Title
Three Statistical Methods for the Social Sciences

Author
Miratrix, Luke Weisman

Publication Date
2012

Peer reviewed|Thesis/dissertation

Three Statistical Methods for the Social Sciences

by

Luke Weisman Miratrix

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 Bin Yu, Co-chair
Professor Jasjeet Sekhon, Co-chair
Professor Terry Speed
Professor Laurent El Ghaoui

Spring 2012

Three Statistical Methods for the Social Sciences

Abstract

Three Statistical Methods for the Social Sciences

by

Luke Weisman Miratrix

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Co-chair

Professor Jasjeet Sekhon, Co-chair

Social sciences offer particular challenges to statistics due to difficulties such as conducting randomized experiments in this domain, the large variation in humans, the difficulty in collecting complete datasets, and the typically unstructured nature of data at the human scale. New technology allows for increased computation and data recording, which has in turn brought forth new innovations for analysis. Because of these challenges and innovations, statistics in the social sciences is currently thriving and vibrant. This dissertation is an argument for evaluating statistical methodology in the social sciences along four major axes: *validity*, *interpretability*, *transparency*, and *employability*. We illustrate how one might develop methods that achieve these four goals with three case studies.

The first is an analysis of post-stratification, a form of covariate adjustment to evaluate treatment effect. In contrast to recent results showing that regression adjustment can be problematic under the Neyman-Rubin model, we show post-stratification, something that can easily done in, e.g., natural experiments, has a similar precision to a randomized block trail as long as there are not too many strata. The difference is $O(1/n^2)$. Post-stratification thus potentially allows for transparently exploiting predictive covariates and random mechanisms in observational data. This case study illustrates the value of analyzing a simple estimator under weak assumptions, and of finding similarities between different methodological approaches so as to leverage earlier findings to a new domain.

We then present a framework for building statistical tools to extract topic-specific keyphrase summaries of large text corpora (e.g., the New York Times) and a human validation experiment to determine best practices for this approach. These tools, built from high-dimensional, sparse classifiers such as L1-logistic regression and the Lasso, can be used to, for example, translate essential concepts across languages, investigate massive databases of aviation reports, or understand how different topics of interest are covered by various media outlets. This case study demonstrates how more modern methods can be evaluated using external validation in order to demonstrate that they produce meaningful and comprehendible results that can be broadly used.

The third chapter presents the trinomial bound, a new auditing technique for elections rooted in very minimal assumptions. We demonstrated the usability of this technique by, in November 2008, auditing contests in Santa Cruz and Marin counties, California. The audits were risk-limiting, meaning they had a pre-specified minimum chance of requiring a full hand count if the outcomes were wrong. The trinomial bound gave better results than the Stringer bound, a tool common in accounting for analyzing financial audit samples drawn with probability proportional to an error bound. This case study focuses on generating methods that are employable and transparent so as to serve a public need.

Throughout, we argue that, especially in the difficult domain of the social sciences, we must spend extra attention on the first axis of validity. This motivates our using the Neyman-Rubin model for the analysis of post-stratification, our developing an approach for external, model-independent validation for the key-phrase extraction tools, and our minimal assumptions for election auditing.

To Mary Miratrix

For the adventures and the crazy.

Contents

List of Figures

List of Tables

Acknowledgments

My prior education leading up to this work has been rather circuitous. I have gained much from this journey, mainly due to the wonderful advice and mentorship I received at each stage. In particular, I would like to thank my (most recent) advisors Prof. Bin Yu and Prof. Jasjeet Sekhon. Over recent years, in addition to the guidance given on my work, we have had so many great conversations on so many topics, giving me a reading list and future-directions list as long as I could wish for. I am extremely grateful to Prof. Philip Stark for his willingness to take me on and include me in his work when I came knocking on his door. Without his willingness to give me a try, I would not have had the opportunity to write this dissertation at all. When in the SESAME program I learned a great deal and received wonderful support from Prof. Michael Ranney. He encouraged me to take some excellent courses which eventually led me to statistics' door, and gave me a real feel for the difficult and tricky reality of experiments in the social sciences. I would also be remiss in not acknowledging the many mentors from my teaching career prior to UC Berkeley, and my advisor Prof. Randy Davis from MIT when I was a graduate student there. I am also grateful for Joan Heller, and others who also hired me as a statistics consultant, for sticking with me as I learned how to do the work. Getting to apply statistics in practice gave me an overall sense of direction to my studies. Throughout all this there are many others such as Profs. Terry Speed, Laurent El Ghaoui, and Sophia Rabe-Hesketh, who played vital roles of encouragement, and who showed me worlds that I hadn't known existed.

And now, for some more specific acknowledgements for the particular work of this dissertation: Thanks to my advisors Bin Yu and Jasjeet Sekhon for helping me so much with the work on post-stratification and text summarization. For the post-stratification chapter, I also thank Peter Aronow, Winston Lin, Terry Speed, Jonathan Wand, and two anonymous reviewers for helpful comments. For the text-summarization work, I am extremely grateful to my collaborators Prof. Laurent El Ghaoui, Prof. Jinzhu Jia and Brian Gawalt. This work was partially supported by the NSF grant SES-0835531 under the "Cyber-Infrastructure and Discovery program," NSF grant DMS-0907632, ARO grant W911NF-11-1-0114, NSF grant CCF-0939370, and NSF-CMMI grant 30148. For the election work, I am grateful to Philip Stark for his mentorship. For allowing us to try election auditing methods, I am extremely grateful to Marin County Registrar of Voters Elaine Ginnold, Santa Cruz County Clerk Gail Pellerin. I am also grateful to their staffs for their generous cooperation and the considerable time and effort they spent counting ballots by hand. Throughout, I am grateful for the support of a Graduate Research Fellowship from the National Science Foundation. And finally, I want to thank my family and friends for their support and encouragement as I did this work. I take full responsibility for all errors in this dissertation.

Chapter 1

Introduction

The social sciences are becoming increasingly quantitative. A variety of factors are causing this shift. First and foremost, information technology allows for the recording of human activity in a much more widespread and detailed manner than ever before. This gives the researcher much greater access to quantitative information, allowing for analyses that were impossible before. Second, increased computational power allows for styles of inference and analyses that would otherwise be impossible. Third, we have increased allegiance to statistical methods and inference. For example, in education the "No Child Left Behind" act, heavily focused on quantitative outcomes, calls for randomized experiments to assess novel education policy. New professional groups such as SREE (the Society for Research on Educational Effectiveness) have explicit mission statements calling for more quantitative and causally rooted research. Political science now places greater value on field trials and natural experiments.

The data available on almost anything is growing due to technological advances and these new data are often unstructured, qualitative, or otherwise not of any traditional type. For example, text data, which is now easily accessible in vast quantities, has a lack of plausible models, an extremely high-dimensional setting, and many big-data concerns. These new sources of data, coupled with greater computational power, have given rise to many novel statistical methods. So far, however, much of this work has focused on prediction and classification. This is not enough. In the social sciences there is real interest in finding explanatory structure in data, as illustrated by the many questions on how to use large bodies of text to understand the causal impacts of policy decisions voiced at a recent "Data Without Borders" event.[1] For example, one NGO operating in rural India wanted to evaluate the efficacy of their remote-MD program using all the text messages between their doctors and patients. Explanatory efforts require the results of an analyses be interpretable and valid. For the social sciences in particular validity is a huge concern due to their focus on understanding causal mechanisms or underlying truths rather than predicting future outcomes.

[1]Data Without Borders is an event which couples volunteer data analysts with needy NGOs and non-profits

Furthermore, methods need to be accessible. New methods should not be baroque, for then any results based on them will likely be unconvincing because of the high barrier to understanding their logical chains of reasoning. If new methods are overly complex or opaque then they are likely to be only used, if at all, in internal communities that other scientists, in effect, will ignore. Yes, the individual characters of different types of data require specifically tailored approaches, but this cannot be taken to an extreme. We need an approach to research that exploits the data and computational resources of the modern world while remaining faithful to classical statistics' values of meaning and interpretability successfully communicated.

Given these concerns, we propose four major axes of evaluation for a statistical method: its *validity*, its *interpretability*, its *transparency*, and its *employability*. *Validity* encompasses, for example, appropriately measuring uncertainty in practice as compared to theory. *Interpretability* reflects the ability of the researcher to connect statistical findings to scientific meaning. *Transparency* is a measure of how easily other researchers would be able to follow and evaluate the work, and *employability* is a measure of how accessible it is for researchers to implement or use the method. These four axes, coming from a rich tradition of classical statistics, can serve as a guide for creating high-quality statistical methods in the increasingly complex modern world. This is the focus of this dissertation.

1.1 Three Illustrative Projects

Good science coupled with pragmatism requires methods for looking at, testing, and presenting data in a manner that does not grievously harm the quality of conclusions drawn. We illustrate this coupling with three projects in three chapters that, we hope, achieve the four goals outlined above. These projects are rooted in modern-day concerns: adjusting treatment effect estimates in natural experiments motivated investigating post-stratification under the Neyman-Rubin model; social scientists grappling with large text corpora led to a novel form of summarization using sparse, high-dimensional regression techniques; and the concerns about election security from the 2000 US election motivated new methods for auditing elections built from first principles in probability. Below, we summarize these projects before discussing how they illustrate an overall philosophy for doing applied statistics.

Post-stratification

Chapter 2, taken from Miratrix, Sekhon and Yu (2012)[54], is on non-parametric adjustment of randomized experiments. Randomized experiments have long been the gold standard of statistical inference and are a key building block for proving the causal effects of, e.g., medical drugs, voter turnout drives, or novel education programs. Our interest in randomized experiments comes from looking for ways to analyze observational data. One approach to observational data is to first identify a random mechanism at play and then conduct an analysis based on that randomization. This creates a pseudo- (or natural) experiment, and

makes the assumptions behind the analysis more explicit and therefore transparent, thus increasing the legitimacy of any conclusions drawn. For example, if a scholarship were given to all students scoring above a fixed threshold on a test, one might reasonably assume that the students just shy of the threshold and those just above it were essentially equivalent populations. Such observational data could be modeled and analyzed as if it were an experiment where the students near the cut-off score were randomly assigned scholarships. This practice, called Regression-Discontinuity, has recently become popular in political science.

To increase power and so reduce cost, researchers will often analyze experiments with regression. For natural experiments, researchers adjust for similar reasons. This can be problematic: such adjustments often require making false assumptions with unclear consequences. An alternative to regression adjustment, inspired by the sampling literature, is post-stratification. Post-stratification is stratifying experimental units, estimating effects within the strata, and averaging appropriately to obtain an overall treatment effect after an experiment is complete. Post-stratification is in spirit akin to blocking, a pre-randomization technique used to increase power and ensure balance. With post-stratification, however, the number of treated units in the strata can vary. We proved post-stratification to be actually quite comparable to blocking by showing the difference in these approaches' variance to be of order $1/n^2$, with n being the number of experimental units. This means that researchers can analyze a natural experiment—where they have no control over the randomization—"as if" they had appropriately blocked it by design. However we also found that in finite sample situations post-stratification can substantially hurt precision if the number of strata is large and the stratification variable poorly chosen. Understanding this trade-off is key for researchers attempting to appropriately analyze experimental results.

Statistics in the media

News media heavily impacts policy through influencing public perception of world events. It is important to understand how coverage of important topics changes over time, and how different sources compare. For example, examining how different media sources portrayed Muslim countries, and how that changed throughout the Iraq war, could give insight into the shifts in public approval for the war. Information technology makes it possible to accumulate vast amounts of text, e.g., the entire output of many media source over years, but analyzing these data requires novel, computationally-intensive techniques.

The StatNews project, headed by Prof. El Ghaoui and Prof. Yu, builds statistical tools and visualization aids to aid such endeavors. These tools rely on topic-specific summarization of large text corpora, which is the focus of Chapter 3, adapted from Miratrix, Jia, Gawalt, Yu, and El Ghaoui (2011)[56]. This style of summarization had not been widely investigated, and traditional summarization techniques do not seem to easily translate. To solve this problem we used sparse regression techniques (e.g., ℓ_1-penalized logistic regression or the Lasso) to generate relevant lists of key phrases that are topic specific. The sparsity ensures resulting summaries are human readable and interpretable. It also allows for the severely high-dimensional setting of text, where any phrase is a covariate. However, as

text is not generated by any plausible, tractable random model, we used a human validation experiment to evaluate the different summarizers. We then redesigned this experiment into a randomized, multi-factor trial to examine how different choices in how to summarize impacted final outcomes.

Our summarization approach uses model-intensive sparse regression, but no theoretical results necessarily hold. In our view, externally validating this approach without regard to the model whatsoever is the only way to demonstrate the legitimacy of the approach. This chapter most fully illustrates how one might use modern methods in the spirit of classical statistics.

Election auditing

Election integrity is a crucial part of a working democracy: we need to be sure that perceived winners are actual winners. Statistical election audits are audits that have a pre-specified minimal chance of noticing if this is not the case. They would typically be conducted by hand-counting ballots in public view. Election auditing has recently gained importance due to votes now being tallied either entirely electronically or via automated machines. Due to fiscal constraints, we prefer audit methods that keep the total number of hand-counted ballots small while preserving this minimal chance of error detection.

But we need to be careful in what we assume. In election auditing, any assumptions made could, in principle, be exploited by an adversary. We need to account for possible fraud as well as random error. Furthermore, transparency is key—the public should ideally understand the process and believe in its effectiveness so as to bolster faith in the electoral process.

In Chapter 4, originally published as Miratrix and Stark (2009)[55] and reprinted with permission from IEEE, we develop the trinomial bound using techniques adapted from the financial auditing literature and then prove that it both controls risk in all circumstances and greatly reduces auditing load in most circumstances. Trinomial bound audits are based only on the ability to accurately count a small sample of precincts by hand after selecting them at random from the entire race. No further assumptions are necessary, and thus these audits are not easily manipulable. We demonstrated that this method was viable by auditing two races in California in conjunction with election officials (we certified both races). This work has also shaped policy debates on legislative reform for election auditing in several states.

Conclusion

This dissertation aims to demonstrate with these three different case studies that a simple structure proven to be correct gives maximal transparency and legitimacy to an analysis. When a simple structure is impossible, then external validation can help avoid spurious conclusions. In either case, for a statistical analysis to be compelling it needs to be transparent; conclusions are legitimate only if they are unlikely to be artifacts of the analysis but they are convincing only if that fact is made clear. If possible, a method should be accessible to

a broad audience both so many can understand resulting analyses and so they can conduct their own.

This dissertation proposes the four qualities of *validity*, *interpretability*, *transparency*, and *employability* be guideposts to aid applied statisticians as they tackle modern problems. Achieving all four qualities requires deep thinking and careful analysis. Creating and understanding a simple method is often, unfortunately, a very complex task. Making a complex method accessible is typically even more daunting. But, hopefully by attending to these four axes, statisticians can generate novel and useful methods for doing statistical work in the social sciences. Ideally, this work shows how a few small steps can be taken in this direction.

Chapter 2

Post-Stratification

2.1 Introduction

One of the most important tools for determining the causal effect of some action is the randomized experiment, where a researcher randomly divides units into groups and applies different treatments to each group. Randomized experiments are the "gold standard" for causal inference because, assuming proper implementation of the experiment, if a difference in outcomes is found, the only possible explanations are a significant treatment effect or random chance. Math gives a handle on the chance, which allows for principled inference about the treatment effect. In the most basic analysis, a simple difference in means is used to estimate the overall sample average treatment effect (SATE), defined as the average difference in the units' outcomes if all were treated as compared to their average outcomes if they were not. This framework and estimator were analyzed by Neyman in 1923[1] under what is now called the Neyman or Neyman-Rubin model of potential outcomes [37]. Under this model, one need make few assumptions not guaranteed by the randomization itself.

Since each additional observation in an experiment sometimes comes at considerable cost, it is desirable to find more efficient estimators than the simple difference-in-means estimator to measure treatment effects. Blocking, which is when experimenters first stratify their units and then randomize treatment within pre-defined blocks, can greatly reduce variance compared to the simple-difference estimator if the strata differ from each other. See "A Useful Method" in [22] for an early overview, [89] for an analysis and comparison with ANOVA, or [43]. However, because blocking must be conducted before randomization, it is often not feasible due to practical considerations or lack of foresight. Sometimes randomization may even be entirely out of the researcher's control, such as with so-called natural experiments. When blocking was not done, researchers often adjust for covariates after randomization. For example, [63] studied a sample of clinical trials analyses and found that 72% of these articles used covariate adjustment. [45] analyzed the experimental results in three major political science journals and found that 74% to 95% of the articles relied on adjustment. Post-

[1] See the English translation by [78].

stratification is one simple form of adjustment where the researcher stratifies experimental units with a pretreatment variable after the experiment is complete, estimates treatment effects within the strata, and then uses a weighted average of these strata estimates for the overall average treatment effect estimate. This is the estimator we focus on.

In this chapter, we use the Neyman-Rubin model to compare post-stratification both to blocking and to using no adjustment. Neyman's framework does not require assumptions of a constant treatment effect or of identically or independently distributed disturbances, assumptions typically made when considering adjustment to experimental data without this framework [e.g., 52]. This avenue for a robust analysis, revitalized by Rubin in the 1970s [68], has recently had much appeal. See, for example, work on general experiments [45], matched pairs [41], or matched pairs of clusters [42].[2] Also see Neyman's own treatment of blocking in the appendix of [60]. Our estimator is equivalent to one from a fully saturated OLS regression. [25, 26] analyzes the regression adjusted estimator under the Neyman-Rubin model without treatment by strata interactions and finds that the asymptotic variance might be larger than if no correction were made. [50] extends Freedman's results and shows that when a treatment by covariate interaction is included in the regression, adjustment cannot increase the asymptotic variance. We analyze the exact, finite sample properties of this saturated estimator. [44] analyzes estimating the treatment effect of a larger population, assuming the given sample being experimented on is randomly drawn from it. However, because in most randomized trials the sample is not taken at random from the larger population of interest, we focus on estimating the treatment effect of the sample. [87] and [47] propose other adjustment methods that also rely on weak assumptions and that have the important advantage of working naturally with continuous or multiple covariates. Due to different sets of assumptions and methods of analysis, these estimators have important differences from each other. See Section 2.6 for further discussion.

We derive the variances for post-stratification and simple difference-in-means estimators under many possible randomization schemes. We show that the difference between the variance of the post-stratified estimator and that of a blocked experiment is on the order of $1/n^2$ with a constant primarily dependent on the proportion of units treated. Post-stratification is comparable to blocking. Like blocking, post-stratification can greatly reduce variance over using a simple difference-in-means estimate. However, in small samples post-stratification can substantially hurt precision, especially if the number of strata is large and the stratification variable poorly chosen.

After randomization, researchers can observe the proportion of units actually treated in each strata. We extend our results by deriving variance formula for the post-stratified and simple-difference estimators conditioned on these observed proportions. These conditional formula help explain why the variances of the estimators can differ markedly with a prognostic covariate: the difference comes from the potential for bias in the simple-difference estimator when there is large imbalance (i.e., when these proportions are far from what is expected). Interestingly, if the stratification variable is not predictive of outcomes the

[2] See [73] for a historical review of the Neyman-Rubin model.

conditional MSE of the simple-difference estimator usually remains the same or even goes down with greater imbalance, but the conditional MSE of the adjusted estimator increases. Adjusting for a poorly chosen covariate has real cost in finite samples.

The rest of the chapter is organized as follows: In the next section, we set up the Neyman-Rubin model, describe the estimators, and then derive the estimators' variances. In Section 2.3 we show that post-stratification and blocking have similar characteristics in many circumstances. In Section 2.4, we present our formula for the estimators' variances conditioned on the observed number of treated units in the strata and discuss their implications. We then align our results with those of [44] in Section 2.5 by extending our findings to the super-population model and discussing the similarities and differences of the two viewpoints. We compare post-stratification to other forms of adjustment in Section 2.6, focusing on how these different approaches use different assumptions. In Section 2.7, we apply our method to the real data example of a large, randomized medical trial to assess post-stratification's efficacy in a real-world example. We also make a hypothetical example from this data set to illustrate how an imbalanced randomization outcome can induce bias which the post-stratified estimator can adjust for. Section 2.8 concludes.

2.2 The Estimators and Their Variances

We consider the Neyman-Rubin model with two treatments and n units. For example consider a randomized clinical trial with n people, half given a drug and the other half given a placebo. Let $y_i(1) \in \mathbb{R}$ be unit i's outcome if it were treated, and $y_i(0)$ its outcome if it were not. These are the *potential outcomes* of unit i. For each unit, we can only observe either $y_i(1)$ or $y_i(0)$ depending on whether we treat it or not. We make the assumption that treatment assignment for any particular unit has no impact on the potential outcomes of any other unit (this is typically called the stable-unit treatment value assumption or SUTVA). In the drug example this means the decision to give the drug to one patient would have no impact on the outcome of any other patient. The treatment effect t_i for unit i is then the difference in potential outcomes, $t_i \equiv y_i(1) - y_i(0)$, which is deterministic.

Although these t_i are the quantities of interest, we cannot in general estimate them because we cannot observe both potential outcomes of any unit i and because the t_i generally differ by unit. The average across a population of units, however, is estimable. Neyman [78] considered the overall Sample Average Treatment Effect, or SATE:

$$\tau \equiv \frac{1}{n}\sum_{i=1}^{n}[y_i(1) - y_i(0)]$$

To conduct an experiment, randomize units into treatment and observe outcomes. Many choices of randomization are possible. The observed outcome is going to be one of the two potential outcomes, and which one depends on the treatment given. Random assignment gives a treatment assignment vector $T = (T_1, \ldots, T_n)$ with $T_i \in \{0, 1\}$ being an indicator

variable of whether unit i was treated or not. T's distribution depends on how the randomization was conducted. After the experiment is complete, we obtain the observed outcomes Y, with $Y_i = T_i y_i(1) + (1 - T_i)y_i(0)$. The observed outcomes are random—but only due to the randomization used. The $y_i(\ell)$ and t_i are all fixed. Neyman considered a *balanced complete randomization*:

Definition 2.2.1 (Complete Randomization of n Units)**.** Given a fixed $p \in (0, 1)$ such that $0 < pn < n$ is an integer, a *Complete Randomization* is a simple random sample of pn units selected for treatment with the remainder left as controls. If $p = 0.5$ (and n is even) the randomization is *balanced* in that there are the same number of treated units as control units.

The classic unadjusted estimator $\hat{\tau}_{sd}$ is the observed *simple difference* in the means of the treatment and control groups:

$$\begin{aligned} \hat{\tau}_{sd} &= \frac{1}{W(1)} \sum_{i=1}^{n} T_i Y_i - \frac{1}{W(0)} \sum_{i=1}^{n} (1 - T_i) Y_i \\ &= \sum_{i=1}^{n} \frac{T_i}{W(1)} y_i(1) - \sum_{i=1}^{n} \frac{(1 - T_i)}{W(0)} y_i(0), \end{aligned}$$

where $W(1) = \sum_i T_i$ is the total number of treated units, $W(0)$ is total control, and $W(1) + W(0) = n$. For Neyman's balanced complete randomization, $W(1) = W(0) = n/2$. For other randomizations schemes the $W(\ell)$ are potentially random.

We analyze the properties of the estimators based on the randomization scheme used—this is the source of randomness. Fisher proposed a similar strategy for testing the "sharp null" hypothesis of no effect (where $y_i(0) = y_i(1)$ for $i = 1, \ldots, n$); under this view, all outcomes are known and the observed difference in means is compared to its exact, known distribution under this sharp null. Neyman, in contrast, *estimates* the variance of the difference in means, allowing for the unknown counterfactual outcomes of the units to vary. These different approaches have different strengths and weaknesses that we do not here discuss. We follow this second approach.

Neyman showed that the variance of $\hat{\tau}_{sd}$ is

$$\mathrm{Var}[\hat{\tau}_{sd}] = \frac{2}{n} \mathbb{E}\left[s_1^2 + s_0^2\right] - \frac{1}{n} S^2 \tag{2.1}$$

where s_ℓ^2 are the sample variances of the observed outcomes for each group, S^2 is the variance of the n treatment effects t_i, and the expectation is over all possible assignments under balanced complete randomization. We extend this work by considering an estimator that (ideally) exploits some pretreatment covariate b using post-stratification in order to reduce variance.

The Post-Stratified Estimator of SATE

Stratification is when an experimenter divides the experimental units into K strata according to some categorical covariate b with $b_i \in \mathcal{B} \equiv \{1, \ldots, K\}$, $i = 1, \ldots, n$. Each stratum k contains $n_k = \#\{i : b_i = k\}$ units. For example, in a cancer drug trial we might have the strata being different stages of cancer. If the strata are associated with outcome, an experimenter can adjust a treatment effect estimate to remove the impact of random variability in the proportions of units treated. This is the idea behind post-stratification. The b_i are observed for all units and are not affected by treatment. The strata defined by the levels of b have stratum-specific SATE_k:

$$\tau_k \equiv \frac{1}{n_k} \sum_{i:b_i=k} [y_i(1) - y_i(0)] \qquad k = 1, \ldots, K.$$

The overall SATE can then be expressed as a weighted average of these SATE_ks:

$$\tau \;\; = \;\; \sum_{k \in \mathcal{B}} \frac{n_k}{n} \tau_k. \tag{2.2}$$

We can view the strata as K mini-experiments. Let $W_k(1) = \sum_{i:b_i=k} T_i$ be the number of treated units in stratum k, and $W_k(0)$ be the number of control units. We can use a simple-difference estimator for each stratum to estimate the SATE_ks:

$$\hat{\tau}_k \;\; = \;\; \sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1) - \sum_{i:b_i=k} \frac{(1 - T_i)}{W_k(0)} y_i(0), \tag{2.3}$$

A post-stratification estimator is an appropriately weighted estimate of these strata-level estimates:

$$\hat{\tau}_{ps} \;\; \equiv \;\; \sum_{k \in \mathcal{B}} \frac{n_k}{n} \hat{\tau}_k. \tag{2.4}$$

These weights echo the weighted sum of SATE_ks in Equation 2.2. Because b and n are known and fixed, the weights are also known and fixed. We derive the variance of $\hat{\tau}_{ps}$ in this chapter.

Technically, this estimator is undefined if $W_k(1) = 0$ or $W_k(0) = 0$ for any $k \in 1, \ldots, K$. Similarly, τ_{sd} is undefined if $W(1) = 0$ or $W(0) = 0$. We therefore calculate all means and variances conditioned on $\mathcal{D}$, the event that $\hat{\tau}_{ps}$ is defined, i.e., that each stratum has at least one unit assigned to treatment and one to control. This is fairly natural: if the number of units in each stratum is not too small the probability of $\mathcal{D}$ is close to 1 and the conditioned estimator is similar to an appropriately defined unconditioned estimator. See Section 2.2.

Different experimental designs and randomizations give different distributions on the treatment assignment vector T and all resulting estimators. Some distributions on T would cause bias. We disallow those. Define the *Treatment Assignment Pattern* for stratum k as the ordered vector $(T_i : i \in \{1, \ldots, n : b_i = k\})$. We assume that the randomization used has *Assignment Symmetry*:

Definition 2.2.2 (Assignment Symmetry)**.** A randomization is *Assignment Symmetric* if the following two properties hold:

1. *Equiprobable Treatment Assignment Patterns*
   All $\binom{n_k}{W_k(1)}$ ways to treat $W_k(1)$ units in stratum k are equiprobable, given $W_k(1)$.
2. *Independent Treatment Assignment Patterns*
   For all strata j, k, with $j \neq k$, the treatment assignment pattern in stratum j is independent of the treatment assignment pattern in stratum k, given $W_j(1)$ and $W_k(1)$.

Complete randomization and Bernoulli assignment (where independent p-coin flips determine treatment for each unit) satisfy Assignment Symmetry. So does blocking, where strata are randomized independently. Furthermore, given a distribution on T that satisfies Assignment Symmetry, conditioning on $\mathcal{D}$ maintains Assignment Symmetry (as do many other reasonable conditionings, such as having at least x units in both treatment and control, and so on). See the supplementary material for a more formal argument. Cluster randomization or randomization where units have unequal treatment probabilities do not, in general, have Assignment Symmetry. In our technical results, we assume that (1) the randomization is Assignment Symmetric and (2) we are conditioning on $\mathcal{D}$, the set of possible assignments where $\hat{\tau}_{ps}$ is defined.

The post-stratification estimator and the simple-difference estimator are used when the initial random assignment ignores the stratification variable b. In a blocked experiment, the estimator used is $\hat{\tau}_{ps}$, but the randomization is done within the strata defined by b. All three of these options are unbiased. We are interested in their relative variances. We express the variances of these estimators with respect to the sample's (unknown) means, variances and covariance of potential outcomes divided into between-strata variation and within-stratum variation. The within-stratum variances and covariances are, for $k = 1, \ldots, K$:

$$\sigma_k^2(\ell) = \frac{1}{n_k - 1} \sum_{i:b_i=k} \left[y_i(\ell) - \bar{y}_k(\ell)\right]^2 \qquad \ell = 0, 1$$

and

$$\gamma_k(1,0) = \frac{1}{n_k - 1} \sum_{i:b_i=k} \left[y_i(1) - \bar{y}_k(1)\right] \left[y_i(0) - \bar{y}_k(0)\right],$$

where $\bar{y}_k(\ell)$ denotes the mean of $y_i(\ell)$ for all units in stratum k. Like many authors, we use $n_k - 1$ rather than n_k for convenience and cleaner formula. The $(1, 0)$ in $\gamma_k(1, 0)$ indicates that this framework could be extended to multiple treatments.

The between-stratum variance and covariance are weighted variances and covariances of the strata means:

$$\bar{\sigma}^2(\ell) = \frac{1}{n - 1} \sum_{k=1}^{K} n_k \left[\bar{y}_k(\ell) - \bar{y}(\ell)\right]^2 \qquad \ell = 0, 1$$

and

$$\bar{\gamma}(1,0) = \frac{1}{n-1}\sum_{k=1}^{K} n_k \left[\bar{y}_k(1) - \bar{y}(1)\right]\left[\bar{y}_k(0) - \bar{y}(0)\right].$$

The population-wide $\sigma^2(\ell)$ and $\gamma(1,0)$ are analogously defined. They can also be expressed as weighted sums of the component pieces. We also refer to the *correlation of potential outcomes* r, where $r \equiv \gamma(1,0)/\sigma(0)\sigma(1)$ and the strata-level correlations, $r_k \equiv \gamma_k(1,0)/\sigma_k(0)\sigma_k(1)$. An overall constant treatment effect gives $r = 1$, $\sigma(0) = \sigma(1)$, $r_k = 1$ for all k and $\sigma_k(0) = \sigma_k(1)$ for all k.

We are ready to state our main results:

Theorem 2.2.1. *The strata-level estimators $\hat{\tau}_k$ are unbiased, i.e.*

$$\mathbb{E}[\hat{\tau}_k] = \tau_k \qquad k = 1, \ldots, K$$

and their variances are

$$Var[\hat{\tau}_k] = \frac{1}{n_k}\left[\beta_{1k}\sigma_k^2(1) + \beta_{0k}\sigma_k^2(0) + 2\gamma_k(1,0)\right] \tag{2.5}$$

with $\beta_{1k} = \mathbb{E}[W_k(0)/W_k(1)|\mathcal{D}]$, the expected ratio of the number of units in control to the number of units treated in stratum k, and $\beta_{0k} = \mathbb{E}[W_k(1)/W_k(0)|\mathcal{D}]$, the reverse.

Theorem 2.2.2. *The post-stratification estimator $\hat{\tau}_{ps}$ is unbiased:*

$$\mathbb{E}[\hat{\tau}_{ps}|\mathcal{D}] = \mathbb{E}\left[\sum_k \frac{n_k}{n}\hat{\tau}_k\right] = \sum_k \frac{n_k}{n}\,\mathbb{E}[\hat{\tau}_k] = \sum_k \frac{n_k}{n}\tau_k = \tau.$$

Its variance is

$$Var[\hat{\tau}_{ps}|\mathcal{D}] = \frac{1}{n}\sum_k \frac{n_k}{n}\left[\beta_{1k}\sigma_k^2(1) + \beta_{0k}\sigma_k^2(0) + 2\gamma_k(1,0)\right]. \tag{2.6}$$

See Appendix A for a proof. In essence we expand the sums, use iterated expectation, and evaluate the means and variances of the treatment indicator random variable. Assignment Symmetry allows for the final sum. Techniques used are similar to those found in many papers classic (e.g., [60, 85]) and recent (e.g., [43]).

Consider the whole sample as a single stratum and use Theorem 2.2.1 to immediately get:

Corollary 2.2.3. *The unadjusted simple-difference estimator $\hat{\tau}_{sd}$ is unbiased, i.e. $\mathbb{E}[\hat{\tau}_{sd}] = \tau$. Its variance is*

$$Var[\hat{\tau}_{sd}|\mathcal{D}] = \frac{1}{n}\left[\beta_1\sigma^2(1) + \beta_0\sigma^2(0) + 2\gamma(1,0)\right], \tag{2.7}$$

where $\beta_1 \equiv \mathbb{E}[W(0)/W(1)|\mathcal{D}]$ *and* $\beta_0 \equiv \mathbb{E}[W(1)/W(0)|\mathcal{D}]$. *In terms of strata-level variances, its variance is*

$$\begin{aligned} Var[\hat{\tau}_{sd}|\mathcal{D}] &= \frac{1}{n}\left[\beta_1\bar{\sigma}^2(1) + \beta_0\bar{\sigma}^2(0) + 2\bar{\gamma}(1,0)\right] + \\ &\quad \frac{1}{n}\sum_k \frac{n_k - 1}{n-1}\left[\beta_1\sigma_k^2(1) + \beta_0\sigma_k^2(0) + 2\gamma_k(1,0)\right]. \end{aligned} \tag{2.8}$$

For completely randomized experiments with np units treated, $\beta_1 = (1-p)/p$ and $\beta_0 = p/(1-p)$. For a balanced completely randomized experiment, Equation 2.7 is the result presented in [78]—see Equation 2.1; the expectation of the sample variance is the overall variance. Then $\beta_\ell = 1$ and

$$\begin{aligned} \mathrm{Var}[\hat{\tau}_{sd}] &= \frac{1}{n}\left(\sigma^2(1) + \sigma^2(0) + 2\gamma(1,0)\right) \\ &= \frac{2}{n}\left(\sigma^2(1) + \sigma^2(0)\right) - \frac{1}{n}\left(\sigma^2(1) + \sigma^2(0) - 2\gamma(1,0)\right) \\ &= \frac{2}{n}\left(\sigma^2(1) + \sigma^2(0)\right) - \frac{1}{n}\mathrm{Var}[y_i(1) - y_i(0)]. \end{aligned}$$

Remarks. β_{1k} is the expectation of $W_k(0)/W_k(1)$, the ratio of control units to treated units in stratum k. For large n_k, this ratio is close to the ratio $\mathbb{E}[W_k(0)]\,/\,\mathbb{E}[W_k(1)]$ since the $W_k(\ell)$ will not vary much relative to their size. For small n_k, however, they will vary more, which tends to result in β_{1k} being noticeably larger than $\mathbb{E}[W_k(0)]\,/\,\mathbb{E}[W_k(1)]$. This is at root of how the overall variance of post-stratification differs from blocking. This is discussed more formally later on and in Appendix A.

For $\ell = 0, 1$ the $\beta_{\ell k}$'s are usually larger than β_ℓ, being expectations of different variables with different distributions. For example in a balanced completely randomized experiment $\beta_1 = 1$ but $\beta_{1k} > 1$ for $k = 1, \ldots, K$ since $W_k(1)$ is random.

All the β's depend on both the randomization and the conditioning on $\mathcal{D}$, and thus the variances from both Equation 2.8 and Equation 2.6 can change (markedly) under different randomization scenarios. As a simple illustration, consider a complete randomization of a 40 unit sample with a constant treatment effect and four strata of equal size. Let all the $\sigma_k(\ell) = 1$ and all $r_k = 1$. If $p = 0.5$, then $\beta_1 = \beta_0 = 1$ and the variance is about 0.15. If $p = 2/3$ then $\beta_1 = 1/2$ and $\beta_0 = 2$. Equation 2.8 holds in both cases, but the variance in the second case will be about 10% larger due to the larger β_0. There are fewer control units, so the estimate of the control outcome is more uncertain. The gain in certainty for the treatment units does not compensate enough. For $p = 0.5$, $\beta_{1k} = \beta_{0k} \approx 1.21$. The post-stratified variance is about 0.11. For $p = 2/3$, $\beta_{1k} \approx 2.44$ and $\beta_{0k} \approx 0.61$. The average is about 1.52. The variance is about 14% larger than the $p = 0.5$ case. Generally speaking, the relative variances of different experimental setups are represented in the β's.

The correlation of potential outcomes, $\gamma_k(1,0)$, can radically impact the variance. If they are maximally negative, the variance can be zero or nearly zero. If they are maximally

positive (as in the case of a constant treatment effect), the variance can be twice what it would be if the outcomes were uncorrelated.

Comparing the Estimators. Both $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ are unbiased, so their MSEs are the same as their variances. To compare $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ take the difference of their variances:

$$\begin{aligned}\mathrm{Var}[\hat{\tau}_{sd}] - \mathrm{Var}[\hat{\tau}_{ps}] = &\left\{ \frac{1}{n}\left(\beta_1\bar{\sigma}^2(1) + \beta_0\bar{\sigma}^2(0) + 2\bar{\gamma}(1,0)\right)\right\} - \\ &\left\{ \frac{1}{n}\sum_{k=1}^{K}\left[\left(\frac{n_k}{n}\beta_{1k} - \frac{n_k-1}{n-1}\beta_1\right)\sigma_k^2(1) + \left(\frac{n_k}{n}\beta_{0k} - \frac{n_k-1}{n-1}\beta_0\right)\sigma_k^2(0)\right] + \right. \\ &\left. \frac{2}{n^2}\sum_{k=1}^{K}\frac{n-n_k}{n-1}\gamma_k(1,0)\right\}. \end{aligned} \tag{2.9}$$

Equation 2.9 breaks down into two parts as indicated by the curly brackets. The first part, $\beta_1\bar{\sigma}^2(1)+\beta_0\bar{\sigma}^2(0)+2\bar{\gamma}(1,0)$, is the between-strata variation. It measures how much the mean potential outcomes vary across strata and captures how well the stratification variable separates out different units, on average. The larger the separation, the more to gain by post-stratification. The second part, consisting of the bottom two lines of Equation 2.9, represents the cost paid by post-stratification due to, primarily, the chance of random imbalance in treatment. This second part is non-positive and is a penalty except in some cases where the proportion of units treated is extremely close to 0 or 1 or is radically different across strata.

If the between-strata variation is larger than the cost paid then Equation 2.9 is positive and it is good to post-stratify. If Equation 2.9 is negative then it is bad to post-stratify. It can be positive or negative depending on the parameters of the population. In particular, if there is no between-strata difference in the mean potential outcomes, then the terms on the first line of Equation 2.9 are 0, and post-stratification hurts. Post-stratification is not necessarily a good idea when compared to doing no adjustment at all.

To assess the magnitude of the penalty paid compared to the gain, multiply Equation 2.9 by n. The first term, representing the between-strata variation, is now a constant, and the scaled gain converges to it as n grows:

Theorem 2.2.4. *Take an experiment with n units randomized under either complete randomization or Bernoulli assignment. Let p be the expected proportion of units treated. Without loss of generality, assume $0.5 \le p < 1$. Let $f = \min\{n_k/n : k = 1,\ldots,K\}$ be the proportional size of the smallest stratum. Let $\sigma^2_{max} = \max_{k,\ell}\sigma_k^2(\ell)$ be the largest variance of all the strata. Similarly define γ_{max}. Then the scaled cost term is bounded:*

$$\left| n\left(Var[\hat{\tau}_{sd}] - Var[\hat{\tau}_{ps}]\right) - \beta_1\bar{\sigma}^2(1) - \beta_0\bar{\sigma}^2(0) - 2\bar{\gamma}(1,0)\right| \le C\frac{1}{n} + O(\frac{1}{n^2})$$

with

$$C = \left(\frac{8}{f(1-p)^2} + \frac{2p}{1-p}\right)\sigma^2_{max} + 2K\gamma_{max}.$$

See Appendix A for the derivation. Theorem 2.2.4 shows us that the second part of Equation 2.9, the harm, diminishes quickly.

If the number of strata K grows with n, as is often the case when coarsening a continuous covariate, the story can change. The second and third lines of Equation 2.9 are sums over K elements. The larger the number of strata K, the more terms in the sums and the greater the potential penalty for stratification, unless the $\sigma_k^2(\ell)$'s shrink in proportion as K grows. For an unrelated covariate, they will not tend to do so. To illustrate, we made a sequence of experiments increasing in size with a continuous covariate z unrelated to outcome. For each experiment with n units, we constructed b by cutting z into $K = n/10$ chunks. Post-stratification was about 15% worse, in this case, than the simple-difference estimator regardless of n. See our supplementary materials for details and other illustrative examples. Theorem 2.2.4 captures dependence on the number of strata through f, the proportional size of the smallest strata. If $f \propto 1/K$ then the difference will be $O(K/n)$. For example, if K grows at rate $O(\log n)$, then the scaled difference will be $O(\log n/n)$, nearly $O(1/n)$.

Overall, post-stratifying on variables not heavily related to outcome is unlikely to be worthwhile and can be harmful. Post-stratifying on variables that do relate to outcome will likely result in large between-strata variation and thus a large reduction in variance as compared to a simple-difference estimator. More strata are not necessarily better, however. Simulations suggest that there is often a law of diminishing returns. For example, we made a simulated experiment with $n = 200$ units with a continuous covariate z related to outcome. We then made b by cutting z up into K chunks for $K = 1, \ldots, 20$. As K increased from 1 there was a sharp drop in variance and then, as the cost due to post-stratification increased, the variance leveled off and then climbed. In this case, $K = 5$ was ideal. We did a similar simulation for a covariate z unrelated to outcome. Now, regardless of K, the $\sigma_k^2(\ell)$ were all about the same and the between-strata variation fairly low. As K grew, the overall variance climbed. In many cases a few moderate-sized strata give a dramatic reduction in variance, but having more strata beyond that has little impact, and can even lead to an increase in $\hat{\tau}_{ps}$'s variance. Please see our supplementary material for details.

Estimation. Equation 2.6 and Equation 2.8 are the actual variances of the estimators. In practice, the variance of an estimator, i.e., the squared standard error, would have to itself be estimated. Unfortunately, however, it is usually not possible to consistently estimate the standard errors of difference-in-means estimators due to so-called identifiability issues as these standard errors depend on $\gamma_k(1, 0)$, the typically un-estimable correlations of the potential outcomes of the units being experimented on (see [78]). One approach to consistently estimate these standard errors is to impose structure to render this correlation estimable or known; [66], for example, demonstrate that quite strong assumptions have to be made to obtain an unbiased estimator for the variance of $\hat{\tau}_{sd}$. It is straightforward, however, to make a non-trivial conservative estimate of this variance by assuming the correlation is maximal. Sometimes there can be nice tricks—see, for example, [1], who estimate these parameters for matched-pairs by looking at pairs of pairs matched on covariates—but generally bounding

the standard error is the best one can do.

This chapter compares the actual variances of the estimators. Estimating these variances is an area for future work, involving these identifiability issues and degrees-of-freedom issues as well. It is quite possible that, in small samples, the increased uncertainty in estimating the many variances composing the standard error of the post-stratification estimator would overwhelm any potential gains.

That being said, all terms except the $\gamma_k(1,0)$ in Equation 2.9 are estimable with standard sample variance, covariance, and mean formula. In particular, $\bar{\gamma}(1,0)$ is estimable. By then making the conservative assumption that the $\gamma_k(1,0)$ are maximal (i.e., that $r_k = 1$ for all k so $\gamma_k(1,0) = \sigma(1)\sigma(0)$), we can estimate a lower-bound on the gain. Furthermore, by then dividing by a similar upper bound on the standard error of the simple-difference estimator, we can give a lower-bound on the percentage reduction in variance due to post-stratification. We illustrate this when we analyze an experiment in Section 2.7.

Not Conditioning on $\mathcal{D}$ Changes Little

Our results are conditioned on $\mathcal{D}$, the set of assignments such that $W_k(\ell) \neq 0$ for all $k = 1, \ldots K$ and $\ell = 0, 1$. This, it turns out, results in variances only slightly different from not conditioning on $\mathcal{D}$.

Define the estimator $\hat{\tau}_{ps}$ so that $\hat{\tau}_{ps} = 0$ if $\neg\mathcal{D}$ occurs, i.e. $W_k(\ell) = 0$ for some k, ℓ. Other choices of how to define the estimator when $\neg\mathcal{D}$ occurs are possible, including letting $\hat{\tau}_{ps} = \hat{\tau}_{sd}$—the point is that this choice does not much matter. In our case $\mathbb{E}[\hat{\tau}_{ps}] = \tau\mathbf{P}\mathcal{D}$. The estimate of the treatment is shrunk by $\mathbf{P}\mathcal{D}$ towards 0. It is biased by $\tau\mathbf{P}\neg\mathcal{D}$. The variance is

$$\mathrm{Var}[\hat{\tau}_{ps}] = \mathrm{Var}[\hat{\tau}_{ps}|\mathcal{D}]\,\mathbf{P}\mathcal{D} + \tau^2\mathbf{P}\neg\mathcal{D}\mathbf{P}\mathcal{D}$$

and the MSE is

$$MSE\,[\hat{\tau}_{ps}] = \mathbb{E}\big[(\hat{\tau}_{ps} - \tau)^2\big] = \mathrm{Var}[\hat{\tau}_{ps}|\mathcal{D}]\,\mathbf{P}\mathcal{D} + \tau^2\mathbf{P}\neg\mathcal{D}.$$

Not conditioning on $\mathcal{D}$ introduces a bias term and some extra variance terms. All these terms are small if $\mathbf{P}\neg\mathcal{D}$ is near 0, which it is: $\mathbf{P}\neg\mathcal{D}$ is $O(ne^{-n})$ (see second part of Appendix A). Not conditioning on $\mathcal{D}$, then, gives substantively the same conclusions as conditioning on $\mathcal{D}$, but the formulae are a bit more unwieldy. Conditioning on the set of randomizations where $\hat{\tau}_{ps}$ is defined is more natural.

2.3 Comparing Blocking to Post-Stratification

Let the *assignment split* W of a random assignment be the number of treated units in the strata:

$$W \equiv (W_1(1), \ldots, W_K(1))$$

A *randomized block trial* ensures that W is constant because we randomize within strata, ensuring a pre-specified number of units are treated in each. This randomization is Assignment Symmetric (Def 2.2.2) and under it the probability of being defined, $\mathcal{D}$, is 1. For blocking, the standard estimate of the treatment effect has the same expression as $\hat{\tau}_{ps}$, but the $W_k(\ell)$s are all fixed. If all blocks have the same proportion treated (i.e., $W_k(1)/n_k = W(1)/n$ for all k), this coincides with $\hat{\tau}_{sd}$.

Because W is constant

$$\beta_{1k} = \mathbb{E}\left[\frac{W_k(0)}{W_k(1)}\right] = \frac{W_k(0)}{W_k(1)} = \frac{1-p_k}{p_k}, \tag{2.10}$$

where p_k is the proportion of units assigned to treatment in stratum k. Similarly, $\beta_{0k} = p_k/(1-p_k)$. Letting the subscript "blk" denote this randomization, plug Equation 2.10 into Equation 2.6 to get the variance of a blocked experiment:

$$\mathrm{Var}_{blk}\left[\hat{\tau}_{ps}\right] = \frac{1}{n}\sum_k \frac{n_k}{n}\left(\frac{1-p_k}{p_k}\sigma_k^2(1) + \frac{p_k}{1-p_k}\sigma_k^2(0) + 2\gamma_k(1,0)\right). \tag{2.11}$$

Post-stratification is similar to blocking, and the post-stratified estimator's variance tends to be close to that of a blocked experiment. Taking the difference between Equation 2.6 and Equation 2.11 gives

$$\mathrm{Var}[\hat{\tau}_{ps}|\mathcal{D}] - \mathrm{Var}_{blk}\left[\hat{\tau}_{ps}\right] = \frac{1}{n}\sum_k \frac{n_k}{n}\left[\left(\beta_{1k} - \frac{1-p_k}{p_k}\right)\sigma_k^2(1) + \left(\beta_{0k} - \frac{p_k}{1-p_k}\right)\sigma_k^2(0)\right]. \tag{2.12}$$

The $\gamma_k(1,0)$ cancelled; Equation 2.12 is identifiable and therefore estimable.

Randomization without regard to b can have block imbalance due to ill luck: W is random. The resulting cost in variance of post-stratification over blocking is represented by the $\beta_{1k} - (1-p_k)/p_k$ terms in Equation 2.12. This cost is small, as shown by Theorem 2.3.1:

Theorem 2.3.1. *Take a post-stratified estimator for a completely randomized or Bernoulli assigned experiment. Use the assumptions and definitions of Theorem 2.2.4. Assume the common case for blocking of $p_k = p$ for $k = 1, \ldots, K$. Then*

$$n\Big(\mathit{Var}[\hat{\tau}_{ps}|\mathcal{D}] - \mathit{Var}_{blk}\left[\hat{\tau}_{ps}\right]\Big) \leq \frac{8}{(1-p)^2}\frac{1}{f}\sigma_{max}^2\frac{1}{n} + O(e^{-fn}).$$

See second part of Appendix A for the derivation.

Theorem 2.3.1 bounds how much worse post-stratification can be as compared to blocking. The scaled difference is on the order of $1/n$. The difference in variance is order $1/n^2$. Generally speaking, post-stratification is similar to blocking in terms of efficiency. The more strata, however, the worse this comparison becomes due to the increased chance of severe imbalance with consequential increased uncertainty in the stratum-level estimates. Many strata are generally not helpful and can be harmful if b is not prognostic.

A note on blocking. Plug Equation 2.10 into the gain equation (Equation 2.9) to immediately see under what circumstances blocking has a larger variance than the simple-difference estimator for a completely randomized experiment:

$$\begin{aligned}\mathrm{Var}[\hat{\tau}_{sd}] - \mathrm{Var}_{blk}\left[\hat{\tau}_{ps}\right] &= \frac{1}{n}\left(\frac{1-p}{p}\bar{\sigma}^2(1) + \frac{p}{1-p}\bar{\sigma}^2(0) + 2\bar{\gamma}(1,0)\right) - \\ &\quad \frac{1}{n^2}\sum_k \frac{n-n_k}{n-1}\left(\frac{1-p}{p}\sigma_k^2(1) + \frac{p}{1-p}\sigma_k^2(0) + 2\gamma_k(1,0)\right) \quad (2.13)\end{aligned}$$

If $p = 0.5$, this is identical to the results in the appendix of [43]. In the worst case where there is no between-strata variation, the first term of Equation 2.13 is 0 and so the overall difference is $O(K/n^2)$. The penalty for blocking is small, even for moderate-sized experiments, assuming the number of strata does not grow with n. ([60] noticed this in a footnote of his appendix where he derived the variance of a blocked experiment.) If the first term is not zero, then it will dominate for large enough n, i.e. blocking will give a more precise estimate. For more general randomizations, Equation 2.9 still holds but the β's differ. The difference in variances is still $O(1/n^2)$.

2.4 Conditioning on the Assignment Split W

By conditioning on the assignment split W we can break down the expressions for variance to better understand when $\hat{\tau}_{ps}$ outperforms $\hat{\tau}_{sd}$. For $\hat{\tau}_{**}$ with $** = ps$ or sd we have

$$\mathrm{Var}[\hat{\tau}_{**}|\mathcal{D}] = \mathrm{MSE}[\hat{\tau}_{**}|\mathcal{D}] = \mathbb{E}_W[\mathrm{MSE}[\hat{\tau}_{**}|W]\,|\mathcal{D}] = \sum_{w\in\mathcal{W}} \mathrm{MSE}[\hat{\tau}_{**}|W=w]\,\mathbf{P}\{W=w|\mathcal{D}\}$$

with $\mathcal{W}$ being the set of all allowed splits where $\hat{\tau}_{ps}$ is defined. The overall MSE is a weighted average of the conditional MSE, with the weights being the probability of the given possible splits W. This will give us insight into when $\mathrm{Var}[\hat{\tau}_{sd}]$ is large.

Conditioning on the split W maintains Assignment Symmetry and sets

$$\beta_{\ell k} = \frac{W_k(1-\ell)}{W_k(\ell)} \text{ for } k \in 1,\ldots,K$$

and $\beta_\ell = W(1-\ell)/W(\ell)$. For $\hat{\tau}_{ps}$ we immediately obtain

$$\mathrm{Var}[\tau_{ps}|W] = \frac{1}{n}\sum_k \frac{n_k}{n}\left(\frac{W_k(0)}{W_k(1)}\sigma_k^2(1) + \frac{W_k(1)}{W_k(0)}\sigma_k^2(0) + 2\gamma_k(1,0)\right). \quad (2.14)$$

Under conditioning $\hat{\tau}_{ps}$ is still unbiased and so the conditional MSE is the conditional variance. $\hat{\tau}_{sd}$, however, can now be *biased* with a conditional MSE larger than the conditional variance if the extra bias term is nonzero. Theorem 2.4.1 show the bias and conditional variance of $\hat{\tau}_{sd}$:

Theorem 2.4.1. *The bias of $\hat{\tau}_{sd}$ conditioned on W is*

$$\mathbb{E}[\hat{\tau}_{sd}|W] - \tau = \sum_{k \in \mathcal{B}} \left[\left(\frac{W_k(1)}{W(1)} - \frac{n_k}{n} \right) \bar{y}_k(1) - \left(\frac{W_k(0)}{W(0)} - \frac{n_k}{n} \right) \bar{y}_k(0) \right],$$

which is not 0 in general, even with a constant treatment effect. $\hat{\tau}_{sd}$'s variance conditioned on W is

$$Var[\hat{\tau}_{sd}|W] = \sum_{k \in \mathcal{B}} \frac{W_{1k} W_{0k}}{n_k} \left(\frac{1}{W_1^2} \sigma_k^2(1) + \frac{1}{W_0^2} \sigma_k^2(0) + \frac{2}{W_1 W_0} \gamma_k(1,0) \right).$$

See first part of Appendix A for a sketch of these two derivations. They come from an argument similar to the proof for the variance of $\hat{\tau}_{ps}$, but with additional weighting terms.

The conditional MSE of $\hat{\tau}_{sd}$ has no nice formula that we are aware, and is simply the sum of the variance and the squared bias:

$$MSE[\hat{\tau}_{sd}|W] = \text{Var}[\hat{\tau}_{sd}|W] + (\mathbb{E}[\hat{\tau}_{sd}|W] - \tau)^2 \tag{2.15}$$

In a typical blocked experiment, W would be fixed at W^{blk} where $W_k^{blk} = n_k p$ for $k = 1, \ldots, K$. For complete randomization, $\mathbb{E}[W] = W^{blk}$. We can now gain insight into the difference between the simple-difference and post-stratified estimators. If W equals W^{blk}, then the conditional variance formula for both estimators reduce to that of blocking, i.e., Equation 2.14 and Equation 2.15 reduce to Equation 2.11. For $\hat{\tau}_{ps}$, the overall variance for each strata is a weighted sum of $W_k(0)/W_k(1)$ and $W_k(1)/W_k(0)$. The more unbalanced these terms, the larger the sum. Therefore the more W deviates from W^{blk}—i.e., the more *imbalanced* the assignment is—the larger the post-stratified variance formula will tend to be. The simple-difference estimator, on the other hand, tends to have smaller variance as W deviates further from W^{blk} due to the greater restrictions on the potential random assignments.

$\hat{\tau}_{ps}$ has no bias under conditioning, but $\hat{\tau}_{sd}$ does if b is prognostic, and this bias can radically inflate the MSE. This bias increases with greater imbalance. Overall, then, as imbalance increases, the variance (and MSE) of $\hat{\tau}_{ps}$ moderately increases. On the other hand, for $\hat{\tau}_{sd}$ the variance can moderately decrease but the bias sharply increases, giving an overall MSE that can grow quite large.

Because the overall MSE of these estimators is a weighted average of the conditional MSEs, and because under perfect balance the conditional MSEs are the same, we know any differences in the unconditional variance (i.e., MSE) between $\hat{\tau}_{sd}$ and $\hat{\tau}_{ps}$ comes from what happens when there is bad imbalance. $\hat{\tau}_{sd}$ has a much higher variance than $\hat{\tau}_{ps}$ when there is potential for large bias. Its variance is smaller when there is not. With post-stratification, we pay for unbiasedness with a bit of extra variance—we are making a different bias-variance tradeoff than with simple-difference.

The split W is directly observable and gives hints to the experimenter as to the success, or failure, of the randomization. Unbalanced splits tell us we have less certainty while balanced

splits are comforting. For example, take a hypothetical balanced completely randomized experiment with $n = 32$ subjects, half men and half women. Consider the case where only one man ends up in treatment as compared to 8 men. In the former case, a single man gives the entire estimate for average treatment outcome for men and a single woman gives the entire estimate for average control outcome for women. This seems *very* unreliable. In the latter case, each of the four mean outcomes are estimated with 8 subjects, which seems more reliable. Our estimates of uncertainty should take this observed split W into account, and we can take it into account by using the conditional MSE rather than overall MSE when estimating uncertainty. The conditional MSE estimates how close one's actual experimental estimate is likely to be from the SATE. The overall MSE estimates how close such estimates will generally be to the SATE over many trials.

This idea of using all observed information is not new. When sampling to find the mean of a population, [38] argue that, for estimators adjusted using post-stratification, variance estimates should be conditioned on the distribution of units in the strata as this gives a more relevant estimate of uncertainty. [86] sharpens this argument by presenting it as one of prediction. Under this view, it becomes more clear what should be conditioned on and what not. In particular, if an estimator is conditionally unbiased when conditioned on an ancillary statistic, then conditioning on the ancillary statistic increases precision. This is precisely the case when conditioning the above estimators on the observed split, assuming Assignment Symmetry. Similarly, in the case of sampling, [72] compare variance estimators for the sample totals that incorporate the mean of measured covariates as compared to the population to get what they argue are more appropriate estimates. [63] extends [76] and examines conditioning on the imbalance of a continuous covariate in ANCOVA. They show that not correcting for imbalance (as measured as a standardized difference in means) gives one inconsistent control on the error rate when testing for an overall treatment effect.

2.5 Extension to an Infinite-Population Model

The presented results apply to estimating the treatment effect for a specific sample of units, but there is often a larger population of interest. One approach is to consider the sample to be a random draw from this larger population, which introduces an additional component of randomness capturing how the SATE varies about the Population Average Treatment Effect, or PATE. See [44]. But if the sample has not been so drawn, using this PATE model might not be appropriate. The SATE perspective should instead be used, with additional work to then generalize the results. See [32] or [43]. Regardless, under the PATE approach, the variances of all the estimators increase, but the substance of this chapter's findings remain.

Let f_k, $k = 1, \ldots, K$, be the proportion of the population in stratum k. The PATE can then be broken down by strata:

$$\tau^* = \sum_{k=1}^{K} f_k \tau_k^*$$

with τ_k^* being the population average treatment effect in stratum k. Let the sample $\mathcal{S}$ be a stratified draw from this population holding the proportion of units in the sample to f_k (i.e. $n_k/n = f_k$ for $k = 1, \ldots, K$). (See below for different types of draws from the population.) τ, the SATE, is random, depending on $\mathcal{S}$. Due to the size of the population, the sampling is close to being with replacement. Alternatively, the sample could be generated by independent draws from a collection of K distributions, one for each stratum. Let $\sigma_k^2(\ell)^*, \gamma_k^2(1,0)^*$, etc., be population parameters. Then the PATE-level MSE of $\hat{\tau}_{ps}$ is

$$\text{Var}[\hat{\tau}_{ps}] = \frac{1}{n} \sum_k f_k \left[(\beta_{1k} + 1)\sigma_k^2(1)^* + (\beta_{0k} + 1)\sigma_k^2(0)^* \right]. \tag{2.16}$$

See Appendix A for the derivation. [44] has a similar formula for the two-strata case. Compare to Equation 2.6: All the correlation of potential outcomes terms $\gamma_k(1,0)$ vanish when moving to PATE. This is due to a perfect trade-off: the more they are correlated, the harder to estimate the SATE τ for the sample, but the easier it is to draw a sample with a SATE τ close to the overall PATE τ^*.

The simple-difference estimator. For the simple-difference estimator, use Equation 2.16 with $K = 1$ to get

$$\text{Var}[\hat{\tau}_{sd}] = \frac{1}{n} \left[(\beta_1 + 1)\sigma^2(1)^* + (\beta_0 + 1)\sigma^2(0)^* \right]. \tag{2.17}$$

Now let $\bar{\sigma}^2(\ell)^*$ be a weighted sum of the squared differences of the strata means to the overall mean:

$$\bar{\sigma}^2(\ell)^* = \sum_{k=1}^{K} f_k \left(\bar{y}_k^*(\ell) - \bar{y}^*(\ell) \right)^2.$$

The population variances then decompose into $\bar{\sigma}^2(\ell)^*$ and strata-level terms:

$$\sigma^2(\ell)^* = \bar{\sigma}^2(\ell)^* + \sum_{k=1}^{K} f_k \sigma_k^2(\ell)^*.$$

Plug this decomposition into Equation 2.17 to get

$$\text{Var}[\hat{\tau}_{sd}] = \frac{1}{n} \left[(\beta_1 + 1) \left(\bar{\sigma}^2(1)^* + \sum_{k=1}^{K} f_k \sigma_k^2(1)^* \right) + (\beta_0 + 1) \left(\bar{\sigma}^2(0)^* + \sum_{k=1}^{K} f_k \sigma_k^2(0)^* \right) \right]$$

Variance gain from post-stratification. For comparing the simple-difference to the post-stratified estimator at the PATE level, take the difference of Equation 2.17 and Equation 2.16 to get

$$\begin{aligned} \text{Var}[\hat{\tau}_{sd}] - \text{Var}[\hat{\tau}_{ps}] = & \frac{1}{n}(\beta_1 + 1)\bar{\sigma}^2(1)^* + \frac{1}{n}(\beta_0 + 1)\bar{\sigma}^2(0)^* \\ & - \frac{1}{n} \sum_{k=1}^{K} f_k \left[(\beta_{1k} - \beta_1)\sigma_k^2(1)^* + (\beta_{0k} - \beta_0)\sigma_k^2(0)^* \right]. \end{aligned}$$

Similar to the SATE view, we again have a gain component (the first line) and a cost (the second line). For Binomial assignment and complete randomization, $\beta_\ell \leq \beta_{\ell k}$ for all k, making the cost nonnegative. There are no longer terms for the correlation of potential outcomes, and therefore this gain formula is directly estimable. The cost is generally smaller than for the SATE model due to the missing $\gamma_k(1,0)$ terms.

The variance of blocking under PATE. For equal-proportion blocking, $W_k(1) = pn_k$ and $W_k(0) = (1-p)n_k$. Using this and $\beta_{\ell k} + 1 = \mathbb{E}[n_k/W_k(\ell)]$, the PATE-level MSE for a blocked experiment is then

$$\mathrm{Var}[\hat{\tau}_{ps}] = \frac{1}{n}\sum_k \frac{n_k}{n}\left[\frac{1}{p}\sigma_k^2(1)^* + \frac{1}{1-p}\sigma_k^2(0)^*\right]$$

For comparing complete randomization (with pn units assigned to treatment) to blocked experiments, plug in the β's. The $\beta_\ell - \beta_{\ell k}$ terms all cancel, leaving

$$\mathrm{Var}[\hat{\tau}_{sd}] - \mathrm{Var}[\hat{\tau}_{ps}] = \frac{1}{n}\frac{1}{p}\bar{\sigma}^2(1)^* + \frac{1}{n}\frac{1}{1-p}\bar{\sigma}^2(0)^* \geq 0$$

Unlike from the SATE perspective, blocking can never hurt from the PATE perspective.

Not conditioning on the n_k. Allowing the n_k to vary introduces some complexity, but the gain formula remain unchanged. If the population proportions are known, but the sample is a completely random draw from the population, the natural post-stratified estimate of the PATE would use the population weights f_k. These weights can be carried through and no problems result. Another approach is to estimate the f_k with n_k/n in the sample. In this latter case, we first condition on the seen vector $N \equiv n_1, \ldots n_k$ and define a τ^N based on N. Conditioned on N, both $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ are unbiased for estimating τ^N, and we can use the above formula with n_k/n instead of f_k. Now use the tower-property of expectations and variances. This results in an extra variance of a multinomial to capture how τ^N varies about τ as N varies. The variances of both the estimators will each be inflated by this extra term, which therefore cancels when looking at the difference.

2.6 Comparisons with Other Methods

Post-stratification is a simple adjustment method that exploits a baseline categorical covariate to ideally reduce the variance of a SATE estimate. Other methods allow for continuous or multiple covariates and are more general. The method that is appropriate for a given application depends on the exact assumptions one is willing to make.

Recently, [25, 26] studied the most common form of adjustment—linear regression—under the Neyman-Rubin model. Under this model, Freedman, for an experimental setting, showed that traditional OLS (in particular ANCOVA) is biased (although it is asymptotically

unbiased), that the asymptotic variance can be larger than with no adjustment, and worse, that the standard estimate of this variance can be quite off, even asymptotically. Freedman's results differ from those in traditional textbooks because, in part, he uses the Neyman-Rubin model with its focus on SATE. Subsequently, [50] expanded these results and showed that OLS with all interactions cannot be asymptotically less efficient than using no adjustment, and further, that Huber-White sandwich estimators of the standard error are asymptotically appropriate. These papers focus primarily on continuous covariates rather than categorical, but their results are general. Our post-stratified estimator is identical to a fully saturated ordinary linear regression with the strata as dummy variables and all strata by treatment interactions—i.e., a two-way ANOVA analysis with interactions. Therefore, our results apply to this regression estimator, and, in turn, all of Lin's asymptotic results apply to our $\hat{\tau}_{ps}$.

[87] propose a semi-parametric method where the researcher independently models the response curve for the treatment group and the control group and then adjusts the estimated average treatment effect with a function of these two curves. This approach is particularly appealing in that concerns about data mining and pre-test problems are not an issue—i.e., researchers can search over a wide class of models looking for the best fit for each arm (assuming they don't look at the consequent estimated treatment effects). With an analysis assuming only the randomization and the infinite super population model, Tsiatis et. al show that asymptotically such estimators are efficient. This semi-parametric approach can accommodate covariates of multiple types: because the focus is modeling the two response curves, there is basically no limit to what information can be incorporated.

A method that does not have the super population assumption is the inference method for testing for treatment effect proposed by Koch and coauthors [e.g., 46, 47]. Koch observed that under the Fisherian sharp null of no treatment effect, one can directly compute the covariance matrix of the treatment indicator and any covariates. Therefore, using the fact that under randomization the expected difference of the covariates should be 0, one can estimate how far the observed mean difference in outcomes is from expected using a χ^2 approximation. (One could also use a permutation approach to get an exact P-value.) However, rejecting Fisher's sharp null, distinct from the null of no difference in average treatment effect, does not necessarily demonstrate an overall average impact. Nonetheless, this approach is very promising. Koch et. al also show that with an additional super population assumption one can use these methods to generate confidence intervals for average treatment effect.

[52] compare post-stratification to blocking using an additive linear population model and a sampling framework, implicitly using potential outcomes for some results. They consider linear contrasts of multiple treatments as the outcome of interest, which is more general than this chapter, but also impose assumptions on the population such as constant variance and, implicitly, a constant treatment effect. Using asymptotics, they isolate the main terms of the estimators' variance and drop lower order ones.

Relative to post-stratification, there are three concerns with these other adjustment methods. First, many of these methods make the assumption of the infinite population sampling model discussed in Section 2.5 (which is equivalent to any model that has independent, random errors, e.g., regression). The consequences of violating this assumption can be unclear.

Therefore, one may prefer estimating sample treatment effects, and then generalizing beyond the given experimental sample using methods such as those of [32]. Second, methods within the SATE framework that depend on a Fisherian sharp null for testing for a treatment effect have certain limitations. In some circumstances, this null may be considered restrictive and generating confidence intervals can be tricky without assuming a strong treatment effect model such as additivity. Third, asymptotic analyses may not apply when analyzing small- or mid-sized experiments, and experiments with such samples sizes is where the need for adjustment is the greatest.

Notwithstanding these concerns, if one is in a context where these concerns do not hold, or one has done work showing that the impact of them is minor, these alternative methods of adjustment depend on relatively weak assumptions and also allow for continuous covariates and multiple covariates—a distinct advantage over post-stratification. These other methods, due to their additional modeling assumptions, may be more efficient as well. Different estimators may be more or less appropriate depending on the assumptions one is willing to make and the covariates one has.

Post-stratification is close in conceptual spirit to blocking. This chapter shows that this conceptual relationship bears out. Blocking, however, is a stronger approach because it requires the choice of which covariates to adjust for to be determined prior to randomization. Blocking has the profound benefit that it forces the analyst to decide how covariates are incorporated to improve efficiency before any outcomes are observed. Therefore, blocking eliminates the possibility of searching over post-adjustment models until one is happy with the results. The importance of this feature of blocking is difficult to overstate. Blocking is, however, not always possible. In medical trials when patients are entered serially, for example, randomization has to be done independently. Natural experiments, where randomization is due to processes outside the researchers' control, are another example particularly of interest in the social sciences. In these cases, post-stratification can give much the same advantages with much the same simplicity. But again, as "Student" (W. S. Gosset) observed, "there is great disadvantage in correcting any figures for position [of plots in agricultural experiments], inasmuch as it savors of cooking, and besides the corrected figures do not represent anything real. It is better to arrange in the first place so that no correction is needed [85]."

2.7 PAC Data Illustration

We apply our methods to evaluating Pulmonary Artery Catheterization (PAC), an invasive and controversial cardiac monitoring device, that was, until recently, widely used in the management of critically ill patients [14, 21]. Controversy arose regarding the use of PAC when a non-random study using propensity score matching found that PAC insertion for critically ill patients was associated with increased costs and mortality [12]. Other observational studies came to similar conclusions leading to reduced PAC use [11]. However, an RCT (PAC-Man) found no difference in mortality between PAC and no-PAC groups [33], which substantiated

the concern that the observational results were subject to selection bias [69].

PAC-Man has 1013 subjects, half treated. The outcome variable investigated here is "qalys" or quality-adjusted life years. Higher values indicate, generally, longer life and higher quality of life. Death at time of possible PAC insertion or shortly after receives a value of 0. Living two years in full health would be a 2. There is a lot of fluctuation in these data. There is a large point mass at 0 (33% of the patients) and a long tail.

Unfortunately, the RCT itself had observed covariate imbalance in predicted probability of death, a powerful predictor of the outcome, which calls into question the reliability of the simple-difference estimate of the treatment effect. More low-risk patients were assigned to receive treatment, which could induce a perceived treatment effect even if none were present. Post-stratification could help with this potential bias and decrease the variance of the estimate of treatment effect. To estimate the treatment effect using post-stratification we first divide the continuous probability of death covariate into K K-tiles. We then estimate the treatment effect within the resulting strata and average appropriately.

This analysis is simplified for the purposes of illustration. We are only looking at one of the outcomes and have dropped several potentially important covariates for the sake of clarity. Statistics on the strata for $K = 4$ are listed on Table 2.1. A higher proportion of subjects in the first two groups were treated than one would expect given the randomization. Imbalance in the first group, with its high average outcome, could heavily influence the overall treatment effect estimate of $\hat{\tau}_{sd}$.

Strata	# Tx	# Co	$SD_k(1)$	$SD_k(0)$	$\hat{y}_k(1)$	$\hat{y}_k(0)$	$\hat{\tau}_k$
Low Risk	136	118	5.80	5.68	5.57	5.41	0.15
Moderate Risk	142	111	3.42	4.17	1.69	2.70	-1.01
High Risk	106	147	3.60	3.75	1.97	2.36	-0.39
Extreme Risk	122	131	3.41	3.10	1.37	1.19	0.18
Overall	506	507	4.56	4.48	2.72	2.84	-0.13

Table 2.1: Strata-Level Statistics for the PAC Illustration

We estimate the minimum gain in precision due to post-stratification by calculating point estimates of all the within- and between-strata variances and the between-strata covariance and plugging these values into Equation 2.9. We are not taking the variability of these estimates into account. By assuming the strata r_k are maximal, i.e., $r_k = 1$ for all k, we estimate a lower bound on the reduction in variance due to post-stratification. The β's are estimated by numerical simulation of the randomization process (with 50,000 trials) and are therefore exact up to uncertainty in this Monte Carlo experiment; these values do not depend on the population characteristics and so there is no sampling variability here. We show the resulting estimates for several different stratifications. For $K = 4$, we estimate the percent reduction of variance, $100\% \times (\mathrm{Var}[\hat{\tau}_{ps}] - \mathrm{Var}[\hat{\tau}_{sd}])/\mathrm{Var}[\hat{\tau}_{sd}]$, to be no less than 12%.

If the true r_k were less than 1, the benefit would be greater. More strata appear somewhat superior, but gains level off rather quickly. See Table 2.2.

The estimate of treatment effect changes markedly under post-stratification. The estimates $\hat{\tau}_{ps}$ hover around -0.28 for $K = 4$ and higher, as compared to the -0.13 from the simple-difference estimator. The post-stratified estimator appears to be correcting the bias from random imbalance in treatment assignment.

We can also estimate the MSE for both the simple-difference and post-stratified estimator conditioned on the imbalance by plugging point estimates for the population parameters into Equation 2.15 and Equation 2.14. We again assume the correlations r_k are maximal. We estimate bias by plugging in the estimated $\hat{y}_k(\ell)$ for mean potential outcomes of the strata. These results are the last columns of Table 2.2; the percentage gain in this case is higher primarily due to the correction of the bias term from the imbalance. When conditioning on the imbalance W, the estimated MSE (i.e., variance) of the post-stratified estimator is slightly higher than the *variance* of the simple-difference estimator, but is substantially lower than its overall MSE. This is due to the bias correction. Because the true variances and the r_k for strata are unknown, these gains are estimates only. They do, however, illustrate the potential value of post-stratification. Measuring the uncertainty of these estimates is an area of future work.

			Uncond. Variance			MSE Conditioned on W				
K	$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	%	MSE$\hat{\tau}_{ps}$	var$\hat{\tau}_{sd}$	bias$\hat{\tau}_{sd}$	MSE$\hat{\tau}_{sd}$	%
2	-0.34	-0.13	0.077	0.081	5%	0.077	0.076	0.207	0.118	35%
4	-0.27	-0.13	0.071	0.081	12%	0.072	0.070	0.137	0.089	19%
10	-0.25	-0.13	0.070	0.081	14%	0.071	0.069	0.119	0.083	15%
15	-0.24	-0.13	0.070	0.081	14%	0.070	0.067	0.115	0.081	13%
30	-0.28	-0.13	0.069	0.081	15%	0.068	0.064	0.148	0.086	21%
50	-0.32	-0.13	0.068	0.081	15%	0.066	0.061	0.190	0.097	32%

Table 2.2: Estimated Standard Errors for PAC. Table shows both conditioned and unconditioned estimates for different numbers of strata.

Matched Pairs Estimation. We can also estimate the gains by building a fake set of potential outcomes by matching treated units to control units on observed covariates. We match as described in [74]. We then consider each matched pair a single unit with two potential outcomes. We use this synthetic set to calculate the variances of the estimators using the formula from Section 2.2.

Matching treatment to controls and controls to treatment gives 1013 observations with all potential outcomes "known." The correlation of potential outcomes is 0.21 across all strata. $\tau = -0.031$. The unconditional variance for the simple-difference and post-stratified estimators are 0.048 and 0.038, respectively. The percent reduction in variance due to post-stratification is 19.6%.

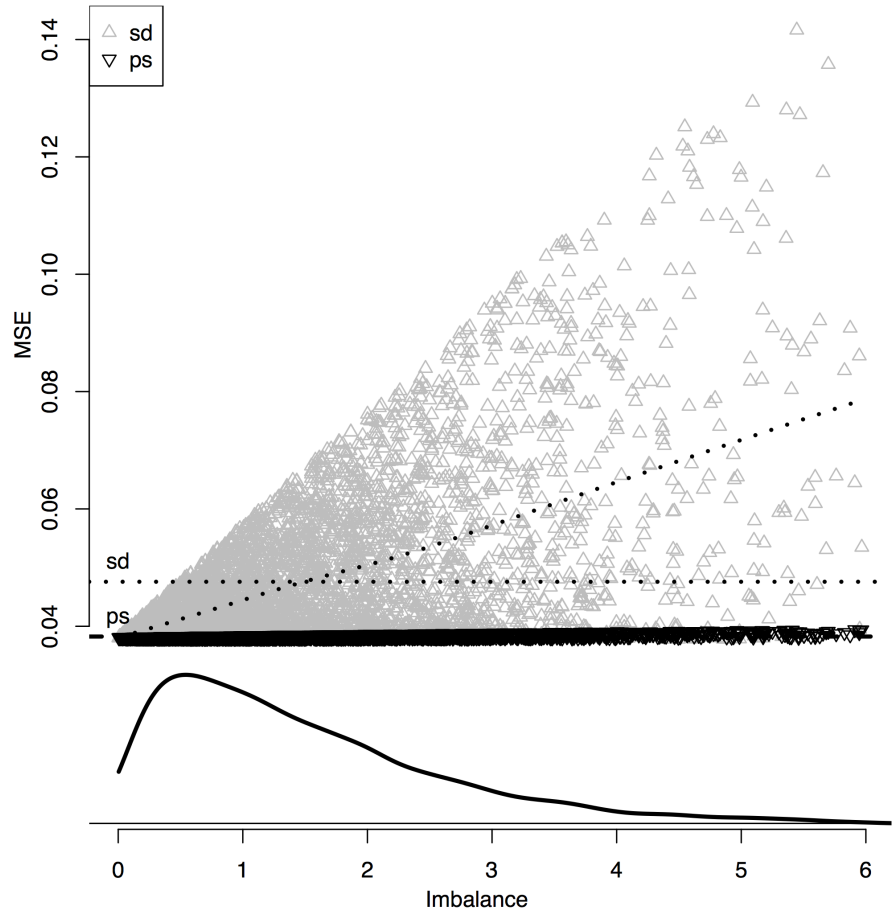

Figure 2.1: PAC MSE Conditioned on Imbalance. Uses constructed matched PAC dataset. Points indicate the conditional MSE of $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ given various specific splits of W. x-axis is the imbalance score for the split. Curved dashed lines interpolate point clouds. Horizontal dashed lines mark unconditional variances for the two estimators. The curve at bottom is the density of the imbalance statistic.

We can use this data set to further explore the impact of conditioning. Assume the treatment probability is $p = 0.5$ and repeatedly randomly assign a treatment vector and compute the resulting conditional variance. Also compute the "imbalance score" for the treatment vector with a chi-squared statistic:

$$\text{Imbalance} \equiv \sum_k \frac{(W_k(1) - pn_k)^2}{pn_k}$$

This procedure produces Figure 2.1. As imbalance increases, the MSE (variance) of $\hat{\tau}_{ps}$ steadily, but slowly, increases as well. The MSE of $\hat{\tau}_{ps}$ is quite resistant to large imbalance.

This is not the case for $\hat{\tau}_{sd}$, however. Generally, high imbalance means high conditional MSE. This is due to the bias term which can get exceedingly large if there is imbalance between different heterogeneous strata. Also, for a given imbalance, the simple-difference estimator can vary widely depending on whether stratum-level bias terms are canceling out or not. This variability is not apparent for the post-stratified estimator, where only the number of units treated drives the variance; the post-stratified points cluster closely to their trend line.

The curve at the bottom shows the density of the realized imbalance score: there is a good chance of a fairly even split with low imbalance. In these cases, the variance of $\hat{\tau}_{sd}$ is smaller than the unconditional formula would suggest. If the randomization turns out to be "typical" the unconditional variance formula would be conservative. If the imbalance is large, however, the unconditional variance may be overly optimistic. This chance of large imbalance with large bias is why the unconditioned MSE of $\hat{\tau}_{sd}$ is larger than that of $\hat{\tau}_{ps}$.

The observed imbalance for the actual assignment was about 2.37. The conditional MSE is 0.083 for $\hat{\tau}_{sd}$ and 0.039 for $\hat{\tau}_{ps}$, a 53% reduction in variance. The conditional MSE for the simple-difference estimator is 75% larger than its unconditional MSE due to the bias induced by the imbalance. We would be overly optimistic if we were to use $\text{Var}[\hat{\tau}_{sd}]$ as a measure of certainty, given the observed, quite imbalanced, split W. For the post-stratified estimator, however, the conditional variance is only about 1% higher than the unconditional; the degree of imbalance is not meaningfully impacting the precision. Generally, with post-stratification, the choice of using an unconditional or conditional formula is less of a concern.

Discussion. The PAC RCT has a strong predictor of outcome. Using it to post-stratify substantially increases the precision of the treatment effect estimate. Furthermore, post-stratification mitigates the bias induced by an unlucky randomization. When concerned about imbalance, it is important to calculate conditional standard errors—not doing so could give overly optimistic estimates of precision. This is especially true when using the simple-difference estimator. The matched-pairs investigation shows this starkly; $\hat{\tau}_{sd}$'s conditional MSE is 75% larger than the unconditional.

2.8 Discussion

Post-stratification is a viable approach to experimental design in circumstances where blocking is not feasible. If the stratification variable is determined beforehand, post-stratification is nearly as efficient as a randomized block trial would have been: the difference in variances between post-stratification and blocking is a small $O(1/n^2)$. However, the more strata, the larger the potential penalty for post-stratification. There is no guarantee of gains.

Conditioning on the observed distribution of treatment across strata allows for a more appropriate assessment of precision. Most often the observed balance will be good, even in moderate-sized experiments, and the conditional variance of both the post-stratified and simple-difference estimator will be smaller than estimated by the unconditional formula. However, when balance is poor, the conditional variance of the estimators, especially for the

simple-difference estimator, may be far larger than what the unconditional formula would suggest. Furthermore, in the unbalanced case, if a truly prognostic covariate is available post-stratification can significantly improve the precision of one's estimate. For a covariate unrelated to outcome, however, a simple-difference estimator can be superior.

When viewing a post-stratified or a blocked estimate as an estimate of the PATE, under the assumption that the sample is a random, independent, draw from a larger population, the potential for decreased precision is reduced. However, in most cases the sample in a randomized trial is not such a random draw. We therefore advocate for viewing the estimators as estimating the SATE, not the PATE.

Problems arise when stratification is determined after treatment assignment. The results of this chapter assume that the stratification is based on a fixed and defined covariate b. However, in practice covariate selection is often done after-the-fact in part because, as is pointed out by [63], it is often quite difficult to know which of a set of covariates are significantly prognostic *a priori*. But variable selection invites fishing expeditions, which undermine the credibility of any findings. Doing variable selection in a principled manner is still notoriously difficult, and is often poorly implemented; [63], for example, found that many clinical trial analyses select variables inappropriately. [87] summarize the controversy in the literature and, in an attempt to move away from strong modeling, and to allow for free model selection, propose a semiparametic approach as a solution.

[3] suggests that, at minimum, all potential covariates for an experiment be listed in the original protocol. Call these z. In our framework, variable-selection is then to *build* a stratification b from z *and* T after having randomized units into treatment and control. b (now B) is random as it depends on T. Questions immediately arise: how does one define the variance of the estimator? Can substantial bias be introduced by the strata-building process? The key to these questions likely depends on appropriately conditioning on both the final, observed, strata and the process of constructing B. This is an important area of future work.

Chapter 3

Validating Text-Summarization Methods

3.1 Introduction

Joseph Pulitzer wrote in his last will and testament, "[Journalism] is a noble profession and one of unequaled importance for its influence upon the minds and morals of the people." Faced with an overwhelming amount of world news, concerned citizens, media analysts and decision-makers alike would greatly benefit from scalable, efficient methods that extract compact summaries of how subjects of interest are covered in text corpora. These compact summaries could be used by the interested people for screening corpora of news articles before detailed reading or further investigation.

We propose a novel approach to perform automatic, subject-specific summarization of news articles that applies to general text corpora. Our approach allows researchers to easily and quickly explore large volumes of news articles. These methods readily generalize to other types of documents. For example, [19] identified potentially dangerous aspects of specific airports by using these methods on pilots' flight logs. We are currently implementing our approach within an on-line toolkit, SnapDragon,[1] which will soon be available to researchers interested in screening corpora of documents relative to a subject matter.

News media significantly drives the course of world events. By choosing which events to report and the manner in which to report them, the media affects the sentiments of readers and through them the wider world. Exposure to news can drive change in society [53, 61, 39], and, even when controlling for topic, can vary in tone, emphasis, and style [8, 29, 15]. Our focus on news media is motivated by a crucial need in a democracy: to understand precisely where and how these media variations occur.

Currently, media analysis is often conducted by hand coding [88, 16, 64]. Hand-coding manually reduces complex text-data to a handful of quantitative variables, allowing for statistical analysis such as regression or simple tests of difference. It is prohibitively labor-

[1] http://statnews2.eecs.berkeley.edu/snapdragon

intensive [39]. In personal correspondence, Denham relayed that each article took roughly fifteen minutes to analyze, suggesting about 28 hours of time for their full text corpus of 115 articles.

In the last five years we have seen the emergence of a computational social science field connecting statistics and machine learning to anthropology, sociology, public policy, and more [48]. Many organizations have introduced automated summary methods: Google news trends, Twitter's trending queries, Crimson Hexagon's brand analysis and others all use computation to make sense of the vast volumes of text now publicly generated. These approaches, discussed below, illustrate the potential of computation for news media analysis.

Summarization by extraction. There are two major approaches to text analysis, key-phrase extraction (listing key-phrases for the document such as in [67, 75, 24, 10]) and sentence extraction (identifying and presenting the "most relevant" sentences of a document as a summary, such as in [35, 30, 59]). Both these approaches score potential key-phrases or sentences found in the text and then select the highest scorers as the summary. This line of research has primarily focused on summarizing individual documents, with one summary for every document in a corpus.

However, when there are multiple documents, even a short summary of each document adds up quickly. Content can be buried in a sea of summaries if most documents are not directly related to the subject of interest. If many documents are similar, the collection of summaries becomes redundant. Moreover, if the subject of interest is usually mentioned in a secondary capacity, it might be missing entirely from the summaries. To address some of these problems, [30] worked on summarizing multiple documents at once to remove redundancy. Under their system, sentences are scored and selected sequentially, with future sentences penalized by similarity to previously selected sentences. In this system, the documents need to be first clustered by overall topic.

[35] fits a latent topic model (similar to LDA, discussed below) for subject-specific summarization of documents. Here the subject is represented as a set of documents and a short narrative of the desired content. All units of text are projected into a latent topic space that is learned from the data independent of the subject and then sentences are extracted by a scoring procedure by comparing the similarity of the latent representations of the sentences to the subject. Although we also summarize an entire collection of documents as they pertain to a specific subject of interest, we do not use a latent space representation of the data. In [57], the authors merge all text into two super-documents and then score individual words based on their differing rates of appearance, normalized by their overall frequency. We analyze the corpus through individual document units.

Summarization via topic modeling. Some analysis algorithms take text information as input and produce a model, usually generative, fit to the data. The model itself captures structure in the data, and this structure can be viewed as a summary. A popular example is the latent Dirichlet allocation [7], which posits that each word observed in the text is

standing in for a hidden, latent "topic" variable. These models are complex and dense, with all the words playing a role in all the topics, but one can still take the most prominent words in a topic as the summary.

[9] had humans evaluate the internal cohesion of learned topics. Respondents were asked to identify "impostor" words inserted into lists of words representing a topic. This showed these approaches as producing cogent and reasonable topics. Supervised versions [6] of these methods can be used to summarize a subject of interest.

Although these methods are computationally expensive and produce dense models requiring truncation for interpretability, they are powerful indications of the capabilities of computer-assisted summarization. These methods analyze the corpus as a whole and model how the documents cover a modest number of organically grown topics. We opt instead for a more directed process of summarizing a particular, specified subject (out of possible millions).

Other automated approaches. Google Trend charts are calculated by comparing the number of times a subject appears in the news outlets that Google compiles to the overall volume of news for a specified time period. Even this simple approach can show how subjects enter and leave public discourse across time. Twitter's trending topics appears to operate similarly, although it selects the hottest topics by those which are gaining in frequency most quickly. Although neither of these tools summarize a specified subject, they are similar in spirit to the normalized simpler methods (co-occur and correlation screen) that we will introduce and investigate in this chapter.

[39] extrapolates from a potentially non-random sample of hand-coded documents to estimate the proportion of documents in several pre-defined categories. This can be used for sentiment analysis (e.g., estimating the proportion of blogs showing approval for some specified public figure). We instead identify key-phrases most associated with a given subject. These key-phrases could then be analyzed directly for sentiment, thus reducing the amount of hand-coding required. Their work is behind Crimson Hexagon, a company currently offering brand analysis to several companies.

In a similar spirit, we believe there is opportunity to answer the question, "What is being said in the news regarding China?" or, more generally, "What is discussed in this corpus of documents regarding subject A?" using machine learning techniques.

Our predictive and sparse approach with human evaluation. In this chapter, we propose to use statistical machine learning tools such as Lasso that are fast, sparse, and different from those described earlier to produce short and interpretable summaries. Our proposal is desirable because media analysis (or general document summarization) tools need to encourage exploration, allowing researchers to easily examine how different topics are portrayed in a corpus or how this portrayal evolves over time or compares across different corpora.

Given a corpus of documents (e.g., the New York Times International Section articles in 2009) as well as a subject of interest (e.g.,subject China as represented as a short list of words *China, Chinas and Chinese*), we establish the predictive framework by first *automatically* labeling documents into positive and negative examples by, for example, determining if they contain words from the subject list. Counts of words/phrases not on the subject list then form the predictor vectors for the documents. After normalizing these vectors, we use scalable, reproducible prediction and classification techniques to identify a small set of words and phrases that best predict the subject as it appears. This overall arc reduces corpora of many millions of words into a few representative key-phrases that constitute how the given *subject* is *treated* in the corpus.

To *validate* these summaries we cannot use traditional machine learning approaches, however, since traditional measures have no guarantee of correlating with actual meaning. We therefore compare the different summary approaches with a human survey. We find that sparse methods such as Lasso indeed produce higher quality summaries than many currently used, simpler, methods. Moreover, we conducted usability testing to investigate how different preprocessing techniques differ in quality of resulting lists in the user survey. We found the choice of preprocessing important, especially with simpler summarization methods. The sparse methods such as Lasso, however, are more robust to potential mistakes made in the data preparation step.

To illustrate, consider how the New York Times treated China (represented as *"china, chinas, chinese"*) in the international section in 2009. One of our summary methods, L1LR, yields a short list of terms: *"beijing, contributed research, global, hu jintao, imports, of xinjiang, peoples liberation army, shanghai, sichuan province, staterun, tibet, trade, uighurs, wen jiabao, xinhua news agency"*. This succinct summary captures main relevant personalities (e.g., Wen Jiabao, Hu Jintao), associated countries and areas (e.g., Uighurs, Tibet), entities (Xinhua news), and topics (trade, imports, global [activity], state-run [organizations]). The presence of "contributed research" indicates that other people, in addition to the authors, contributed to many of the Times articles on China. These terms inform interested readers including China experts about how China is being treated by the New York Times and suggest directions for further reading on topics such as Sichuan, Xinjiang, Tibet, and trade. Table 3.2 contains four other sample summaries.

The rest of this chapter is organized as follows. Section 3.2 describes our proposal including the predictive framework and the preprocessing choices used in the study. Section 3.2 proposes different key-phrase selectors (e.g.,Lasso, co-occurrence) one might use to generate the final summary. We then describe the validation experiment designed to examine the performance of the summarizers built from the choices in Sections 3.3. Results of this experiment are shown in Section 3.4. Section 3.5 concludes.

3.2 Our Approach: Predictive, Fast, and Sparse

Our approach is based on a predictive binary classification framework. In a typical binary classification scenario, data units (e.g.,news articles or paragraphs) belong to two classes and features of a data unit are used to predict its class membership. Classification of text documents using the phrases in those documents as features is familiar and well-studied [28, 91].

We turn subject-specific (e.g., China) summarization into a binary classification problem by forming two classes, that is, we automatically label news articles (or other document units) in a corpus as subject-related and irrelevant. See Section 3.2, where we discuss several different ways of labeling. We then use a predictive classifier to generate a summary list. To be precise, we take those words and phrases most important for classification as a summary of the subject relative to the corpus.

A predictive framework consists of n units, each with a class label $y_i \in \{-1, +1\}$ and a collection of p possible features that can be used to predict this class label. Each unit $i \in \mathcal{I} \equiv \{1, \ldots, n\}$ is attributed a value x_{ij} for each feature $j \in \mathcal{J} \equiv \{1, \ldots, p\}$. These x_{ij} form a $n \times p$ matrix X. The n units are blocks of text taken from the corpus (e.g., entire articles or individual paragraphs), the class labels y_i indicate whether document unit i contains content on a subject of interest, and the features are all the possible key-phrases that could be used to summarize the subject. As mentioned earlier, we consider several ways of automatically labeling or assigning class memberships based on the document unit itself in Section 3.2. Matrix X and vector y can be built in several ways. We build X by reweighting the elements of a document-term matrix C:

Definition 3.2.1. A document-term matrix C sets

$$C_{ij} := \text{ The number of times key-phrase } j \text{ appears in document } i$$

This is often called the *bag-of-phrases model*: each document is represented as a vector with the jth element being the total number of times that the specific phrase j appears in the document. Stack these row vectors to make the matrix $C \in \mathbb{R}^{n \times p}$ of counts. C has one row for each document and one column for each phrase. C tends to be highly sparse: most entries are 0.

To transform raw text into this vector space, convert it to a collection of individual text document units, establish a dictionary of possible phrases, and count how often each of the dictionary's phrases appear in each of the document units. Once this is completed, the summarizing process consists of 3 major steps:

1. Reweight: build X from C;
2. Label: build y by identifying which document units in the corpus likely treat the specified subject;
3. Select: extract a list of phrases that ideally summarize the subject.

How the document units are *labelled*, how the document units are *vectorized*, and how phrases are *selected* can all be done in different ways. Different choices for these steps result in different summarizers, some better than others. We describe these steps and choices fully in Sections 3.2 and 3.2.

Iraq	Russia	Germany	Mexico
american	a medvedev	angela merkel	and border protection
and afghanistan	caucasus	berlin	antonio betancourt
baghdad	europe	chancellor angela	cancn
brigade	gas	european	chihuahua
combat	georgia	france and	denise grady
gen	interfax news agency	frankfurt	drug cartels
in afghanistan	iran	group of mostly	guadalajara
invasion	moscow	hamburg	influenza
nuri	nuclear	marwa alsherbini	oaxaca
pentagon	president dmitri	matchfixing	outbreak
saddam	republics	minister karltheodor zu	president felipe
sergeant	sergei	munich	sinaloa
sunni	soviet	nazi	swine
troops	vladimir	world war	texas
war and who			tijuana

Table 3.1: Four Sample Summaries of Four Different Countries. The method used, a count rule with a threshold of 2, the Lasso for feature selection, and tf-idf reweighting of features, was one of the best identified for article-unit analysis by our validation experiment.

Data pre-processing

In this section, we describe in detail how we pre-process a corpus of documents into a vectorized space so that the predictive classification approach can be employed. Our description is for a general document corpus, but at times we use examples from news article summarization to ground the general description and to show the need to consider context.

Choosing the document units

We divide the raw text into units of analysis and determine which of those units have relevant information about the subject, and summarize based on common features found in these units. The granularity with which the text is partitioned may then have some impact on the resulting summaries. In particular, we hypothesized that using smaller, lower-word-count units of text should produce more detail-oriented summaries, while using larger units

will highlight key-phrases dealing more with the larger themes discussed when the subject of interest is mentioned.

We tested this hypothesis by comparing summarizers that analyze at the article level to those which analyze at the component-paragraphs level. Interestingly, we found no large differences. See Section 3.4.

Identifying potential summarizing phrases

To build the document-term matrix C we first identify all possible phrases that could be part of a summary. This list of possibilities constitute our *dictionary*. Building this dictionary begins with asking, "Which text phrases are acceptably descriptive?" Sometimes the answer to this question suggests a manually-defined dictionary: if summaries should only list, e.g., countries then the dictionary would be easy to assemble by hand.

In many situations, however, the dictionary of terms should be kept large, and possibly be drawn from the corpus itself. Different decisions—Should capitalization matter? Are punctuation marks terms? Can terms include numerals?—yield dictionaries varying widely in size and utility. Terms could be further disambiguated by many natural language tools, e.g., part-of-speech tagging, which would again increase dictionary size. Any automated system for forming a dictionary will entail at least some uncertainty in term identification.

We elected to use a large dictionary containing all phrases of up to three words in length. We generated our dictionary by first removing all numerals and punctuation, then case-folding (converting all text to lowercase). We then segmented each document into overlapping phrases, consisting of all single words, bigrams and trigrams in that document unit. Some text analysts stem their phrases, reducing words to a core root prefix, e.g., truncating *"terror,"* *"terrorist,"* and *"terrorism"* to *"terror"*. But we do not stem. There is a semantic difference if a particular subject is associated with *"canadians"*, the citizenry versus *"canada"* the country. For our corpus of articles from the New York Times International Section in 2009, this identified in 4.5 million distinct phrases. A first-pass pruning, removing all phrases appearing fewer than six times, resulted in a dictionary of $p = 216,626$ distinct phrases.

Representing subjects as lists of words/phrases and automatically labeling documents

We train a classifier to predict document labels, $y_i \in \{-1, +1\}$, with their vectors of phrase features, $x_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$. In the labeling step we automatically build the label vector $y = (y_1, \ldots, y_n)$ by deciding whether each document unit in the corpus should be considered a positive class example or a negative class example.

Establishing the class labels for a news document corpus is sometimes straightforward. For instance, for comparing 1989 articles about mental illness to those from 1999 as Wahl et. al did, the labels would be simple: the documents from the opposing years go in opposite classes. We build y by identifying which of the document units treat the subject of interest

with "treat" being precisely defined later. For small enough n, y could be built by hand. For corpora too large to admit manual labeling, we need reasonable automatic labeling. Ideally this need not be a perfect identification—noise in labeling should not have undue impact on the resulting summaries.

In the large sense, a subject is a concept of interest that an investigator might have. We represent a subject with a small list of words or phrases, e.g., the subject of China would be well represented by the set {*"china," "chinas," "chinese"*}. Specifically, let the *subject* $Q \subset \mathcal{J}$ be a list of words or phrases selected by the user to capture a subject of interest.

count-m. We consider two general automatic labeling techniques. The first technique, count-m, marks document unit i as treating a subject if related phrases appear frequently enough, as given by:

Definition 3.2.2. Count-m labeling labels document unit i as:

$$y_i = 2 \cdot \mathbb{1}\{r_i \geq m\} - 1$$

where $\mathbb{1}\{\cdot\}$ is the indicator function and $r_i \equiv \sum_{j \in Q} c_{ij}$ is the total number of subject-specific phrases in unit i.

hardcount-m. The second automatic labeling technique, hardcount-m, drops any document i with $0 < r_i < m$ from the data set instead of labeling it with -1. The hardcount method considers those documents too ambiguous to be useful as negative class examples. It produces the same positive example set as count-m. We hypothesized that dropping the ambiguous document units would heighten the contrast in content between the two classes, and thus lead to superior summaries. It did not. See Section 3.4.

We generate y this way because we are interested in how a subject is treated in a corpus. Other approaches are possible. For example, y might identify which articles were written in a specific date range; this would then lead to a summary of what was covered in that date range.

Reweighting and removing features

It is well known that baseline word frequencies impact information retrieval methods, and so often raw counts are adjusted to account for commonality and rarity of terms [e.g., 57, 71]. In the predictive framework, this adjustment is done in the construction of the feature matrix X. We consider three different constructions of X, all built on the bag-of-phrases representation. [71] examined a variety of weighting approaches for document retrieval in a multi-factor experiment and found choice of approach to be quite important; we take a similar comparative approach for our task of summarizing a corpus of documents.

We first remove the columns corresponding to the phrases in subject Q from the set of possible features $\mathcal{J}$ to prevent the summary from being trivial and circular. We also remove sub-phrases and super-phrases. For example, if Q is {*"united states"*} then candidate

summary phrases *"united"*, *"states"*, *"of the united"*, and *"states of america"* would all be removed. The removal is easily automated.

Our baseline then is to simply drop stop words (words determined a priori as too uninformative to merit inclusion). Our second approach is to rescale each phrase vector (column of C) to have unity L^2 norm. Our third is an implementation of the tf-idf technique [71, 70], rescaling the bag-of-phrases components so that appearances of rarer phrases are considered more important than common ones.

Stop Words. Stop words are low information words such as *"and,"* or *"the"*, typically appearing with high frequency. Stop words may be context dependent. For example, in US international news *"united states"* or *"country"* might be considered high frequency and low information. High-frequency words have higher variance and effective weight in many methods, causing them to be erroneously selected as features due to sample noise. To deal with these nuisance words, many text-processing methods use a fixed, hand-built stop-word list and preemptively remove all features on that list from consideration [e.g., 91, 40, 28].

This somewhat ad-hoc method does not adapt automatically to the individual character of a given corpus and presents many difficulties. Switching to a corpus of a different language would require new stop word lists. When considering phrases instead of single words, the stop word list is not naturally or easily extended. For example, simply dropping phrases containing any stop word is problematic: it would be a mistake to label *"shock* and *awe"* uninteresting. On the other hand, there are very common candidate phrases that are entirely made up of stop words, e.g., *"of the,"* so just culling the single word phrases is unlikely to be sufficient. See [57] for further critique.

L^2-rescaling. As an alternative, appropriately adjusting the document vectors can act in lieu of a stop-word list by reducing the variance and weight of the high-frequency features. We use the corpus to estimate baseline appearance rates for each feature and then adjust the matrix C by a function of these rates; this core idea is discussed by [57].

We define L^2-rescaling to be:

Definition 3.2.3. X is a *L^2-rescaled* version of C if each column of C is rescaled to have unit length under the L^2 norm. I.e.:

$$x_{ij} = \frac{c_{ij}}{\sqrt{z_j}}, \text{ where } z_j \equiv \sum_{i=1}^{n} c_{ij}^2$$

tf-idf Weighting. An alternate rescaling comes from the popular tf-idf heuristic [70], which attempts to de-emphasize commonly occurring terms while also trying to account for each document's length.

Definition 3.2.4. X is a *tf-idf weighted* version of C if

$$x_{ij} := \frac{c_{ij}}{q_i} \log\left(\frac{n}{d_j}\right)$$

where $q_i \equiv \sum_{j=1}^{p} c_{ij}$ is the sum of the counts of all key-phrases in document i and $d_j \equiv \sum_{i=1}^{n} \mathbb{1}\{c_{ij} > 0\}$ is the number of documents in which term j appears at least once.

Under tf-idf, words which appear in a large proportion of documents are shrunk considerably in their representation in X. Words which appear in all n documents, such as *"the"*, are zeroed out entirely. A potential advantage of tf-idf is that it might ease comparisons between documents of different lengths because term counts are rescaled by the total count of terms in the document.

To illustrate the advantages of reweighting, we generated four summaries from the L1LR feature selection method with all combinations of L^2-rescaling and stop-word removal (see supplementary material). Without reweighting, if no stop words are deleted the list is dominated by high-frequency, low-content words such as *"of,"* *"and,"* and *"was"*. The just stop-word removal list is only a bit better. It contains generic words such as *"mr,"* *"percent,"* and *"world."* The rescaled list does not contain these words. This is a common problem with stop-word lists: they get rid of the worst offenders, but do not solve the overall problem. The list from stop-word removal together with L^2-rescaling and the list from just L^2-rescaling are the same—the rescaling, in this case, has rendered the stop-word list irrelevant.

We hypothesized that feature weighting is more transparent and reproducible than stop-word removal and that it results in superior summaries when compared to stop-word removal. With the human validation experiment, we compared using L^2-rescaling, tf-idf weighting, and stop-word removal as the pre-processing step for each of our feature selectors and found that humans indeed prefer lists coming from the reweighting methods.

Four Feature Selection Methods

Classical prediction yields models that give each feature a non-zero weight. The models are thus hard to interpret when there are many features, as is typically the case with text analysis (our data set contains more than 200,000 potential phrases). We, however, want to ensure that the number of phrases selected is small so the researcher can easily read and consider the entire summary. Short summaries are quick to digest, and thus are easily comparable. Such summaries might even be automatically generated for a corpus in one language and then translated to another, thus easing comparison of media coverage from different nationalities and allowing insight into foreign language news [13]. Fundamentally, a small set of features is more easily evaluated by human researchers.

Given the features X and document labels y for a subject, we extract the columns of X that constitute the final summary. We seek a subset of phrases $\mathcal{K} \subseteq \mathcal{J}$ with cardinality as close as possible to, but no larger than, a target k, the desired summary length. We typically use $k = 15$ phrases, but 30 or 50 might also be desirable. The higher the value of

k, the more detailed and complex the summary. We require the selected phrases be *distinct* meaning that we don't count sub-phrases. For example, if *"united states"* and *"united"* are both selected, we drop *"united"*.

The constraint of short summaries makes the summarization problem a sparse feature selection problem, as studied in, e.g., [23, 49, 90]. *Sparse* methods, such as L^1-penalized regression, naturally select a small subset of the available features (in our case candidate key-phrases) as being relevant predictors.

In other domains, L^1-regularized methods are useful for sparse model selection; they can identify which of a large set of mostly irrelevant features are associated with some outcome. In our domain there is no reasonable underlying model that is indeed sparse; we expect different phrases to be more or less relevant, but few to be completely and utterly irrelevant. Nevertheless, we still employ the sparse methods to take advantage of their feature selection aspects, hoping that the most important features will be selected first.

We examine four methods for extraction, detailed below. Two of them, Co-occurrence and Correlation Screening, are scoring schemes where each feature gets scored independently and the top-scoring features are taken as a summary. This is similar to traditional key-phrase extraction techniques and to other methods currently used to generate word clouds and other text visualizations. The other two (the Lasso and L1LR) are L^1 regularized least squares linear regression and logistic regression, respectively. Table 3.2 displays four summaries for China, one from each feature selector: choice matters greatly. Co-occurrence and L1-penalized logistic regression (L1LR), are familiar schemes from previous work [27].

Co-occurrence

Co-occurrence is our simplest, baseline, method. The idea is to take those phrases that appear most often (or have greatest weight) in the positively marked text as the summary. This method is often used in, e.g., newspaper charts showing the trends of major words over a year (such as Google News Trends[2]) or word or tag clouds (created at sites such as Wordle[3]).

By the feature selection step we have two labelled document subsets, $\mathcal{I}^+ = \{i \in \mathcal{I} | y_i = +1\}$, of cardinality $\#\mathcal{I}^+$, and $\mathcal{I}^- = \{i \in \mathcal{I} | y_i = -1\}$, of cardinality $\#\mathcal{I}^-$. The relevance score s_j of feature j for all $j \in \mathcal{J}$ is then:

$$s_j = \frac{1}{\#\mathcal{I}^+} \sum_{i \in I^+} x_{ij}$$

s_j is the average weight of the phrase in the positively marked examples.

For some k', let $\bar{s}$ be the $(k'+1)$th highest value found in the set $\{s_j \mid j \in \mathcal{J}\}$. Build $K = \{j \in \mathcal{J} \ : \ s_j > \bar{s}\}$, the set of (up to) k' phrases with the highest average weight across the positive examples. Any phrases tied with the $(k'+1)$th highest value are dropped,

[2] `http://www.google.com/trends`

[3] `http://www.wordle.net/`

	Co-occurrence	Correlation	L1LR	Lasso
1	and	beijing and	asian	asian
2	by	beijings	beijing	beijing
3	contrib. research	contrib. research	contrib. research	contrib. research
4	for	from beijing	euna lee	exports
5	global	global	global	global
6	has	in beijing	hong kong	hong kong
7	hu jintao	li	jintao	jintao
8	in beijing	minister wen jiabao	north korea	north korea
9	its	president hu jintao	shanghai	shanghai
10	of	prime minister wen	staterun	tibet
11	that	shanghai	uighurs	uighurs
12	the	the beijing	wen jiabao	wen jiabao
13	to	tibet	xinhua	xinhua
14	xinhua	xinhua the		
15	year	zhang		

Table 3.2: A Comparison of the Four Feature Selection Methods. Four sample summaries of news coverage of China. (Documents labeled via count-2 on articles, X from L^2-rescaling.) Note superior quality of Lasso and L1LR on the right.

sometimes giving a list shorter than k'. The size of K after subphrases are removed can be even less. Let the initial value of k' be k, the actual desired length. Now adjust k' upwards until just before the summary of *distinct* phrases is longer than k. We are then taking the $k' \geq k$ top phrases and removing the sub-phrases to produce k or fewer distinct phrases in the final summary.

If $X = C$, i.e. it is not weighted, then s_j is the average number of times feature j appears in $\mathcal{I}^+$, and this method selects those phrases that appear most frequently in the positive examples. The weighting step may, however, reduce the Co-occurrence score for common words that appear frequently in both the positive and negative examples. This is especially true if, as is usually the case, there are many more negative examples than positive ones. Appropriate weighting can radically increase this method's performance.

Correlation Screening

Correlation Screening selects features with the largest absolute correlation with the subject labeling y. It is a fast method that independently selects phrases that tend to appear in the positively marked text and not in the negatively marked text. Score each feature as:

$$s_j = |cor(x_j, y)|$$

Now select the k highest-scoring, distinct features as described for Co-occurrence, above.

L1-penalized linear regression (the Lasso)

The Lasso is an L^1-penalized version of linear regression and is the first of two feature selection methods examined in this chapter that address our model-sparsity-for-interpretability constraint explicitly. Imposing an L^1 penalty on a least-squares problem regularizes the vector of coefficients, allowing for optimal model fit in high-dimensional ($p > n$) regression settings. Furthermore, L^1 penalties typically result in sparse feature-vectors, which is desirable in our context. The Lasso takes advantage of the correlation structure of the features to, in principle, avoid selecting highly correlated terms. For an overview of the Lasso and other sparse methods see, e.g., *The Elements of Statistical Learning* [34].

The Lasso is defined as:

$$(\hat{\beta}(\lambda), \hat{\gamma}) := \arg \min_{\beta,\gamma} \sum_{i=1}^{m} \left\| y - x_i^T \beta - \gamma \right\|^2 + \lambda \sum_j |\beta_j|. \tag{3.1}$$

The penalty term λ governs the number of non-zero elements of β. We use a non-penalized intercept, γ, in our model. Penalizing the intercept would shrink the estimated ratio of number of positive example documents to the number of negative example documents to 1. This is not desirable; the number of positive examples is far less than 50%, as shown in Table 1 in the supplementary materials, and in any case is not a parameter which needs estimation for our summaries. We solve this convex optimization problem with a modified version of the BBR algorithm [28] described further in Section 3.2.

L1-penalized logistic regression (L1LR)

Similar to the Lasso, L1-penalized logistic regression (L1LR) is typically used to obtain a sparse feature set for predicting the log-odds of an outcome variable being either $+1$ or -1. It is widely studied in the classification literature, including text classification [see 28, 40, 91]. We define the model as:

$$(\hat{\beta}(\lambda), \hat{\gamma}) := \arg \min_{\beta,\gamma} - \sum_{i=1}^{m} \log \left(1 + \exp[-y_i(x_i^T \beta + \gamma)]\right) + \lambda \sum_j |\beta_j|. \tag{3.2}$$

The penalty term λ again governs the number of non-zero elements of β. As with Lasso, we again do not penalize the intercept. We implement L1LR with a modified form of the BBR algorithm.

Implementation and computational cost

Computational costs primarily depend on the size and sparsity of X. We store X as a list of tuples, each tuple being a row and column index and value of a non-zero element. This list is sorted so it is quick to identify all elements in a matrix column. This data structure saves both in storage and computational cost.

The complexity for the tf-idf and L^2 rescaling methods are $O(Z)$, with Z being the number of nonzero elements in X, because we only have to re-weight the nonzero elements and the weights are only calculated from the nonzero elements. Stop-word elimination is also $O(Z)$.

The running times of the four feature selection methods differ widely. For Correlation Screening and Co-occurrence, the complexity is $O(Z)$. The Lasso and L1LR depend on solving convex optimization problems. We implemented them using a modified form of the BBR algorithm [28]. The BBR algorithm is a coordinate descent algorithm for solving L^1 penalized logistic regressions with penalized (or no) intercept. It cycles through all the columns of X, iteratively computing the optimal $\hat{\beta}_j$ for feature j holding the other $\hat{\beta}_k$ fixed. We modified the BBR algorithm such that 1) we can solve the Lasso with it; 2) we do not penalize the intercept; and 3) the implementation exploits the sparsity of X. Not penalizing the intercept preserves sparsity, even if we chose to center the columns of X.

For each iteration of the modified BBR, we first calculate the optimum intercept $\hat{\gamma}$ given the current value of $\hat{\beta}$ and then cycle through the features, calculating the update to β_j of $\Delta\beta_j$ for $j = 1, \ldots, p$. The complexity for calculating $\Delta\beta_j$ is $O(Z_j)$, where Z_j is the number of nonzero elements in the jth column of X. We then have a complexity cost of $O(Z)$ for each full cycle through the features. We stop when the decrease in the loss function after a full cycle through β_1 to β_p is sufficiently small. For both the Lasso and Logistic regression, we need only a few iterations to obtain the final solution. When not exploiting the sparse matrix, the complexity of a coordinate decent iteration is $O(n \times p)$. When $Z \ll n \times p$, our implementation saves a lot of computation cost.

For both the Lasso and L1LR, higher values of λ result in the selection of fewer features. A sufficiently high λ will return a β with zero weight for all phrases, selecting no phrase features, and $\lambda = 0$ reverts the problem to ordinary regression, leading to some weight put on all phrases in most circumstances. By doing a binary search between these two extremes, we quickly find a value of λ for which $\beta(\lambda)$ has the desired k distinct phrases with non-zero weight.

Computation speed comparisons. We timed the various methods to compare them given our data set. The average times to summarize a given subject for each method, not including the time to load, label, and rescale the data, are on Table 3.3. As expected, Co-occurrence and Correlation Screening are roughly the same speed. The data-preparation steps by themselves (not including loading the data into memory) average a total of 15 seconds, more expensive by far than the feature selection for the simple methods (although we did not optimize the labeling of y or the dropping the subject-related features from X).

The Lasso is currently about 9 times slower and L1LR is more than 100 times slower than the baseline Co-occurrence using current optimization techniques, but these techniques are evolving fast. For example, one current area of research, safe feature elimination, allows for quickly and safely pruning many irrelevant features before fitting, leading to substantial speed-ups [18]. This pre-processing step allows for a huge reduction in the number of features

when the penalty parameter is high, which is precisely the regime where the desired list length is short.

The difference between the Lasso and L1LR may seem surprising given the running time analysis above. L1LR is slower for two main reasons: for a given λ L1LR requires more iterations to converge on average, and for each iteration, L1LR takes more time due to, in part, more complex mathematical operations.

We implemented the sparse regression algorithms and feature correlation functions in C; the overall package is in Matlab. L1LR is the slowest method, although further optimization is possible. The speed cost for Lasso, especially when considering the overhead of labeling and rescaling, is fairly minor, as shown on the third column of Table 3.3.

	Phrase selection (sec)	Total time (sec)	Percent increase
Co-occurrence	1.0	20.3	
Correlation Screen	1.0	20.3	0%
The Lasso	9.3	28.7	+41%
L1LR	104.9	124.2	+511%

Table 3.3: Computational Speed Chart. Average running times for the four feature selection methods over all subjects considered. Second column includes time to generate y and adjust X. Final column is percentage increase in total time over Co-occurrence, the baseline method.

Comments

The primary advantages of Co-occurrence and Correlation Screening is that they are fast, scalable, and easily distributed across multiple platforms for parallel processing. Unfortunately, as they score each feature independently from the others, they cannot take advantage of any dependence between features to aid summarization, for example, to remove redundant phrases. The Lasso and L1LR, potentially, do. The down side is sparse methods can be more computationally intensive.

Final summaries consist of a target of k *distinct* key-phrases. The feature-selectors are adjusted to provide enough phrases such that once sub-phrases (e.g., *"united"* in *"united states"*) are removed, the list is k phrases long. This removal step, similar to stop-word removal, is somewhat ad hoc. It would be preferable to have methods that naturally select distinct phrases that do not substantially overlap. Sparse methods are such methods; they do not need to take advantage of this step, supporting the heuristic knowledge that L^1-penalization tends to avoid selecting highly correlated features. With tf-idf, an average of about one phrase is dropped for Lasso and L1LR. The independent feature selection methods, however, tend to drop many phrases. See Section 1 of the supplementary material.

3.3 Human Experiment

Four sample summaries of the coverage of four different countries are shown in Table 3.2. Sometimes fragments are selected as stand-ins for complete phrases, e.g., *"President Felipe [Calderòn]."* These summaries inform us as to which aspects of these countries are of most concern to the New York Times in 2009: even now, Nazis and the World Wars are tied to Germany. Iraq and Afghanistan are also tied closely. Gen[erals] and combat are the major focus in Iraq. The coverage of Mexico revolves around the swine flu, drug cartels, and concerns about the border. Russia had a run-in with Europe about gas, and nuclear involvement with Iran. These summaries provide some pointers as to future directions of more in-depth analysis. They came from a specific combination of choices for the reweighting, labeling, and feature selection steps. But are these summaries better, or worse, than the summaries from a different summarizer?

Comparing the efficacy of different summarizers requires systematic evaluation. To do this, many researchers use corpora with existing summaries (e.g., using the human-encoded key-phrases in academic journals such as in [24]), or corpora that already have human-generated summaries (such as the TIPSTER dataset used in [59]). Others have humans generate summaries for a sample of individual documents and compare the summarizer's output to this human baseline. We, however, summarize many documents, and so we cannot use an annotated evaluation corpus or summaries of individual documents.

In the machine-learning world, numerical measures such as prediction accuracy or model fit are often used to compare different techniques, and has been successful in many applications. While we hypothesize that prediction accuracy should correlate with summary quality to a certain extent, there are no results to demonstrate this. Indeed some researchers found that it does not always [27, 9]. Because of this, evaluating summarizer performance with numerical measures is not robust to critique.

Although time consuming, people can tell how well a summary relates to a subject, as the hand-coding practice in media analysis shows. Because final outcomes of interest are governed by human opinion, the only way to validate that a summarizer is achieving its purpose is via a study where humans assess summary quality. We therefore design and conduct such a study. Our study has three main aims: to verify that features used for classification are indeed good key-phrases, to help learn what aspects of the summarizers seem most important in extracting the key meaning of a corpus, and to determine which feature selection methods are most robust to different choices of pre-processing (choice of granularity, labeling of document units, and rescaling of X).

We compare our four feature selection methods under a variety of labeling and reweighting choices in a crossed, randomized experiment where non-experts read both original documents and our summaries and judge the quality and relevance of the output. Even though we expect individuals' judgements to vary, we can average the responses across a collection of respondents and thus get a measure of overall, generally shared opinion.

We carried out our survey in conjunction with the XLab, a campus lab dedicated to helping researchers conduct human experiments. We recruited 36 respondents (undergraduates

at a major university) from the lab's respondent pool via a generic, nonspecific message stating that there was a study that would take up to one hour of time. While these experiments are expensive and time consuming, they are necessary for evaluating text summarization tools.

Description of text corpus

For our investigation we used the International Section of the New York Times for the 2009 year. Articles were scraped from the newspaper's RSS feed,[4] and the HTML markup was stripped from the text. We obtained 130,266 paragraphs of text comprising 9,560 articles. The New York Times, upon occasion, will edit an article and repost it under a different headline and link; these multiple versions of the articles remain in the data set. By looking for similar articles as measured by a small angle between their feature vectors in the document-term matrix C, we estimate that around 400 articles (4–5%) have near-duplicates.

The number of paragraphs in an article ranges from 1 to 38. Typical articles[5] have about 16 paragraphs (with an Inter-Quartile Range (IQR) of 11 to 20 paragraphs). However, about 15% of the articles, the "World Briefing" articles, are a special variety that contain only one long paragraph.[6] Among the more typical, non-"World Briefing" articles, the distribution of article length as number of paragraphs is bell-shaped and unimodal. The longer articles, with a median length of 664 words, have much shorter paragraphs (median of 38 words), generally, than the "Word Briefing" single-paragraph articles (median of 87 words).

Generating the sample of summaries

A choice of each of the options described above gives a unique summarizer. We evaluated 96 different summarizers built from these factors:

1. We evaluated summarizers that analyzed the data at the article-unit and paragraph-unit level (see Section 3.2).

2. When performing paragraph-unit analysis, we labeled document units using count-1, count-2, and hardcount-2. For the article-unit analysis we considered these three, plus count-3 and hardcount-3 (see Section 3.2).

3. We considered tf-idf weighting, L^2 rescaling, and simple stop-word removal (see Section 3.2).

4. We considered four feature-selection techniques (see Section 3.2).

[4] `feed://feeds.nytimes.com/nyt/rss/World`

[5] See, e.g., `http://www.nytimes.com/2011/03/ 04/world/americas/04mexico.html`

[6] See, e.g., `http://www.nytimes.com/2011/03/03/ world/americas/03briefs-cuba.html`

Choices of unit of analysis, labeling, the three preprocessing options, and the four feature-selection methods give 96 different summarizers indexed by these combinations of factors.

We compared the efficacy of these combinations by having respondents assess the quality of several different summaries generated by each summarizer. We applied each summarizer to the set of all articles in the New York Times International Section from 2009 for 15 different countries of interest. These countries are listed in Table 1 in the supplementary materials. We only considered countries with reasonable representation in the corpus. After identifying these countries, we hand-specified a phrase set Q for each county by including any plurals and possessives of the country and any common names for the country's people. Using these 15 subjects on each of the 96 summarizers, we calculated 1,440 summaries.

Supplementary Table 1 also includes the number of positively marked examples under all the count-m labeling schemes we used for both article- and paragraph-unit analysis. The "article-1" header is the most generous labeling: any article that mentions any of the words associated with the subject one or more times is marked as treating the subject. Even under this, positive examples are scarce; it is clear we are attempting to summarize something that does not constitute a large portion of the text.

The survey and respondents

For our survey, paid respondents were convened in a large room of kiosks where they could be asked questions in a focused, controlled environment. . Each respondent sat at a computer and was given a series of questions over the course of an hour. Respondents assessed a series of summaries and articles presented in 6 blocks of 8 questions each. Each block considered a single (randomly selected) subject from our list of 15. Within a block, respondents were first asked to read four articles and rate their relevance to the specified subject. Respondents were then asked to read and rate four summaries of that subject randomly chosen from the subject's library of 96. Respondents could not go back to previous questions.

Before the survey all respondents did a sample series of four practice questions and were then asked if they had any questions as to how to score or rate articles and summaries.

Evaluating article topicality. To insure that respondents had a high probability of seeing several articles actually relevant to the subject being investigated, the articles presented in a block were selected with a weighted sampling scheme with weights proportional to the number of times the block's country's name was mentioned. We monitored the success of this scheme (and collected data about the quality of the automatic labelers) by asking the respondents to evaluate each shown article's relevance to the specified subject on a 1 to 7 scale.

With 4 or higher scored as relevant, respondents saw at least 2 articles (out of the 4) on the subject of interest about 75% of the time. With 3 or higher, the number of blocks with at least two relevant articles rises to 91%. We attempt to summarize how a subject is treated overall, including how it is treated in articles in which the subject is only a secondary consideration. For example, an article focused on world energy production may

discuss Russia. Hence, even a modest score of 3 or 4 is a likely indication that the article has subject-related content.

Only the first 120 words of each article were shown; consultation with journalists suggests this would not have a detrimental impact on content presented, as a traditional newspaper article's "inverted pyramid" structure moves from the most important information to more minute details as it progresses [65].

Evaluating summaries. A simple random sample of 4 summaries were presented after the four articles. Each summary was presented on its own screen. The respondents were asked to assess each summary in four respects:

1. Content: How does this list capture the content about [subject] in the text you just read? (1-7 scale, 7 being fully captured)

2. Relevance: How many words irrelevant or unrelated to [subject] does this list contain? (1-7 scale, 7 being no irrelevance)

3. Redundancy: How many redundant or repeated words does this list contain? (1-7 scale, 7 being no redundancies)

4. Specificity: Given what you just read, would you say this list is probably too general or too specific a summary of how [subject] was covered by the newspaper in 2009? (response options: too general, about right, too specific, not relevant, and can't tell)

All respondents finished their full survey, and fewer than 1% of the questions were skipped. Time to completion ranged from 14 to 41 minutes, with a mean completion time of 27 minutes.

3.4 Human Survey Results

We primarily examined an aggregate "quality" score, taken as the mean of the three main outcomes (Content, Relevance, and Redundancy). Figure 3.1 shows the raw mean aggregate outcomes for the article-unit and paragraph-unit data. We immediately see that the Lasso and L1LR perform better than Co-Occurrence and Correlation Screen, and Tf-idf is a good choice of rescaling for articles, L^2-rescaling for paragraphs. Labeling does not seem to matter for the articles, but does for paragraphs. Clearly, the unit of analysis interacts with the other three factors, and so we conduct further analysis of the article-unit and paragraph-unit data separately. Section 3.4 has overall comparisons.

We analyze the data by fitting the respondents' responses to the summarizer characteristics using linear regression. The full model includes terms for respondent, subject, unit type, rescaling used, labeling used, and feature selector used, as well as all interaction terms for the latter four features.

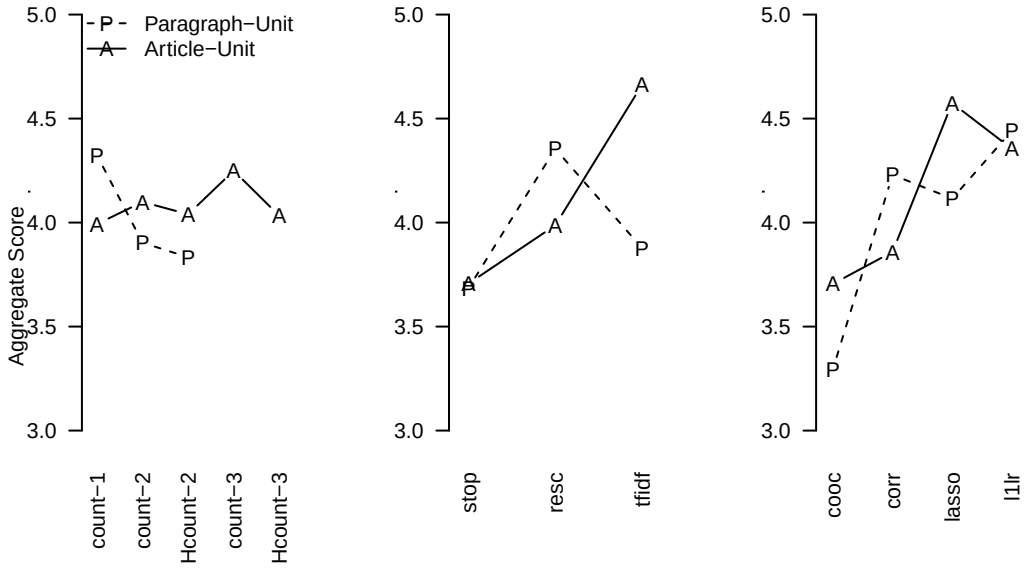

Figure 3.1: Aggregate Results of Human Survey. Outcome is aggregate score based on the raw data. There are major differences between article-unit analysis and paragraph-unit analysis when considering the impact of choices in preprocessing.

In all models, there are large respondent and subject effects. Some subjects were more easily summarized than others, and some respondents were more critical than others. Interactions between the four summarizer factors are (unsurprisingly) present ($df = 33, F = 4.14, \log P \approx -13$ under ANOVA). There are significant three-way interactions between unit, feature-selector, and rescaling ($P \approx 0.03$) and labeling, feature-selector, and rescaling ($P \approx 0.03$). Interaction plots (Figure 3.1) suggest that the sizes of these interactions are large, making interpretation of the marginal differences for each factor potentially misleading. Table 3.4 shows all significant two-way interactions and main effects for the full model, as well as for models run on the article-unit and paragraph-unit data separately.

Article unit analysis

Interactions between factors make interpretation difficult, but overall, Lasso is a good summarizer that is resistant to preprocessing choices. Interestingly, the simplest method, Co-occurrence, is on par with Lasso under tf-idf.

The left column of Figure 3.2 shows plots of the three two-way interactions between feature selector, labeling scheme, and rescaling method for the article-unit data. There is a strong interaction between rescaling and feature-selection method ($df = 6, F = 8.07, \log P \approx -8$, top-left plot), and no evidence of a labeling by feature-selection interaction or a labeling by rescaling interaction. Model-adjusted plots (not shown) akin to Figure 3.2 do not

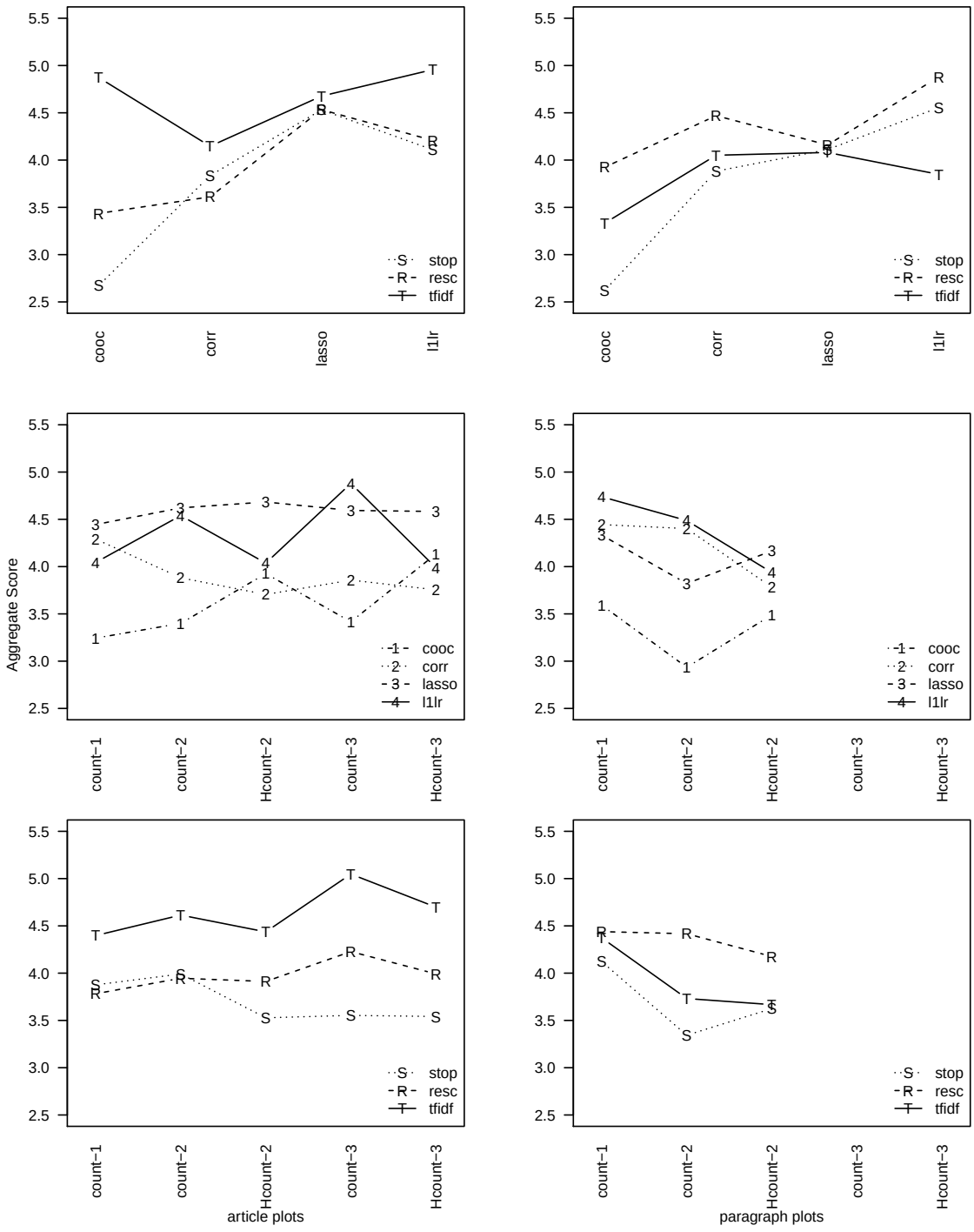

Figure 3.2: Aggregate Quality Plots. Pairwise interactions of feature selector, labeling, and rescaling technique. Left-hand side are for article-unit summarizers, right for paragraph-unit. See testing results for which interactions are significant.

	All data				Article-unit			Paragraph-unit		
Factor	Unit	Feat	Lab	Resc	Feat	Lab	Resc	Feat	Lab	Resc
Unit	**.**	-2	-1	-7						
Feature Selection		**-15**	.	-10	**-10**	.	-7	**-6**	.	-2
Labeling			**.**	.		**.**	.		**-1**	.
Rescaling				**-14**			**-15**			**-3**

Table 3.4: Main Effects and Interactions of Factors. Main effects along diagonal in bold. A number denotes a significant main effect or pairwise interaction for aggregate scores, and is the base-10 log of the P-value. "." denotes lack of significance. "All data" is all data in a single model. "Article-unit" and "paragraph-unit" indicate models run on only those data for summarizers operating at that level of granularity.

differ substantially in character. Table 3.4 show all significant main effects and pairwise interactions. There is no significant three-way interaction.

Lasso is the most consistent method, maintaining high scores under almost all combinations of the other two factors. In Figure 3.2, note how Lasso has a tight cluster of means regardless of rescaling used in the top-left plot and how Lasso's outcomes are high and consistent across all labeling in the middle-left plot. Though L1LR or Co-occurrence may be slightly superior to Lasso when the data has been vectorized according to tf-idf, they are not greatly so, and, regardless, both these methods seem fragile, varying a great deal in their outcomes based on the text preprocessing choices. Note, for example, how vulnerable the Co-occurrence feature-selection method is to choice of rescaling.

Tf-idf seems to be the best overall rescaling technique, consistently coming out ahead regardless of choice of labeling or feature-selection method. Note how its curve is higher than the rescaling and stop-word curves in both the top- and bottom-left plots in Figure 3.2. Under tf-idf, all the methods seem comparable. Alternatively put, tf-idf brings otherwise poor feature selectors up to the level of the better selectors.

Adjusting P-values with Tukey's honest significant difference and calculating all pairwise contrasts for each of the three factors show which choices are overall good performers, ignoring interactions. For each factor, we fit a model with no interaction terms for the factor of interest and then performed pairwise testing, adjusting the P-values to control familywise error rate. See Table 3.5 for the resulting rankings of the factor levels. Co-occurrence and Correlation Screening are significantly worse than L1LR and Lasso (correlation vs. L1LR gives $t = 3.46, P < 0.005$). The labeling method options are indistinguishable. The rescaling method options are ordered with tf-idf significantly better than rescaling ($t = 5.08, \log P \approx -4$), which in turn is better than stop-word removal ($t = 2.45, P < 0.05$).

Data Included	Order (article)	Order (paragraph)
All	cooc, corr < L1LR, Lasso stop < resc < tf-idf	cooc < corr, Lasso, L1LR tfidf, stop < resc
tf-idf only	no differences	no differences
L^2 only	cooc < L1LR, Lasso; corr < Lasso	no differences
stop only	cooc < corr, L1LR, Lasso; corr < Lasso	cooc < Lasso, L1LR
cooc only	stop < resc < tf-idf	stop < resc
corr only	stop < tf-idf	no differences
Lasso only	no differences	no differences
L1LR only	no differences	tf-idf < resc

Table 3.5: Quality of Feature Selectors. This table compares the significance of the separation of the feature selection methods on the margin. Order is always from lowest to highest estimated quality. A "<" denotes a significant separation. All P-values corrected for multiple pairwise testing. The last seven lines are lower power due to subsetting the data.

Paragraph unit analysis

For the paragraph-unit summarizers, the story is similar. Lasso is again the most stable to various pre-processing decisions, but does not have as strong a showing under some of the labeling choices. Co-occurrence is again the most unstable. L1LR and Correlation Screening outperform Lasso under some configurations. The main difference from the article-unit data is that tf-idf is a poor choice of rescaling and L^2-rescaling is the best choice.

The right column of Figure 3.2 shows the interactions between the three factors. There is again a significant interaction between rescaling and method ($df = 6, F = 3.25, P < 0.005$, top-plot). This time, however, it is not entirely due to Co-occurrence being sensitive to rescaling. Co-occurrence is still sensitive, but correlation and L1LR are as well. Stop-word removal does quite well for L1LR and Lasso, suggesting that rescaling is less relevant for shorter document units.

Co-occurrence is significantly worse than the other three on the margin (Co-occurrence vs. Correlation Screening gives an adjusted pairwise test with $t = 4.11, P < 0.0005$), but the other three are indistinguishable. Labeling matters significantly ($df = 2, F = 5.23, P < 0.01$), with count-1 doing better in the margin than count-2 and hardcount-2. The higher threshold is likely removing too many substantive paragraphs from the set of positive examples. See Table 1 in the supplementary materials—around 75% of the examples are dropped by moving from count-1 to count-2.

Analysis of subscores

The above analysis considers the aggregate score across (1) specific *Content* captured, (2) *Redundancy* of phrases in the list, and (3) general *Relevance* of phrases in the list to the subject. We also performed the above analyses for each of the three sub-scores separately.

Overall conclusions mainly hold, with a few important exceptions. The paragraph-unit results are especially variable, suggesting that paragraph-unit analysis requires more fine-tuning to get good results than article-unit.

The Lasso and L1LR maintain word-lists that have few repeats (good Redundancy scores), but their information capture degrades when given quite short units of text. This partially explains the weaker performance of Lasso in the aggregate scores for the paragraph-unit. For the paragraph unit summarizers, L^2-Rescaling is clearly superior for Relevance and Content scores, but inferior to tf-idf for Redundancy.

The marginal Redundancy scores for feature selection method are extremely differentiated, with L1LR and Lasso both scoring high and Co-occurrence and Correlation Screening scoring quite low. Correlation Screening's poor Redundancy score substantially reduces its aggregate score. This might be solved by a different, more sophisticated, pruning technique. Indeed, given Correlation Screening's quite high scores for Relevance and Content, fixing the Redundancy problem could result in a good, fast summarizer that may well outperform the penalized regression methods.

Summary of Overall Results

The feature selectors interact differently with labeling and rescaling under the two different units of analyses. While the overall summary quality was no different between these two varieties of summarizer, interaction plots suggest labeling is important, with count-2 being more appropriate for articles and count-1 being more appropriate for paragraph units (see top plots of Figure 3.1). This is unsurprising: a count of 1 vs. 2 means a lot more in a single paragraph than an entire article.

Preprocessing choice is a real concern. While stop-word removal and L^2-rescaling seem relatively consistent across both units of analysis, tf-idf works much worse, overall, for the paragraph unit summarizers than with articles. This is probably due to the short length of the paragraph causing rescaling by term frequency to have large and varying impact. It might also have to do with tf-idf correctly adjusting for the length of the short "World-Briefing" articles. Under Lasso, however, these decisions seem less important, regardless of unit size.

Comparing the performance of the feature selectors is difficult due to the different nature of interactions for paragraph and article units. That said, Lasso consistently performed well. For the article-unit it performed near the top. For the paragraph-unit it did better than most but was not as definitively superior. L1LR, if appropriately staged, also performs well.

We hypothesized that paragraph-unit analysis would generate more specific summaries and article-unit more general. This does not seem to be the case; in analyzing the results for the fourth question on generality vs. specificity of the summaries (not shown), there was no major difference found between article-unit and paragraph-unit summarizers.

It is on the surface surprising that the Lasso often outperformed L1LR as L1LR fits a model that is more appropriate for the binary outcome of the labeling. The Lasso has a L^2-loss, which is sensitive to outliers, while L1LR's logistic curve is less sensitive. However, the

design matrix X, especially under rescaling, is heavily restricted. All entries are nonnegative and few are large. This may limit the opportunity for individual entries in the L^2 loss to have a significant impact, ameliorating the major drawback of the Lasso.

There is no evidence that dropping units that mention the subject below a given threshold (the hardcount labeling technique) is a good idea. Indeed, it appears to be a bad one. The pattern of a quality dip between count-n and hardcount-n appears both in the paragraph- and article-unit results. Perhaps articles that mention a subject only once are important negative examples. The sub-scores offer no further clarity on this point.

3.5 Discussion

News media significantly impact our day to day lives and the direction of public policy. Analyzing the news, however, is a complicated task. The labor intensity of hand coding either leads to small-scale studies, or great expense. This and the amount of news available to a typical reader strongly motivate automated methods to help with media analysis.

We proposed a sparse predictive framework for extracting meaningful summaries of specific subjects from document corpora including corpora of news articles. We constructed different summarizers based on our proposed approach by combining different options of data preparation scheme, document-granularity, labeling choice, and feature selection method. A human validation experiment was strongly advocated and carried out for comparing these different summarizers among themselves and with human understanding.

Based on the human experiment, we conclude that the features selected using a prediction framework do generally form informative key-phrase summaries for subjects of interest for the corpus of New York Times international section articles. These summaries are superior to those from simpler methods of the kind currently in wide use such as Co-occurrence. The Lasso is a good overall feature selector that seems robust to how the data is vectorized and labeled and it is computationally scalable. L1LR, a natural fit model-wise, can perform well if preprocessing is done correctly. However, it is computationally expensive. Data preparation is important: the vector representation of the data should incorporate some reweighting of the phrase appearance counts. Tf-idf is a good overall choice unless the document units are small (e.g., paragraphs, and, presumably, headlines, online comments, and tweets) in which case an L^2 scaling should be used. We also learned that limiting phrases to three words or fewer is a potential problem; we encountered it, for example, when dealing with political leaders frequently mentioned with title (as in "Secretary of State Hillary Clinton"). [40] proposed a greedy descent approach for L1LR that allows for arbitrary-length key-phrases, but it currently does not allow for intercept or reweighting. However, it potentially could. Alternatively, natural language tools such as parts of speech tagging could pull out such names as distinct features. These approaches are currently under investigation.

Our framework provides a general tool for summarization of corpora of documents relative to a subject matter. Using this tool, researchers can easily explore a corpus of documents with an eye to understanding a concise portrayal of any subject they desire. We are now

in the process of working with social scientists to use this tool to carry out research in a substantive social science field. Our proposed approach is being implemented within an on-line toolkit, SnapDragon.[7] This toolkit is intended for researchers interested in quickly assessing how corpora of documents cover specified subjects.

[7]http://statnews2.eecs.berkeley.edu/snapdragon

Chapter 4

Election Auditing With the Trinomial Bound

4.1 Introduction

Electronic voting machines and vote tabulation software are complex and opaque, raising concerns about their reliability and vulnerability. Audits can provide a measure of "software independence," controlling the risk that errors—whatever their source—cause the apparent outcome to differ from the outcome a full hand count would show [81, 79, 80, 83]. Several states have laws mandating election audits and others are considering such laws [31].[1] It is crucial to ensure that the audit trail is accurate, durable and complete from its creation through the audit. If there is no audit trail, there can be no audit. If there is an audit trail, but no audit, there is no assurance of accuracy. If there is an audit trail and an audit, but the audit trail does not reflect the electoral outcome, there is still no assurance.

Henceforth, we assume that the audit trail is complete and accurate. When we say "the apparent outcome is correct" we mean the apparent outcome is the same outcome that a full hand count of the audit trail would show. "The apparent outcome is wrong" means a full hand count would show a different outcome.

An election outcome can be checked by hand counting the entire audit trail. This, however, is expensive and time-consuming, and unnecessary unless the outcome is wrong. The goal of a *statistical* audit, which compares a hand count of a random sample of batches of ballots to the audit trail for those batches, is to ensure that the outcome is correct without a full hand count—unless the outcome is wrong. If the outcome is wrong, a full hand count is needed to set the record straight. A *risk-limiting* audit has a minimum pre-specified chance, $1-\alpha$, of requiring a full hand count whenever the apparent outcome is wrong.[2] The *risk*, α, is the largest possible chance that there will not be a full hand count when the outcome is wrong, no matter what caused the discrepancies between the apparent outcome and the audit

[1] See also `http://www.verifiedvoting.org/article.php?id=5816` (last visited 18 February 2009).

[2] See `http://www.electionaudits.org/bp-risklimiting` (last visited 19 February 2009).

trail. (We assume that $\alpha < 1$; otherwise, an audit would be unnecessary.) The guaranteed minimum chance of a full hand count when the outcome is wrong is $1 - \alpha$.

In statistical language, a risk-limiting audit is a significance-level α test of the null hypothesis "the outcome is wrong" against the alternative hypothesis "the outcome is right." Commonly, tests are formulated so that the null hypothesis that things are "good"; here, it is that things are "bad." The reason is that, in the Neyman-Pearson paradigm, the chance of incorrectly rejecting the null hypothesis is controlled to be at most α. We want to control the chance that an incorrect outcome will go undetected, i.e., the chance that there is not a full hand count when there should be.

Not rejecting the null hypothesis entails a full hand count. A good test simultaneously limits the chance of incorrectly rejecting the null hypothesis to at most α and has high power. That is, a good test has chance at least $1 - \alpha$ of requiring a full hand count when the outcome is wrong, and is very likely to conclude that the outcome is right, with a minimum of hand counting, when the outcome is indeed right.

The outcome can be right even when there are some errors, and audits of voter-marked paper ballots generally find errors at a rate of a few tenths of a percent.[3] For a test to have good power, it needs to have a large probability of rejecting the null hypothesis even when some errors are observed, provided the outcome of the race is right. The issue is whether, in light of the errors found in the sample, there is still compelling statistical evidence that the outcome of the race is correct.

Audits compare hand counts of a random sample of batches to reported totals for those batches.[4] The sampling design used in this chapter is sampling with probability proportional to an error bound (PPEB) [2, 83]. Suppose the error in batch p can be no larger than u_p. Let $U = \sum_p u_p$ be the total of all the error bounds. In PPEB, there are n independent draws from the set of N batches. In each draw, the chance of selecting batch p is u_p/U. This makes it more likely that batches that can conceal more error will be audited.

Sampling proportional to an error bound is common in financial auditing, where it is called *dollar unit sampling* or *monetary unit sampling* (MUS) [17]. A standard problem in financial auditing is to find an upper confidence bound for the total overstatement of a set of accounts. Each account has a "book value" in dollars; the real value—the value an audit would reveal—might be lower. The overstatement is the book value minus the real value. The overstatement can be no larger than the book value. Thus, book value is an error bound and MUS is PPEB.

[3] We have seen much better accuracy than this, for instance, in the audit of the race in Marin county described here, and in a November 2008 audit in Yolo County, CA, we participated in. If something goes wrong—a ballot definition error, miscalibrated scanner, bug, or fraud—errors can be much larger. Direct-recording electronic voting machines (DREs) should be perfectly accurate, and any errors in DRE results are cause for alarm and should be thoroughly investigated.

[4] The design of the sample matters for the probability calculations and for efficiency. Some methods, such as SAFE [51] use a simple random sample of batches. Others use stratified simple random samples [81, 79, 80]. States, including California and Minnesota, require drawing random samples stratified by county; batches are ballots for a single precinct. Stratifying on the method of voting—by mail, early, in-precinct or provisional—can have logistical advantages.

Methods used to analyze MUS data generally convert the overstatement to *taint*, which is the overstatement divided by the book value. For instance, if an account with a book value of \$1,000 has an audited value of \$900, the overstatement is \$100 and the taint is $\$100/\$1,000 = 0.1$, i.e., ten cents per dollar.

Working with taint in PPEB samples has theoretical advantages; see [5, 4, 20, 58, 83]. The expected taint of each PPEB draw is the overall error in the population divided by the total of the error bounds for the population. Moreover, the observed taints are independent and identically distributed. Those features make it straightforward to use the taint in a PPEB sample to find an upper confidence bound on the total overstatement error.

There is an extensive literature on confidence bounds for overstatement from PPEB samples [17]. Apparently, [84] developed the first such confidence bound, based on nesting binomial confidence bounds. That bound turns out to be quite conservative in practice; the multinomial bound of [20, 58] is sharper. See section 4.5. The multinomial bound bins the taint into pennies (zero cents per dollar, one cent per dollar, ..., 100 cents per dollar), and uses the multinomial distribution of the counts in each bin to make a confidence bound on the population taint by inverting hypothesis tests. [5, 4] develop a different improvement of the bound in [84]. [83] shows how some common probability inequalities can be used with the taint in a PPEB sample to test hypotheses about the overall error. Those tests can be converted into confidence bounds as well.

We present here a simplified variant of the multinomial bound, the trinomial bound. It divides the taint into three bins and constructs an upper confidence bound for the expected taint by inverting a set of hypothesis tests. The acceptance regions for the trinomial bound differ from those of the multinomial bound.[5] For the kind of data that typically arise in election audits, computing the trinomial bound is straightforward.[6] The trinomial confidence bound for the taint can be small even when some errors are observed. When that happens, the audit stops short of a full hand count and the risk is still limited to at most α.

We used the trinomial bound to audit two November 2008 races, one in Santa Cruz County and one in Marin County, California. Table 4.1 summarizes the election results. The Santa Cruz County contest was for County Supervisor in the 1st District. The competitive candidates were John Leopold and Betty Danner. According to the semi-official results provided to us by the Santa Cruz County Clerk's office, Leopold won with votes on 45% of the 26,655 ballots. Danner received the votes on 37% of the ballots. The remaining ballots were undervoted, overvoted, or had votes for minor candidates.[7]

The Marin County race was for Measure B, county-wide contest that required a simple

[5]The multinomial bound bases the hypothesis tests on "step-down sets," which partially order the set of possible outcomes. We order outcomes by sample mean of the binned taints, which is more intuitive. Using the sample mean to order outcomes for the 101-bin multinomial would be combinatorially complex, but since the trinomial has only three bins it turns out to be simple.

[6]The Kaplan-Markov bound [83] seems to be comparable, but easier to compute; there has been no extensive comparison so far.

[7]In calculating the confidence bound on the error, the audit took every ballot into account, not just the ballots with votes.

majority. According to the semi-official results, provided to us by the Marin County Registrar of Voters office, 121,295 ballots were cast in the race. 51% of the ballots recorded "yes" votes; 35% said "no." The remaining 14% had undervotes or overvotes.

Both audits were designed to limit the risk to $\alpha = 0.25$. That is, the chance of a full hand count was at least 75% if the outcome was wrong. Both outcomes were confirmed without a full hand count.

County	Total Ballots	Winner	Loser	Margin	Precincts	Batches
Santa Cruz	26,655	45%	37%	8%	76	105
Marin	121,295	51%	35%	16%	189	544

Country	Batches Audited	# Ballots Audited	% Ballots Audited
Santa Cruz	16	7,105	27%
Marin	14	3,347	3%

Table 4.1: Summary of the Two Races Audited. The main contenders in the race for Santa Cruz County Supervisor, 1st District, were John Leopold and Betty Danner. Leopold apparently won. Marin Measure B required a simple majority. It apparently passed. The audits confirmed both outcomes.

This chapter is organized as follows. Section 4.2 reviews notation and points to other work for details. Section 4.3 develops the trinomial confidence bound and a method for selecting the bins and the sample size. Section 4.4 explains how the trinomial bound was used to audit contests in Marin and Santa Cruz counties and presents the audit results. Section 4.5 compares the trinomial bound to the Stringer bound. Section 4.6 presents conclusions.

4.2 Notation and Assumptions

We generally follow the notation in [79, 80, 83]. There are K candidates; voters may vote for up to $f \geq 1$ of them (the contest has f winners). There are N batches of ballots, indexed by p. There are v_{kp} votes reported for candidate k in batch p. There are actually a_{kp} votes cast for candidate k in batch p. The total vote reported for candidate k is $V_k = \sum_p v_{kp}$, the sum of the votes reported for candidate k in the N batches. The total actual vote for candidate k is $A_k = \sum_p a_{kp}$. The set $\mathcal{W}$ comprises the indices of the apparent winners so $\#\mathcal{W} = f$. The set $\mathcal{L}$ comprises the indices of the apparent losers, so $\#\mathcal{L} = K - f$.

If $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$ then

$$V_{w\ell} \equiv V_w - V_\ell > 0. \tag{4.1}$$

The outcome of the election is right if for every $w \in \mathcal{W}$ and $\ell \in \mathcal{L}$,

$$A_{w\ell} \equiv A_w - A_\ell > 0. \tag{4.2}$$

Define

$$e_{w\ell p} \equiv \frac{v_{wp} - v_{\ell p} - (a_{wp} - a_{\ell p})}{V_{w\ell}}. \tag{4.3}$$

That is the amount by which error in batch p overstated the margin between candidate w and candidate ℓ, expressed as a fraction of the reported margin between them.

If the outcome of the race is wrong, there is some pair $w \in \mathcal{W}$, $\ell \in \mathcal{L}$ for which

$$\sum_p e_{w\ell p} \geq 1. \tag{4.4}$$

Define

$$e_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} e_{w\ell p}. \tag{4.5}$$

[79] shows that a sufficient condition for the outcome to be correct is

$$E \equiv \sum_p e_p < 1. \tag{4.6}$$

This condition is sufficient but not necessary; tightening the condition could yield better tests.

We want to draw a statistical inference about E from a random sample of batches, making a bare minimum of assumptions about $\{e_p\}$. We do assume that we have a bound b_p on the total number of ballots in batch p. [79] shows that from such a bound we can deduce that

$$e_p \leq u_p \equiv \max_{w \in \mathcal{W}, \ell \in \mathcal{L}} \frac{b_p + v_{wp} - v_{\ell p}}{V_{w\ell}}. \tag{4.7}$$

Let

$$U \equiv \sum_p u_p. \tag{4.8}$$

We call e_p the overstatement error in batch p, E the overstatement error, u_p the maximum overstatement error in batch p, and U the maximum overstatement error.

The sample is selected as follows: We draw n times independently (with replacement) from the set of N batches. In each draw, the probability of selecting batch p is u_p/U. This is called a PPEB (probability proportional to an error bound) sample [2]; it is equivalent to monetary unit sampling and dollar unit sampling in financial auditing [17].

This chapter gives a method to compute an upper $1-\alpha$ confidence bound E_α^+ for E from a PPEB sample. One general strategy for risk-limiting audits, described in [81, 79, 80], is to test the hypothesis that the outcome is wrong sequentially: The auditor draws a sample, then assesses whether there is sufficiently strong evidence that the outcome is correct. If

there is, the audit stops. If there is not, the audit sample is enlarged and the new evidence is assessed. Eventually, either there is strong evidence that the outcome is right, or there will have been a full hand count.

Stage s of a sequential audit can be viewed as a test at significance level α_s. In this chapter, we focus on a single stage: The hypothesis that the outcome is wrong is rejected at significance level α_s if $E^+_{\alpha_s} < 1$. That might be the only stage of an audit that takes a sample then either stops or conducts a full hand count, or it might be one of the stages of a multi-stage audit that could expand the sample once or more before demanding a full hand count.

The two audits we conducted using the new method were single-stage audits. We drew an initial sample of n batches and calculated an upper 75% confidence bound for E from the errors the hand counts uncovered in those batches. If that upper confidence bound, $E^+_{0.25}$, had been greater than 1, the elections officials would have conducted complete hand counts.

4.3 The Trinomial Confidence Bound

Our method for constructing a $1-\alpha$ upper confidence bound E^+_α for E is similar to the multinomial bound with clustering [20, 58].

The *taint* t_p of batch p is the ratio of the actual overstatement in batch p to the maximum overstatement in batch p:

$$t_p \equiv \frac{e_p}{u_p} \le 1. \tag{4.9}$$

Now,

$$E = \sum_p e_p = \sum_p \frac{e_p}{u_p} u_p = \sum_p t_p u_p. \tag{4.10}$$

Suppose we draw a PPEB sample of size n. Let T_j denote the taint of the jth draw. Then the expected value of T_j is

$$\mathbb{E}[T]_j = \sum_p t_p u_p / U = E/U. \tag{4.11}$$

Multiplication by U transforms an upper $1-\alpha$ confidence bound for $\mathbb{E}[T]_j$ into an upper $1-\alpha$ confidence bound for E. See also [83].

Let $d \in (0,1)$. Define

$$Y_j \equiv \begin{cases} 0, & T_j \le 0 \\ d, & 0 < T_j \le d \\ 1, & T_j > d. \end{cases} \tag{4.12}$$

For any $d \in (0,1)$,[8] Y_j is stochastically larger than T_j (i.e., $\Pr[Y_j \ge T_j] = 1$), so

$$\mathbb{E}[T]_j \le \mathbb{E}[Y]_j. \tag{4.13}$$

[8]Some papers on the multinomial bound in financial auditing suggest that d can be chosen after the data are collected. We have seen no proof that post hoc selection of d results in a valid confidence bound. We select d before the data are collected.

Let

$$\begin{aligned} \pi_0 &\equiv \Pr[Y_j = 0], \\ \pi_d &\equiv \Pr[Y_j = d], \text{ and} \\ \pi_1 &\equiv \Pr[Y_j = 1]; \end{aligned}$$

and let $\pi \equiv (\pi_0, \pi_d, \pi_1)$. Define $\mu = \mu(d) \equiv (0, d, 1)$. Then

$$\mathbb{E}[Y]_j = 0\pi_0 + d\pi_d + 1\pi_1 = \mu \cdot \pi. \tag{4.14}$$

Define

$$Z \equiv (\#\{j : Y_j = 0\}, \#\{j : Y_j = d\}, \#\{j : Y_j = 1\}). \tag{4.15}$$

This is a random 3-vector. Its first component is the number of observed taints that are no bigger than zero; its second is the number of observed taints that are strictly positive but no bigger than d; and its third is the number of observed taints that exceed d. It has a trinomial distribution with category probabilities π.

We will use Z to find a set $S_\alpha(Z)$ such that

$$\Pr_\pi[S_\alpha(Z) \ni \pi] \geq 1 - \alpha. \tag{4.16}$$

That is, $S_\alpha(Z)$ is a $1 - \alpha$ confidence set for π. Then

$$t_\alpha^+ \equiv \max_{\gamma \in S_\alpha(Z)} \mu \cdot \gamma \tag{4.17}$$

is the upper endpoint of a $1 - \alpha$ upper confidence interval for $\mathbb{E}[Y]_j$ and hence for $\mathbb{E}[T]_j$. It follows that Ut_α^+ is the upper endpoint of a $1 - \alpha$ upper confidence interval for E.

We construct $S_\alpha(Z)$ by inverting hypothesis tests about π. We are ultimately interested in inferring that $\mu \cdot \pi$ is not large, so it makes sense to reject the hypothesis $\pi = \gamma$ when

$$\mu \cdot Z \leq z_\gamma, \tag{4.18}$$

with

$$z_\gamma = z_\gamma(\alpha) \equiv \max \left\{ z : \Pr_\gamma[\mu \cdot Z \leq z] \leq \alpha \right\}, \tag{4.19}$$

so that the test has level α.

The test statistic $\mu \cdot Z$ orders the possible values of Z by the sample mean of the values of Y_j from which Z was constructed.[9] To find a confidence bound for $\mathbb{E}[T]_j$, we invert the

[9]This test statistic generally results in a different test from the "step-down set" acceptance region used by [20, 58].

hypothesis tests to find the confidence set $S_\alpha(z)$ of trinomial category probabilities γ for which the hypothesis $\pi = \gamma$ would not be rejected if we observed $Z = z$. That set is

$$\begin{aligned} S_\alpha(z) &\equiv \{\gamma = (\gamma_0, \gamma_d, \gamma_1) \in \Re^3 : \gamma \geq 0, \gamma_0 + \gamma_d + \gamma_1 = 1, \\ &\qquad \text{and } \mu \cdot z > z_\gamma\} \\ &= \{\gamma = (\gamma_0, \gamma_d, \gamma_1) \in \Re^3 : \gamma \geq 0, \gamma_0 + \gamma_d + \gamma_1 = 1, \\ &\qquad \text{and } \Pr_\gamma[\mu \cdot Z \leq \mu \cdot z] > \alpha\}. \end{aligned} \tag{4.20}$$

The corresponding confidence bound for $\mathbb{E}[T]_j$ is the largest value of $\mu \cdot \gamma$ over $\gamma \in S_\alpha(z)$:

$$t_\alpha^+ = t_\alpha^+(z) = \max_{\gamma:\, \Pr_\gamma[\mu \cdot Z \leq \mu \cdot z] > \alpha} \mu \cdot \gamma. \tag{4.21}$$

We now characterize the solution to the optimization problem (4.21) in some useful cases. If $\mu \cdot Z/n \geq 1/U$, we certainly will not be able to conclude that $E < 1$. The question is how much smaller than $1/U$ the "sample mean" $\mu \cdot Z/n$ must be to provide strong evidence that $E < 1$. Because $\alpha < 1$ by assumption,

$$\Pr_{(0,0,1)}[\mu \cdot Z \leq \mu \cdot z] < \alpha \tag{4.22}$$

unless $z = (0, 0, n)$. If $z = (0, 0, n)$, $t_\alpha^+ = 1$.

If $z \neq (0, 0, n)$, then the maximum in (4.21) is attained for some γ for which $\Pr_\gamma[\mu \cdot Z \leq \mu \cdot z] = \alpha$.[10] Suppose no observed taints are greater than d and $k < 1/d$ taints are strictly positive. Then $z = (n - k, k, 0)$ and

$$\begin{aligned} \Pr_\gamma[\mu \cdot Z \leq \mu \cdot z] &= \sum_{j=0}^{k} \Pr_\gamma[Z = (n - j, j, 0)] \\ &= \sum_{j=0}^{k} \binom{n}{j} \gamma_0^{n-j} \gamma_d^j. \end{aligned} \tag{4.23}$$

Hence,

$$\begin{aligned} t_\alpha^+ &= 1 + \max_{\gamma_0, \gamma_d} \{(d-1)\gamma_d - \gamma_0 : \gamma_0, \gamma_d \geq 0,\ \gamma_0 + \gamma_d \leq 1, \\ &\qquad \text{and } \sum_{j=0}^{k} \binom{n}{j} \gamma_0^{n-j} \gamma_d^j = \alpha\}. \end{aligned} \tag{4.24}$$

The two-dimensional optimization problem (4.24) can be solved using an ascent method or by searching. The R package "elec," available through CRAN (`http://cran.r-project.org`), implements the computation.

[10] To see this, note (i) that $\mu \cdot \gamma$ increases continuously and monotonically as mass is moved either from γ_0 to γ_1 or from γ_d to γ_1, and (ii) that $\Pr_\gamma[\mu \cdot Z \leq \mu \cdot z]$ decreases monotonically and continuously as mass is moved either from γ_0 to γ_1 or from γ_d to γ_1. Suppose the maximum in 4.21 were attained for some $\delta = (\delta_0, \delta_d, \delta_1)$ with $\Pr_\delta[\mu \cdot Z \leq \mu \cdot z] > \alpha$. By assumption, $\delta \neq (0, 0, 1)$. Hence, either $\delta_0 > 0$ or $\delta_d > 0$. Moving an infinitesimal amount mass from either of those components to δ_1 increases $\mu \cdot \delta$ and decreases $\Pr_\delta[\mu \cdot Z \leq \mu \cdot z]$. Hence, δ cannot be optimal.

Selecting n and d

No matter what values we select for n and d, the upper confidence bound for E will be conservative. However, if we choose n very small or d very large, the audit will not be able to provide strong evidence that $E < 1$, even when the outcome of the election is correct. The confidence bound E_α^+ will be greater than 1, and the audit will progress—either to the next stage or to a full hand count. On the other hand, setting n large entails a lot of auditing in the first stage, perhaps more than necessary to confirm the outcome when the outcome is in fact correct.

We select d and n iteratively, using simulation to estimate the power of the test against a "realistic" alternative hypothesis under which there is error, but not enough error to alter the outcome of the contest. In the alternative, the error is randomly distributed. Batches are tainted with probability τ, independently. If batch p is tainted, it has an overstatement of (up to) η votes, and the error is $\min\{\eta/V_{w\ell}, u_p\}$. The amount of taint that the η votes represents thus depends on the batch. For batches with small u_p, an overstatement of η votes is a large taint, while for batches with large u_p it is a small taint. Because the chance of drawing batch p is smaller for batches with small u_p, it is less likely that the sample will include the larger taints.

We adjust d and n iteratively until the chance is approximately $1 - \beta$ that the $1 - \alpha$ trinomial confidence bound for E is less than one. The chance is estimated by simulation. The confidence level is always at least $1 - \alpha$. Adjusting n and d only affects the power.

In the simulations to select d and n for the Marin and Santa Cruz County audits, which were conducted at level $\alpha = 0.25$, we used $\tau = 0.05$, $\eta = 10$ votes and $1 - \beta = 0.9$. These choices resulted in using $d = 0.047$, $n = 19$ for Santa Cruz and $d = 0.038$, $n = 14$ for Marin.

4.4 November 2008 Audits in Marin and Santa Cruz Counties

In November 2008 we audited races in Marin and Santa Cruz counties, CA, using the trinomial bound,[11] as follows: The elections officials provided us the semi-official results $\{v_{kp}\}$ and the number of ballots cast in each batch, which we took as $\{b_p\}$. From $\{v_{kp}\}$ and $\{b_p\}$ we calculated $\{u_p\}$ and U. We selected the number of draws n as described in section 4.3.

The elections officials rolled dice to generate 6-digit seeds which they sent to us.[12] We used the seeds in the R implementation of the Mersenne Twister algorithm to make n PPEB draws to select batches for audit. The batches selected were counted by hand by members of the staffs of the Santa Cruz County Clerk's office and the Marin County Registrar of Voters office. They reported the hand-count results to us. We calculated confidence bounds for E from the observed discrepancies and U using the trinomial bound. In both cases, the 75% upper confidence bounds were less than 1, so no further counting was required.

[11] We audited a race in Yolo County, CA, using a different method.

[12] The Santa Cruz seed was 541,227; the Marin seed was 568,964.

Section 4.4 describes the Santa Cruz County audit in some detail. Section 4.4 summarizes the Marin County audit.

Santa Cruz County Supervisor, 1st District

There were 152 batches containing 0 to 855 ballots (median 66). The maximum potential error per batch ranged from $u_p = 0\%$ to 49% of the margin: Some individual batches could hide enough error to account for nearly half the margin. The distribution of the $\{u_p\}$ was heavily skewed to the right. The total possible margin overstatement across all batches was $U = 13.46$.

As described in section 4.3, we used $d = 0.047$ and $n = 19$ in this audit. Since the draws are independent, they need not yield distinct batches. The expected number of distinct batches in 19 PPEB draws is

$$\sum_p \left(1 - \left(1 - \frac{u_p}{U}\right)^n\right) = 16.3 \tag{4.25}$$

and the expected number of ballots in the sample is

$$\sum_p b_p \left(1 - \left(1 - \frac{u_p}{U}\right)^n\right) = 7,214. \tag{4.26}$$

A simple random sample would have required a much larger audit to control the risk to the same level.[13] The 19 draws produced 16 distinct batches containing 7,105 ballots in all. Even with PPEB, a high proportion of ballots needed to be audited, which is typical for small races. The sample size needed to control the risk does not depend directly on the size of the race. Wide variations in the error bounds $\{u_p\}$ also contribute to the need for a larger sample. Table 4.2 gives the audit results.

While analyzing the data we learned that, although the audit data included provisional ballots, the original totals on which we had based the audit did not.[14] This increased the number of ballots in several audited batches and changed the margins in some of them. The audit also showed a difference of one in the number of ballots in some VBM batches. We attribute that difference to ballots that needed special treatment. To ensure that the audit remained statistically conservative, we treated every change to the reported margins—including changes produced by provisional ballots—as error in the reported counts, i.e., as error uncovered by the audit.[15] The change in b_p, the number of ballots in a batch, affects u_p.

[13] For example, the method in [81, 80] would have required a simple random sample of $n = 38$ batches, with the expectation of counting 13,017 ballots, on the order of twice the effort required by the trinomial bound with PPEB sampling.

[14] Apparently 806 provisional ballots had been cast in the race in all. Among the audited batches, precinct 1005 had 37; 1007 had 30; 1019 had 32; 1060 had 11; and 1101 had 39.

[15] It would also have been conservative to treat all the provisional ballots as error, but we had no way to separate the votes for the provisional and original ballots, so it was impossible to isolate the error in the original counts.

			Leopold		Danner				
Batch ID	b_p	u_p	Reported	Actual	Reported	Actual	MOV	t_p	Times
1002 VBM	573	0.28	251	252	227	227	-1	-0.002	1
1005 PCT	556	0.32	292	304	166	170	-8	-0.012	1
1005 VBM	436	0.23	208	208	150	150	0	0	1
1007 PCT	692	0.40	367	382	205	216	-4	-0.005	1
1007 VBM	630	0.33	311	311	240	240	0	0	1
1013 VBM	557	0.28	261	261	216	216	0	0	2
1017 VBM	399	0.21	191	191	139	139	0	0	1
1019 PCT	448	0.25	218	223	128	137	4	0.007	1
1019 VBM	378	0.20	186	186	128	128	0	0	1
1027 VBM	232	0.11	107	107	98	98	0	0	1
1028 VBM	365	0.15	136	137	174	174	-1	-0.003	1
1037 VBM	758	0.33	261	261	309	309	0	0	2
1053 VBM	18	0.01	10	10	4	4	0	0	1
1060 PCT	322	0.17	142	145	105	108	0	0	2
1073 VBM	20	0.01	11	11	3	4	1	0.036	1
1101 PCT	721	0.35	312	321	275	279	-5	-0.007	1

Table 4.2: Santa Cruz Audit Data. The major contestants in the contest for Supervisor, 1st District, were Leopold and Danner. Sixteen batches were sampled, three of them twice. The number of ballots initially reported for the batch is b_p. The upper bound on the taint in batch p is u_p. In each PPEB draw, the probability of selecting batch p is proportional to u_p. MOV is number of votes by which error increased the apparent margin for Danner. The taint t_p is the observed overstatement of the margin in the batch divided by the maximum possible overstatement of the margin in the batch. "Times" is the number of times the batch was selected in 19 PPEB draws. Two positive taints were found, both less than $d = 0.047$.

If u_p is still an upper bound on e_p, the audit remains valid. Since the bound u_p is extremely conservative (calculated by assuming that *all* the votes in batch p are actually for the loser) and there are so few provisional ballots in all, it is implausible that $e_p > b_p$ in any batch.

The largest observed taint, 0.036, was a 1-vote overstatement in a tiny precinct. The largest absolute overstatement, 4 votes, was in a much larger precinct; that taint was only 0.007. "Error" was as large as 8 votes in some batches, an atypically high rate for voter-marked optically scanned ballots. As far as we can tell, this discrepancy was due to miscommunication, not an error in the counts per se. This experience underscores the importance of clear communication among the auditors and elections officials and their staff.

Apparently, the majority of the provisional ballots in the sample were for the winner, so including them among the ballots in the audited batches only strengthened the evidence that the outcome was right. Despite treating changes caused by including provisional ballots

as errors, only two batches had margin overstatements, both less than $d = 0.047$. (If any of the three batches that were drawn twice had positive taint, the taint of that batch would count twice.)

The trinomial observation was thus $z = (17, 2, 0)$. The calculation of the trinomial confidence bound is illustrated in Figure 4.1. The upper confidence bound for $\mathbb{E}[T]_j$ is $t^+_{0.25} = 0.072$, which yields the upper confidence bound

$$E^+_{0.25} = Ut^+_{0.25} = 13.46 \times 0.072 = 0.97 < 1. \tag{4.27}$$

This allowed us to reject the hypothesis that the outcome was wrong and stop the audit without a full manual count.[16]

Marin County Measure B

Table 4.1 summarizes the results of the race. In Marin, "decks" of VBM ballots are run through the scanner as a group. Decks usually contain about 250 ballots, sometimes from several precincts. To collect all the ballots for a single precinct could require sorting through several decks of ballots. This is laborious and prone to error; for a race as large as Measure B, the effort is prohibitive. For this reason, we used the decks as batches.

There was a complication. While the total number of ballots b_p in each deck is known, the number of votes for each candidate or position is not. (The vote tabulation software would not generate such subtotals without extensive hand editing.) To calculate a rigorous upper bound u_p for decks, we made extremely conservative assumptions: $v_{wp} = b_p$ but $a_{\ell p} = b_p$. That is, to find an upper bound on the margin overstatement in batch p we assumed that every ballot was reported as cast for the apparent winner, but that in reality every ballot was cast for the reported loser. That leads to the bound $u_p = 2b_p$. While this is extremely conservative, the resulting sample size was still manageable. The sample size n was larger than it would have been had we known v_{wp} and $v_{\ell p}$, but that was balanced by the labor saved in not having to generate vote totals for the decks manually.[17] The bound would have been effectively much more conservative if only a subset of the ballots in a deck included Measure B, but Measure B was county-wide.

There were 544 batches in all—189 batches of ballots cast in precinct and 355 decks. Using (small) decks as batches reduces the expected workload because the more batches there are, the smaller the size of each. The number of draws required does not depend directly on the number of batches in the population, so dividing the ballots into many small batches usually leads to less counting than dividing the population into fewer large batches.

The total error bound was $U = 9.78$. The distribution of error bounds was roughly bell-shaped, with a spike at 0.025 because many decks were about the same size (roughly

[16] On the basis of the trinomial bound, the P-value of the hypothesis that the outcome is wrong is 0.24.

[17] If the vote tabulation software had been able to report v_{wp} and $v_{\ell p}$ for each deck, we would not have had to use such a conservative bound. Data export from vote tabulation systems is a serious bottleneck for election auditing.

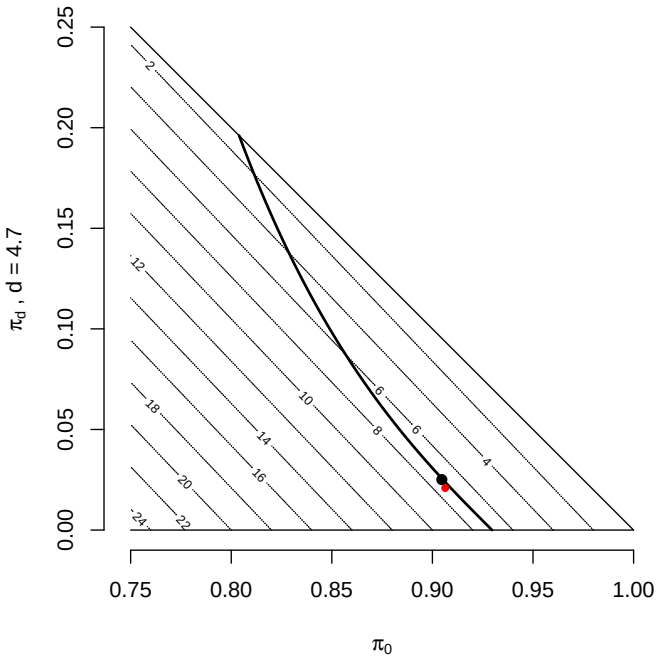

Figure 4.1: The optimization problem over trinomial category probabilities for the Santa Cruz audit.

The heavy line is the set $\{\gamma : \Pr_\gamma[\mu \cdot Z \leq \mu \cdot z] = \alpha = 0.25\}$. The parallel lines are the contours of $100 \times \mu \cdot \gamma$. The points to the right of the heavy line comprise the confidence set. The heavy dot is the category probability vector with the largest value of $\mu \cdot \gamma$ among parameters in the confidence set. For this contest, $U = 13.46$, so the audit can stop if the confidence set excludes $1/13.46 \approx 0.074$, corresponding to a contour line at 7.4 in the units of this figure.

250 ballots each). In this election no batch could hold error of more than 3% of the margin. In contrast, in the Santa Cruz race some batches could hold errors of up to 48% of the margin.

As described in section 4.3, we chose $d = 0.038$ and $n = 14$ draws, which were expected to yield 13.8 distinct batches and 3,424 ballots. The expected number of batches is close to the number of draws because the error bounds u_p are reasonably uniform and no u_p is very large, in contrast to the bounds in Santa Cruz. With simple random sampling, the audit would have required roughly 22 batches to control the risk to the same level ($\alpha = 0.25$). The expected number of ballots to audit would have been about 4,900, 44% more than with PPEB and the trinomial bound.

Batch ID	b_p	Yes	No	u_p
D-31	91	50	33	0.009
D-43	108	59	40	0.011
D-104	40	16	16	0.004
D-191	217	137	57	0.022
D-255	246	156	67	0.025
D-286	258	144	88	0.026
D-301	245	129	88	0.025
D-339	248	134	80	0.025
IP-1002	316	151	110	0.018
IP-1017	362	186	133	0.021
IP-3013	277	125	102	0.015
IP-3014	498	256	152	0.030
IP-3017	318	154	111	0.018
IP-3020	123	64	39	0.007

Table 4.3: Marin Audit Results. Reported votes in the audit sample in Marin. Fourteen batches were selected using PPEB from 355 decks of vote-by-mail ballots (the 8 batch IDs beginning with "D") and 189 batches of ballots cast in precincts (the 6 batch IDs beginning with "IP"). b_p is the total number of ballots in the batch. Yes and No are the votes for and against the measure. u_p is the bound on error in that batch, given b_p and the reported totals. The audit found no errors.

Once the decks to audit were selected, subtotals for those decks were produced, in order to have semi-official figures to audit. This involved replicating the data base and generating a special report for each audited precinct by manually deleting every batch but one and generating a report for the remaining batch, an arduous and error-prone procedure. Those subtotals were then audited by hand-counting paper ballots. Table 4.3 lists the reported votes in the 14 batches in the sample, which included 3347 ballots. Remarkably, the audit found no errors. The vector of trinomial counts was thus $z = (14, 0, 0)$. The 75% confidence bound for taint was $t^+_{0.25} = 0.094$, and the 75% confidence bound for E was

$$E^+_{0.25} = 0.0943 \times 9.78 = 0.922 < 1, \tag{4.28}$$

so the audit stopped without a full hand count. The corresponding P-value was about 0.22.

Late problems in Marin County

We discovered in late July 2009, long after the end of the canvass period, that while Marin County had not found any discrepancies in any audited batches, the totals they audited were not identical to the totals on which we had based the audit calculations. In Marin County, voters in precincts with fewer than 250 registered voters are required to vote by mail, and

VBM ballots are reported as if they were IP ballots. For larger precincts, the IP results were final by 7 November, but for precincts with fewer than 250 registered voters, the "nominal" IP results were not final until 14 November: It takes longer for the VBM ballots to be sorted and tallied.[18] We based our audit calculations on the IP results in the 7 November statement of vote, understanding—incorrectly—that those were final. They were final for larger precincts, but not for VBM-only precincts. Marin County audited the 14 November statement of vote. Again, this emphasizes the importance of clear communication between auditors and elections officials, and shows the value of pilot studies.

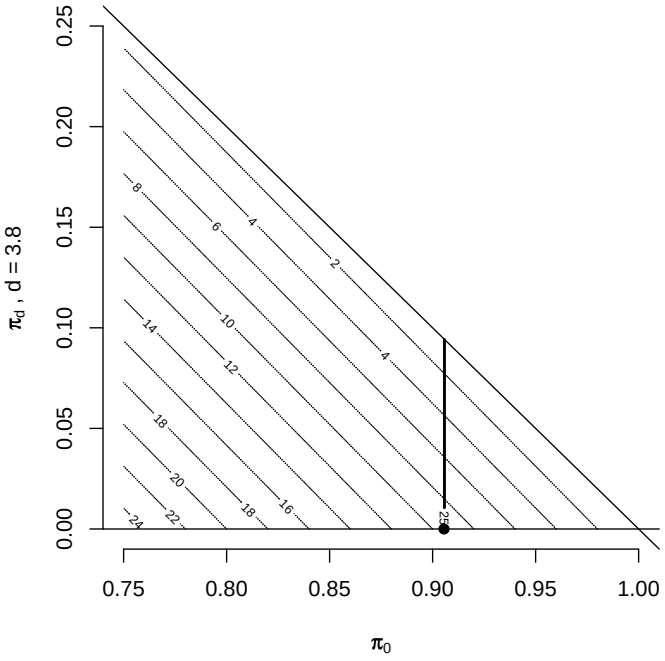

Figure 4.2: The optimization problem over trinomial category probabilities for the Marin audit. The heavy line is the set $\{\gamma : \Pr_\gamma[\mu \cdot Z \leq \mu \cdot z] = \alpha = 0.25\}$. The parallel lines are contours of $100 \times \mu \cdot \gamma$. The confidence set consists of the points to the right of the heavy line. The point is the category probability vector in the confidence set with the largest value of $\mu \cdot \gamma$. Because no errors were found, the maximum lies on the boundary. For this contest, $U = 9.78$, so the audit can stop if the confidence set excludes $1/9.78 \approx 0.102$, corresponding to a contour line at 10.2 in the units of this figure.

[18]The VBM ballots for VBM-only precincts get special treatment: They are segregated from the other VBM ballots and sorted by precinct.

4.5 Comparison with the Stringer Bound

The Stringer bound [84] has long been used in financial auditing to find an upper confidence bound on the overstatement of a group of accounts using a PPEB sample. It is generally—though not always—quite conservative, more so than the multinomial bound [62]. If there are M non-zero taints, $t_1 > \cdots > t_M$, the Stringer bound is

$$t^+_{S,\alpha} \equiv \pi^+_\alpha(0) + \sum_{j=1}^{M} \left[\pi^+_\alpha(j) - \pi^+_\alpha(j-1)\right] t_j, \tag{4.29}$$

where $\pi^+_\alpha(k)$ is the exact $1-\alpha$ upper confidence bound for π from datum $X \sim \text{Bin}(n, \pi)$ when the observed value of X is k.

Table 4.4 compares the 75% upper confidence bound for E based on the Stringer bound and the trinomial bound for the Santa Cruz and Marin audit data. For the Santa Cruz data, the Stringer bound is larger but still below 1, so it would have permitted the audit to stop. When all the taints are non-positive, as they are for the Marin data, the Stringer bound equals the trinomial bound. The Kaplan-Markov bound [83] can be sharper, especially if there are negative taints.

County	n	positive taints	Stringer	Trinomial
Santa Cruz	19	0.036, 0.007	0.984	0.956
Marin	14	none	0.922	0.922

Table 4.4: 75% upper confidence bounds for E. The Stringer bound is larger than the trinomial bound for the Santa Cruz race, but both are below one: The audit could stop if either method were used. The trinomial and Stringer bounds are equal and less than one for the Marin race.

4.6 Conclusion

We used a novel method to audit two November 2008 contests in California, one in Santa Cruz County and one in Marin County. The audits were conducted in a way that guaranteed at least a 75% chance of a full hand count if the outcome of the contest were wrong. Neither audit resulted in a full hand count.

The method we used, the trinomial bound, constructs an upper confidence bound for the total overstatement error E in the race. For the apparent outcome of the race to be wrong, it is necessary that $E \geq 1$. Hence, if the confidence bound for E is less than 1, the audit can stop. If the confidence bound is 1 or greater, there is a full manual count. This results in a risk-limiting audit, i.e., an audit with a guaranteed minimum chance of a full manual count whenever the apparent outcome is wrong.

The trinomial bound relies on a sample drawn with probability proportional to a bound on the overstatement error in each batch of ballots (PPEB sampling), a technique long used in financial auditing, but new to election auditing [2]. There are other ways of using PPEB samples to draw inferences about E [83, 82]. The trinomial bound constructs a confidence set for the category probabilities for a trinomial variable from the *taints* observed in the PPEB sample, then projects and scales that confidence set to find a confidence bound for E.

The audit in Marin county posed unusual logistic challenges because ballots were not sorted by precinct. We used batches defined by "decks" of ballots that were fed through scanners as a group. The inability of the vote tabulation software to produce batch subtotals made it necessary then to use extremely conservative bounds on the possible error in each batch: twice the number of ballots.

Election audits face considerable logistic challenges. The time and effort of counting votes by hand is one. The lack of good "data plumbing" is another. Current vote tabulation systems do not seem to export data in formats that are convenient for audits, necessitating hours of error-prone hand editing. Elections officials and legislators interested in promoting post-election audits could help by demanding this functionality. Embracing standard data formats would also help considerably.

Chapter 5

Appendix A: Proofs and Examples for Post-Stratification

Conditioning on $\mathcal{D}$ Maintains Assignment Symmetry

Assume the original randomization is Assignment Symmetric. The event $\mathcal{D}$ of $\hat{\tau}_{ps}$ being defined is a function of W, the vector of number of treated units in the strata:

$$\mathbf{1}_{\mathcal{D}} = f(W) \equiv \prod_{k=1}^{K} \mathbf{1}_{\{W_k>0\}} \mathbf{1}_{\{W_k<n_k\}}$$

Treatment assignment pattern T_k is independent of pattern T_i given W, so since $\mathcal{D}$ is a function of W, T_k is independent of T_i given $W, \mathcal{D}$: conditioning on $\mathcal{D}$ maintains independence of treatment assignment patterns.

Now let Ω_w be the space of possible values of W and consider two assignment patterns s and t in stratum k. We have

$$\mathbf{P}\{T_k = s|W = w\} = \mathbf{P}\{T_k = s|W_k = w_k\} = \mathbf{P}\{T_k = t|W_k = w_k\} = \mathbf{P}\{T_k = t|W = w\}$$

due to the unconditioned Assignment Symmetry. Then

$$\begin{aligned} \mathbf{P}\{T_k = s|W_k = \ell, \mathcal{D}\} &= \frac{1}{Z} \sum_{w\in\Omega_W} \mathbf{P}\{T_k = s|W = w\}\, \mathbf{1}_{\{w_k=\ell\}} \mathbf{1}_{\{f(w)=1\}} \mathbf{P}\{W = w\} \\ &= \mathbf{P}\{T_k = t|W_k = \ell, \mathcal{D}\} \end{aligned}$$

with $Z = \sum \mathbf{1}_{\{w_k=\ell\}} \mathbf{1}_{\{f(w)=1\}} \mathbf{P}\{W = w\}$. Therefore. conditioning on $\mathcal{D}$ maintains equiprobable treatment assignment patterns.

5.1 Derivation of Theorem 2.2.1, the Variance Formula

Under Assignment Symmetry the chance of any given unit being treated is $W_k(1)/n_k$ so

$$\mathbb{E}[T_i|W_k(1)] = \frac{W_k(1)}{n_k}$$

for unit i in stratum k. Then

$$\mathbb{E}\left[\frac{T_i}{W_k(1)}\right] = \mathbb{E}\,\mathbb{E}\left[\frac{T_i}{W_k(1)}|W_k(1)\right] = \mathbb{E}\left[\frac{1}{n_k}\right] = \frac{1}{n_k}.$$

Rearrange $\beta_{1k} \equiv \mathbb{E}[W_k(0)/W_k(1)] = n_k\,\mathbb{E}[1/W_k(1)] - 1$ to get $\mathbb{E}[1/W_k(1)] = (\beta_{1k}+1)/n_k$ and

$$\mathbb{E}\left[\frac{T_i^2}{W_k^2(1)}\right] = \mathbb{E}\,\mathbb{E}\left[\frac{T_i}{W_k^2(1)}|W_k(1)\right] = \frac{1}{n_k}\,\mathbb{E}\left[\frac{1}{W_k(1)}\right] = \frac{\beta_{1k}+1}{n_k^2}.$$

These derivations are easier if we use $\alpha_{1k} \equiv \mathbb{E}[1/W_k(1)]$, but the β's are more interpretable and lead to nicer final formula. To continue, Assignment Symmetry gives

$$\begin{aligned}
\mathbb{E}[T_iT_j|W_k(1)=w] &= \mathbf{P}\{T_i=1 \wedge T_i=1|W_k(1)=w\} \\
&= \frac{\binom{n_k-2}{w-2}}{\binom{n_k}{w}} = \frac{(n_k-2)!}{(w-2)!(n_k-w)!}\cdot\frac{w!(n_k-w)!}{n_k!} \\
&= \frac{w(w-1)}{n_k(n_k-1)}
\end{aligned}$$

so

$$\mathbb{E}\left[\frac{T_iT_i}{W_k^2(1)}\right] = \mathbb{E}\left[\frac{W_k(1)(W_k(1)-1)}{W_k^2(1)}\right]\cdot\frac{1}{n_k(n_k-1)} = \frac{-\beta_{1k}+n-1}{n_k^2(n_k-1)}.$$

There are analogous formula for the control unit terms. Similarly,

$$\mathbb{E}\left[\frac{T_i(1-T_i)}{W_k(1)W_k(0)}\right] = \mathbb{E}\left[\frac{W_k(1)(n_k-W_k(1))}{W_k(1)W_k(0)}\right]\cdot\frac{1}{n_k(n_k-1)} = \frac{1}{n_k(n_k-1)}.$$

We use these relationships to compute means and variances for the strata-level estimators.

Unbiasedness. The strata-level estimators are unbiased:

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_k] &= \mathbb{E}\left[\sum_{i:b_i=k}\frac{T_i}{W_k(1)}y_i(1) - \sum_{i:b_i=k}\frac{1-T_i}{W_k(0)}y_i(0)\right] \\
&= \sum_{i:b_i=k}\mathbb{E}\left[\frac{T_i}{W_k(1)}\right]y_i(1) - \sum_{i:b_i=k}\mathbb{E}\left[\frac{1-T_i}{W_k(0)}\right]y_i(0) \\
&= \sum_{i:b_i=k}\frac{1}{n_k}y_i(1) - \sum_{i:b_i=k}\frac{1}{n_k}y_i(0) = \tau_k.
\end{aligned}$$

Variance. $\text{Var}[\hat{\tau}_k] = \mathbb{E}[\hat{\tau}_k^2] - \tau^2$. $\mathbb{E}[\hat{\tau}_k^2]$ breaks down into three big terms:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_k^2] &= \underbrace{\mathbb{E}\left[\sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1)\right]^2}_{(a)} \\ &- \underbrace{2\,\mathbb{E}\left[\left(\sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1)\right)\left(\sum_{i:b_i=k} \frac{1-T_i}{W_k(0)} y_i(0)\right)\right]}_{(b)} + \underbrace{\mathbb{E}\left[\sum_{i:b_i=k} \frac{1-T_i}{W_k(0)} y_i(0)\right]^2}_{(c)}.\end{aligned}$$

Simplify the three parts of the above. For part (a):

$$\begin{aligned}(a) &= \mathbb{E}\left[\sum_{i:b_i=k} \frac{T_i^2}{n_k^2(1)} y_i^2(1) + \sum_{i\neq j} \frac{T_i T_i}{n_k^2(1)} y_i(1) y_i(1)\right] \\ &= \sum_{i:b_i=k} \mathbb{E}\left[\frac{T_i^2}{n_k^2(1)}\right] y_i^2(1) + \sum_{i\neq j} \mathbb{E}\left[\frac{T_i T_i}{n_k^2(1)}\right] y_i(1) y_i(1) \\ &= \frac{\beta_{1k}+1}{n_k^2} \sum_{i:b_i=k} y_i^2(1) + \frac{-\beta_{1k}+n_k-1}{n_k^2(n_k-1)} \sum_{i\neq j} y_i(1) y_i(1).\end{aligned}$$

Part (c) is similar. The cross-terms are:

$$\begin{aligned}(b) &= 2\,\mathbb{E}\left[\sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1) \frac{1-T_i}{W_k(0)} y_i(0)\right] + 2\,\mathbb{E}\left[\sum_{i\neq j} \frac{T_i}{W_k(1)} y_i(1) \frac{1-T_i}{W_k(0)} y_i(0)\right] \\ &= 0 + 2\sum_{i\neq j} \mathbb{E}\left[\frac{T_i}{W_k(1)} \frac{1-T_i}{W_k(0)}\right] y_i(1) y_i(0) \\ &= \frac{2}{n_k(n_k-1)} \sum_{i\neq j} y_i(1) y_i(0).\end{aligned}$$

The first term vanishes since $T_i(1-T_i) = 0$ always.

These are the three parts of the expectation of the square. We have related components in τ_k^2 when you expand the square:

$$\tau_k^2 = \underbrace{\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(1)\right)^2}_{(a')} - \underbrace{2\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(1)\right)\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(0)\right)}_{(b')} + \underbrace{\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(0)\right)^2}_{(c')}.$$

The variance is $\text{Var}[\hat{\tau}_k] = (a)-(a')-(b)+(b')+(c)-(c')$, a sum of several ugly differences. Expanding (a') and plugging in gives the first difference:

$$\begin{aligned}
(a)-(a') &= \frac{\beta_{1k}+1}{n_k^2}\sum_{i:b_i=k} y_i^2(1) + \frac{-\beta_{1k}+n_k-1}{n_k^2(n_k-1)}\sum_{i\neq j} y_i(1)y_i(1) \\
&\quad -\frac{1}{n_k^2}\sum_{i:b_i=k} y_i^2(1) - \frac{1}{n_k^2}\sum_{i\neq j} y_i(1)y_i(1) \\
&= \left(\frac{\beta_{1k}+1}{n_k} - \frac{1}{n_k^2}\right)\sum_{i:b_i=k} y_i^2(1) + \left(\frac{-\beta_{1k}+n_k-1}{n_k^2(n_k-1)} - \frac{1}{n_k^2}\right)\sum_{i\neq j} y_i(1)y_i(1) \\
&= \frac{\beta_{1k}}{n_k}\left[\frac{1}{n_k}\sum_{i:b_i=k} y_i^2(1) - \frac{1}{n_k(n_k-1)}\sum_{i\neq j} y_i(1)y_i(1)\right] \\
&= \frac{\beta_{1k}}{n_k}\sigma_k^2(1).
\end{aligned}$$

$(c)-(c')$ is similar. The cross terms are:

$$\begin{aligned}
(b)-(b') &= \frac{2}{n_k(n_k-1)}\sum_{i\neq j} y_i(1)y_i(0) - \frac{2}{n_k^2}\sum_{i:b_i=k} y_i(1)y_i(0) - \frac{2}{n_k^2}\sum_{i\neq j} y_i(1)y_i(0) \\
&= \left(\frac{2}{n_k(n_k-1)} - \frac{2}{n_k^2}\right)\sum_{i\neq j} y_i(1)y_i(0) - \frac{2}{n_k^2}\sum_{i:b_i=k} y_i(1)y_i(0) \\
&= -\frac{2}{n_k}\left[\frac{1}{n_k}\sum_{i:b_i=k} y_i(1)y_i(0) - \frac{1}{n_k(n_k-1)}\sum_{i\neq j} y_i(1)y_i(0)\right] \\
&= -\frac{2}{n_k}\gamma_k(0,1).
\end{aligned}$$

Sum the above to get Equation 2.5.

Theorem 2.2.2 The mean is immediate. For the variance, observe:

$$\begin{aligned}
\text{Var}[\hat{\tau}_{ps}] &= \mathbb{E}\left[\sum_{k=1}^{K}\frac{n_k}{n}\left(\hat{\tau}_k-\tau_k\right)^2\right] \\
&= \sum_{k=1}^{K}\left(\frac{n_k}{n}\right)^2\mathbb{E}\left[\left(\hat{\tau}_k-\tau_k\right)^2\right] + \sum_{k\neq r}\frac{n_k n_r}{n^2}\mathbb{E}\left[\left(\hat{\tau}_k-\tau_k\right)\left(\hat{\tau}_r-\tau_r\right)\right].
\end{aligned}$$

The first sum is what we want. The second is 0 since, using the tower property and Assignment Symmetry

$$\mathbb{E}[\mathbb{E}[\left(\hat{\tau}_k-\tau_k\right)\left(\hat{\tau}_r-\tau_r\right)|W]] = \mathbb{E}[\mathbb{E}[\left(\hat{\tau}_k-\tau_k\right)|W]\,\mathbb{E}[\left(\hat{\tau}_r-\tau_r\right)|W]] = \mathbb{E}[0\cdot 0] = 0.$$

Theorem 2.4.1. Calculate the MSE of $\hat{\tau}_{sd}$ conditioned on the split W with a slight modification to the above derivation. Define a new estimator that is a weighted difference in means:

$$\hat{\alpha}_k \equiv A_k \sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1) - B_k \sum_{i:b_i=k} \frac{1-T_i}{W_k(0)} y_i(0)$$

with A_k, B_k constant. $\hat{\alpha}_k$ is an unbiased estimator of the difference in means weighted by A_k and B_k:

$$\mathbb{E}[\hat{\alpha}_k] = \mathbb{E}\left[A_k \sum_{i:b_i=b} \frac{T_i}{W_k(1)} y_i(1) - B_k \sum_{i:b_i=b} \frac{T_k}{W_i(0)} y_i(0) \right] = A_k \bar{y}_k(1) - B_k \bar{y}_k(0).$$

Now follow the derivation of the variance of $\hat{\tau}_k$ propagating A_k and B_k through. These are constant and they come out, giving

$$\mathrm{Var}[\hat{\alpha}_k] = \frac{1}{n_k} \left[A_k^2 \beta_{1k} \sigma_k^2(1) + B_k^2 \beta_{0k} \sigma_k^2(0) + 2 A_k B_k \gamma_k(1,0) \right].$$

Expand $\hat{\tau}_{sd}$ into strata terms:

$$\hat{\tau}_{sd} = \sum_{k=1}^{K} \left[\frac{W_{1k}}{W_1} \sum_{i:b_i=k} \frac{T_i}{W_{1k}} y_i(1) - \frac{W_{0k}}{W_0} \sum_{i:b_i=k} \frac{1-T_i}{W_{0k}} y_i(0) \right] = \sum_{k=1}^{K} \hat{\alpha}_k$$

with $A_k = W_{1k}/W_1$ and $B_k = W_{0k}/W_0$. Conditioning on W makes the A_k and the B_k constants, $\beta_{1k} = W_{0k}/W_{1k}$, and $\beta_{0k} = W_{1k}/W_{0k}$. Assignment symmetry ensures that, conditional on W, the stratum assignment patterns are independent, so the $\hat{\alpha}_k$ are as well, and the variances then add:

$$\mathrm{Var}[\hat{\tau}_{sd}|W] = \sum_{k=1}^{K} \mathrm{Var}[\hat{\alpha}_k|W].$$

The bias is $\mathbb{E}[\hat{\tau}_{sd}|W] - \tau$ with

$$\mathbb{E}[\hat{\tau}_{sd}|W] = \sum_{k=1}^{K} \mathbb{E}[\hat{\alpha}_k|W] = \sum_{k=1}^{K} A_k \bar{y}_k(1) - B_k \bar{y}_k(0).$$

Expand τ as in Equation 2.2 and rearrange terms.

Extending to PATE. First, decompose the variance:

$$\mathrm{Var}[\hat{\tau}_{ps}|\mathcal{D}] = \mathbb{E}_{\mathcal{S}}\left[\mathrm{Var}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]\,|\mathcal{D}\right] + \mathrm{Var}_{\mathcal{S}}\left[\mathbb{E}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]\,|\mathcal{D}\right]$$

The first term is simply the expectation of Equation 2.6, the SATE variance formula. Since $\mathcal{S}$ is random, so are the $\sigma_k^2(\ell)$, etc. The expectation of these quantities over $\mathcal{S}$ gives the

population parameters as they are unbiased estimators. The β's are all constant, and $\mathcal{D}$ is independent of $\mathcal{S}$. Therefore $\mathbb{E}_{\mathcal{S}}[X|\mathcal{D}] = \mathbb{E}_{\mathcal{S}}[X]$ and:

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[\text{Var}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]\,|\mathcal{D}] &= \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_k \frac{n_k}{n}\left[\beta_{1k}\sigma_k^2(1) + \beta_{0k}\sigma_k^2(0) + 2\gamma_k(1,0)\right]|\mathcal{D}\right] \\ &= \frac{1}{n}\sum_k \frac{n_k}{n}\left[\beta_{1k}\sigma_k^2(1)^* + \beta_{0k}\sigma_k^2(0)^* + 2\gamma_k(1,0)^*\right]. \qquad (5.1)\end{aligned}$$

The second term is

$$\begin{aligned}\text{Var}[\mathbb{E}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]] &= \text{Var}[\tau] \\ &= \text{Var}\left[\sum_{k=1}^K \frac{n_k}{n}\tau_k\right] \\ &= \frac{n_k^2}{n^2}\sum_{k=1}^K \text{Var}[\bar{y}_{k1} - \bar{y}_{k0}] \\ &= \frac{n_k^2}{n^2}\sum_{k=1}^K \frac{1}{n_k}\left[\sigma_k^2(1)^* + \sigma_k^2(0)^* - 2\gamma_k(1,0)^*\right]. \qquad (5.2)\end{aligned}$$

Sum Equation 5.1 and Equation 5.2 to get the PATE-level MSE.

5.2 Proofs of the Bounds on Variance Differences

$\beta_{\ell k}$ can be approximated by $\mathbb{E}[W_k(1-\ell)] \,/\, \mathbb{E}[W_k(\ell)]$. For example, in the complete randomization case $\beta_{1k} \approx (1-p)/p$. Generally, the β's are larger than their approximations. They can be less, but only by a small amount. For complete randomization and Bernoulli assignment, the difference between the β's and their approximations is bounded by the following theorem:

Theorem 5.2.1. *Take an experiment with n units randomized under either complete randomization or Bernoulli assignment. Let p be the expected proportion of units treated. Let $\mathcal{D}$ be the event that $\hat{\tau}_{ps}$ is defined. Let $p_{max} = max(p, 1-p)$ andn_{min} be the smallest strata size. Then $\beta_{1k} - (1-p)/p$ is bounded above:*

$$\begin{aligned}\beta_{1k} - \frac{1-p}{p} &\leq \frac{4}{p^2}\frac{1}{n_k} - \frac{1}{p}\frac{1}{n_k+1} + \max\left[\left(\frac{n_k}{2} - \frac{4}{p^2 n_k}\right)e^{-\frac{p^2}{2}n_k}, 0\right] + 2n_k K\left(p_{max}\right)^{n_{min}} \\ &= \frac{4}{p^2}\frac{1}{n_k} + O(n_k e^{-n_{min}}).\end{aligned}$$

Furthermore, it is tightly bounded below:

$$\beta_{1k} - \frac{1-p}{p} \geq -\frac{2}{p}(1-p)^{n_k} - 2n_k K(p_{max})^{n_{min}} = -O(n_k e^{-n_{min}}).$$

Similar results apply for the β_{0k} and β_ℓ.

Proof. Start without conditioning on $\mathcal{D}$. $W_{1k} = \sum T_i$ with $T_i \in \{0, 1\}$. For Bernoulli assignment, the T_i are i.i.d Bernoulli variables with probability p of being 1. For completely randomized experiments, the W_{1k} are distributed according to a hypergeometric distribution, i.e., as the number of white balls drawn in n_k draws without replacement from an urn of n balls with np white balls. Regardless, $\mathbb{E}[W_{1k}] = n_k p$.

Define $Y_{n_k} \equiv (n_k/W_{1k}) \times \mathbf{1}_{\{W_{1k}>0\}}$. Due to the indicator function, $Y_{n_k} \leq n_k$. Given $\mathcal{D}$, the event that *all* strata-level estimators are well-defined, $Y_{n_k} = n_k/W_{1k}$ so

$$\beta_1 - \frac{1-p}{p} = \mathbb{E}\left[\frac{W_{0k}}{W_{1k}}|\mathcal{D}\right] - \frac{1-p}{p} = \mathbb{E}\left[\frac{n_k}{W_{1k}}|\mathcal{D}\right] - \frac{1}{p} = \mathbb{E}[Y_{n_k}|\mathcal{D}] - \frac{1}{p}.$$

We first show the probability of $\neg\mathcal{D}$ is very small, which will allow for approximating the expectation of the conditioned Y_{n_k} with the unconditioned. If n_{min} is the size of the smallest strata, then

$$\begin{aligned}
\mathbf{P}\neg\mathcal{D} &\leq \sum_{k=1}^{K} \mathbf{P}\{W_{1k} = 0 \text{ or } W_{0k} = 0\} \\
&\leq 2K \max_{\ell=0,1;k=1,\ldots,K} \mathbf{P}\{W_{\ell k} = 0\} \\
&\leq 2K \left(p_{max}\right)^{n_{min}}.
\end{aligned}$$

Expand the expected value of Y as

$$\mathbb{E}[Y_{n_k}] = \mathbb{E}[Y_{n_k}|\mathcal{D}]\,\mathbf{P}\mathcal{D} + \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\,\mathbf{P}\neg\mathcal{D}.$$

Use this and the bound $Y_{n_k} \leq n_k$ to get

$$\begin{aligned}
\left|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}]\right| &= \left|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}|\mathcal{D}]\,\mathbf{P}\mathcal{D} - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\,\mathbf{P}\neg\mathcal{D}\right| \\
&= \left|\mathbb{E}[Y_{n_k}|\mathcal{D}]\,(1 - \mathbf{P}\mathcal{D}) - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\,\mathbf{P}\neg\mathcal{D}\right| \\
&= \left|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\right|\mathbf{P}\neg\mathcal{D} \\
&\leq n\mathbf{P}\neg\mathcal{D} = 2nK\left(p_{max}\right)^{n_{min}} \qquad (5.3)
\end{aligned}$$

This shows that $\mathbb{E}[Y_{n_k}|\mathcal{D}]$ is quite close to $\mathbb{E}[Y_{n_k}]$, i.e.

$$\mathbb{E}[Y_{n_k}] - \frac{1}{p} - 2nK\left(p_{max}\right)^{n_{min}} \leq \beta_1 - \frac{1-p}{p} \leq \mathbb{E}[Y_{n_k}] - \frac{1}{p} + 2nK\left(p_{max}\right)^{n_{min}}.$$

Now we need the following lemma to get a handle on $\mathbb{E}[Y_{n_k}]$:

Lemma 5.2.2. *Let W be a Binomial (n,p) random variable or a hypergeometric (n,w,N) random variable, i.e., a sample of size n from coin flips with probability of heads p or an urn with $N = nc$ balls, $c > 1$, of which $w = ncp$ are white. Then for $Y = (n/W)\mathbf{1}_{\{W>0\}}$:*

$$-\frac{2}{p}(1-p)^n \le \mathbb{E}[Y] - \frac{1}{p} \le \frac{4}{p^2}\frac{1}{n} - \frac{1}{p}\frac{1}{n+1} + \max\left[\left(\frac{n}{2} - \frac{4}{p^2 n}\right)\exp\left(-\frac{p^2}{2}n\right), 0\right].$$

See below for proof. Use Lemma 5.2.2 on $\mathbb{E}[Y_{n_k}]$. This gives our stated bounds.

□

Proof of 5.2.2

Proof. First we derive the lower bound on the expectations. For ease of notation, define $\mathbb{E}[X;\mathcal{A}]$ as the expectation of $X\mathbf{1}_{\{A\}}$. For both Bernoulli assignment or complete randomization,

$$np = \mathbb{E}[W] = \mathbb{E}[W; W>0] = \mathbb{E}[W|W>0]\,\mathbf{P}\{W>0\}.$$

Also, $\mathbf{P}\{W=0\} \le (1-p)^n$. For a random variable $X > 0$, $\mathbb{E}[1/X] \ge 1/\mathbb{E}[X]$. Therefore

$$\begin{aligned}
\mathbb{E}\Big[\frac{n}{W}; W>0\Big] &= \mathbb{E}\Big[\frac{n}{W}|W>0\Big]\,\mathbf{P}\{W>0\} \\
&\ge \frac{n}{\mathbb{E}[W;W>0]}\mathbf{P}\{W>0\}^2 \\
&= \frac{1}{p} + \frac{1}{p}\left(\mathbf{P}\{W>0\}^2 - 1\right) \\
&\ge \frac{1}{p} - \frac{2}{p}(1-p)^n
\end{aligned}$$

For the upper bound, expand $\mathbb{E}[n/W; W>0]$ into two terms and analyze each term. Namely, we will show that $\mathbb{E}[n/W; W>0] = I + D$ with

$$I \equiv \mathbb{E}\left[\frac{n}{W+1}; W>0\right] \le \frac{1}{p} - \frac{1}{n+1}\frac{1}{p}$$

and

$$\begin{aligned}
D &\equiv \mathbb{E}\left[\frac{n}{W} - \frac{n}{W+1}; W>0\right] = \mathbb{E}\left[\frac{n}{W(W+1)}; W>0\right] \\
&\le \min_{0<\alpha_n<p}\left\{\frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2 + \max\left[\left(\frac{n}{2} - \frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2\right)\exp(-2n\alpha_n^2), 0\right]\right\}
\end{aligned}$$

Instead of minimizing the bound across the possible values of α_n, we can simply fix $\alpha_n = p/2$ to obtain a looser, but more intelligible, bound:

$$D \le \frac{4}{p^2}\frac{1}{n} + \max\left[\left(\frac{n}{2} - \frac{4}{np^2}\right)\exp(-\frac{p^2}{2}n), 0\right].$$

We show D first. Let α_n be in $(0, p)$. Then:

$$\begin{aligned} D &= \mathbb{E}\left[\frac{n}{W(W+1)}; p - \frac{W}{n} < \alpha_n\right] + \mathbb{E}\left[\frac{n}{W(W+1)}\mathbf{1}_{\{W>0\}}; p - \frac{W}{n} \geq \alpha_n\right]. \\ &\leq \frac{1}{n}\,\mathbb{E}\left[\left(\frac{n}{W}\right)^2; p - \alpha_n < \frac{W}{n}\right] + \frac{n}{2}\mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\}. \end{aligned}$$

The $n/2$ is because $W > 0$ implies $W \geq 1$.

[36] famously bounded the tail probabilities of sums of independent random variables, allowing us to control the probability of W/n being far from p. He also, in section 6 of the same work, generalized his bound to the hypergeometric. We use both of these results:

$$\mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\} \leq \exp(-2n\alpha_n^2).$$

Because $0 < p - \alpha_n < W/n$ we have

$$\begin{aligned} D &\leq \frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2\left(1 - \mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\}\right) + \frac{n}{2}\mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\} \\ &\leq \frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2 + \max\left[\left(\frac{n}{2} - \frac{1}{n}\frac{1}{(p-\alpha_n)^2}\right)\exp(-2n\alpha_n^2), 0\right]. \end{aligned}$$

The $\max(\cdot,\cdot)$ comes from the choice of α_n possibly making $n/2 - 1/n(p-\alpha_n)^2 < 0$ which would invert the Hoeffding bound. We instead conservatively set this quantity to 0.

To evaluate I, consider the Binomial case first. Express the expectations as a sum and re-index the sum and add in the first two terms to get the sum of the distribution of a $(n+1, p)$ binomial variable:

$$\begin{aligned} I \equiv \mathbb{E}\left[\frac{n}{W+1}; W > 0\right] &= \sum_{k=1}^{n} \frac{n}{k+1}\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= \frac{n}{n+1}\frac{1}{p}\left(\sum_{k=0}^{n+1}\frac{(n+1)!}{k!(n+1-k)!}p^k(1-p)^{n+1-k}\right. \\ &\qquad\qquad \left. - (1-p)^{n+1} - (n+1)p(1-p)^n\right) \\ &= \frac{n}{n+1}\frac{1}{p}\Big(1 - (1-p)^{n+1} - (n+1)p(1-p)^n\Big) \\ &= \frac{1}{p} - \frac{1}{n+1}\frac{1}{p} - \frac{n}{n+1}\frac{1}{p}(np+1)(1-p)^n. \end{aligned}$$

This is exact for the Binomial case. To extend to complete randomization, we use a further result from Hoeffding. Hoeffding showed that, for a continuous, convex function $f(x)$,

$\mathbb{E}_{srs}[f(W)] \leq \mathbb{E}_{bin}[f(W)]$. Let $f(x)$ be $n/(x+1)$. $f(x)$ is continuous, convex for $x \geq 0$. Furthermore for Binomial W

$$\mathbb{E}\left[\frac{n}{W+1}; W > 0\right] + n(1-p)^n = \mathbb{E}[f(W)]$$

as $n/(W+1) = f(W)$ for all W. So

$$\mathbb{E}_{srs}[\frac{n}{W+1}; W > 0] \leq \mathbb{E}_{srs}[f(W)] \leq \mathbb{E}_{bin}[f(W)] = \mathbb{E}_{bin}[\frac{n}{W+1}; W > 0] + n(1-p)^n$$

Thus we gain an extra (small) $n(1-p)^n$ term to bound I, but this term is more than offset by the negative term $n/(n+1) \times (np+1)/p \times (1-p)^n$ and so we drop both.

To get the overall bound, sum the bounds for I and D. □

Remarks: As a side note, [77] improves Hoeffding's bound for sampling without replacement, implying that the rate of the βs convergence is faster under complete randomization than for Bernoulli.

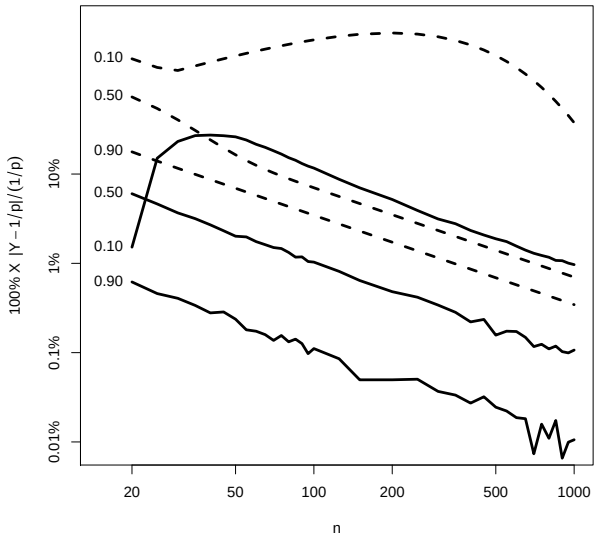

Figure 5.1: log-log plot comparing actual percent difference $100\% \times (\mathbb{E}[Y] - 1/p)/(1/p)$, Y as defined in Lemma 5.2.2 to the given bound. Three probabilities of assignment shown: $p = 0.1, 0.5$, and 0.9. Actual differences computed with Monte Carlo. Y generated with Bernoulli distribution.

Remark on Lemma 5.2.2. Numerical calculation shows the constants of the $1/n$ term are overly large, but the rate of $1/n$ appears to be correct. Figure 5.1 show a log-log plot of the actual percent increase of Y's over $1/p$ for several values of p and n along with the calculated bounds. When the exponential term becomes negligible, the bound appears to be about 4, 7, and 31 times bigger for $p = 0.1, 0.5$, and 0.9 respectively, i.e., the constants on the $1/n$ term are overstated by this much. For low p, the exponential terms can remain for quite some time in the bound and there is significant bias in actuality due to high chance of 0 units assigned to treatment. The log-log slope is -1 suggesting the $1/n$ relationship.

Proof of Theorem 2.2.4. Assume the conditions stated for Theorem 2.2.4 and consider Equation 2.9. Replace all σs and γs with σ^2_{max} and γ^2_{max}. Replace all $\beta_{\ell 0}$ with $\tilde{\beta}_0$, the largest such β for some stratum k. Same for $\tilde{\beta}_1$. Collapse the sums to get

$$\text{scaled cost} \leq \left(\tilde{\beta}_0 - \frac{n-K}{n-1}\beta_0\right)\sigma^2_{max} + \left(\tilde{\beta}_1 - \frac{n-K}{n-1}\beta_1\right)\sigma^2_{max} + 2\frac{K-1}{n-1}\gamma_{max}.$$

Then,

$$\begin{aligned}\left|\tilde{\beta}_0 - \frac{n-K}{n-1}\beta_0\right| &\leq \left|\tilde{\beta}_0 - \beta_0\right| + \left|\frac{n-K}{n-1}\beta_0 - \beta_0\right| \\ &\leq \frac{4}{(1-p)^2}\frac{1}{fn} + \frac{K-1}{n-1}\frac{p}{1-p} + O(\frac{1}{n^2}).\end{aligned}$$

Because the lower bound is so tight, we don't need to double the bound from Theorem 5.2.1 for bounding the difference $\left|\tilde{\beta} - \beta_0\right|$. Because the β_1 expression will be smaller at the end, we can simply double the β_0 expression. This gives the bound.

Proof of Theorem 2.3.1. This is handled the same way as for Theorem 2.2.4, but is more direct.

5.3 Toy Examples of Gain and Loss

In this section we provide a few small examples to demonstrate the potential for gain or loss due to post-stratification. Each of the following scenarios specify a particular collection of potential outcomes, and Table 5.1 shows the resulting variances of the unadjusted estimator and post-stratified estimator (both conditioned on $\mathcal{D}$). In all cases we assume complete randomization with $100p\%$ units treated, with p as stated on the table. Table 5.1 also shows the variance if the randomization were done via blocking.

We plug the parameters defined by the stated population into the variance formulas presented in Chapter 2. We numerically compute the βs by conducting the described randomization $50,000$ times and computing the mean βs for those randomizations where all strata estimators were defined (i.e., we condition on $\mathcal{D}$). The results on Table 5.1 are exact

up to the uncertainty in computing the βs. Bernoulli randomization gives near-identical results (since the β's are near identical). Directly estimating variance with a monte-carlo of point estimates also gives identical results up to sampling error, further validating the formula as correct.

						variances			% gain/loss		
	n	K	p	$\mathbf{P}\mathcal{D}$	τ	$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	blk	blk:ps	sw:ps	sw:blk
I.A	40	4	0.50	99.9%	1.00	1.01	1.36	0.92	-7%	26%	33%
I.B	40	4	0.50	99.9%	1.00	1.01	0.85	0.92	-11%	-20%	-8%
I.C	40	4	0.30	93.7%	1.00	1.28	1.62	1.09	-12%	21%	33%
I.D	40	4	0.50	99.9%	1.00	1.01	2.24	0.92	-4%	55%	59%
I.E	40	4	0.50	99.9%	1.00	1.01	0.91	0.92	-10%	-11%	-1%
II.A	100	4	0.50	99.9%	1.91	0.39	0.63	0.37	-3%	39%	42%
II.B	100	4	0.30	97.2%	1.91	0.52	0.80	0.47	-5%	35%	41%
III.A	200	2	0.50	100%	0.94	0.24	0.30	0.24	0%	21%	21%
III.B	200	5	0.50	100%	0.94	0.21	0.30	0.21	-2%	28%	30%
III.C	200	10	0.50	100%	0.94	0.22	0.30	0.21	-4%	25%	29%
III.D	200	20	0.50	97%	0.94	0.24	0.30	0.21	-10%	20%	30%
III.E	200	25	0.50	84.4%	0.94	0.25	0.30	0.21	-13%	18%	30%
IV.A	200	2	0.50	100%	0.94	0.30	0.30	0.30	-1%	0%	0%
IV.B	200	5	0.50	100%	0.94	0.31	0.30	0.30	-2%	-2%	0%
IV.C	200	10	0.50	100%	0.94	0.32	0.30	0.30	-5%	-7%	-2%
IV.D	200	20	0.50	97%	0.94	0.35	0.30	0.31	-14%	-17%	-3%
IV.E	200	25	0.50	84.4%	0.94	0.37	0.30	0.31	-19%	-23%	-4%
V.A	100	4	0.50	99.9%	1.80	0.32	0.55	0.30	-4%	42%	45%
V.B	200	4	0.50	100%	1.80	0.15	0.28	0.15	-2%	45%	47%
V.C	400	4	0.50	100%	1.80	0.07	0.14	0.07	-1%	47%	47%
V.D	800	4	0.50	100%	1.80	0.04	0.07	0.04	0%	47%	48%

Table 5.1: Variances of Estimators for Several Scenarios. K is the number of strata. $\mathbf{P}\mathcal{D}$ is the probability of $\hat{\tau}_{ps}$ being defined, estimated by simulation. τ is actual SATE. The percentages are calculated as $100\% \times \Delta / \text{Var}[\hat{\tau}_{sd}]$ with Δ being the specified difference between variances.

The families of scenarios are as follows:

I) We first consider a simple experiment with four strata, A, B, C, and D, with 10 units each.

   A) In the first scenario, the units in strata B, C and D are replicates of A shifted up by +2, +4, and +6, respectively. There is a constant treatment effect of +1. There is substantial between-strata variation, and therefore post-stratification is beneficial.

The left plot in Figure 5.2 displays the relationship between potential outcomes and strata. This is the idealized constant-treatment effect situation where stratification separates units of different types.

B) As Scenario I.A, but now the units in B, C, and D are simple replicates of A's units not shifted. There is no difference between strata and so we see the full price paid by spurious post-stratification. It is easy in this small experiment for a random imbalance to occur. An imbalance overweights some of the units, making it easier to reach extreme values for estimated treatment effect. This results in a larger variance. In this scenario blocking is also a poor choice, incurring a small cost.

C) As Scenario I.A, but with probability of treatment $p = 0.3$. The small proportion treated makes it easier to have very few units estimating the average treatment effect in a stratum (or overall). All estimators' variances increase, but post-stratification still comes out ahead of simple difference.

D) Now we have differing treatment effects of $-3, 0, +2$, and $+5$ for the four strata. We no longer have an overall constant treatment effect: different strata respond to treatment differently. Here the between-strata correlation of potential outcomes is near 1.00. This makes post-stratification work very well.

E) A reverse of Scenario 1.D, we now have differing treatment effects of $+5, +2, 0, -3$ for the four strata. The trend of the group means is opposite to the trend of outcomes within groups, which causes problems. The between-strata correlation of potential outcomes is -0.95. See right plot on Figure 5.2. The between-strata term is small due to this negative correlation. Negative correlations are good for randomization because it means that if a randomly high unit is put into treatment, a randomly high unit will probably be put into control as well to compensate. Post-stratification does not take advantage of this, and thus does more poorly than the unadjusted estimator, which does. Blocking also does not fair well in this case for the same reasons.

In all the above, which are for a small-sized experiment, post-stratification is somewhat close to blocking.

II) A slightly more complex experiment with unequal strata sizes. A has 60 units, B has 15, C has 15 and D has 10. We drew the $y_k(0)$ for A from a $N(5, 5)$ population, B from a $N(3, 10)$, C from a $N(7, 15)$ and D from a $N(3, 15)$, where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The treatment effects for all units, drawn from a $unif(-1, 5)$ distribution, were added to the control outcomes. The presented results are the variances of the estimators under different randomizations of a *single sample* drawn from this described population.

A) Equal treatment proportions of $p = 0.5$. Post-stratification helps. It is also close to blocking.

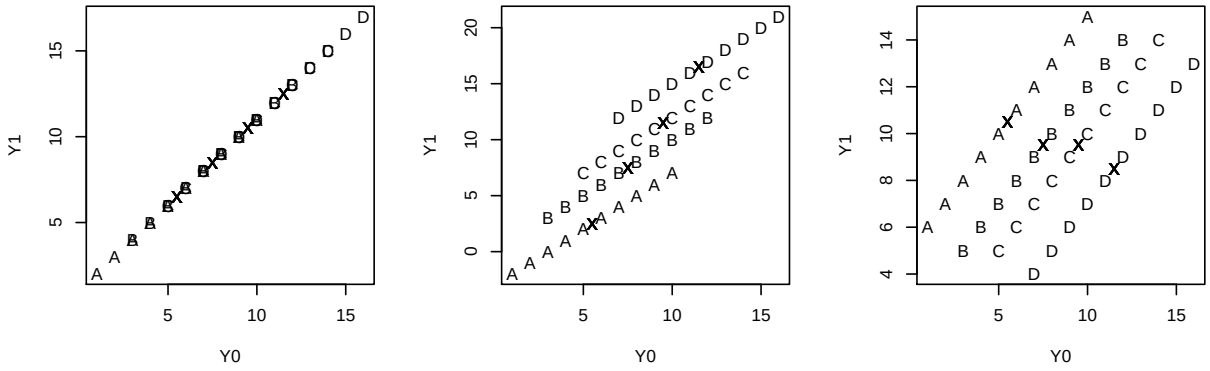

Figure 5.2: From left to right, the potential outcomes for scenarios I.A, I.D, and I.E. Strata membership on the plots are denoted A through D. "X" denotes strata means.

B) $p = 0.3$. The efficacy declines slightly due to the increased chance of imbalance. Blocking does not suffer as much.

III) A set of experiments with a continuous covariate z evenly spaced on the interval $[0, 100]$ which we then partition into K strata of equal sizes. We vary K to see the impact of finer stratification. The control outcome for unit i is distributed as $y_i(0) \sim N(\sqrt{z_i}, 9)$ and the treatment outcome as $y_i(1) \sim N(y_i(0) + 1, 1)$. About 5 strata seems ideal although even two strata is far better than doing nothing. Too many strata and we see less benefit, plus a large increase in the chance of an undefined estimator.

IV) As III, but now z is useless. We generate this set by permuting the observed z from III, breaking any connection between the covariate and the outcomes. $\hat{\tau}_{sd}$ is completely unaffected. As the number of strata increase, things worsen for post-stratification due to the increased chance of an accidental imbalance giving a single unit a great deal of weight. Blocking also suffers, but not by nearly as much.

V) In this set of experiments, the set-up being the same as for Experiment II, we first generated an initial set of data, and then replicated the units within the strata to increase n. The number of strata is thus held constant and the treatment effect, covariances and variances for subsequent experiments remain essentially unchanged. As n grows, the percentage increase in variance of $\hat{\tau}_{ps}$ over blocking converges to 0 at rate $1/n$, and thus the percentage gain over $\hat{\tau}_{sd}$ converges to a fixed relative improvement in precision over the unadjusted estimate.

Discussion. Generally speaking, post-stratification is similar to blocking in terms of efficiency. The more strata, however, the worse this comparison becomes due to the increased

chance of severe imbalance with consequential increase high uncertainty in the stratum-level estimates. Post-stratification's overall efficacy depends on how much larger the between-stratum variation is compared to the penalty paid by giving some observations greater weight due to random assignment imbalance. Having many strata is generally not helpful and can be harmful if b is not prognostic. A moderate number of strata seems to offer protection from this: compare $K = 5$ for scenarios III and IV.

Examining Conditional Variance

To illustrate how the variance of the estimators conditioned on the split W varies, we repeatedly conduct a randomization for a specific sample and calculate the conditional MSE for both estimators given the generated split as shown in the latter half of Section 7 of Chapter 2. These simulations demonstrate that if b is indeed prognostic, then the MSE of $\hat{\tau}_{ps}$ is far lower than that of $\hat{\tau}_{sd}$, and this difference increases with degree of imbalance. However, if b is not prognostic, then the reverse trend is evident. The post-stratified estimator does worse in the very circumstance when people might use it: to adjust for a seen imbalance in the randomization. It is not necessarily beneficial to adjust—the variable used for adjustment must be selected with care.

The left side of Figure 5.3 shows 5000 such calculations for Scenario III.B, presented above. With low imbalance, the variance of $\hat{\tau}_{ps}$ is even smaller than the unconditional formula would suggest. But as imbalance deteriorates, the variance of $\hat{\tau}_{ps}$ increases.

Compared to $\hat{\tau}_{ps}$, the simple-difference estimator $\hat{\tau}_{sd}$ is vulnerable to poor splits. Generally, high imbalance means high conditional MSE. This is due to the bias term which can get exceedingly large if there is imbalance between different heterogeneous strata. We see a similar trend to the analogous PAC-Man example.

If b is not prognostic, however, the story changes. The experimental units in Scenario IV.B, shown on right of Figure 5.3, are the same as for Scenario III.B, but the elements of the covariate vector b have been shuffled to break b's prognostic ability. Because the units are the same, the unconditional variance of $\hat{\tau}_{sd}$ is the same as well. Because b is no longer prognostic, post-stratification does not help, as illustrated by the elevated unconditional and conditional trend lines. The post-stratified estimator still worsens with greater imbalance as it did before because the cost of imbalance comes from the number of observations in the treatment and control groups, something unrelated to b. The simple-difference estimator, however, often can even improve with large imbalance. This is due to imbalance ensuring a greater comparability of treatment and control units—if it were known that b was not connected to the potential outcomes then it would actually be most ideal to treat all of some strata and none of the others.

In other scenarios (not shown) these trends are repeated. Furthermore, when there are few strata, the imbalance tends to be low (e.g., Scenarios I and II, or III with small K) with a heavily right skewed distribution of conditional variance—most of the time there is a good balance and low conditional variance, but there is a low chance of a bad split and a high conditional variance. In such circumstances, there is very a good chance that

the conditional variance of a post-stratified experiment is even closer to its corresponding blocked experiment than one would initially expect from Equation 12 in the main document. Also in such circumstances the pattern of the MSE of $\hat{\tau}_{sd}$ worsening for prognostic b and improving for unrelated b as imbalance increases is even more apparent.

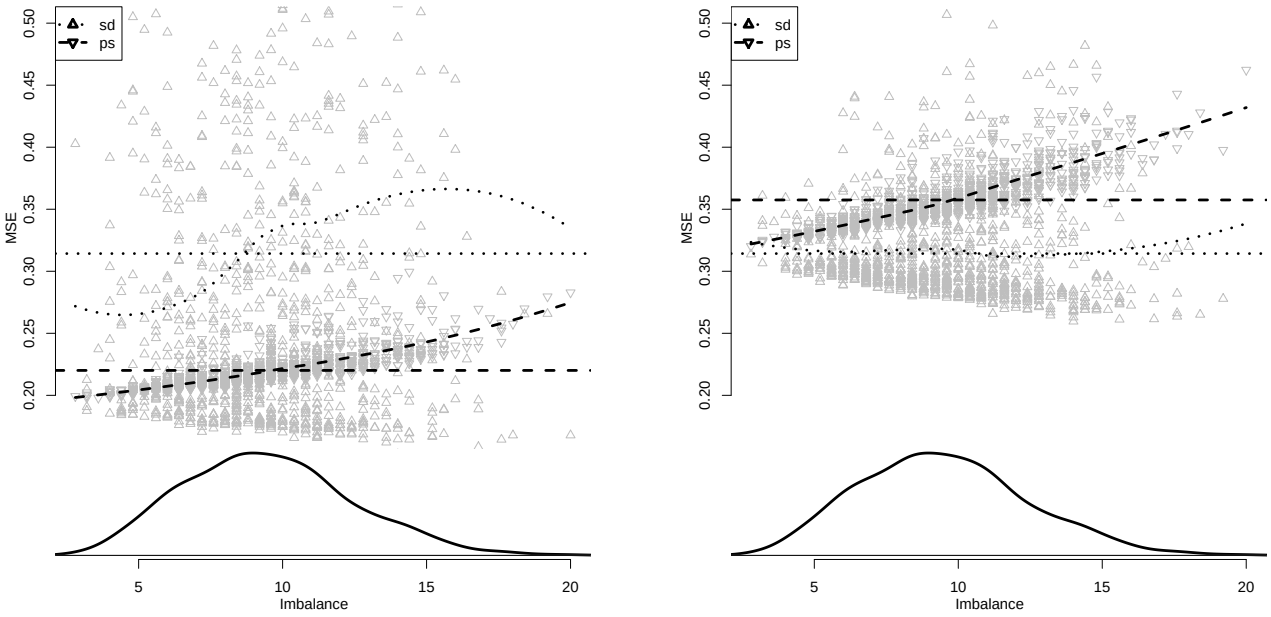

Figure 5.3: Conditional Variance of Scenario III.D (left) and Scenario IV.D (right). Points indicate the conditional MSE of $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ given various specific splits of W. x-axis is the imbalance score for the split. Curved dashed lines interpolate point clouds. Horizontal dashed lines mark unconditional variances for the two estimators. The curves at bottom are the densities of the imbalance statistic.

Chapter 6

Appendix B: Further Details on Text Summarization

6.1 Supplementary Tables

Table 6.1 shows the countries used in the human validation experiment along with the number of positive examples found for each under different labeling schemes. Table 6.2 gives an example of lists generated with and without stop-word removal, demonstrating how regularization achieves the impact of stop-word removal and also removes "second-order" stop words as well, giving better results in general.

6.2 The Impact of Selecting Distinct Phrases

Final summaries consist of a target of k *distinct* key-phrases. The feature-selectors are adjusted to provide enough phrases such that once sub-phrases (e.g., *"united"* in *"united states"*) are removed, the list is k phrases long. This removal step, similar to stop-word removal, is somewhat ad hoc. It would be preferable to have methods that naturally select distinct phrases that do not substantially overlap. Sparse methods have some protection against selecting highly correlated features, and thus they might not need this cleaning step as sub-phrases tend to be highly correlated with parent phrases, with correlations often exceeding 0.8. To investigate this, we examined the average value of $k' - k$, the difference of the length of the summary without sub-phrases removed to the length with this removal. Results are shown in Table 6.3. The sparse methods indeed do not need to take advantage of this step, supporting the heuristic knowledge that L^1-penalization tends to avoid selecting correlated features. Under tf-idf, only a little over 1 phrase, on average, is dropped. The independent feature selection methods, however, drop many phrases on average.

For Correlation Screening, this difference is because sub-phrases are often extremely highly correlated with parent phrases—if a given phrase is highly correlated with the outcome, then any sub-phrase or parent phrase is likely to also be highly correlated. This

	article-1		article-2		article-3		paragraph-1		paragraph-2	
subject	#	%	#	%	#	%	#	%	#	%
china	1436	15%	970	10%	800	8%	6455	5%	2026	1.6%
iran	1387	15%	906	9%	715	7%	4875	4%	1621	1.2%
iraq	1139	12%	710	7%	562	6%	4806	4%	1184	0.9%
afghanistan	1133	12%	729	8%	592	6%	4774	4%	659	0.5%
israel	1126	12%	591	6%	388	4%	4478	3%	1537	1.2%
pakistan	989	10%	650	7%	555	6%	4454	3%	1384	1.1%
russia	981	10%	699	7%	590	6%	4288	3%	1168	0.9%
france	867	9%	419	4%	291	3%	2815	2%	586	0.4%
india	848	9%	613	6%	537	6%	2368	2%	559	0.4%
germany	788	8%	387	4%	284	3%	2333	2%	459	0.4%
japan	566	6%	273	3%	195	2%	1780	1%	406	0.3%
mexico	413	4%	238	2%	189	2%	1475	1%	392	0.3%
south korea	382	4%	208	2%	136	1%	1254	1%	251	0.2%
egypt	361	4%	231	2%	194	2%	1070	1%	230	0.2%
turkey	281	3%	125	1%	96	1%	797	1%	197	0.2%

Table 6.1: Our Experiment's Subjects With Sizes of Positive Example Sets. "#" denotes number and "%" denotes portion of units positively marked. A greater proportion of units are marked positive in the article-unit analysis. Generally, only a small portion of articles are considered topical for a given subject.

problem is especially common with the names of political leaders, e.g., Prime Minister Wen Jiabao in the second column of Table 3.2. Correlation Screening is virtually unusable without dropping sub-phrases and expanding the list to the desired length.

The amount of sub-phrase reduction in Co-occurrence-derived summaries strongly depends on the reweighting method used. Under stop-word removal there is little reduction since many of the selected phrases are combinations of non-overlapping stop-words, such as *"of the," or "to the,"* where the individual component stop-words have been removed prior to summarization. Under L^2-rescaling, the typically common stop-word combinations no longer score highly, and problems similar to those seen in the Correlation Screening results arise: groups of parent- and sub-phrases score similarly, requiring sub-phrase pruning to improve list quality.

	stop-word only	stop word and rescaling	no adjustment	rescaling only
1	afghanistan	asian	afghanistan	asian
2	beijing	beijing	and	beijing
3	companies	contributed research	beijing	contributed research
4	countries	euna lee	countries	euna lee
5	economic	global	global	global
6	global	hong kong	has	hong kong
7	hong	jintao	his	jintao
8	military	north korea	its	north korea
9	mr	shanghai	mr	shanghai
10	north	staterun	north	staterun
11	percent	uighurs	of	uighurs
12	the united states	wen jiabao	the united	wen jiabao
13	uighurs	xinhua	to	xinhua
14	world		united states	
15	year		was	

Table 6.2: Comparative Effects of Reweighting Methods. Four summaries of "China" from all combinations of L^2-rescaling and stop-word removal. The phrase-selection method used is L1LR with count-2 labeling on full articles.

	Reweighting Method		
Feat. Sel. Method	stop-word	L^2-rescaling	tf-idf rescaling
Co-occurrence	2.7	12.8	7.3
Correlation	12.9	12.9	12.7
L1LR	0.5	3.9	1.2
Lasso	0.6	3.7	1.2

Table 6.3: Phrase Reduction for the Four Feature Selectors. Each entry shows the mean number of sub-phrases dropped, on average, for all varieties of summarizer with specified rescaling and feature-selection method for a target summary length of $k = 15$ phrases. For example, under tf-idf we need to generate a full list of 16.2 phrases with L1LR, on average, to achieve a final list length of 15 phrases. The sparse methods do not need much pruning. Correlation Screening selects highly related sub-phrases and therefore requires much pruning.

Bibliography

[1] Alberto Abadie and Guido W. Imbens. *Estimation of the Conditional Variance in Paired Experiments*. 2007.

[2] J.A. Aslam, R.A. Popa, and R.L. Rivest. "On Auditing Elections When Precincts Have Different Sizes". In: *2008 USENIX/ACCURATE Electronic Voting Technology Workshop, San Jose, CA, 28–29 July*. 2008.

[3] Michael L. Beach and Paul Meier. "Choosing Covariates in the Analysis of Clinical Trials". In: *Controlled Clinical Trials* 10 (1989), pp. 1615–1735.

[4] P.J. Bickel. "Correction: Inference and auditing: The Stringer bound". In: *Intl. Stat. Rev.* 61 (1993), p. 487.

[5] P.J. Bickel. "Inference and auditing: The Stringer bound". In: *Intl. Stat. Rev.* 60 (1992), pp. 197–209.

[6] David Blei and Jon McAuliffe. "Supervised Topic Models". In: *Advances in Neural Information Processing Systems 20*. Ed. by J.C. Platt et al. Cambridge, MA: MIT Press, 2008, pp. 121–128.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation". In: *J. Mach. Learn. Res.* 3 (2003), pp. 993–1022. ISSN: 1532-4435.

[8] Regina P. Branton and Johanna Dunaway. "Slanted Newspaper Coverage of Immigration: The Importance of Economics and Geography". In: *Policy Studies Journal* 37.2 (2009), pp. 257–273. ISSN: 1541-0072.

[9] Jonathan Chang et al. "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio et al. 2009, pp. 288–296.

[10] Jilin Chen et al. "Diverse Topic Phrase Extraction from Text Collection". In: *WWW 2006*. Edinburgh, UK, 2006.

[11] DR Chittock et al. "Severity of illness and risk of death associated with pulmonary artery catheter use". In: *Critical Care Medicine* 32 (2004), pp. 911–915.

[12] AF Connors et al. "The effectiveness of right heart catheterization in the initial care of critically ill patients". In: *Journal of the American Medical Association* 276 (1996), pp. 889–897.

[13] Xinyu Dai et al. "SBA-term: Sparse Bilingual Association for Terms". In: *Fifth IEEE International Conference on Semantic Computing*. Palo Alto, CA, USA, 2011.

[14] JE Dalen. "The Pulmonary Artery Catheter—Friend, Foe, or Accomplice?" In: *Journal of the American Medical Association* 286 (2001), pp. 348–350.

[15] D D'Alessio and M Allen. "Media bias in presidential elections: a meta-analysis". In: *Journal of Communication* 50.4 (2000), pp. 133–156. ISSN: 1460-2466.

[16] Bryan E. Denham. "Hero or Hypocrite?" In: *International Review for the Sociology of Sport* 39.2 (2004), pp. 167–185.

[17] 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*. Washington, D.C.: National Academy Press, 1988, p. 91.

[18] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. "Safe feature elimination for the LASSO". In: (2011). Submitted, April 2011.

[19] Laurent El Ghaoui et al. "Sparse Machine Learning Methods for Understanding Large Text Corpora: Application to Flight Reports". In: *Conference on Intelligent Data Understanding*. Oct. 2011.

[20] S.E. Fienberg, J. Neter, and R.A. Leitch. "Estimating total overstatement error in accounting populations". In: *J. Am. Stat. Assoc.* 72 (1977), pp. 295–302.

[21] S Finfer and A Delaney. "Pulmonary artery catheters as currently used, do not benefit patients". In: *British Medical Journal* 333 (2006), pp. 930–1.

[22] R. A. Fisher. "The Arrangement of Field Experiments". In: *Journal of the Ministry of Agriculture of Great Britain* 33 (1926), pp. 503–513.

[23] George Forman. "An extensive empirical study of feature selection metrics for text classification". In: *J. Mach. Learn. Res.* 3 (2003), pp. 1289–1305. ISSN: 1532-4435.

[24] Eibe Frank et al. "Domain-specific keyphrase extraction". In: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*. California: Morgan Kaufmann, 1999, pp. 668–673.

[25] David A. Freedman. "On regression adjustments in experiments with several treatments". In: *The annals of applied statistics* 2.1 (2008), pp. 176–196.

[26] David A. Freedman. "On regression adjustments to experimental data". In: *Advances in Applied Mathematics* 40 (2008), pp. 180–193.

[27] Brian Gawalt et al. "Discovering word associations in news media via feature selection and sparse classication". In: *MIR '10*. Proceedings of the International Conference on Multimedia Information Retrieval. Philadelphia, Pennsylvania, USA, 2010, pp. 211–220.

[28] Alexander Genkin, David D Lewis, and David Madigan. "Large-Scale Bayesian Logistic Regression for Text Categorization". In: *Technometrics* 49.3 (2007), pp. 291–304.

[29] Martin Gilens and Craig Hertzman. "Corporate Ownership and News Bias: Newspaper Coverage of the 1996 Telecommunications Act". In: *The Journal of Politics* 62.02 (2000), pp. 369–386.

[30] Jade Goldstein et al. "Multi-document summarization by sentence extraction". In: *NAACL-ANLP 2000 Workshop on Automatic summarization*. 2000, pp. 40–48.

[31] J. L. Hall et al. "Implementing risk-limiting post-election audits in California". In: *Proc. 2009 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09)*. USENIX. Montreal, Canada, 2009.

[32] Erin Hartman et al. *From SATE to PATT: Combining Experimental with Observational Studies*. 2011.

[33] S. Harvey et al. "An assessment of the clinical effectiveness of pulmonary artery catherters in patient management in intensive care (PAC-Man): a randomized controlled trial". In: *Lancet* 366 (2005), pp. 472–77.

[34] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. unknown: Springer, 2003.

[35] Leonhard Hennig. "Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis". In: *Recent Advances in Natural Language Processing*. 2009.

[36] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30.

[37] Paul W. Holland. "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396 (1986), pp. 945–960.

[38] D. Holt and T. M. F. Smith. "Post Stratification". In: *J. R. Statistic Society* 142.1 (1979), pp. 33–46.

[39] Daniel Hopkins and Gary King. "A Method of Automated Nonparametric Content Analysis for Social Science". In: *American Journal of Political Science* 54.1 (2010), 229–247.

[40] Georgiana Ifrim, Gkhan Bakir, and Gerhard Weikum. "Fast Logistic Regression for Text Categorization with Variable-Length N-grams". In: *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008, pp. 354–362.

[41] Kosuke Imai. "Variance identification and efficiency analysis in randomized experiments under the matched-pair design". In: *Statistics in Medicine* 27 (2008), pp. 4857–4873.

[42] Kosuke Imai, Gary King, and Clayton Nall. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation". In: *Statistical Science* 24.1 (2009), pp. 29–53.

[43] Kosuke Imai, Gary King, and Elizabeth A. Stuart. "Misunderstandings between experimentalists and observationalists about causal inference". In: *J. R. Statistic Society* 171.Part 2 (2008), pp. 481–502.

[44] Guido W. Imbens. *Experimental Design for Unit and Cluster Randomized Trials*. 2011.

[45] Luke Keele, Corrine McConnaughy, and Ismail White. "Adjusting Experimental Data". In: *Experiments in Political Science*. 2009.

[46] G.G. Koch et al. "A review of some statistical methods for covariance analysis of categorical data". In: *Biometrics* (1982), pp. 563–595.

[47] G.G. Koch et al. "Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them". In: *Statistics in Medicine* 17.15-16 (1998), pp. 1863–1892.

[48] D. Lazer et al. "Life in the network: the coming age of computational social science". In: *Science (New York, NY)* 323.5915 (2009), p. 721.

[49] L. Lee and S. Chen. "New Methods for Text Categorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents". In: *Lecture Notes in Computer Science* 4031 (2006), p. 1280.

[50] Winston Lin. "Agnostic Notes on Regression Adjustment to Experimental Data". Working Paper. 2010.

[51] J. McCarthy et al. "Percentage-Based versus Statistical-Power-Based Vote Tabulation Audits". In: *The American Statistician* 62 (2008), pp. 11–16.

[52] Richard McHugh and John Matts. "Post-stratification in the randomized clinical trial". In: *Biometrics* 39 (1983), pp. 217–225.

[53] Joanne M. Miller and Jon A. Krosnick. "News Media Impact on the Ingredients of Presidential Evaluations: Politically Knowledgeable Citizens Are Guided by a Trusted Source". In: *American Journal of Political Science* 44.2 (2000), pp. 301–315.

[54] Luke W. Miratrix, Jasjeet S. Sekhon, and Bin Yu. "Adjusting treatment effect estimates by post-stratification in randomized experiments". In: *JRSS Series B (submitted)* (2012).

[55] Luke W. Miratrix and P.B. Stark. "Election Audits using a Trinomial Bound". In: *IEEE Transactions on Information Forensics and Security* 4 (2009), pp. 974–981.

[56] Luke W. Miratrix et al. *What is in the news on a subject: automatic and sparse summarization of large document corpora*. UC Berkeley Dept. of Statistics Technical Report #801. 2011.

[57] Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". In: *Political Analysis* 16.4 (2008), pp. 372–403.

[58] J. Neter, R.A. Leitch, and S.E. Fienberg. "Dollar unit sampling: Multinomial bounds for total overstatement and understatement errors". In: *The Accounting Review* 53 (1978), pp. 77–93.

[59] Joel Neto, Alex Freitas, and Celso Kaestner. "Automatic Text Summarization Using a Machine Learning Approach". In: *Advances in Artificial Intelligence* 2507 (2002), pp. 205–215.

[60] J. Neyman, K. Iwaszkiewicz, and S. Kolodziejczyk. "Statistical problems in agricultural experimentation (with Discussion)." In: *Supplement of Journal of the Royal Statistical Society* 2 (1935), 107 180.

[61] Erik C. Nisbet and Teresa A. Myers. "Challenging the State: Transnational TV and Political Identity in the Middle East". In: *Political Communication* 27.4 (2010), pp. 347–366.

[62] R. Plante, J. Neter, and R.A. Leitch. "Comparative performance of multinomial, cell and Stringer bounds". In: *Auditing: A Journal of Practice & Theory* 5 (1985), pp. 40–56.

[63] Stuart J. Pocock et al. "Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: curent practice and problems". In: *Statistics in Medicine* 21 (2002), pp. 2917–2930.

[64] Amy E. Potter. "Voodoo, Zombies, and Mermaids: U.S. newspaper coverage of Haiti". In: *Geographical Review* 99.2 (2009), pp. 208 –230.

[65] H. Pottker. "News and its communicative quality: the inverted pyramid–when and why did it appear?" In: *Journalism Studies* 4 (2003), 501–511(11).

[66] Charles S. Reichardt and Harray F. Gollob. "Justifying the use and increasing the power of a t Test for a randomized experiment with a convenience sample". In: *Psyschological Methods* 4.1 (1999), pp. 117–128.

[67] Stuart Rose et al. "Automatic keyword extraction from individual documents". In: *Text Mining: Applications and Theory*. Ed. by Michael W. Berry and Jacob Kogan. unknown: John Wiley and Sons, Ltd, 2010.

[68] D. B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5 (1974), p. 688.

[69] Y Sakr et al. "Sepsis Occurrence in Acutely Ill Patients Investigators. Use of the pulmonary artery catheter is not associated with worse outcome in the ICU". In: *Chest* 128.4 (2005), pp. 2722–31.

[70] G. Salton. "Developments in automatic text retrieval". In: *Science* 253.5023 (1991), pp. 974–980.

[71] G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval". In: *Information processing and management* 24.5 (1988), pp. 513–523.

[72] Carl-Erik Särndal, Bengt Swensson, and Jan H. Wretman. "The weighted residual technique for estimating the variance of the general regression estimator of the finite population total". In: *Biometrika* 76.3 (1989), pp. 527–537.

[73] J. S. Sekhon. "Opiates for the matches: Matching methods for causal inference". In: *Annual Review of Political Science* 12 (2009), pp. 487–508.

[74] Jasjeet S. Sekhon and Richard D. Grieve. "A Matching Method for Improving Covariate in Cost-Effectiveness Analysis". In: *Health Economics* (2011). Forthcoming.

[75] P. Senellart and V. D. Blondel. "Automatic Discovery of Similar Words". In: *Survey of Text Mining II*. Springer, 2008.

[76] J. Senn S. "Covariate imbalance and random allocation in clinical trials". In: *Statistics in Medicine* 8 (1989), pp. 467–475.

[77] R. J. Serfling. "Probability Inequalities for the sum in sampling without replacement". In: *The Annals of Statistics* 2.1 (1974), pp. 39–48.

[78] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9". In: *Statistical Science* 5.4 ([1923] 1990), pp. 465–472.

[79] P.B. Stark. "A sharper discrepancy measure for post-election audits". In: *Ann. Appl. Stat.* 2 (2008), pp. 982–985.

[80] P.B. Stark. "CAST: Canvass audits by sampling and testing". In: *IEEE Transactions on Information Forensics and Security, Special Issue on Electronic Voting* 4 (2009), pp. 708–717.

[81] P.B. Stark. "Conservative Statistical Post-Election Audits". In: *Ann. Appl. Stat.* 2 (2008), pp. 550–581.

[82] P.B. Stark. *Efficient post-election audits of multiple contests: 2009 California tests*. 2009 Conference on Empirical Legal Studies. 2009.

[83] P.B. Stark. "Risk-limiting post-election audits: P-values from common probability inequalities". In: *IEEE Transactions on Information Forensics and Security* 4 (2009), pp. 1005–1014.

[84] K.W. Stringer. "Practical aspects of statistical sampling in auditing". In: *Proceedings of the Business and Economic Statistics Section*. American Statistical Association. Washington, D.C., 1963, pp. 405–411.

[85] Student. "On Testing Varieties of Cereals". In: *Biometrika* 3/4 (1923), pp. 271–293.

[86] R. Sundberg. "Conditional statistical inference and quantification of relevance". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1 (2003), pp. 299–315.

[87] Anastasios A. Tsiatis et al. "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach". In: *Statistics in Medicine* 27 (2008), pp. 4658–4677.

[88] Otto E. Wahl, Amy Wood, and Renee Richards. "Newspaper Coverage of Mental Illness: Is It Changing?" In: *Psychiatric Rehabilitation Skills* 6(1) (2002), pp. 9–31.

[89] M. B. Wilk. "The Randomization Analysis of a Generalized Randomized Block Design". In: *Biometrika* 42.1/2 (1955), pp. 70–79.

[90] Y. Yang and I. O. Pendersen. "A comparative study on feature selection in text categorization". In: *ICML-97, 14th International Conference on Machine Learning*. Nashville, US, 1997, pp. 412–420.

[91] Tong Zhang and Frank J. Oles. "Text Categorization Based on Regularized Linear Classfiication Methods". In: *Information Retrieval* 4 (2001), pp. 5–31.