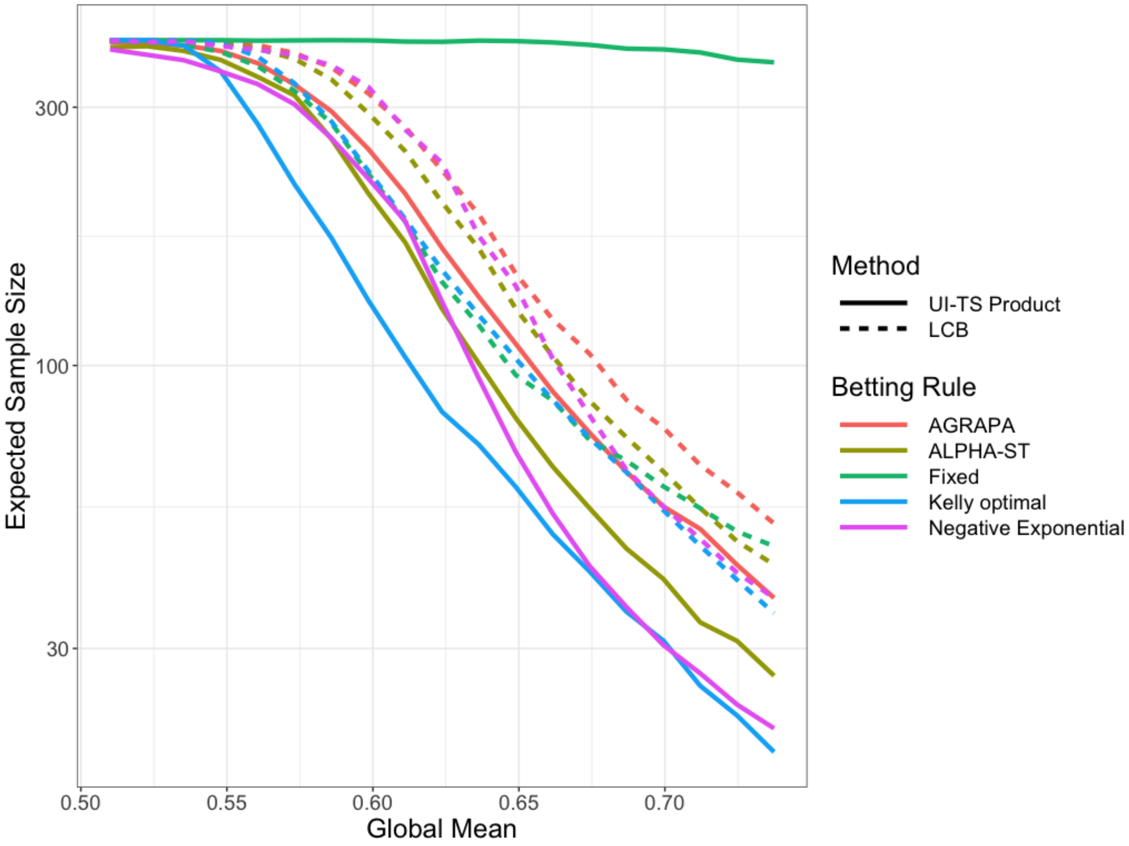


Figure 7.5: Expected global sample sizes $\mathbb{E}[n_\tau]$ (y-axis; $\log_{10}$ scale) for the global null $H_0 : \mu < 0.5$ for Bernoulli populations with various true means (x-axis) and gaps between strata (columns). The LCB and UI-TS methods (linetypes) were used with various settings for the bets (line colors). The allocation strategy is round robin. Expected sample sizes are taken as the empirical average sample size to stop at level $\alpha = 0.05$ over 1000 simulations. All methods assumed sampling was with replacement, but sample sizes were capped at 400, so the expected sample size estimates may be biased downwards. LCB = lower confidence bound; UI-TS = union-intersection test statistic.

size (especially when used with aGRAPA bets), but the UI-TS had comparable performance when used with negative exponential or aGRAPA bets. When the stratum means were spread apart and the alternative was close to the null, the UI-TS was slightly sharper than unstratified sampling. This is in keeping with the classical theory of stratification in terms of MSE: stratification improves efficiency when between-stratum variation is large compared to within-stratum variation [Neyman, 1934, Cochran, 1977].

7.8 Conclusions

UI-TSs are considerably more efficient than the simple approach of combining confidence bounds for sequential stratified inference. Our results apply to a large range of applied problems, including auditing, measurement, survey sampling, and randomized controlled trials. We extended the criterion of Kelly optimality to stratified sampling in order to construct efficient tests. However, because it targets stopping time, the Kelly optimal UI-TS is not necessarily the most efficient in terms of sample size when selections can vary across $\boldsymbol{\eta} \in \mathcal{C}$. We navigated this issue by examining heuristic $\boldsymbol{\eta}$-oblivious selection strategies: round robin or "greedy" selection, targeting the growth of the smallest I-TSM. These rules had comparable performance; we generally found that the bets mattered much more to efficiency than the selections.

That said, finding the Kelly-optimal $\boldsymbol{\eta}$-oblivious selections or the $\boldsymbol{\eta}$-aware selections that minimize the expected sample size (rather than the expected stopping time) is an important open challenge for sequential stratified inference. We conjecture that the former problem will be easier to solve than the latter. Another interesting direction for research is to examine the performance of TSMs targeting regret rather than Kelly-optimality [Orabona and Jun, 2022] and leveraging optimal portfolio theory [Cover and Thomas, 2006]. We primarily considered predictable betting methods to construct efficient UI-TSs. When a prior on the alternative is available, the method of mixtures can be used to construct the GRO E-value for unstratified sampling [Grünwald et al., 2023]. The correctness of the prior determines the efficiency but not the frequentist validity of the test. Generalizing the method of mixtures to stratified inference may lead to efficient UI-TSs when the alternative is decribed by a prior. Finally, it may be fruitful to consider combining functions beyond the product when combining UI-TSs. While product combining is dominant for E-values larger than 1 [Vovk and Wang, 2021], this can be a difficult condition to ensure when bets cannot be arbitrarily $\boldsymbol{\eta}$-aware (e.g., in large K problems). Alternative combining functions (like Fisher combining or average combining) may allow for consistent inference in these cases without much additional fuss [Spertus and Stark, 2022]. Indeed, Cho et al.

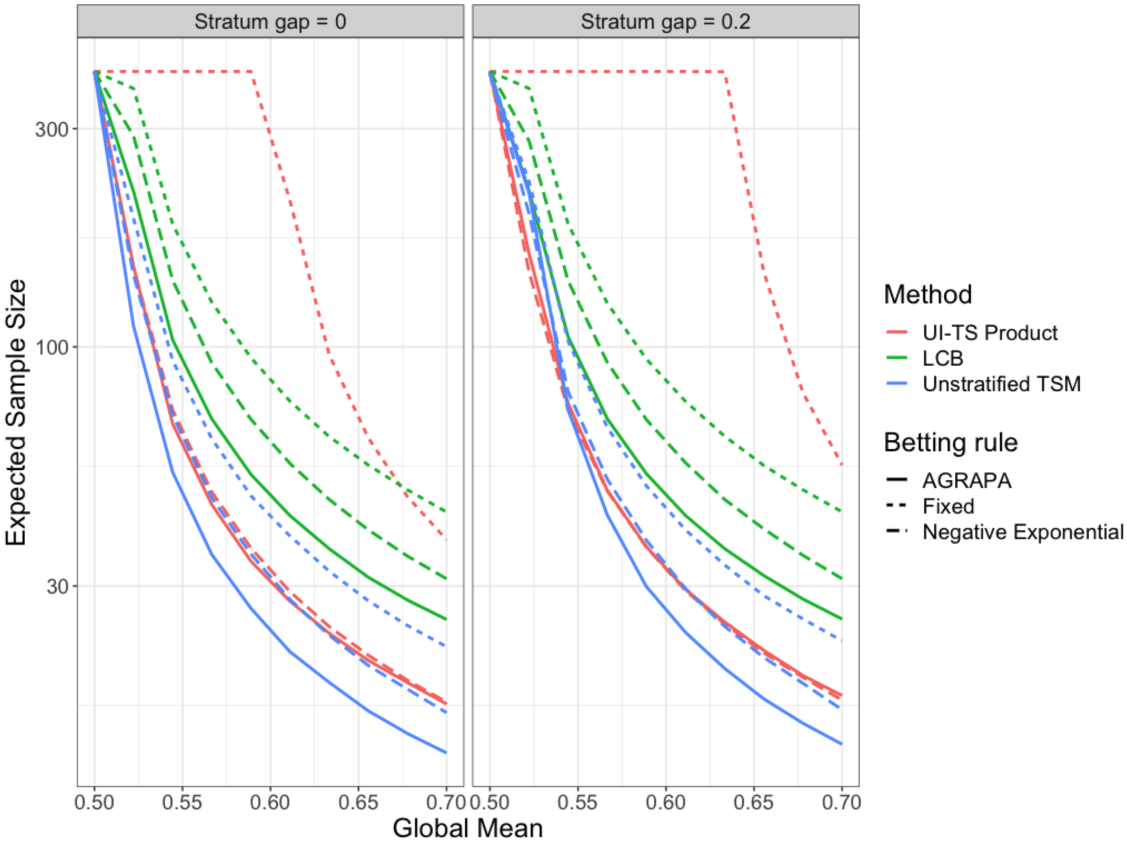


Figure 7.6: Expected sample sizes (y-axis; $\log_{10}$ scale) of various tests (line colors and types) of the null $H_0 : \mu \leq 1/2$ for 2-stratum populations drawn from truncated Gaussian distributions with varying global means (x-axis) and gaps between stratum-wise means (columns). Populations consist of $N_1 = N_2 = 200$ units within each stratum, drawn from truncated Gaussian distributions with standard deviation $\sigma = 0.05$. Stratum selections were round robin for all stratified methods. LCB = lower confidence bound; TSM = test supermartingale; UI-TS = union-of-intersections test statistic

[2024] show that average combining can lead to a test of power 1 and computationally tractable optimization over $\mathcal{C}$. However, in initial simulations comparing different combining functions, we found product combining tended to dominate except when the betting rule and selection rules were chosen very poorly (i.e. if bets were fixed and selections were round robin). Further comparing strategies for constructing UI-TSs is an interesting area for future research.

Bibliography

W. Amelung and W. Zech. Minimisation of organic matter disruption during particle-size fractionation of grassland epipedons. *Geoderma*, 92(1):73–85, September 1999. ISSN 0016-7061. doi: 10.1016/S0016-7061(99)00023-3. URL https://www.sciencedirect.com/science/article/pii/S0016706199000233.

T.W. Anderson. Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. *Bulletin of the International Statistical Institute*, 43:249–251, 1967.

T.W. Anderson. Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. *US Dept of the Navy*, 1969. doi: 10.21236/AD0696676.

A.W. Appel and P.B. Stark. Evidence-based elections: Create a meaningful paper trail, then audit. *Georgetown Law Technology Review*, 4.2:523–541, 2020. https://georgetownlawtechreview.org/wp-content/uploads/2020/07/4.2-p523-541-Appel-Stark.pdf.

A.W. Appel, R. DeMillo, and P.B. Stark. Ballot-marking devices cannot assure the will of the voters. *Election Law Journal, Rules, Politics, and Policy*, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3375755.

A. Arlotto and M.J. Steele. Beardwood–Halton–Hammersley theorem for stationary ergodic sequences: A counterexample. *Annals of Applied Probability*, 26(4):2141–2168, August 2016. ISSN 1050-5164, 2168-8737. doi: 10.1214/15-AAP1142. URL https://projecteuclid.org/euclid.aoap/1472745454. Publisher: Institute of Mathematical Statistics.

P. Aronow and J. Middleton. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1, 01 2013. doi: 10.1515/jci-2012-0009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

R.R. Bahadur and L.J. Savage. The Nonexistence of Certain Statistical Procedures in Nonparametric Problems. *The Annals of Mathematical Statistics*, 27(4):1115 – 1122, 1956. doi: 10.1214/aoms/1177728077. URL https://doi.org/10.1214/aoms/1177728077.

X. Bai, Y. Huang, W. Ren, M. Coyne, P.-A. Jacinthe, B. Tao, D. Hui, J. Yang, and C. Matocha. Responses of soil carbon sequestration to climate-smart agriculture practices: A meta-analysis. *Global Change Biology*, 25:2591–2606, 2019. doi: 10.1111/gcb.14658.

J.M. Baker, T.E. Ochsner, R.T. Venterea, and T.J. Griffis. Tillage and soil carbon sequestration—what do we really know? *Agriculture, Ecosystems & Environment*, 118:1–5, 2007. doi: 10.1016/j.agee.2006.05.014.

P. Baker and M. Haberman. In Torrent of Falsehoods, Trump Claims Election Is Being Stolen. *The New York Times*, November 2020. ISSN 0362-4331. URL https://www.nytimes.com/2020/11/05/us/politics/trump-presidency.html.

M.A. Bates and R. Glennerster. The generalizability puzzle. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm, 2017. Accessed: 2024-04-19.

J.P. Beem-Miller, A.Y.Y. Kong, S. Ogle, and D. Wolfe. Sampling for soil carbon stock assessment in rocky agricultural soils. *Soil Science Society of America Journal*, 80: 1411–1423, 2016. doi: 10.2136/sssaj2015.11.0405.

N. Begill, A. Don, and C. Poeplau. No detectable upper limit of mineral-associated organic carbon in temperate agricultural soils. *Global Change Biology*, 29(16):4662–4669, 2023. ISSN 1365-2486. doi: 10.1111/gcb.16804. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.16804. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.16804.

V. Bellon-Maurel and A. McBratney. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. *Soil Biology and Biochemistry*, 43(7):1398–1410, July 2011. ISSN 0038-0717. doi: 10.1016/j.soilbio.2011.02.019. URL http://www.sciencedirect.com/science/article/pii/S0038071711001106.

R.L. Berger and D.D. Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994. ISSN 01621459. URL `http://www.jstor.org/stable/2290928`.

R.A. Berk and D.A. Freedman. Statistical assumptions as empirical commitments. *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, pages 234–254, 2003.

D.A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Springer, October 1985. ISBN 0412248107.

P.J. Bickel. Correction: Inference and auditing: The Stringer bound. *Intl. Stat. Rev.*, 61:487, 1993.

A. Bispo, L. Andersen, D.A. Angers, M. Bernoux, M. Brossard, L. Cécillon, R.N.J. Comans, J. Harmsen, K. Jonassen, F. Lamé, C. Lhuillery, S. Maly, E. Martin, A.E. Mcelnea, H. Sakai, Y. Watabe, and T.K. Eglin. Accounting for carbon stocks in soils and measuring ghgs emission fluxes from soils: Do we have the necessary standards? *Frontiers in Environmental Science*, 5, 2017. doi: 10.3389/fenvs.2017.00041.

S. Bittinger, M. Yates, and M. Phillips. Statistical sampling and extrapolation: Due process challenges in the false claims act litigation and medicare appeals arenas, 2022. URL `https://www.americanbar.org/groups/health_law/publications/health_lawyer_home/december-2022/statistical-sampling-and-extrapolation/`.

M. Blom, P.J. Stuckey, and V. Teague. RAIRE: Risk-limiting audits for IRV elections. `https://arxiv.org/abs/1903.08804`, 2019.

F. Blyth, S. Perryman, P. Poulton, M. Glendining, and A. Gregory. Highfield ley-arable experiment cropping sequence 1949-2023. Electronic Rothamsted Archive, Rothamsted Research, 2023. URL `https://doi.org/10.23637/rrn1-HLAcrop4923-01`.

D. A. Bossio, S. C. Cook-Patton, P. W. Ellis, J. Fargione, J. Sanderman, P. Smith, S. Wood, R. J. Zomer, M. von Unger, I. M. Emmer, and B. W. Griscom. The role of soil carbon in natural climate solutions. *Nature Sustainability*, 3(5):391–398, May 2020. ISSN 2398-9629. doi: 10.1038/s41893-020-0491-z. URL `https://www.nature.com/articles/s41893-020-0491-z`. Publisher: Nature Publishing Group.

L. Breiman. Optimal gambling systems for favorable games. *Berkeley Symposium on Mathematical Statistics and Probability*, 4:65–78, 1961.

D. J. Brus. Using regression models in design-based estimation of spatial means of soil properties. *European Journal of Soil Science*, 51(1):159–172, 2000. doi: https://doi.org/10.1046/j.1365-2389.2000.00277.x. URL `https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2389.2000.00277.x`.

D.J. Brus and J.J. de Gruijter. Design-based generalized least squares estimation of status and trend of soil properties from monitoring data. *Geoderma*, 164:172–180, 2011. doi: 10.1016/j.geoderma.2011.06.001.

D.J. Brus, J.J. de Gruijter, and J.W. van Groenigen. Chapter 14: Designing Spatial Coverage Samples Using the k-means Clustering Algorithm. In P. Lagacherie, A. B. McBratney, and M. Voltz, editors, *Developments in Soil Science*, volume 31 of *Digital Soil Mapping*, pages 183–192. Elsevier, January 2006. doi: 10.1016/S0166-2481(06)31014-8. URL `http://www.sciencedirect.com/science/article/pii/S0166248106310148`.

G. Cardano. Liber de ludo alea. *Opera Omnia, Vol. 1*, 1966 (originally 1525). Reprint of 1663 Lyon edition.

C.J. Carey, J. Weverka, R. DiGaudio, T. Gardali, and E.L. Porzig. Exploring variability in rangeland soil organic carbon stocks across california (usa) using a voluntary monitoring network. *Geoderma Regional*, 22:e00304, 2020.

D. Caughey, A. Dafoe, X. Li, and L. Miratrix. Randomisation inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5): 1471–1491, November 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad080. URL `https://doi.org/10.1093/jrsssb/qkad080`.

D. Chaitlin. Sidney Powell shares 270-page binder of documents buttressing election fraud claims, December 2020. URL `https://www.washingtonexaminer.com/news/sidney-powell-shares-election-fraud-claims`. Section: News.

B. Chakraborty and S.A. Murphy. Dynamic Treatment Regimes. *Annual Review of Statistics and Its Application*, 1:447–464, 2014. ISSN 2326-8298. doi: 10.1146/annurev-statistics-022513-115553.

A. Chappell, J. Sanderman, M. Thomas, A. Read, and C. Leslie. The dynamics of soil redistribution and the implications for soil organic carbon accounting in agricultural south-eastern australia. *Global Change Biology*, 18:2081–2088, 2012. doi: 10.1111/j.1365-2486.2012.02682.x.

A. Chatterjee, R. Lal, L. Wielopolski, M. Z. Martin, and M. H. Ebinger. Evaluation of Different Soil Carbon Determination Methods. *Critical Reviews in Plant Sciences*, 28(3):164–178, April 2009a. ISSN 0735-2689. doi: 10.1080/07352680902776556. URL `https://doi.org/10.1080/07352680902776556`.

A. Chatterjee, R. Lal, L. Wielopolski, M.Z. Martin, and M.H. Ebinger. Evaluation of different soil carbon determination methods. *CRC Critical Reviews in Plant Sciences*, 28:164–178, 2009b. doi: 10.1080/07352680902776556.

C. Chenu, D.A. Angers, P. Barré, D. Derrien, D. Arrouays, and J. Balesdent. Increasing organic stocks in agricultural soils: Knowledge gaps and potential innovations. *Soil and Tillage Research*, 188:41–52, May 2019. ISSN 0167-1987. doi: 10.1016/j.still.2018.04.011. URL `https://www.sciencedirect.com/science/article/pii/S0167198718303738`.

B. Cho, K. Gan, and N. Kallus. Peeking with PEAK: Sequential, Nonparametric Composite Hypothesis Tests for Means of Multiple Data Streams, February 2024. URL `http://arxiv.org/abs/2402.06122`. arXiv:2402.06122 [cs, stat].

W.G. Cochran. *Sampling Techniques*. John Wiley & Sons, Inc., New York, 3rd edition, 1977.

S.R. Cole and E.A. Stuart. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, 172 (1):107–115, July 2010. ISSN 1476-6256. doi: 10.1093/aje/kwq084.

R.T. Conant, C.E.P. Cerri, B.B. Osborne, and K. Paustian. Grassland management impacts on soil carbon stocks: a new synthesis. *Ecological Applications*, 27:662–668, 2017. doi: 10.1002/eap.1473.

M.F. Cotrufo, M.G. Ranalli, M.L. Haddix, J. Six, and E. Lugato. Soil carbon storage informed by particulate and mineral-associated organic matter. *Nature Geoscience*, 12(12):989–994, December 2019. ISSN 1752-0908. doi: 10.1038/s41561-019-0484-6. URL `https://www.nature.com/articles/s41561-019-0484-6`. Publisher: Nature Publishing Group.

A.A. Cournot. *Exposition of the Theory of Chances and Probabilities*. NG Verlag, 1843. URL `https://arxiv.org/pdf/1902.02781`. Translated by Oscar Sheynin (2013).

T.M. Cover and J.A. Thomas. *Elements of Information Theory, 2nd Edition*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, 2006. ISBN 0471241954. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471241954`.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975. ISSN 00063444. URL `http://www.jstor.org/stable/2335362`.

I.J. Dahabreh, S.E. Robertson, E.J. Tchetgen, E.A. Stuart, and M.A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, June 2019. ISSN 1541-0420. doi: 10.1111/biom.13009.

F.N. David. Studies in the history of probability and statistics i. dicing and gaming (a note on the history of probability). *Biometrika*, 42:1–15, 1955.

M. Davis, B. Alves, D. Karlen, K. Kline, M. Galdos, and D. Abulebdeh. Review of soil organic carbon measurement protocols: A us and brazil comparison and recommendation. *Sustainability*, 10:53, 2017. doi: 10.3390/su10010053.

J.J. de Gruijter, D.J. Brus, M.F.P. Bierkens, and M. Knotters. *Sampling for Natural Resource Monitoring*. Springer-Verlag, Berlin Heidelberg, 2006. ISBN 978-3-540-22486-0. doi: 10.1007/3-540-33161-1. URL `https://www.springer.com/gp/book/9783540224860`.

J.J. de Gruijter, A.B. McBratney, B. Minasny, I. Wheeler, B.P. Malone, and U. Stockmann. Farm-scale soil carbon auditing. *Geoderma*, 265:120–130, 2016. doi: 10.1016/j.geoderma.2015.11.010.

B. De Vos, B. Vandecasteele, J. Deckers, and B. Muys. Capability of Loss-on-Ignition as a Predictor of Total Organic Carbon in Non-Calcareous Forest Soils. *Communications in Soil Science and Plant Analysis*, 36(19-20):2899–2921, October 2005. ISSN 0010-3624. doi: 10.1080/00103620500306080. URL `https://doi.org/10.1080/00103620500306080`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00103620500306080.

S.M. Devine, A.T. O'Geen, H. Liu, Y. Jin, H.E. Dahlke, R.E. Larsen, and R.A. Dahlgren. Terrain attributes and forage productivity predict catchment-scale soil organic carbon stocks. *Geoderma*, 368:114286, 2020. doi: 10.1016/j.geoderma.2020.114286.

P. Diaconis and F. Mosteller. Methods for studying coincidences. *Journal of the American Statistical Association*, 84(408):853–861, 1989. doi: 10.1080/01621459.1989.10478847. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478847.

P. Diggle and P.J. Ribeiro. *Model-based Geostatistics*. Springer Series in Statistics. Springer-Verlag, New York, 2007. ISBN 978-0-387-32907-9. doi: 10.1007/978-0-387-48536-2. URL https://www.springer.com/gp/book/9780387329079.

P. Ding, A. Feller, and L. Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(3):655–671, 2016. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/24775356.

P. Ding, A. Feller, and L. Miratrix. Decomposing Treatment Effect Variation. *Journal of the American Statistical Association*, 114(525):304–317, January 2019. ISSN 0162-1459. doi: 10.1080/01621459.2017.1407322. URL https://doi.org/10.1080/01621459.2017.1407322. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1407322.

B. Efron and D. Feldman. Compliance as an Explanatory Variable in Clinical Trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991. ISSN 0162-1459. doi: 10.2307/2289707. URL https://www.jstor.org/stable/2289707. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

N. Egami and E. Hartman. Elements of External Validity: Framework, Design, and Analysis. *American Political Science Review*, 117(3):1070–1088, August 2023. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055422000880. URL https://www.cambridge.org/core/journals/american-political-science-review/article/elements-of-external-validity-framework-design-and-analysis/2D0914404C84B3F169732FF1D5E39420.

A. Ek, P.B. Stark, P.J. Stuckey, and D. Vukcevic. *Adaptively Weighted Audits of Instant-Runoff Voting Elections: AWAIRE*, page 35–51. Springer Nature Switzerland, 2023. ISBN 9783031437564. doi: 10.1007/978-3-031-43756-4_3. URL http://dx.doi.org/10.1007/978-3-031-43756-4_3.

B.H. Ellert, H.H. Janzen, and T. Entz. Assessment of a method to measure temporal change in soil carbon storage. *Soil Science Society of America Journal*, 66:1687–1695, 2002. doi: 10.2136/sssaj2002.1687.

J.R. England and R.A. Viscarra Rossel. Proximal sensing for soil carbon accounting. *SOIL*, 4(2):101–122, May 2018. ISSN 2199-3971. doi: https://doi.org/10.5194/soil-4-101-2018. URL `https://www.soil-journal.net/4/101/2018/`.

FAO. Measuring and modelling soil carbon stocks and stock changes in livestock production systems: Guidelines for assessment, 2020. URL `http://www.fao.org/3/ca2934en/CA2934EN.pdf`.

S.E. Fienberg and J.M. Tanur. Reconsidering the Fundamental Contributions of Fisher and Neyman on Experimentation and Sampling. *International Statistical Review / Revue Internationale de Statistique*, 64(3):237–253, 1996. ISSN 0306-7734. doi: 10.2307/1403784. URL `https://www.jstor.org/stable/1403784`. Publisher: [Wiley, International Statistical Institute (ISI)].

S.E. Fienberg, J. Neter, and R.A. Leitch. Estimating total overstatement error in accounting populations. *J. Am. Stat. Assoc.*, 72:295–302, 1977.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922. ISSN 02643952. URL `http://www.jstor.org/stable/91208`.

R.A. Fisher. *Statistical methods for research workers.* Oliver and Boyd, 1925.

R.A. Fisher. *The design of experiments.* Oliver and Boyd, 1935.

A. Franzluebbers. Soil organic carbon sequestration and agricultural greenhouse gas emissions in the southeastern usa. *Soil & Tillage Research*, 83:120–147, 2005. doi: 10.1016/j.still.2005.02.012.

D.A. Freedman. On regression adjustments in experiments with several treatments. *Annals of Applied Statistics*, 2(1):176–196, March 2008a. ISSN 1932-6157, 1941-7330. doi: 10.1214/07-AOAS143. URL `https://projecteuclid.org/euclid.aoas/1206367817`. Publisher: Institute of Mathematical Statistics.

D.A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, February 2008b. ISSN 0196-8858. doi: 10.1016/j.aam.2006.12.003. URL `https://www.sciencedirect.com/science/article/pii/S019688580700005X`.

K. Georgiou, R.B. Jackson, O. Vindušková, R.Z. Abramoff, A. Ahlström, W. Feng, J.W. Harden, A.F.A. Pellegrini, H.W. Polley, J.L. Soong, W.J. Riley, and M.S. Torn. Global stocks and capacity of mineral-associated soil organic carbon. *Nature Communications*, 13(1):3797, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31540-9. URL `https://www.nature.com/articles/s41467-022-31540-9`. Publisher: Nature Publishing Group.

G. Gigerenzer, Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Krüger. *The Empire of Chance*. Cambridge University Press, 1989.

Amanda K. Glazer, Jacob V. Spertus, and Philip B. Stark. More Style, Less Work: Card-style Data Decrease Risk-limiting Audit Sample Sizes. *Digital Threats: Research and Practice*, 2(4):1–15, December 2021. ISSN 2692-1626, 2576-5337. doi: 10.1145/3457907. URL `https://dl.acm.org/doi/10.1145/3457907`.

E. Goidts, B. Van Wesemael, and M. Crucifix. Magnitude and sources of uncertainties in soil organic carbon (soc) stock assessments at various scales. *European Journal of Soil Science*, 60:723–739, 2009. doi: 10.1111/j.1365-2389.2009.01157.x.

P.I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer-Verlag, New York, 3 edition, 2005. ISBN 978-0-387-20279-2. doi: 10.1007/b138696. URL `https://www.springer.com/gp/book/9780387202792`.

E. Gorham. Northern Peatlands: Role in the Carbon Cycle and Probable Responses to Climatic Warming. *Ecological Applications*, 1(2):182–195, 1991. ISSN 1939-5582. doi: 10.2307/1941811. URL `https://onlinelibrary.wiley.com/doi/abs/10.2307/1941811`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/1941811.

H. Gosnell, N. Robinson-Maness, and S. Charnley. Profiting from the sale of carbon offsets: A case study of the trigg ranch. *Rangelands*, 33:25–29, 2011. doi: 10.2111/1551-501X-33.5.25.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society, B: Statistical Methodology*, in press, 2023.

I. Hacking. *The Emergence of Probability*. Cambridge University Press, Cambridge, second edition, 2006.

J.L. Hall, L.W. Miratrix, P.B. Stark, M. Briones, E. Ginnold, F. Oakley, M. Peaden, G. Pellerin, T. Stanionis, and T. Webber. Implementing risk-limiting post-election audits in California. In *Proc. 2009 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09)*, Montreal, Canada, August 2009. USENIX. URL `http://www.usenix.org/event/evtwote09/tech/full_papers/hall.pdf`.

M.H. Hansen, W.N. Hurwitz, and W.G. Madow. *Sample survey methods and theory, volume 1: method and applications*. Wiley, 1953.

J. Hassink. The capacity of soils to preserve organic C and N by their association with clay and silt particles. *Plant and Soil*, 191(1):77–87, April 1997. ISSN 1573-5036. doi: 10.1023/A:1004213929699. URL `https://doi.org/10.1023/A:1004213929699`.

M.J. Higgins, R.L. Rivest, and P.B. Stark. Sharper p-values for stratified post-election audits. *Statistics, Politics, and Policy*, 2(1), 2011. URL `http://www.bepress.com/spp/vol2/iss1/7`.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.*, 58:13–30, 1963.

T. Holland-Letz and A. Kopp-Schneider. Optimal experimental designs for dose–response studies with continuous endpoints. *Archives of Toxicology*, 89 (11):2059–2068, 2015. ISSN 0340-5761. doi: 10.1007/s00204-014-1335-2. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4655015/`.

P.S. Homann, P. Sollins, M. Fiorella, T. Thorson, and J.S. Kern. Regional soil organic carbon storage estimates for western oregon by multiple approaches. *Soil Science Society of America Journal*, 62:789–796, 1998. doi: 10.2136/sssaj1998.03615995006200030036x.

L. Howard, R.L. Rivest, and P.B. Stark. A review of robust post-election audits: Various methods of risk-limiting audits and Bayesian audits. Technical report, Brennan Center for Justice, 2019. `https://www.brennancenter.org/sites/default/files/2019-11/2019_011_RLA_Analysis_FINAL_0.pdf`.

S.R. Howard, A. Ramdas, J. McAuliffe, and J.S. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), apr 2021. doi: 10.1214/20-aos1991. URL `https://doi.org/10.1214%2F20-aos1991`.

J. Hsu. *Multiple Comparisons: Theory and Methods*. Chapman and Hall, London, 1996.

S.H. Hurlbert. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*, 54(2):187–211, 1984. ISSN 0012-9615. doi: 10.2307/1942661. URL https://www.jstor.org/stable/1942661. Publisher: Ecological Society of America.

G.W. Imbens and D.B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. URL https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB.

Indigo Agriculture. Additionality. https://www.indigoag.com, 2024. Accessed: 2024-04-19.

International Energy Agency. Net Zero Roadmap: A Global Pathway to Keep the 1.5 °C Goal in Reach – Analysis, September 2023. URL https://www.iea.org/reports/net-zero-roadmap-a-global-pathway-to-keep-the-15-0c-goal-in-reach.

A.A. Jackson Hammond, M. Motew, C.D. Brummitt, M.L. DuBuisson, G. Pinjuv, D.V. Harburg, E.E. Campbell, A.A. Kumar, P. Stanley, et al. Implementing the soil enrichment protocol at scale: opportunities for an agricultural carbon market. *Frontiers in Climate*, 3, 2021. doi: 10.3389/fclim.2021.686440. URL https://doi.org/10.3389/fclim.2021.686440.

R. Jandl, M. Rodeghiero, C. Martinez, M.F. Cotrufo, F. Bampa, B. van Wesemael, R.B. Harrison, I.A. Guerrini, D.D. Richter, L. Rustad, K. Lorenz, A. Chabbi, and F. Miglietta. Current status, uncertainty and future needs in soil organic carbon monitoring. *Science of The Total Environment*, 468–469:376–383, 2014. doi: 10.1016/j.scitotenv.2013.08.026.

J.F.W. Johnston. *Experimental Agriculture: Being The Results Of Past, And Suggestions For Future Experiments In Scientific And Practical Agriculture.* William Blackwell and Sons, 1849. ISBN 978-1-4368-4219-8.

D.L. Jones, J. Rousk, G. Edwards-Jones, T.H. DeLuca, and D.V. Murphy. Biochar-mediated changes in soil quality and plant growth in a three-year field trial. *Soil Biology and Biochemistry*, 45:113–124, 2012. doi: 10.1016/j.soilbio.2011.10.012.

C.G. Jung. *Synchronicity: An Acausal Connecting Principle.* Princeton University Press, Princeton, 2010.

C. Kahn. Half of Republicans say Biden won because of a 'rigged' election: Reuters/Ipsos poll. *Reuters*, November 2020. URL `https://www.reuters.com/article/us-usa-election-poll-idUSKBN27Y1AJ`.

H.M. Kaplan. A method of one-sided nonparametric inference for the mean of a nonnegative population. *The American Statistician*, 41:157–158, 1987.

N. Karmarkar. A new polynomial time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

M. Kasy and A. Sautmann. Adaptive Treatment Assignment in Experiments for Policy Choice. *Econometrica*, 89(1):113–132, 2021. ISSN 1468-0262. doi: 10.3982/ECTA17527. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17527`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17527.

J. L. Kelly Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926, 1956. doi: https://doi.org/10.1002/j.1538-7305.1956.tb03809.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1956.tb03809.x`.

L. Kish. *Survey Sampling*. John Wiley & Sons, Inc., New York, 1965.

K. Kosmelj, A. Cedilnik, and P. Kalan. Comparison of a two-stage sampling design and its composite sample alternative: An application to soil studies. *Environmental and Ecological Statistics*, 8(2):109–119, June 2001. ISSN 1573-3009. doi: 10.1023/A:1011378431085. URL `https://doi.org/10.1023/A:1011378431085`.

A.N. Kravchenko and G.P. Robertson. Whole-profile soil carbon stocks: The danger of assuming too much from analyses of too little. *Soil Science Society of America Journal*, 75(1):235–240, 2011.

D.G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, December 1951. ISSN 0038-223X. URL `https://journals.co.za/content/saimm/52/6/AJA0038223X_4792`. Publisher: Southern African Institute of Mining and Metallurgy.

Y. Kuzyakov and E. Blagodatskaya. Microbial hotspots and hot moments in soil: Concept & review. *Soil Biology and Biochemistry*, 83:184–199, 2015. doi: 10.1016/j.soilbio.2015.01.025.

S.R. Künzel, J.S. Sekhon, P.J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, March 2019. doi: 10.1073/pnas.1804597116. URL `https://www.pnas.org/doi/full/10.1073/pnas.1804597116`. Publisher: Proceedings of the National Academy of Sciences.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL `https://www.sciencedirect.com/science/article/pii/0196885885900028`.

R. Lal. Regenerative agriculture for food and climate. *Journal of Soil and Water Conservation*, 75(5):123A–124A, September 2020. ISSN 0022-4561, 1941-3300. doi: 10.2489/jswc.2020.0620A. URL `https://www.jswconline.org/content/75/5/123A`. Publisher: Soil and Water Conservation Society Section: A Section.

R. Lal and B. A. Stewart. *Soil and Climate*. CRC Press, September 2018. ISBN 978-0-429-48726-2. doi: 10.1201/b21225. URL `https://www-taylorfrancis-com.libproxy.berkeley.edu/books/e/9780429487262`.

R.M. Lark. Some considerations on aggregate sample supports for soil inventory and monitoring. *European Journal of Soil Science*, 63(1):86–95, 2012. ISSN 1365-2389. doi: 10.1111/j.1365-2389.2011.01415.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2389.2011.01415.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2389.2011.01415.x.

C. Le Quéré, R.M. Andrew, P. Friedlingstein, S. Sitch, J. Hauck, J. Pongratz, P.A. Pickers, J.I. Korsbakken, G.P. Peters, J.G. Canadell, A. Arneth, V.K. Arora, L. Barbero, A. Bastos, L. Bopp, F. Chevallier, L.P. Chini, P. Ciais, S.C. Doney, T. Gkritzalis, D.S. Goll, I. Harris, V. Haverd, F.M. Hoffman, M. Hoppema, R.A. Houghton, G. Hurtt, T. Ilyina, A.K. Jain, T. Johannessen, C.D. Jones, E. Kato, R.F. Keeling, K.K. Goldewijk, P. Landschützer, N. Lefèvre, S. Lienert, Z. Liu, D. Lombardozzi, N. Metzl, D.R. Munro, J.E.M.S. Nabel, S. Nakaoka, C. Neill, A. Olsen, T. Ono, P. Patra, A. Peregon, W. Peters, P. Peylin, B. Pfeil, D. Pierrot, B. Poulter, G. Rehder, L. Resplandy, E. Robertson, M. Rocher, C. Rödenbeck, U. Schuster, J. Schwinger, R. Séférian, I. Skjelvan, T. Steinhoff, A. Sutton, P.P. Tans, H. Tian, B. Tilbrook, F.N. Tubiello, I.T. van der Laan-Luijkx, G.R. van der Werf, N. Viovy, A.P. Walker, A.J. Wiltshire, R. Wright, S. Zaehle, and B. Zheng. Global Carbon Budget 2018. *Earth System Science Data*, 10(4):2141–2194, December 2018. ISSN 1866-3508. doi: 10.5194/essd-10-2141-2018. URL https://essd.copernicus.org/articles/10/2141/2018/. Publisher: Copernicus GmbH.

E. Learned-Miller and P. Thomas. A new confidence interval for the mean of a bounded random variable. *University of Massachusetts*, 2019.

E. L. Lehmann. The fisher, neyman-pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249, 1993. ISSN 01621459. URL http://www.jstor.org/stable/2291263.

E.L. Lehmann. *Elements of large-sample theory*. Springer, 1999.

E.L. Lehmann. *Fisher, Neyman, and the creation of classical statistics*. Springer Science + Business Media, 2011.

E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.

J. Lehmann, J. Kinyangi, and D. Solomon. Organic matter stabilization in soil microaggregates: implications from spatial heterogeneity of organic carbon contents and carbon forms. *Biogeochemistry*, 85:45–57, 2007. doi: 10.1007/s10533-007-9105-3.

A. Levine. Donald Trump's Favorite Voting Machines, September 2020. URL http://washingtonmonthly.com/2020/09/23/donald-trumps-favorite-voting-machines/.

X. Li and P. Ding. Rerandomization and Regression Adjustment. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):241–268, February 2020. ISSN 1369-7412. doi: 10.1111/rssb.12353. URL https://doi.org/10.1111/rssb.12353.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295 – 318, 2013. doi: 10.1214/12-AOAS583. URL https://doi.org/10.1214/12-AOAS583.

M. Lindeman and P.B. Stark. A gentle introduction to risk-limiting audits. *IEEE Security and Privacy*, 10:42–49, 2012.

Z. Luo, E. Wang, and O.J. Sun. Can no-tillage stimulate carbon sequestration in agricultural soils? A meta-analysis of paired experiments. *Agriculture, Ecosystems & Environment*, 139(1):224–231, October 2010. ISSN 0167-8809.

doi: 10.1016/j.agee.2010.08.006. URL http://www.sciencedirect.com/science/article/pii/S0167880910002094.

D.A. MacKenzie. *Statistics in Britain, 1865-1930: The Social Construction of Scientific Knowledge*. Edinburgh University Press, 1981.

S. Majumder, S. Neogi, T. Dutta, M.A. Powel, and P. Banik. The impact of biochar on soil carbon sequestration: Meta-analytical approach to evaluating environmental and economic advantages. *Journal of Environmental Management*, 250:109466, 2019. doi: 10.1016/j.jenvman.2019.109466.

R. Marcus, E. Peritz, and K.R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335748.

E. Marin-Spiotta, N.T. Chaopricha, A.F. Plante, A.F. Diefendorf, C.W. Mueller, A.S. Grandy, and J.A. Mason. Long-term stabilization of deep soil carbon by fire and burial during early holocene climate change. *Nature Geoscience*, 7:428–432, 2014. doi: 10.1038/ngeo2169.

C. Mathers, C.K. Black, B.D. Segal, R.B. Gurung, Y. Zhang, M.J. Easter, S. Williams, M. Motew, E.E. Campbell, C.D. Brummitt, K. Paustian, and A.A. Kumar. Validating DayCent-CR for cropland soil carbon offset reporting at a national scale. *Geoderma*, 438:116647, October 2023. ISSN 0016-7061. doi: 10.1016/j.geoderma.2023.116647. URL https://www.sciencedirect.com/science/article/pii/S0016706123003245.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization, 2009.

B.A. Miller, S. Koszinski, W. Hierold, H. Rogasik, B. Schroder, K. Van Oost, M. Wehrhan, and M. Sommer. Towards mapping soil carbon landscapes: Issues of sampling scale and transferability. *Soil and Tillage Research*, 156:194–208, 2016. doi: 10.1016/j.still.2015.07.004.

K. Minami. Methane from rice production. *Fertilizer research*, 37(3):167–179, October 1994. ISSN 1573-0867. doi: 10.1007/BF00748935. URL https://doi.org/10.1007/BF00748935.

B. Minasny, B.P. Malone, A.B. McBratney, et al. Soil carbon 4 per mille. *Geoderma*, 292:59–86, 2017.

C. Natali, G. Bianchini, and P. Carlino. Thermal stability of soil carbon pools: Inferences on soil nature and evolution. *Thermochimica Acta*, 683:178478, 2020. doi: 10.1016/j.tca.2019.178478.

National Academies of Sciences, Engineering, and Medicine. *Securing the Vote: Protecting American Democracy.* The National Academies Press, Washington, DC, 2018. ISBN 978-0-309-47647-8. doi: 10.17226/25120. URL `https://www.nap.edu/catalog/25120/securing-the-vote-protecting-american-democracy`.

A.K. Nayak, M.M. Rahman, R. Naidu, B. Dhal, C.K. Swain, A.D. Nayak, R. Tripathi, M. Shahid, M.R. Islam, and H. Pathak. Current and emerging methodologies for estimating carbon sequestration in agricultural soils: A review. *The Science of the Total Environment*, 665:890–912, May 2019. ISSN 1879-1026. doi: 10.1016/j.scitotenv.2019.02.125.

M. Necpalova, R.P. Anex, A.N. Kravchenko, L.J. Abendroth, S.J. Del Grosso, W.A. Dick, M.J. Helmers, D. Herzmann, J.G. Lauer, E.D. Nafziger, et al. What does it take to detect a change in soil carbon stock? a regional comparison of minimum detectable difference and experiment duration in the north central united states. *Journal of Soil and Water Conservation*, 69:517–531, 2014. doi: 10.2489/jswc.69.6.517.

M. Necpálová, R.P. Anex, M.N. Fienen, S.J. Del Grosso, M.J. Castellano, J.E. Sawyer, J. Iqbal, J.L. Pantoja, and D.W. Barker. Understanding the DayCent model: Calibration, sensitivity, and identifiability through inverse modeling. *Environmental Modelling & Software*, 66:110–130, 2015. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2014.12.011. URL `https://www.sciencedirect.com/science/article/pii/S1364815214003685`.

J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczki*, 10:1–51, 1923.

J. Neyman. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934. ISSN 0952-8385. doi: 10.2307/2342192. URL `https://www.jstor.org/stable/2342192`. Publisher: [Wiley, Royal Statistical Society].

J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. Ser. A*, 231:289–337, 1933. ISSN 02643952. URL `http://www.jstor.org/stable/91247`.

T.Q. Nguyen, C. Ebnesajjad, S.R. Cole, and E.A. Stuart. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, 11(1):225–247, March 2017. ISSN 1932-6157, 1941-7330. doi: 10.1214/16-AOAS1001. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-11/issue-1/Sensitivity-analysis-for-an-unobserved-moderator-in-RCT-to-target/10.1214/16-AOAS1001.full. Publisher: Institute of Mathematical Statistics.

N. Nitta. Rice farming to restore soil. https://www.earthisland.org/journal/index.php/articles/entry/rice-farming-to-restore-soil/#, 2022. Accessed: 2024-04-19.

E.E. Oldfield, A.J. Eagle, R.L. Rudek, J. Sanderman, and D.R. Gordon. *Agricultural soil carbon credits: Making sense of protocols for carbon sequestration and net greenhouse gas removals*. Environmental Defense Fund, New York, New York, 2021.

F. Orabona and K.S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio, 2022.

K. Ottoboni, P.B. Stark, M. Lindeman, and N. McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (SUITE). In *Electronic Voting. E-Vote-ID 2018. Lecture Notes in Computer Science*. Springer, 2018. https://link.springer.com/chapter/10.1007/978-3-030-00419-4_12.

S.M. O'Rourke and N.M. Holden. Optical sensing and chemometric analysis of soil organic carbon – a cost effective alternative to conventional laboratory methods? *Soil Use and Management*, 27(2):143–155, 2011. ISSN 1475-2743. doi: 10.1111/j.1475-2743.2011.00337.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-2743.2011.00337.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-2743.2011.00337.x.

J. Padarian, B. Minasny, and A. B. McBratney. Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional*, 16:e00198, March 2019. ISSN 2352-0094. doi: 10.1016/j.geodrs.2018.e00198. URL http://www.sciencedirect.com/science/article/pii/S2352009418302785.

P. Pallmann, A.W. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight, L.V. Hampson, J. Holmes, A.P. Mander, L. Odondi, M.R. Sydes, S.S. Villar, J.M.S. Wason, C.J. Weir, G.M. Wheeler, C. Yap, and T. Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1):29, February 2018. ISSN 1741-7015. doi: 10.1186/s12916-018-1017-7. URL `https://doi.org/10.1186/s12916-018-1017-7`.

Panel on Nonstandard Mixtures of Distributions. *Statistical models and analysis in auditing: A study of statistical models and methods for analyzing nonstandard mixtures of distributions in auditing.* National Academy Press, Washington, D.C., 1988.

W. J. Parton. The CENTURY model. In David S. Powlson, Pete Smith, and Jo U. Smith, editors, *Evaluation of Soil Organic Matter Models*, pages 283–291, Berlin, Heidelberg, 1996. Springer. ISBN 978-3-642-61094-3. doi: 10.1007/978-3-642-61094-3_23.

G.P. Patil, S.D. Gore, and C. Taillie. *Composite Sampling: A Novel Method to Accomplish Observational Economy in Environmental Studies.* Environmental and Ecological Statistics. Springer US, 2011. ISBN 978-1-4419-7627-7. doi: 10.1007/978-1-4419-7628-4. URL `https://www.springer.com/gp/book/9781441976277`.

K. Pearson. Mathematical contributions to the theory of evolution iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187, 1895.

F. Pesarin and L. Salmaso. *Permutation tests for complex data: Theory, applications, and software.* John Wiley and Sons, Ltd., West Sussex, UK, 2010a.

F. Pesarin and L. Salmaso. The permutation testing approach: a review. *Statistica*, 70 (4):481–509, December 2010b. ISSN 1973-2201. doi: 10.6092/issn.1973-2201/3599. URL `https://rivista-statistica.unibo.it/article/view/3599`. Number: 4.

K. Pilegaard. Processes regulating nitric oxide emissions from soils. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1621): 20130126, July 2013. ISSN 0962-8436. doi: 10.1098/rstb.2013.0126. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3682746/`.

H.J. Poffenbarger, D.C. Olk, C. Cambardella, J. Kersey, M. Liebman, A. Mallarino, J. Six, and M.J. Castellano. Whole-profile soil organic matter content, composition, and stability under cropping systems that differ in belowground inputs. *Agriculture, Ecosystems & Environment*, 291:106810, 2020. doi: 10.1016/j.agee.2019.106810.

W.M. Post, R.C. Izaurralde, L.K. Mann, and N. Bliss. Monitoring and verifying changes of organic carbon in soil. In *Storing Carbon in Agricultural Soils: A Multi-Purpose Environmental Strategy*, pages 73–99. Springer, Netherlands, Dordrecht, 2001. doi: 10.1007/978-94-017-3089-1_4.

E. Potash, K. Guan, A.J. Margenot, D.K. Lee, A. Boe, M. Douglass, E. Heaton, C. Jang, V. Jin, N. Li, R. Mitchell, N. Namoi, M. Schmer, S. Wang, and C. Zumpf. Multi-site evaluation of stratified and balanced sampling of soil organic carbon stocks in agricultural fields. *Geoderma*, 438:116587, 2023. ISSN 0016-7061. doi: https://doi.org/10.1016/j.geoderma.2023.116587. URL `https://www.sciencedirect.com/science/article/pii/S0016706123002641`.

M.J. Pringle, D.E. Allen, R.C. Dalal, J.E. Payne, D.G. Mayer, P. O'Reagain, and B.P. Marchant. Soil carbon stock in the tropical rangelands of Australia: Effects of soil type and grazing pressure, and determination of sampling requirement. *Geoderma*, 167-168:261–273, November 2011. ISSN 0016-7061. doi: 10.1016/j.geoderma.2011.09.001. URL `http://www.sciencedirect.com/science/article/pii/S0016706111002552`.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576 – 601, 2023. doi: 10.1214/23-STS894. URL `https://doi.org/10.1214/23-STS894`.

J.B. Reeves. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma*, 158(1):3–14, August 2010. ISSN 0016-7061. doi: 10.1016/j.geoderma.2009.04.005. URL `http://www.sciencedirect.com/science/article/pii/S0016706109001220`.

H.E. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

G.P. Robertson, K.M. Klingensmith, M.J. Klug, E.A. Paul, J.R. Crum, and B.G. Ellis. Soil resources, microbial activity, and primary production across an agricultural ecosystem. *Ecological Applications*, 7:158–170, 1997. doi: 10.1890/1051-0761(1997)007[0158:SRMAAP]2.0.CO;2.

J.P. Romano and M. Wolf. Finite sample nonparametric inference and large sample efficiency. *The Annals of Statistics*, 28(3):756 – 778, 2000. doi: 10.1214/aos/1015951997.

P.R. Rosenbaum. *Observational Studies.* Springer, 2002. URL https://link.springer.com/book/10.1007/978-1-4757-3692-2.

R. Ryals and W.L. Silver. Effects of organic matter amendments on net primary productivity and greenhouse gas emissions in annual grasslands. *Ecological Applications: A Publication of the Ecological Society of America*, 23(1):46–59, January 2013. ISSN 1051-0761.

R. Ryals, M. Kaiser, M.S. Torn, A.A. Berhe, and W.L. Silver. Impacts of organic matter amendments on carbon and nitrogen dynamics in grassland soils. *Soil Biology and Biochemistry*, 68:52–61, January 2014. ISSN 0038-0717. doi: 10.1016/j.soilbio.2013.09.011. URL http://www.sciencedirect.com/science/article/pii/S003807171300312X.

N.P.A. Saby, P.H. Bellamy, X. Morvan, D. Arrouays, R.J.A. Jones, F.G.A. Verheijen, M.G. Kibblewhite, A.N.N. Verdoodt, J.B. Üveges, A. Freudenschuß, et al. Will european soil-monitoring networks be able to detect changes in topsoil organic carbon content? *Global Change Biology*, 14:2432–2442, 2008. doi: 10.1111/j.1365-2486.2008.01658.x.

A. Saltelli and M. Giampietro. What is wrong with evidence based policy, and how can it be improved? *Futures*, 91:62–71, 2017.

J. Sanderman and J.A. Baldock. Accounting for soil carbon sequestration in national inventories: a soil scientist's perspective. *Environmental Research Letters*, 5(3): 034003, 2010. doi: 10.1088/1748-9326/5/3/034003.

J. Sanderman, T. Hengl, and G.J. Fiske. Soil carbon debt of 12,000 years of human land use. *Proceedings of the National Academy of Sciences*, 114(36):9575–9580, September 2017. doi: 10.1073/pnas.1706103114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1706103114. Publisher: Proceedings of the National Academy of Sciences.

G.R. Sanford, J.L. Posner, R.D. Jackson, C.J. Kucharik, J.L. Hedtcke, and T. Lin. Soil carbon lost from Mollisols of the North Central U.S.A. with 20 years of agricultural best management practices. *Agriculture, Ecosystems & Environment*, 162:68–76, November 2012. ISSN 0167-8809. doi: 10.1016/j.agee.2012.08.011. URL https://www.sciencedirect.com/science/article/pii/S0167880912003222.

L.J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46(253):55–67, 1951. ISSN 01621459. URL http://www.jstor.org/stable/2280094.

J.C. Scott. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT London, 0 edition edition, February 1999. ISBN 978-0-300-07815-2.

G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184: 407–431, 04 2021. doi: 10.1111/rssa.12647.

G. Shafer and V. Vovk. *Game theoretic foundations for probability and finance*. Wiley Series in Probability and Statistics, 2019.

M.J. Silvapulle. A test in the presence of nuisance parameters. *Journal of the American Statistical Association*, 91(436):1690–1693, 1996. doi: 10.1080/01621459.1996.10476739.

W. Silver, S. Vergara, and A. Mayer. Carbon sequestration and greenhouse gas mitigation potential of composting and soil amendments on california's rangelands. *California's Fourth Climate Change Assessment*, 2018. URL `https://www.energy.ca.gov/media/2059`.

W.L. Silver, R. Ryals, and V. Eviner. Soil carbon pools in california's annual grassland ecosystems. *Rangeland Ecology & Management*, 63:128–136, 2010. doi: 10.2111/REM-D-09-00106.1.

J. Six, S. Doetterl, M. Laub, C.R. Müller, and M. Van de Broek. The six rights of how and when to test for soil C saturation. *SOIL*, 10(1):275–279, April 2024. ISSN 2199-3971. doi: 10.5194/soil-10-275-2024. URL `https://soil.copernicus.org/articles/10/275/2024/`. Publisher: Copernicus GmbH.

E.W. Slessarev, J. Zelikova, J. Hamman, D. Cullenward, and J. Freeman. Depth matters for soil carbon accounting. *CarbonPlan*, 2021. URL `https://carbonplan.org/research/soil-depth-sampling`.

E.W. Slessarev, A. Mayer, C. Kelly, K. Georgiou, J. Pett-Ridge, and E.E. Nuccio. Initial soil organic carbon stocks govern changes in soil carbon: Reality or artifact? *Global Change Biology*, 29(5):1239–1247, 2023. ISSN 1365-2486. doi: 10.1111/gcb.16491. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.16491`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.16491.

A. Slivkins. Introduction to Multi-Armed Bandits, April 2024. URL `http://arxiv.org/abs/1904.07272`. arXiv:1904.07272 [cs, stat].

P. Smith. An overview of the permanence of soil organic carbon stocks: influence of direct human-induced, indirect and natural effects. *European Journal of Soil Science*, 56(5):673–680, 2005. ISSN 1365-2389. doi: 10.1111/j.1365-2389.2005.00708.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2389.2005.00708.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2389.2005.00708.x.

P. Smith, J.F. Soussana, D. Angers, L. Schipper, C. Chenu, D.P. Rasse, N.H. Batjes, F. van Egmond, S. McNeill, M. Kuhnert, C. Arias-Navarro, J.E. Olesen, N. Chirinda, D. Fornara, E. Wollenberg, J. Álvaro-Fuentes, A. Sanz-Cobena, and K. Klumpp. How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, 26(1):219–241, 2020. ISSN 1365-2486. doi: 10.1111/gcb.14815. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14815`.

J.V. Spertus. Optimal sampling and assay for estimating soil organic carbon. *Open Journal of Soil Science*, 11:93–121, 2021. doi: 10.4236/ojss.2021.112006.

J.V. Spertus and P.B. Stark. Sweeter than SUITE: Supermartingale Stratified Union-Intersection Tests of Elections. In Robert Krimmer, Melanie Volkamer, David Duenas-Cid, Peter Rønne, and Micha Germann, editors, *Electronic Voting*, pages 106–121, Cham, 2022. Springer International Publishing. ISBN 978-3-031-15911-4.

P. Stanley, J.V. Spertus, J. Chiartas, P.B. Stark, and T. Bowles. Valid inferences about soil carbon in heterogeneous landscapes. *Geoderma*, 430:116323, February 2023. ISSN 0016-7061. doi: 10.1016/j.geoderma.2022.116323. URL `https://www.sciencedirect.com/science/article/pii/S0016706122006309`.

P.B. Stark. Conservative statistical post-election audits. *Ann. Appl. Stat.*, 2:550–581, 2008a. URL `http://arxiv.org/abs/0807.4005`.

P.B. Stark. A sharper discrepancy measure for post-election audits. *Ann. Appl. Stat.*, 2:982–985, 2008b. URL `http://arxiv.org/abs/0811.1697`.

P.B. Stark. CAST: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security, Special Issue on Electronic Voting*, 4:708–717, 2009a.

P.B. Stark. Risk-limiting post-election audits: P-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, 4:1005–1014, 2009b.

P.B. Stark. Auditing a collection of races simultaneously. Technical report, arXiv.org, 2009c. URL http://arxiv.org/abs/0905.1422v1.

P.B. Stark. Efficient post-election audits of multiple contests: 2009 California tests. http://ssrn.com/abstract=1443314, 2009d. 2009 Conference on Empirical Legal Studies.

P.B. Stark. Risk-limiting postelection audits: Conservative p-values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, 4:1005–1014, 2009e. doi: 10.1109/TIFS.2009.2034190.

P.B. Stark. Super-simple simultaneous single-ballot risk-limiting audits. In *Proceedings of the 2010 Electronic Voting Technology Workshop / Workshop on Trustworthy Elections (EVT/WOTE '10)*. USENIX, 2010. URL http://www.usenix.org/events/evtwote10/tech/full_papers/Stark.pdf.

P.B. Stark. Delayed stratification for timely risk-limiting audits. https://www.stat.berkeley.edu/~stark/Preprints/delayed19.pdf, 2019.

P.B. Stark. Sets of half-average nulls generate risk-limiting audits: SHANGRLA. *Financial Cryptography and Data Security, Lecture Notes in Computer Science*, 12063, 2020. Preprint: http://arxiv.org/abs/1911.10035.

P.B. Stark. ALPHA: Audit that learns from previously hand-audited ballots. *The Annals of Applied Statistics*, 17(1):641 – 679, 2023. doi: 10.1214/22-AOAS1646. URL https://doi.org/10.1214/22-AOAS1646.

P.B. Stark and D.A. Wagner. Evidence-based elections. *IEEE Security and Privacy*, 10:33–41, 2012. https://www.stat.berkeley.edu/~stark/Preprints/evidenceVote12.pdf.

Sonya K. Sterba. Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration. *Multivariate behavioral research*, 44(6):711–740, November 2009. ISSN 0027-3171. doi: 10.1080/00273170903333574. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2856970/.

D.L. Stevens and A.R. Olsen. Spatially Balanced Sampling of Natural Resources. *Journal of the American Statistical Association*, 99(465):262–278, 2004. ISSN 0162-1459. URL http://www.jstor.org/stable/27590371. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

C.E. Stewart, K. Paustian, R.T. Conant, A.F. Plante, and J. Six. Soil carbon saturation: concept, evidence and evaluation. *Biogeochemistry*, 86(1):19–31, October 2007. ISSN 1573-515X. doi: 10.1007/s10533-007-9140-0. URL https://doi.org/10.1007/s10533-007-9140-0.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. ISSN 00063444. URL http://www.jstor.org/stable/2331554.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. Bradford Books, Cambridge, Massachusetts, first edition edition, March 1998. ISBN 978-0-262-19398-6.

S.P. Syswerda, A.T. Corbin, D.L. Mokma, A.N. Kravchenko, and G.P. Robertson. Agricultural management and soil carbon storage in surface vs. deep layers. *Soil Science Society of America Journal*, 75(1):92–101, 2011.

C.E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag, New York, 1992. ISBN 978-0-387-40620-6. URL https://www.springer.com/gp/book/9780387406206.

N.E. Tautges, J.L. Chiartas, A.C.M. Gaudin, A.T. O'Geen, I. Herrera, and K.M. Scow. Deep soil inventories reveal that impacts of cover crops and compost on soil carbon sequestration differ in surface and subsurface soils. *Global Change Biology*, 25:3753–3766, 2019. doi: 10.1111/gcb.14762.

Al. A. Tchouproff. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (chapter i-iii). *Metron*, 2:461–493, 1923.

T. Thamo and D.J. Pannell. Challenges in developing effective policy for soil carbon sequestration: perspectives on additionality, leakage, and permanence. *Climate Policy*, 16(8):973–992, November 2016. ISSN 1469-3062. doi: 10.1080/14693062.2015.1075372. URL https://doi.org/10.1080/14693062.2015.1075372. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14693062.2015.1075372.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285. URL https://doi.org/10.1093/biomet/25.3-4.285.

Y. Tillé and M. Wilhelm. Probability Sampling Designs: Principles for Choice of Design and Balancing. *arXiv:1612.04965 [stat]*, December 2016. URL http://arxiv.org/abs/1612.04965. arXiv: 1612.04965.

K.W. Tsui and S. Weerahandi. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, 84(406):602–607, 1989. ISSN 01621459. URL http://www.jstor.org/stable/2289949.

Stanford University. Growing climate solutions, October 2021. URL https://news.stanford.edu/2021/10/12/growing-climate-solutions/. Section: Science & Technology.

US CMS. Medicare program integrity manual, 2023. URL https://www.cms.gov/regulations-and-guidance/guidance/manuals/downloads/pim83c08.pdf.

US DHS. Medicare contractors were not consistent in how they reviewed extrapolated overpayments in the provider appeals process, 2020. URL https://oig.hhs.gov/oas/reports/region5/51800024.pdf.

J. W. van Groenigen, W. Siderius, and A. Stein. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, 87(3):239–259, January 1999. ISSN 0016-7061. doi: 10.1016/S0016-7061(98)00056-1. URL http://www.sciencedirect.com/science/article/pii/S0016706198000561.

J. Ville. *Étude critique de la notion de collectif*. PhD Thesis, 1939. URL http://eudml.org/doc/192893.

R. A. Viscarra Rossel, R. Webster, M. Zhang, Z. Shen, K. Dixon, Y.-P. Wang, and L. Walden. How much organic carbon could the soil store? The carbon sequestration potential of Australian soil. *Global Change Biology*, 30(1):e17053, 2024. ISSN 1365-2486. doi: 10.1111/gcb.17053. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.17053. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.17053.

R.A. Viscarra Rossel, D.J. Brus, C. Lobsey, Z. Shi, and G. McLachlan. Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. *Geoderma*, 265:152–163, 2016. ISSN 0016-7061. doi: https://doi.org/10.1016/j.geoderma.2015.11.016. URL https://www.sciencedirect.com/science/article/pii/S0016706115301312.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3), June 2021. doi: 10.1214/20-aos2020. URL `https://doi.org/10.1214%2F20-aos2020`.

A.M.. Wadoux, J. Padarian, and B. Minasny. Multi-source data integration for soil mapping using deep learning. *SOIL*, 5(1):107–119, March 2019. ISSN 2199-3971. doi: https://doi.org/10.5194/soil-5-107-2019. URL `https://www.soil-journal.net/5/107/2019/`.

A. Wald. On Cumulative Sums of Random Variables. *The Annals of Mathematical Statistics*, 15(3):283 – 296, 1944. doi: 10.1214/aoms/1177731235. URL `https://doi.org/10.1214/aoms/1177731235`.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, 16:117–186, 1945.

A. Wald. *Sequential Analysis*. Dover Publications, New York, 1947.

A. Wald. *Statistical Decision Functions*. John Wiley and Sons, New York, 1950.

L.R. Walker, D.A. Wardle, R.D. Bardgett, and B.D. Clarkson. The use of chronosequences in studies of ecological succession and soil development. *Journal of Ecology*, 98(4):725–736, 2010. ISSN 1365-2745. doi: 10.1111/j.1365-2745.2010.01664.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2745.2010.01664.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2745.2010.01664.x.

K. Walter, A. Don, B. Tiemeyer, and A. Freibauer. Determining soil bulk density for carbon stock calculations: A systematic method comparison. *Soil Science Society of America Journal*, 80:579–591, 2016a. doi: 10.2136/sssaj2015.11.0407.

K. Walter, A. Don, B. Tiemeyer, and A. Freibauer. Determining Soil Bulk Density for Carbon Stock Calculations: A Systematic Method Comparison. *Soil Science Society of America Journal*, 80(3):579–591, 2016b. ISSN 1435-0661. doi: 10.2136/sssaj2015.11.0407. URL `https://acsess.onlinelibrary.wiley.com/doi/abs/10.2136/sssaj2015.11.0407`. _eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2136/sssaj2015.11.0407.

L. Wasserman. *All of statistics*. Springer, 2004.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 02 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad009. URL `https://doi.org/10.1093/jrsssb/qkad009`. qkad009.

I. Waudby-Smith, P.B. Stark, and A. Ramdas. RiLACS: Risk Limiting Audits via Confidence Sequences. In Robert Krimmer, Melanie Volkamer, David Duenas-Cid, Oksana Kulyk, Peter Rønne, Mihkel Solvak, and Micha Germann, editors, *Electronic Voting*, pages 124–139, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86942-7.

I. Waudby-Smith, L. Wu, A. Ramdas, N. Karampatziakis, and P. Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM / IMS Journal of Data Science*, January 2024. doi: 10.1145/3643693. URL `https://dl.acm.org/doi/10.1145/3643693`.

R. Webster and M. Lark. *Field Sampling for Environmental Science and Management.* Taylor & Francis Group, London, UNITED KINGDOM, 2012. ISBN 978-1-136-47035-6. URL `http://ebookcentral.proquest.com/lib/berkeley-ebooks/detail.action?docID=1024518`.

J.P. Wendell and J. Schmee. Exact inference for proportions from a stratified finite population. *Journal of the American Statistical Association*, 91:825–830, 1996a. doi: 10.1080/01621459.1996.10476950.

J.P. Wendell and J. Schmee. Exact inference for proportions from a stratified finite population. *J. Am. Stat. Assoc.*, 91:825–830, 1996b. URL `links.jstor.org/sici?sici=0162-1459%28199606%2991%3A434%3C825%3AEIFPFA%3E2.0.CO%3B2-D`.

J.W. Wendt and S. Hauser. An equivalent soil mass procedure for monitoring soil organic carbon in multiple soil layers. *European Journal of Soil Science*, 64(1):58–65, 2013. doi: https://doi.org/10.1111/ejss.12002. URL `https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12002`.

W. J. Westlake. Statistical Aspects of Comparative Bioavailability Trials. *Biometrics*, 35(1):273–280, 1979. ISSN 0006-341X. doi: 10.2307/2529949. URL `https://www.jstor.org/stable/2529949`. Publisher: [Wiley, International Biometric Society].

N.K. Wijewardane, Y. Ge, S. Wills, and Z. Libohova. Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. *Soil Science Society of America Journal*, 82(3):722–731, 2018. ISSN 1435-0661. doi: 10.2136/sssaj2017.10.0361. URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.2136/sssaj2017.10.0361.

K.M. Wolter. An Investigation of Some Estimators of Variance for Systematic Sampling. *Journal of the American Statistical Association*, 79(388):781–790, December 1984. ISSN 0162-1459. doi: 10.1080/01621459.1984.10477095. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477095. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1984.10477095.

T. Wright. *Exact Confidence Bounds when Sampling from Small Finite Universes*. Springer-Verlag, New York, 1991.

X. Yan, Z. Cai, S. Wang, and P. Smith. Direct measurement of soil organic carbon content change in the croplands of china. *Global Change Biology*, 17:1487–1496, 2011. doi: 10.1111/j.1365-2486.2010.02286.x.

Y. Zhang, J. M. Lavallee, A. D. Robertson, R. Even, S. M. Ogle, K. Paustian, and M. F. Cotrufo. Simulating measurable ecosystem carbon and nitrogen dynamics with the mechanistically defined mems 2.0 model. *Biogeosciences*, 18(10):3147–3171, 2021. doi: 10.5194/bg-18-3147-2021. URL https://bg.copernicus.org/articles/18/3147/2021/.

Appendix A

Supplementary materials for chapter 2

A.1 Mathematical Framework

We have a plot $\mathcal{P} \subset \mathbb{R}^3$. An element of $\mathcal{P}$ is a 3-tuple (x, y, z). x denotes longitude or x-axis distance from an origin (e.g. the lower left hand corner of a rectangular plot), y denotes latitude or y-axis distance, and z denotes depth.

At (x, y, z) the soil has some concentration of SOC, which we will denote by $c(x, y, z) \in [0, 100]$ with units in percent or equivalently grams SOC per hectogram of soil. Note that sometimes SOC concentration is reported in grams per kilogram. Note also that $c(x, y, z)$ is best conceptualized as an average over a small window centered at (x, y, z). Taking the design-based perspective, we consider $c(x, y, z)$ to be fixed but unknown.

To convert to grams SOC per volume of soil, take $d(x, y, z)$ to be the density of the soil at point (x, y, z), e.g. in grams per cubic centimeter. This is called the "bulk density" in soil science. The amount or *stock* of carbon in a small area centered at point (x, y, z) is thus $c(x, y, z) \times d(x, y, z)$. The total amount or stock of carbon in a plot is:

$$\mathcal{T} = \int_{\mathcal{P}} c(x, y, z) \times d(x, y, z) d\mathcal{P} = \int_0^{x_{\max}} \int_0^{y_{\max}} \int_0^{z_{\max}} c(x, y, z) \times d(x, y, z)\ dz\ dy\ dx$$

Assuming constant bulk density means that $d(x, y, z) = d$ and the total carbon becomes:

$$\mathcal{T} = d \int_{\mathcal{P}} c(x, y, z) d\mathcal{P} = d \times \mu$$

where $\mu := \int_{\mathcal{P}} c(x, y, z) d\mathcal{P}$ is the population average SOC concentration—the key parameter to be estimated through soil sampling. The bulk density d must also be estimated.

The population variance of a plot is formally:

$$\sigma_p^2 = \int_{\mathcal{P}} \left[c(x, y, z) - \mu\right]^2 d\mathcal{P}$$

The population variance is a measure of heterogeneity that is instrumental in determining the precision of estimates of μ.

The mean μ and variance σ_p^2 are estimated using sampled cores. The plot is often sliced into profiles along depth, and positions (x, y) locations are randomly sampled within depth. From here on we will assume we are sampling within a profile and ignore depth. Randomly sampled positions are denoted $\{(X_i, Y_i)\}_{i=1}^n$ and the n corresponding cores are denoted $\{c(X_i, Y_i)\}_{i=1}^n$, or $\{C_1, ..., C_n\}$ when the location is not important. We suppose here that these cores are selected by a UIRS. The properties of the sample mean of cores from a UIRS, $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$, are simple and well-understood: $\bar{C}$ is unbiased ($\mathbb{E}[\bar{C}] = \mu$) and has variance $\mathbb{V}[\bar{C}] = \sigma_p^2/n$.

Given a UIRS $\{C_1, ..., C_n\}$ compositing bins the cores into k groups of size n/k. The groups are denoted by a set of indices $\{g_1, ..., g_k\}$, where $g_1 = \{1, ..., n/k\}$, $g_2 = \{n/k+1, ..., 2n/k\}$, etc. The cores in each group are physically mixed together to form composite samples $\{S_1, ...S_k\}$. Under compositing additivity, we have $S_i = \frac{k}{n}\sum_{j\in g_i} C_j$. Note also that the above notation covers the case where no compositing is done, with $k = n$ and $S_i = C_i$.

Under equal proportions compositing of cores from a UIRS, it follows immediately that the sample mean of the composited cores is an unbiased estimate of μ:

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} S_i\right] = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in g_i}\frac{k}{n}C_j\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} C_i\right] = \mu$$

Furthermore, because the sample mean of the composite samples is equivalent to the sample mean of the constituents, its variance is also

$$\mathbb{V}\left[\frac{1}{k}\sum_{i=1}^{k} S_i\right] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} C_i\right] = \frac{\sigma_p^2}{n}.$$

Assay error is drawn from an unknown distribution with positive support and denoted δ_i. Measured samples are $S_i^* = S_i\delta_i$. We assume assays are unbiased so that $\mathbb{E}(\delta_i) = 1$ and $\mathbb{E}(S_i^*) = S_i$ where the expectation is with respect to the assay error only (not the sampling distribution). We also assume that the assay error has constant variance $\mathbb{V}(\delta_i) = \sigma_\delta^2$, that does not depend on S_i.

Our estimator is the mean of k measured samples composited from n cores: $\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} S_i^*$. Under our assumptions, $\hat{\mu}$ is an unbiased estimator:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} S_i\delta_i\right] = \frac{1}{k}\sum_{i=1}^{k}\left[\sum_{j\in g_i}\frac{k}{n}\mathbb{E}[C_j]\right]\mathbb{E}[\delta_i] = \frac{1}{k}\sum_{i=1}^{k}\left[\sum_{j\in g_i}\frac{k}{n}\mu\right] = \mu. \quad \text{(A.1)}$$

Its variance is:

$$\begin{aligned}
\mathbb{V}\left[\hat{\mu}\right] &= \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{V}[S_i\delta_i] \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\left(\mathbb{V}[S_i]\mathbb{V}[\delta_i] + \mathbb{E}[S_i]^2\mathbb{V}[\delta_i] + \mathbb{E}[\delta_i]^2\mathbb{V}[S_i]\right) \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\left(\frac{k}{n}\sigma^2\sigma_\delta^2 + \mu^2\sigma_\delta^2 + \frac{k}{n}\sigma^2\right) \\
&= \frac{\sigma^2(1+\sigma_\delta^2)}{n} + \frac{\mu^2\sigma_\delta^2}{k}
\end{aligned}$$

A.2 Optimizations

A.2.1 Minimum error with a fixed budget

For a budget B, fixed in advance, we seek the solution to the optimization problem:

$$\begin{aligned}
\mathbb{V}(\hat{\mu})_{\text{opt}} = &\min_{M,P}\min_{n,k} \quad \frac{\sigma_p^2(1+\sigma_\delta^2)}{n} + \frac{\mu^2\sigma_\delta^2}{k} && \text{(A.2)} \\
&\text{s.t.} \qquad \text{cost}_0 + n\cdot\text{cost}_c + k\cdot(\text{cost}_P + \text{cost}_A) \leq B && \text{(A.3)} \\
&\qquad\qquad k \geq 1 && \text{(A.4)} \\
&\qquad\qquad k \leq n && \text{(A.5)}
\end{aligned}$$

where as above cost_0 is the fixed cost, cost_c is the cost of sampling a single core, cost_P is the cost of sample prep, and cost_A is the cost of assay. A and P additionally denote the assay and sample preparation schemes, which affect costs and σ_δ.

For a fixed A and P, the inner optimization problem can be solved in closed form using a Lagrange multiplier for the constraint. The optimal sampling and assay sizes are then:

$$n_{\text{opt}} = (B - \text{cost}_0)\frac{\sigma_p\sqrt{1+\sigma_\delta^2}}{[\sigma_p\sqrt{(1+\sigma_\delta^2)\text{cost}_c} + \mu\sigma_\delta\sqrt{\text{cost}_P + \text{cost}_A}]\sqrt{\text{cost}_c}} \qquad \text{(A.6)}$$

$$k_{\text{opt}} = (B - \text{cost}_0)\frac{\mu\sigma_\delta}{[\sigma_p\sqrt{(1+\sigma_\delta^2)\text{cost}_c} + \mu\sigma_\delta\sqrt{\text{cost}_P + \text{cost}_A}]\sqrt{\text{cost}_P + \text{cost}_A}}. \qquad \text{(A.7)}$$

This solution ignores the constraints $k \geq 1$ and $k \leq n$. If we find $k_{\text{opt}} < 1$, then set $k_{\text{opt}} = 1$ and $n_{\text{opt}} = (B - \text{cost}_0 - \text{cost}_P - \text{cost}_A)/\text{cost}_c$. If we find $k_{\text{opt}} > n^*_{P,M}$ then set $k_{\text{opt}} = n_{\text{opt}} = (B - \text{cost}_0)/(\text{cost}_c + \text{cost}_P + \text{cost}_A)$. To obtain integer solutions while staying under budget, n_{opt} and k_{opt} should be rounded down.

A.2.2 Minimum cost for a given precision

Given a maximum variance V that we can tolerate, we seek the minimum budget over all ways of allocating the budget to samples and assays while achieving that precision. Formally:

$$
\begin{aligned}
B_{\text{opt}} = \min_{M,P} \min_{n,k} \quad & \text{cost}_0 + n \cdot \text{cost}_c + k \cdot (\text{cost}_P + \text{cost}_A) && \text{(A.8)} \\
\text{s.t.} \quad & \frac{\sigma_p^2(1+\sigma_\delta^2)}{n} + \frac{\mu^2\sigma_\delta^2}{k} \leq V && \text{(A.9)} \\
& k \geq 1 && \text{(A.10)} \\
& k \leq n. && \text{(A.11)}
\end{aligned}
$$

The solution of the inner optimization (for fixed M and P) is:

$$
n_{\text{opt}} = \left[\left(\frac{\text{cost}_c \cdot \sigma_p^2(1+\sigma_\delta^2)}{\text{cost}_P + \text{cost}_A} \right)^{1/2} + 1 \right] \frac{\mu^2\sigma_\delta^2}{V} \tag{A.12}
$$

$$
k_{\text{opt}} = \frac{\sigma_p^2(1+\sigma_\delta^2)}{V\left(1 - \left[\left(\frac{\text{cost}_c \cdot \sigma_p^2(1+\sigma_\delta^2)}{\text{cost}_P + \text{cost}_A} \right)^{1/2} + 1 \right]^{-1} \right)}. \tag{A.13}
$$

The constraints $k \geq 1$ and $k \leq n$ are not respected by these solutions. If we find $k_{\text{opt}} < 1$, then set $k_{\text{opt}} = 1$ and $n_{\text{opt}} = \sigma_p^2/V$ (obtained, for example, when there is no assay error). If we find $k_{\text{opt}} \geq n_{\text{opt}}$, then set $k_{\text{opt}} = n_{\text{opt}} = (\sigma_p^2(1+\sigma_\delta^2) + \mu^2\sigma_\delta^2)/V$. To get integer solutions n_{opt} and k_{opt} should be rounded up.

A.3 Estimating σ_δ^2

A.3.1 Replicate assays

Suppose we have r replicated, unbiased assays for the ith sample. The replicates are denoted $\{S^*_{i1}, S^*_{i2}, ..., S^*_{ir}\}$, where $S^*_{ij} = S_i\delta_{ij}$ is the true SOC concentration in composite sample i multiplied by an independent, mean 1 assay error. The sample mean over replicates is $\bar{S}_i = \frac{1}{r}\sum_{j=1}^{r} S^*_{ij}$, which is an unbiased estimate of S_i with variance $\mathbb{V}[\bar{S}_i|S_i] = \mathbb{V}[S^*_{ij}|S_i]/r$. An unbiased estimate of S_i^2 is the squared sample mean minus its variance, i.e. $\bar{S}_i^2 - \frac{1}{r(1-r)}\sum_{j=1}^{r}(S^*_{ij} - \bar{S}_i)^2$. Thus we might estimate σ^2_δ by plugging in the unbiased estimators of $\mathbb{V}[S^*_{ij}|S_i]$ and S_i^2:

$$\hat{\sigma}^2_\delta = \frac{\frac{1}{r-1}\sum_{j=1}^{r}(S^*_{ij} - \bar{S}^*_i)^2}{(\bar{S}^*_i)^2 - \frac{1}{r(r-1)}\sum_{j=1}^{r}(S^*_{ij} - \bar{S}^*_i)^2}. \tag{A.14}$$

This is not necessarily unbiased. If the numerator and denominator were independent, then Jensen's inequality would make the estimate conservative in expectation: $\mathbb{E}[\hat{\sigma}^2_\delta] > \sigma_\delta$. We used this replicated measurement technique to estimate the error of DC-EA.

The assay variance estimate $\hat{\sigma}^2_\delta$ can be computed on any sample that is replicated 2 or more times. One strategy to estimate σ_δ is thus to duplicate every sample ($r = 2$) and then take the average or median, though the variance of estimates may be high. Plotting $\hat{\sigma}^2_{\delta i}$ against S^*_i should indicate potential violations of the constant assay error variance assumption. Another strategy is to replicate a single sample some large number of times, say $r = 30$, but this will not provide information about the constant variance assumption. A good balance is to measure a few samples along a grid of S^*_i values some moderately large number of times, say $r = 5$. Under constant assay error variance, the estimates $\hat{\sigma}^2_{\delta i}$ should be fairly close and there should not be a trend in S^*_i.

A.3.2 Prediction Methods

Suppose we have a method that is calibrated to an unbiased assay (e.g. DC-EA) by regression, like LOI or MIRS. There are two sources of error in the estimate. First, there is the assay error of the calibration assay which can be estimated directly through replication as per Section A.3.1. Second, there is the error in the calibration itself, i.e. prediction error. We discuss two ways to estimate the prediction error. To approximate the total error of a prediction method we recommend simply adding the pieces together.

Prediction methods typically assume an additive error model and estimate the variance of the additive error out of sample. We will call this estimate RMSE_v for validation root mean squared error. Now, if SOC concentration has been first transformed to the log scale, then the model implicitly assumes that error is multiplicative on the original scale, our estimate of the error in prediction is then $\exp(\text{RMSE}_v)$. On the other hand, if SOC is modeled directly (i.e. without a log transformation) then we can approximate the error on a multiplicative scale by assuming that the additive error pertains to the average SOC assay, which suggests dividing by the average assay. Thus we estimate the prediction error by $\frac{\text{RMSE}_v}{\hat{\mu}}$.

In our application, LOI and MIRS were calibrated both to DC-EA assays. For both of these methods, we had an RMSE_v of SOC modeled on the original scale (without a log transform) so we estimated the error of these methods as

$$\hat{\sigma}_{\delta,\text{LOI}} = \hat{\sigma}_{\delta,\text{DC-EA}} + \frac{\text{RMSE}_{v,\text{LOI}}}{\hat{\mu}}$$
$$\hat{\sigma}_{\delta,\text{MIRS}} = \hat{\sigma}_{\delta,\text{DC-EA}} + \frac{\text{RMSE}_{v,\text{MIRS}}}{\hat{\mu}}$$

A.4 Shortest path through random points in $\mathcal{P}$

Recall that the area of plot $\mathcal{P}$ is $\mathcal{A}$. Finding the shortest path through n points is known as the traveling salesman problem. If the n points are generated randomly and independently with density $f(x)$ then the Beardwood-Halton-Hammersley theorem for $\mathbb{R}^2$ says that the length of the shortest path converges to:

$$L_n/\sqrt{n} \to \beta_2 \int_{\mathbb{R}^2} \sqrt{f(x)}dx$$

β_2 is an unknown constant, but analytical bounds and numerical simulations have pegged it at about 0.714 [Arlotto and Steele, 2016]. For a UIRS, $f(x) = \frac{1}{\mathcal{A}}$ on $\mathcal{P}$ and 0 elsewhere. The integral thus evaluates to $\sqrt{\mathcal{A}}$.

Appendix B

Supplementary materials for chapter 6

B.1 Computational details

The following describes details of the allocation simulations in Section 6.4. Within each stratum, we computed null means along an equispaced grid of $(2\max\{N_1, N_2\})$ points[1] for $\theta_1 \in [\varepsilon_1, \theta/w_1 - \varepsilon_1]$ with $\theta_2 = (\theta - w_1\theta_1)/w_2$. The null means were then adjusted to $\beta_1 := \theta_1 + 1 - \bar{A}_1^c$ and $\beta_2 := \theta_1 + 1 - \bar{A}_2^c$. The conditional null means β_{i1} and β_{i2} were computed as:

$$\beta_{ik} = \frac{N_k\beta_k - \sum_{j=1}^{i-1} X_{ik}}{N_k - (i-1)}$$

Tuning parameters for ALPHA-ST were chosen as in Stark [2023, Section 2.5.2] with $d_k = 20$ and the initial estimate τ_{0k} set to $u_k^A = 1$, the expected mean when there is no error in the CVRs. For ALPHA-UB, we set

$$\tau_{ik}^{\text{UB}} := \frac{(d_k\tau_{0k} + \sum_{j=1}^{i-1} X_{jk})/(d_k + i - 1) + f_k u_k/\hat{\sigma}_{ik}^2}{1 + f_k/\hat{\sigma}_{ik}^2}.$$

The first term in the numerator of τ_{ik}^{UB} is truncated shrinkage estimator ALPHA-ST. The second term biases τ_{ik}^{UB} towards u_k with a weight proportional to the inverse running sample variance $\hat{\sigma}_{ik}^2$. The constant of proportionality f_k is a tuning parameter set to $f_k := .01$; higher f_k would bias τ_{ik} towards u_k more aggressively. The variance-dependent bias amounts to betting more when the population variance is low, which it tends to be in comparison audits when the voting system works properly. Truncation keeps τ_{ik} within its allowed range.

For both ALPHA strategies, τ_{ik} was truncated to be in $[\beta_{ik} + \varepsilon_k, u_k(1-\delta)]$, where $\varepsilon_k := 1/2N_k$ was the minimum value of one assorter and $\delta = 2.220446 \times 10^{-16}$ was machine precision. If $\beta_{ik} + \varepsilon_k \geq u_k$, we set the corresponding terms in the supermartingale to 1: that (composite) null is true.

Each stratum selection rule was applied to every supermartingale. For proportional allocation, there was no additional selection: samples were gathered round-robin across strata, omitting any strata that were fully exhausted. For lower-sided testing, the sampling from a stratum ceased when the lower-sided test rejected at level .05. This was implemented by setting all future terms in the supermartingale equal to 1 after rejection. The stratumwise supermartingales were then multiplied to produce $2\max\{N_1, N_2\}$ intersection supermartingales and their minimum (over nulls) was found at each sample size. The reciprocal of this minimized intersection

[1] The cardinality was chosen so that a null mean was computed for every possible (discrete) value of θ_k. A finer grid is unnecessary; a coarser grid may not find the true minimum.

supermartingale was a sequence of P-values corresponding to P_M^* under a particular sample allocation rule. The same strategy, but using Fisher pooling, was used to find P_F^*. The sample size at risk limit $\alpha = 5\%$ is the sample size for which the P-value sequence first hits or crosses 0.05, summed across both strata.

Appendix C

Supplementary materials for chapter 7

C.1 Example: Kelly-optimality for a point mass population

We find the minimum expected stopping time $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau^*]$ when the alternative is $\mathcal{X}_{\boldsymbol{N}} \in \aleph^1_{\boldsymbol{N}} \cap \aleph^\delta_{\boldsymbol{N}}$, where $\aleph^\delta_{\boldsymbol{N}}$ contains all stratified populations of size $\boldsymbol{N}$ with $x_{ki} = \mu_k$ for all k and i. That is, the population distribution consists of point masses within strata. The procedure that achieves that minimum is a Kelly-optimal UI-TS, constructed by applying Lemma 3 and Lemma 4. Since $\tau \leq n_\tau$, $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau^*]$ immediately yields a lower bound on the minimum expected sample size $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[n^*_\tau]$. We assume sampling is with replacement so the Kelly-optimal rules do not depend on time.

To begin, consider the betting rule

$$\lambda_k(\boldsymbol{\eta}) = \frac{1\{\mu_k > \eta_k\}}{\eta_k}.$$

It is clearly Kelly-optimal when the within-stratum distributions are point masses: it bets the maximum amount $1/\eta_k$ if $x_{ik} = \mu_k > \eta_k$, in which case the bet is certain to succeed, and 0 otherwise. Following Lemma 3, the Kelly-optimal selection rule always pulls the strata with the largest log-growth (with ties broken arbitrarily). This gives us an explicit formula for the Kelly-optimal I-TSM at $\boldsymbol{\eta}$:

$$M^*_t(\boldsymbol{\eta}) := \max_k \prod_{i=1}^t [1 + \eta_k^{-1}(\mu_k - \eta_k)] = \max_k \left(\frac{\mu_k}{\eta_k}\right)^t,$$

where we have dropped the indicator because there is always at least one stratum with $\mu_k > \eta_k$ since the alternative is true. The Kelly-optimal UI-TS is thus:

$$M^*_t = \min_{\boldsymbol{\eta} \in \mathcal{C}} \max_k \left(\frac{\mu_k}{\eta_k}\right)^t.$$

Further specialize to $K = 2$, $w_1 = w_2$, and $\eta_0 = 1/2$. Letting $\eta := \eta_1$ so $(1 - \eta) = \eta_2$, we can write

$$M^*_t = \min_{\eta \in [0,1]} \left(\frac{\mu_1}{\eta} \vee \frac{\mu_2}{1-\eta}\right)^t,$$

which is minimized when the two terms inside the parentheses are equal. That is, the minimizer satisfies

$$\frac{\mu_1}{\eta^*} = \frac{\mu_2}{1-\eta^*} \implies \frac{\eta^*}{1-\eta^*} = \frac{\mu_1}{\mu_2}.$$

This nonlinear equation can be solved numerically for any (μ_1, μ_2). The stopping time is the point at which M_t^* crosses $1/\alpha$. This point solves:

$$\left(\frac{\mu_1}{\eta^*}\right)^t = \alpha^{-1}$$

$$\iff$$

$$t(\log \mu_1 - \log \eta^*) = -\log \alpha$$

$$\iff$$

$$t = \frac{\log \alpha}{\log \eta^* - \log \mu_1},$$

and the Kelly-optimal stopping time is $\tau^* = \lceil t \rceil$. In Figure C.1, we plot the stopping time of a Kelly-optimal UI-TS at level $\alpha = 0.05$ over a range of μ_1 and μ_2 in the alternative.

C.2 Proofs

C.2.1 Proof of Lemma 1

Part 1 of Lemma 1, on the separate validity of the lower confidence bounds (LCBs), follows immediately by inverting level α tests constructed from within-stratum betting TSMs. Indeed, L_{kt} is precisely the η_k that yields a within-stratum TSM of size $1/\alpha$: $M_{kt}(L_{kt}) = 1/\alpha$. The corresponding within-stratum P-value is $P_{kt}(L_{kt}) := 1/M_{kt}(L_{kt}) = \alpha$, which implies each L_{kt} is a $(1 - \alpha)$ LCB satisfying the first claim. This topic is covered extensively in Waudby-Smith and Ramdas [2023].

Part 2 of Lemma 2 follows from an application of the closed testing principle. Specifically, recall that the family-wise error rate for a collection of partial hypotheses $\{H_{0k}\}_{k=1}^K$ is controlled by testing every intersection hypothesis corresponding to possible subsets of $\{H_{0k}\}_{k=1}^K$. Take each partial (marginal) hypothesis to be

$$H_{0k} : \mu_k \leq L_k,$$

which corresponds to a TSM of $M_{kt}(L_{kt}) = 1/\alpha$, and note that $\{M_{kt}(L_{kt})\}_{k=1}^K$ are mutually independent under stratified sampling. Every intersection hypothesis can be tested by taking the product of two or more within-stratum TSMs, and the resulting product will always be above $1/\alpha$. For example, the 2-way intersection $H_{0k} \cap H_{0j}$ is tested by $M_{kt}(L_{kt})M_{jt}(L_{jt}) = 1/\alpha^2 > 1/\alpha$. As a result, every higher-order intersection

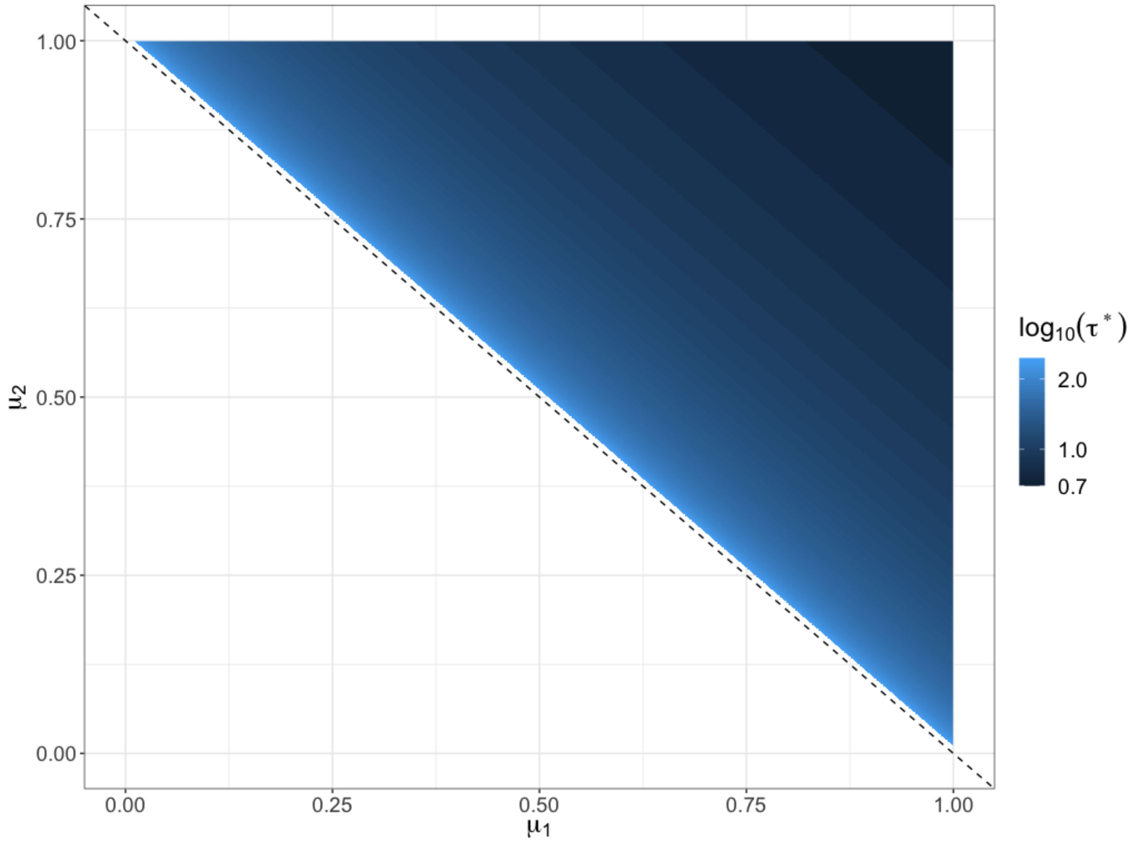


Figure C.1: Kelly-optimal stopping times (color, $\log_{10}$ scale) at level $\alpha = 0.05$ under sampling with replacement from a range of stratified point mass populations with means μ_1 (x-axis) and μ_2 (y-axis). The global null is $H_0 : \mu \leq 1/2$ and the dotted line indicates the boundary of the null $\aleph^0_{\boldsymbol{N}}$. All populations here satisfy $\mu > 1/2 + 0.005$, and so are in the interior of the alternative $\aleph^1_{\boldsymbol{N}}$. The stopping times lower bound the workload of any test since $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[\tau] \leq \mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}}[n_\tau]$.

hypothesis will be rejected, the P-values $\{P_{kt}(L_{kt})\}_{k=1}^K$ are simultaneously valid by closed testing, and the LCBs $\{L_{kt}\}_{k=1}^K$ are simultaneously below the stratum-wise means $\{\mu_k\}_{k=1}^K$ with probability $1 - \alpha$.

C.2.2 Proof of Lemma 3

We can write the expected log-growth as

$$\begin{aligned}
\mathbb{E}\left[\log \Delta M_t(\boldsymbol{\eta})\right] &= \mathbb{E}\left[\log \sum_{k=1}^K 1\{S(t) = k\} Z_{kt}(\eta_k)\right] \\
&= \mathbb{E}\left[\log \prod_{k=1}^K Z_{kt}(\eta_k)^{1\{S(t)=k\}}\right] \\
&= \mathbb{E}\left[\sum_{k=1}^K 1\{S(t) = k\} \log Z_{kt}(\eta_k)\right] \\
&= \sum_{k=1}^K p_{kt}(\eta_k) \mathbb{E}[\log Z_{kt}(\eta_k)]
\end{aligned}$$

Given any allocation strategy $(\boldsymbol{p}_t)_{t\in\mathbb{N}}$, the overall expected log-growth is a convex combination of the expected-log growth within-strata, and is maximized when each $\mathbb{E}[\log Z_{kt}(\eta_k)]$ is maximized. This is ensured by using the Kelly optimal bet $\lambda_k^* = \arg\max_{\lambda \in [0, 1/\eta_k]} \mathbb{E}[\log Z_{kt}(\eta_k)]$ within each stratum. Now, assume we use the bets $\boldsymbol{\lambda}^*$ so that each term $\mathbb{E}[\log Z_{kt}(\eta_k)]$ is maximized. The allocation probabilities

$$p_{kt}^*(\boldsymbol{\eta}) := 1\{k = \arg\max_j \mathbb{E}[\log Z_{jt}(\eta_j)]\}$$

maximize $\mathbb{E}[\log \Delta M_t(\boldsymbol{\eta})]$ because they always choose the stratum with the largest expected log growth.

C.3 Computational properties

In this section, we evaluate the computational tractability of finding M_t. Finding the minimum over $\mathcal{C}$ is a hard problem when the betting and allocation rules can be arbitrarily $\boldsymbol{\eta}$-aware. We examine the possibility of finding a solution by brute force in small populations with discrete support and few strata. We then show how constraining the rules can ease the computational burden and allow M_t to be computed in more general settings.

C.3.1 Brute-force search in discrete, finite populations

Let $\mathcal{U}_k$ denote the set of possible unique values of elements of $\mathcal{X}_k$. In general, $\mathcal{U}_k \subseteq [0,1]$ and $|\mathcal{U}_k|$ may be uncountably infinite (e.g., when $\mathcal{U}_k = [0,1]$). However, when $|\mathcal{U}_k|$ is finite (e.g., when $\mathcal{U}_k$ is binary), μ_k can take only finitely many values. In that case $\mathcal{B}$ is a finite set, and $|\mathcal{B}|$ may even be small enough to enumerate and compute every $M_t(\boldsymbol{\eta})$ by brute force, no matter if $M_t(\boldsymbol{\eta})$ is non-convex, multimodal, non-smooth, etc over $\boldsymbol{\eta}$.

Let Ω_k be the set of all possible means μ_k in a single stratum. In general the number of possible length-N_k bags with elements in $\mathcal{U}_k$ is:

$$\binom{N_k + |\mathcal{U}_k| - 1}{|\mathcal{U}_k| - 1},$$

which follows from Feller's bars and stars argument, partitioning the N_k numbers into $|\mathcal{U}_k|$ (possibly empty) bins according to which value in $\mathcal{U}_k$ each number takes.

However, in many practical applications (e.g., stratified risk-limiting comparison audits), the size of Ω_k is reduced because elements of $\mathcal{U}_k$ have a relatively small least common multiple. For example, if $\mathcal{U}_k = \{0, 0.5, 1\}$ then

$$\Omega_k = \{0, 0.5/N_k, 1.0/N_k, 1.5/N_k, \dots, 1\},$$

and

$$|\Omega_k| = 2N_k + 1.$$

In general, we have

$$\Omega_k \subseteq \left\{0, \frac{1}{c_k N_k}, \frac{2}{c_k N_k}, \frac{3}{c_k N_k}, \dots, 1\right\},$$

for

$$c_k = \text{LCM}\left(\left\{\frac{1}{x} \quad \forall\, x \in \mathcal{U}_k\right\}\right),$$

where LCM represents the least common multiple of the reciprocals. This implies that $|\Omega_k| = c_k N_k + 1$, a significant reduction in $|\Omega_k|$ when the elements of $\mathcal{U}_k$ are multiples of each other.

Now, the size of $\mathcal{C}$ depends on the number of possible *intersection* means $\boldsymbol{\mu}$ lying on the hyperplane $\boldsymbol{\mu} \cdot \boldsymbol{w} = \eta_0$ defined by H_0. The size of $\mathcal{C}$ is loosely bounded by the number of possible within-stratum means

$$|\mathcal{C}| \le \prod_{k=1}^{K} |\Omega_k| \le \prod_{k=1}^{K} (c_k N_k + 1).$$

Going by this upper bound, $\mathcal{C}$ is not feasible to enumerate for large K, N, or $|\mathcal{X}_k|$. In particular, even in the unique case that the strata are of equal size ($N_k = N/K$) and all binary ($\mathcal{X}_k = 2$), the complexity reduces to:

$$|\Omega| = (N/K)^K.$$

Even this is intractable for moderate N and $K > 2$.

C.3.2 Vertex enumeration under $\boldsymbol{\eta}$-oblivious selection and betting

The class of $\boldsymbol{\eta}$-oblivious bets and selections comprise an important class of tuning parameters for a betting UI-TS. When such strategies are used the computational burden is eased substantially because the minimum must occur on a limited set of points. We show this here, and characterize the number of points that must be searched to compute M_t under such strategies.

Suppose $\boldsymbol{\lambda}_t(\boldsymbol{\eta}) := \boldsymbol{\lambda}_t$ and $\boldsymbol{p}_t(\boldsymbol{\eta}) := \boldsymbol{p}_t$ are $\boldsymbol{\eta}$-oblivious. We aim to show that the minimum of $M_t(\boldsymbol{\eta})$ occurs on a *vertex* of the polytope $\mathcal{C}$ representing the set of intersection nulls. We will accomplish this by establishing that $M_t(\boldsymbol{\eta})$ is log-concave in $\boldsymbol{\eta}$. To that end, the first two partial derivatives of $\log M_t(\boldsymbol{\eta})$ are:

$$\begin{aligned}
\frac{\partial}{\partial \eta_k} \log M_t(\boldsymbol{\eta}) &= -\sum_{i=1}^{T_k(t)} \frac{\lambda_{ki}}{1 + \lambda_{ki}(X_{ki} - \eta_k)} \\
\frac{\partial^2}{\partial \eta_k \eta_j} \log M_t(\boldsymbol{\eta}) &= 0 \ \ \forall \ \ k \neq j \\
\frac{\partial^2}{\partial \eta_k^2} \log M_t(\boldsymbol{\eta}) &= -\sum_{i=1}^{T_k(t)} \left[\frac{\lambda_{ki}}{1 + \lambda_{ki}(X_{ki} - \eta_k)} \right]^2.
\end{aligned}$$

All the mixed partials are 0, while the second partials are all semi-negative, and strictly negative as long as $\lambda_{ki} > 0$ for some i and all k. Thus, the Hessian of $\log M_t(\boldsymbol{\eta})$ is negative semi-definite, $\log M_t(\boldsymbol{\eta})$ is concave in $\boldsymbol{\eta}$, and $M_t(\boldsymbol{\eta})$ is log-concave and has a unique *maximum* at $\boldsymbol{\eta}^\dagger$ (since we wish to find the minimum, the direction of convexity is the opposite of what we may hope). $M_t(\boldsymbol{\eta})$ is decreasing in any direction moving away from $\boldsymbol{\eta}^\dagger$. Because we want to minimize $M_t(\boldsymbol{\eta})$, we can do so by moving away from it's maximum. Without knowing the best direction, we know that the farthest we can get from $\boldsymbol{\eta}^\dagger$ while remaining in $\mathcal{C}$ is on a vertex of $\mathcal{C}$. Thus the minimizer occurs in $\mathcal{V}$, the set of vertices of $\mathcal{C}$. If $|\mathcal{V}|$ is small enough, we can compute M_t by enumerating the values of $M_t(\boldsymbol{\eta})$ for all $\boldsymbol{\eta} \in \mathcal{V}$ and choosing the smallest.

The number of vertices of $\mathcal{C}$

We derive $|\mathcal{V}|$ in an important special case: strata of equal size and global null $\eta_0 = 1/2$. Accordingly, suppose $\boldsymbol{w} = \frac{1}{K}\mathbf{1}$. Then, if K is even, the elements of $\mathcal{V}$ are all K-vectors with $K/2$ elements equal to 1 and the rest equal to 0:

$$|\mathcal{V}| = \binom{K}{K/2}.$$

If K is odd, an examplar element of $\mathcal{V}$ is a K-vector with $(K-1)/2$ values equal to 1, exactly one value equal to $1/2 - \frac{K-1}{2K}$, and the rest equal to 0. That is:

$$|\mathcal{V}| = K\binom{K-1}{(K-1)/2}.$$

When the strata vary in size, there can be fewer or more vertices than this. The vertices can be enumerated, e.g., using the python package `pypoman`, which provides rapid enumeration for $K \leq 16$ or so. When $K = 15$, $\eta_0 = 0.5$, and $N_k = N/K$, we must compute $|\mathcal{V}| = 51480$ I-TSMs each of length N. As K grows, the vertex enumeration approach eventually becomes infeasible.

C.3.3 Convexifying $\boldsymbol{\eta}$-aware bets

If we can choose λ_{ki} as a function of η_k, then we can improve the computational tractability of M_t. In particular, if M_t is *convex* in $\boldsymbol{\eta}$ and $T_k(t)$ is fixed over $\boldsymbol{\eta}$, then we can find the minimum of M_t over $\boldsymbol{\eta}$ and derive a sampling strategy to maximize the growth of the worst null. In particular, we consider $\lambda_{ki} = \exp(\bar{X}_{k(i-1)} - \eta_k)$ where $\bar{X}_{k(i-1)}$ is the lagged running sample mean in stratum k. This allows bets to be larger when the sample mean is larger in a particular stratum. We first note that λ_{ki} is a valid bet:

$$0 < \lambda_{ki} = \exp(\bar{X}_{k(i-1)} - \eta_k) \leq \exp(1 - \eta_k) \leq 1/\eta_k$$

for all $\eta_k \in [0, 1]$.

Then, we prove that

$$\ln M_t(\boldsymbol{\eta}) := \sum_{k=1}^{K} \sum_{i=1}^{T_k(t)} \ln(1 + \exp(\bar{X}_{k(i-1)} - \eta_k)(X_{ki} - \eta_k))$$

is convex. We prove this by showing $\lambda_{ki} = \exp(a - b\eta_k)$ for any constants a and $b \geq 1$ makes the objective convex. For simplicity, we denote $h_i(\eta_k) = \lambda_{ki}$. We observe that $h_i'(\eta_k) = -bh_i(\eta_k)$.

To show this, we first compute the Hessian of $\ln M_t$. The first derivative becomes:

$$\begin{aligned}\frac{\partial \ln M_t(\boldsymbol{\eta})}{\partial \eta_k} &= \frac{\partial}{\partial \eta_k} \sum_{i=1}^{T_k(t)} \ln(1 + h_i(\eta_k)(X_{ki} - \eta_k)) \\ &= -\sum_{i=1}^{T_k(t)} \frac{h_i(\eta_k)(b(X_{ki} - \eta_k) + 1)}{1 + h_i(\eta_k)(X_{ki} - \eta_k)}.\end{aligned}$$

We observe that the mixed partials are 0 since $M_t(\eta_k)$ does not depend on any η_j where $j \neq k$. Thus, we only calculate the diagonal values of the Hessian.

$$\frac{\partial^2 \ln M_t(\eta)}{\partial \eta_k^2} = -\frac{\partial}{\partial \eta_k} \sum_{i=1}^{T_k(t)} \frac{h_i(\eta_k)(b(X_{ki} - \eta_k) + 1)}{1 + h_i(\eta_k)(X_{ki} - \eta_k)} = -\sum_{i=1}^{T_k(t)} \frac{gf' - fg'}{g^2},$$

where

$$\begin{aligned}g(\eta_k) &= 1 + h_i(\eta_k)(X_{ki} - \eta_k), \\ g'(\eta_k) &= h_i'(\eta_k)(X_{ki} - \eta_k) - h_i(\eta_k) \\ &= h_i(\eta_k)(-b(X_{ki} - \eta_k) - 1) \\ f(\eta_k) &= h_i(\eta_k)(b(X_{ki} - \eta_k) + 1) \\ f'(\eta_k) &= h_i'(\eta_k)(b(X_{ki} - \eta_k) + 1) + h_i(\eta_k)(-b) \\ &= -bh_i(\eta_k)(b(X_{ki} - \eta_k) + 2)\end{aligned}$$

Now, in order for this to be positive, we need

$$-[gf' - fg'] \geq 0.$$

Without loss of generality, we analyze a single index i of the summation; since this holds for any term, this must also hold for the summation. In particular, the function is convex if

$$\begin{aligned}-[(1 + h_i(\eta_k)(X_{ki} - \eta_k))(-bh_i(\eta_k)(b(X_{ki} - \eta_k) + 2)) \\ - (h_i(\eta_k)(b(X_{ki} - \eta_k) + 1))h_i(\eta_k)(-b(X_{ki} - \eta_k) - 1)] \geq 0.\end{aligned}$$

Simplifying the expression yields

$$\begin{aligned}bh_i(\eta_k)(1 + h_i(\eta_k)(X_{ki} - \eta_k))(b(X_{ki} - \eta_k) + 2) \geq h_i(\eta_k)^2(b(X_{ki} - \eta_k) + 1))^2, \\ \Longleftarrow\ b(1 + h_i(\eta_k)(X_{ki} - \eta_k)) \geq h_i(\eta_k)(b(X_{ki} - \eta_k) + 1), \\ \Longleftrightarrow\ b + bh_i(\eta_k)(X_{ki} - \eta_k) \geq bh_i(\eta_k)(X_{ki} - \eta_k) + h_i(\eta_k), \\ \Longleftrightarrow\ b \geq h_i(\eta_k),\end{aligned}$$

which is always true for any $b \geq 1$ by the bounds on $h_i(\eta_k)$. Thus, since the second partials are non-negative, the Hessian is positive semi-definite for any X_{ki}, η_k and the objective is convex.

C.4 Stratified sequential testing of a simple null

Suppose we wish to test a simple null $H_0^{\mathrm{S}} : \mathcal{X}_{\boldsymbol{N}} = \mathcal{X}_{\boldsymbol{N}}^0$ against a simple alternative $\mathcal{X}_{\boldsymbol{N}} = \mathcal{X}_{\boldsymbol{N}}^1$. While we are generally interested in the composite null $\mathcal{X}_{\boldsymbol{N}} \in \aleph_{\boldsymbol{N}}^0$ in this paper, a simple null provides a simple optimal strategy: the SPRT of Wald [1945].

Specifically, let $\mathcal{X}_k^0$ be the null population in stratum k. Analogously, $\mathcal{X}_k^1$ is the within-stratum alternative. Furthermore, let $f(x, \mathcal{X})$ denote the probability density at x under uniform sampling with or without replacement from $\mathcal{X}$. Following the construction of the SPRT within each stratum, form the terms

$$Z_{ki} := \frac{f(X_{ki}, \mathcal{X}_k^1)}{f(X_{ki}, \mathcal{X}_k^0)}.$$

The running product $\prod_{i=1}^{T_k(t)} Z_{ki}$ measures evidence against the simple null within stratum k, and the product across strata:

$$\prod_{k=1}^{K} \prod_{i=1}^{T_k(t)} Z_{ki}$$

is a martingale when H_0^{S} is true. Furthermore, Wald [1945] shows that this choice of Z_{ki} maximizes $\mathbb{E}_{\mathcal{X}_{\boldsymbol{N}}^1}[\log Z_{ki}]$, i.e., that it is Kelly-optimal. The growth rate optimality of the SPRT is analogous to the power optimality of the likelihood ratio in the fixed-sample case, as in the canonical lemma of Neyman and Pearson [1933].

If the null is simple and the alternative is composite, the SPRT is still valid with denominator $f(X_{ki}, \mathcal{X}_k^0)$. However, there is no uniquely optimal choice for the numerator. One option is to form predictable estimates of the parameters of $\mathcal{X}_k$ and plug them into the numerator [Ramdas et al., 2023]. The method of mixtures leads to a consistent test, and is GRO (in the terminology of Grünwald et al. [2023]) when the mixing distribution is a generative prior, from which the alternative was drawn.