DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

School-Based Data Teams Ask the Darnedest Questions About Statistics: Three Essays in the Epistemology of Statistical Consulting and Teaching

Citation
Parker, Sean Stanley. 2014. School-Based Data Teams Ask the Darnedest Questions About Statistics: Three Essays in the Epistemology of Statistical Consulting and Teaching. Doctoral dissertation, Harvard Graduate School of Education.

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:13383545

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

School-Based Data Teams Ask the Darnedest Questions about Statistics: Three Essays in the Epistemology of Statistical Consulting and Teaching

Sean Parker

Catherine Z. Elgin
Terrence Tivnan
John B. Willett

A Thesis Presented to the Faculty
of the Graduate School of Education of Harvard University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

2014

©2014
Sean Parker
All Rights Reserved

Dedication

To my mother, Noreen Parker

To my father, Edward Parker

To my mother-in-love, Lucy Cooper

You have taught me by example to never stop learning and to never stop giving.

Thank you.

Acknowledgments

This dissertation would have been impossible without the help of my committee, Kate Elgin, Terry Tivnan, and John Willett, and my partner, Betsy Smith, who stuck by me when I was stuck. I am infinitely grateful to my committee. I am more so to my partner.

For proofreading and formatting, I thank Marc Baldwin and the staff of Edit911.

I must thank the data teams for which I have consulted and the students for whom I have taught. They are my inspiration.

Table of Contents

Page

Dedication i

Acknowledgments ii

List of Figures vi

Abstract vii

Chapter 1: How Not to Mislead with True Statistics 1

Data Teams 1

Reflective Equilibrium 3

Accidentally Misleading with True Statistics 10

Some Examples of False Implicature 10

An Introduction to Implicature 13

The Gricean Maxims 18

The Value of Not Misleading in the Context of Data Teams 24

Accuracy, Sincerity, and Conversational Competence as Instrumental Goods 25

Free-Rider Problems of Accuracy, Sincerity, and Conversational Competence 29

Accuracy, Sincerity, and Conversational Competence as Intrinsic Goods 33

Recommendations 38

Relevance 39

Truth 41

Evidential Adequacy 44

Informational Moderation 47

Perspicuity 48

Conclusion 50

Chapter 2: The Sample Mean and the Sample Median as Exemplars 52

Exemplification 54

The Possession Criterion of Exemplification 55

The Reference-by-Possession Criterion of Exemplification 57

The Elucidation Criterion of Exemplification 57

Attenuated vs. Replete Exemplification 58

Attenuated Exemplars in Science 60

Attenuated Exemplars at the Intersection of Science and Policymaking 61

Replete Exemplars in Science 63

Replete Exemplars in Policymaking 64

Moderate vs. Extreme Exemplification 65

Ordinal Schemes and Labeling 65

Ordinal Schemes and Exemplification 67

The Median, Rationally Reconstructed as an Attenuated, Moderate Exemplar 69

What is a Rational Reconstruction? 69

Objectively Comparing Values within a Sample 70

Objectively Comparing Values Between Subsamples 72

Outliers vs. Medians as Bases of Comparison 73

Fictive vs. Non-Fictive Exemplification 77

The Mean as a Fictive Representation 79

The Mean as an Exemplar Representation 82

The Mean as an Elucidating Exemplar 84

Elucidating Sample Values Is Elucidating the World....89

Chapter 3: Why Infinite Populations for Statistical Inference?....95

What Is Sampling Error? 106

A Process-Minded Perspective on Sampling Error 106

A Population-Minded Perspective on Sampling Error 110

Target Counterfactuals and Method Counterfactuals 114

What Are Explanations? 123

Hempel's Account of Explanation....123

Lawlike vs. Accidental Generalizations 125

Sampling Error as an Explanation 130

What Are Populations? 132

Reference and Populations....133

Evaluating Claims about Largely Fictive Populations....138

Why Be Population-Minded? 150

References....154

Vita....158

List of Figures

Figure 1. A table for categorizing possible subjects into three types 97

Figure 2. A table of distinctions in this essay 105

Figure 3. A sampling distribution of proportions from all possible samples ($n = 3$) of actual coin flips ($N = 5$) 119

Figure 4. A sampling distribution of proportions from all possible samples ($n = 3$) of possible coin flips 120

Figure 5. The real-number line upon which sample-mean differences in baseline GPA by program assignment are placed 146

Figure 6. For baseline GPA by program assignment, the sample-mean difference from the factual sample 146

Figure 7. For baseline GPA by program assignment, sample-mean differences from 136 counterfactual samples and the factual sample 147

Figure 8. Estimated program effect sizes from 137 counterfactual samples and the factual sample 147

Abstract

School-Based Data Teams Ask the Darnedest Questions about Statistics: Three Essays in the Epistemology of Statistical Consulting and Teaching

Sean Parker, Doctoral Student

The essays in this thesis attempt to answer the most difficult questions that I have faced as a teacher and consultant for school-based data teams. When we report statistics to our fellow educators, what do we say and what do we leave unsaid? What do averages mean when no student is average? Why do we treat our population of students as infinite when we test for statistical significance? I treat these as important philosophical questions.

In the first essay, I use Paul Grice's philosophical analysis of conversational logic to understand how data teams can accidentally mislead with true statistics, and I use Bernard Williams's philosophical analysis of truthfulness to understand the value, for data teams, of not misleading with statistics. In short, statistical reports can be misleading when they violate the Gricean maxims of conversation (e.g., "be relevant," "be orderly"). I argue that, for data teams, adhering to the Gricean maxims is an intrinsic value, alongside Williams's intrinsic values of Sincerity and Accuracy. I conclude with some recommendations for school-based data teams.

In the second essay, I build on Nelson Goodman and Catherine Z. Elgin's analyses of exemplification to argue that averages (i.e., medians and means) are attenuated, moderate, and sometimes fictive exemplars. As such, medians and means lend themselves to scientific objectivity.

In the third essay, I use Goodman's theory of counterfactuals and Carl Hempel's theory of explanation to articulate why data teams should make statistical inferences to infinite populations that include possible but not actual students. Data teams are generally concerned that their results are explainable by random chance. Random chance, as an explanation, implies lawlike generalizations, which in turn imply counterfactual claims about possible but not actual subjects. By statistically inferring to an infinite population of students, data teams can evaluate those counterfactual claims in order to assess the plausibility of random chance as an explanation for their findings.

Chapter 1: How Not To Mislead with True Statistics

School-based data teams work hard to report true statistics, because the reports of these data teams influence school policy and affect children's lives. True statistics, however, can be misleading, so data teams must work hard not only to report true statistics, but to ensure that they are reporting non-misleading statistics. In this essay, I use Paul Grice's philosophical analysis of conversational logic to understand how data teams can accidentally mislead with true statistics, and I use Bernard Williams's philosophical analysis of truthfulness to understand the value, for data teams, of not misleading with statistics. I conclude with a set of recommendations for data teams, using Grice's maxims to frame my recommendations. In order to frame my research questions, I begin with a brief background discussion of data teams. To answer my research questions, I use the philosophical method of reflective equilibrium, which I will discuss briefly.

Data Teams

Schools across the nation are increasingly forming data teams of teachers and administrators (Boudett & Steele, 2007; Kerr, 2006; Scherer, 2008), with the purpose of better understanding the teaching and learning that are taking place in their schools through the analysis of quantitative educational data on their students. Most schools have abundant quantitative educational data (e.g., demographic, assessment, enrollment, and disciplinary data), but few schools have faculty who possess the quantitative skills to analyze those data productively, leading Ronka, Lachat, Slaughter, and Meltzer (2008) to describe schools as "data rich, but information poor" (p. 8). Data teams strive to turn *data*

into *information*, and then to communicate that information to diverse educational stakeholders.

Boudett, City, and Murnane (2005) recommended a step-by-step, collaborative process for school-based data teams of diverse faculty who have no previous instruction in data analysis. One major function of the data team process is to provide necessary instruction in statistics, so as to help teams to interpret data meaningfully by providing tools for summary and inference. Data teams can receive statistical instruction and then, in turn, become statistical instructors for their school communities with the goal of fostering a culture of analytic inquiry. Data teams generate statistical reports for their school communities and, in the process of presenting those reports, are tasked not only with describing the statistics but also with teaching how to interpret them. Statistical reports can be very informative, but they can also be very misleading. This reality provides the basis for my first research question: How can data teams accidentally mislead with true statistics?

Murnane, Sharkey, and Boudette (2005) argued that data teams use educational data in three ways: *instrumentally*, *symbolically,* and *conceptually*. They use data *instrumentally* to detect signals for placement, promotion, or graduation; they use data *symbolically* as a rhetorical device to promote educational agendas to which they are already committed; and they use data *conceptually* when they analyze the data to gain insight into the teaching and learning process for the sake of school improvement. Citing Deming's (2000) research into improving quality control, Murnane et al. (2005) advocated for the *conceptual* use of data. In this essay, I accept their recommendation and use it to drive my own focus on the conceptual use of data and, consequently, on

statistical reporting conducive to conceptual work, because school improvement is the ultimate goal of data teams. If non-misleading statistics are beneficial for school improvement, then they can have great value. Thus, my second research question asks: What is the value, for data teams, of not misleading with statistics? In other words, how does a data team's non-misleadingness fit into the value system of the school?

Reflective Equilibrium

I address my research questions using the philosophical method of *reflective equilibrium* (Daniels, 2011; Elgin, 1999, 2001; Goodman, 1983). Philosophers use this method to reconcile diverse commitments to facts, principles, standards, methods, categories, goals, and values. A *commitment* is a belief or statement that we may use as a basis for decision making. In this essay, I apply the method of reflective equilibrium in order to reconcile data team commitments, statistical commitments, and philosophical commitments. The answer to each of my research questions is itself a commitment. According to the method of reflective equilibrium, a good answer to a philosophical research question is one that contributes to a *mutually supportive* system of commitments *grounded in practice*. A system of commitments is mutually supportive when each commitment in the system is reasonable in light of the others; the system is grounded in practice when it is reasonable in light of commitments accepted previously by any participating practitioner (Elgin, 1999). The initial commitments need not be scientifically objective or unbiased; they serve simply as starting points for the inquiry. The method of reflective equilibrium is a process of inquiry through which philosophers work toward objectivity and freedom from bias by balancing commitments to increase mutual supportiveness. But the process must start with some set of commitments, and

there is no better place to start than with the commitments of practitioners. The method of reflective equilibrium yields justification because it starts with the commitments of practitioners, which are the best available commitments. As a result, when we assimilate and accommodate our commitments in order to achieve a more mutually supportive system, the improved set of commitments is grounded in the best available starting point and is therefore reasonable. As a practitioner of data teams, statistics, and philosophy, I can and should use my own set of commitments as a starting point for my philosophical essay.

Reflective equilibrium in the context of data teams involves laying out a set of inquiry-starting commitments to facts, principles, standards, methods, categories, goals, and values. Consider, for example, the following set of data team commitments:

- **Facts**: In our sample of middle-school students, the mean grade-point average (GPA) is half a point lower for students who are eligible for free lunch than for their ineligible counterparts, and half a point higher for students who participated in a particular program than for students who did not participate in that program.
- **Principles**: Differences in group averages due solely to sampling error are irrelevant to policy.
- **Standards**: The appropriate alpha level is .05.
- **Methods**: Null-hypothesis significance testing, for practical purposes, can rule out sampling error as the sole explanation of observed differences in group averages.
- **Categories**: A difference in population averages between students who are and are not eligible for free lunch on a valid measure of important skills and concepts

will be categorized as an *SES* (i.e., socioeconomic status) *achievement gap*. A difference in population averages between program participants and non-participants with regard to an outcome will be categorized as a *program effect* if program participation was randomized.

- **Goals**: Identify SES achievement gaps. Identify program effects. Build a culture of data-analytic inquiry.
- **Values**: Identifying, and ultimately closing, SES achievement gaps is necessary for school improvement. We should "recognize the uniqueness and dignity of individuals of differing races, religions, classes, ethnicities, sexual orientations, learning styles and abilities" (Newton Public Schools, 2008).

Once we lay out a set of inquiry-starting commitments, we ask: Which ones are mutually supportive? Which ones are mutually inconsistent or otherwise in tension with one another? If we find tensions, we adapt our commitments. We may reassess a fact, adapt a method, adjust a value, or simply eliminate a commitment that conflicts with others (Goodman, 1983). We do what we must in order to minimize tensions and maximize mutual supportiveness, and in doing so we evolve a new set of commitments, but this new set is grounded in practice because it has evolved from the inquiry-starting commitments that were drawn from practice. Even when we completely reject an inquiry-starting commitment, we consider why it seemed tenable to practitioners. A set of currently tenable statements of facts, principles, standards, methods, categories, goals, and values is in reflective equilibrium when the members of that set are not only mutually supportive, upon consideration, but also plausible in light of the inquiry-starting commitments. This process is iterative, because we can (and should) always be adding

new items to our set of commitments insofar as they are held by practitioners (e.g., practitioners of data teams, of statistics, or of epistemology). The aspiration is to attain a *wide* reflective equilibrium that considers *all* relevant facts, principles, standards, methods, categories, goals, and values held by practitioners (Elgin, 2001).

Reflection can reveal mutual inconsistencies within a set of inquiry-starting commitments. For example, the seemingly relevant fact of a half-point difference in sample means may be deemed irrelevant in light of three other commitments: the principle that differences due to sampling error are meaningless, the standard of a .05 alpha level, and the method of null-hypothesis significance testing. If so, a data team can move toward equilibrium by eliminating that fact from the set of relevant commitments. Relegating a particular fact to irrelevance is not the only option in light of mutual inconsistency, because no member of the set is beyond criticism. Perhaps, for instance, the alpha level of .05 is too stringent a standard, and .10 will better serve the data team's goal of identifying SES achievement gaps and identifying program effects.

Ultimately, we must make the adjustments that maximize mutual supportiveness within the current set of commitments, but we must also keep in mind future and past sets of commitments. We must consider future sets of commitments because our set is ever widening as we add new commitments in our striving for wide reflective equilibrium. If we modify our alpha level, are we doing so universally, or only contingently based on contextual cues? Does this modification promise to be mutually supportive in the future? Of course, whether or not we revise a commitment now, it, like all commitments, remains open to revision in the future. We must also keep in mind the past because, although we left our inquiry-starting commitments behind in favor of our current set of commitments,

our inquiry-starting commitments were tethered directly to practice and we must link our current set of commitments with the initial ones in order to retain the connection with practice. Suppose that we modify our alpha level from .05 to .10; it is still important to ask why .05 seemed tenable in the first place. Perhaps it was a misunderstanding, and if so that would be important information to know so that we can avoid such misunderstandings in future practice. Perhaps there were legitimate considerations that lent tenability to an alpha level of .05, and we need to remember them because they may still be relevant at our new alpha level as well. In the next and final section of this introduction, in addition to presenting overviews of each chapter, I will sketch the inquiry-starting commitments that I will proceed to reflectively equilibrate in each chapter.

In this essay, I do not argue against deeply entrenched principles and values. I do not dispute the principle that correlation does not imply causation, or the value of recognizing the uniqueness and dignity of individuals of differing races, religions, classes, ethnicities, sexual orientations, learning styles, and abilities. This stance is consistent with Quine's (1951) Principle of Minimal Mutilation, which dictates that, when striving for reflective equilibrium, we should tinker with less entrenched commitments before we tinker with core commitments. A common misconception about philosophy is that philosophers question everything all the time. Indeed, some philosophers do question everything some of the time, such as Descartes (1641/1907) in the first meditation of his Mediations on First Philosophy. I, however, am not doing philosophy from my armchair. I am starting in the thick of things as a data team consultant.

Data teams hold diverse commitments ranging from scientific principles to school values. In order to avoid self-contradiction and move toward mutual supportiveness, those diverse commitments must be integrated. Such extensive integration calls for an epistemologist, because epistemologists are trained to work with commitments of unbounded variety. Statisticians reconcile statistical commitments, teachers reconcile pedagogical commitments, lawyers reconcile legal commitments, leaders reconcile strategic commitments, and administrators reconcile bureaucratic commitments, but epistemologists reconcile *all* commitments. The presence of the epistemologist does not make redundant the statistician, the teacher, the lawyer, the leader, or the administrator, because expertise matters. In fact, I draw on my expertise as a data team statistician and team leader to deepen my epistemology, but my epistemological work ranges over and beyond statistics and leadership to encompass a variety of relevant and tenable commitments sufficient for a deep philosophical inquiry.

One can spur deep philosophical inquiry while starting from a handful of seemingly unobjectionable commitments. This handful of commitments is not a random sample or a systematic sample. Rather, the commitments are hand-picked to elucidate problems, clarify questions, and stimulate thought. At the Harvard Divinity School, the following two commitments may be sufficient to spur a philosophical essay: (a) God is good, and (b) evil exists. The divinity-school student would show how the two commitments appear to conflict and would then reconcile them by modifying one or the other or by adding bridging commitments. Since I am in the School of Education, I cull my commitments not from church fathers but from data teams, grounding my epistemological work in current practice, analyzing commitments that are relevant and

initially tenable to me as a data team member right now (Elgin, 1999). I carefully select data team commitments that clarify my research questions.

To resolve conflicts and strengthen mutual supportiveness among data team commitments, I draw on insights from the epistemological literature. My questions are about reporting statistics in the *data team context*. I define this context as the set of commitments that for me as a team member are relevant and tenable, including statistical commitments. By situating statistical commitments as a subset of data team commitments, I can integrate those commitments via the method of reflective equilibrium. This integration involves both assimilation and accommodation. I assimilate by adapting the statistical commitments so that they align with the broader set of data team commitments, but I also accommodate by adapting the broader set of commitments to align with the statistical commitments. This yields a refined set of commitments, which includes answers to my questions about reporting statistics in the data team context.

Some of the answers to my questions may be context-dependent, such that they apply particularly to school-based data teams. Other answers may apply to a broad variety of teams that use statistical models. I have served as a statistical consultant not only for school-based data teams but also for hospital-based data teams and developmental laboratories, and I believe that this essay can inform statistical consulting across contexts. But when context matters, I focus on school-based data teams, because, in my experience, school-based data teams are the least acculturated to quantitative data. As Kuhn (1996) argued, philosophy is needed most when the rules of the scientific game are least established.

Accidentally Misleading with True Statistics

In this section, I ask the question: How can data teams accidentally mislead with true statistics? To answer this question, I use Paul Grice's philosophy of conversational implicature. I argue that data teams can accidentally mislead with true statistics by unintentionally generating false implicatures. I begin with some examples of false implicatures. I then formally define the term *implicature*, discuss the Gricean maxims of conversation from which implicatures derive, and explain how data teams can generate false implicatures by violating the Gricean maxims.

Some Examples of False Implicature

The truth can be misleading. To illustrate, I will begin with a true statement. Then I will add some context and consider reasonable inferences from the true statement in conjunction with the context. It will turn out, however, that the reasonable inferences are false.

A data team is called upon by the school committee to give evidence for the effectiveness of a program. The data team reports, "On average, students who participated in the program achieved statistically significantly higher GPAs than students who did not participate in the program." The school committee reasonably infers from the data team's report that the program *caused* the difference in scores; after all, reporting on the causal impact of the program was the sole purpose of the data team's report, so there would have been no other reason to highlight this finding. In actuality, however, the data team has no knowledge about the causal impact of the program, because the study was observational, not experimental or even quasi-experimental.

The school committee also reasonably infers that the data team used customary scientific standards of evidence. The school committee has only a rough understanding of the term "statistically significantly," but it rightly understands that this is a stamp of scientific approval of the quality of evidence. (The school committee probably does not understand that it is a stamp of very limited approval, strictly with respect to ruling out sampling error as the sole explanation for the finding.) For testing of statistical significance, the customary alpha level is 0.05, so the school committee could reasonably infer, despite the lack of an explicit statement, that the data team used an alpha level of 0.05. In fact, after careful deliberation, the data team decided to use an alpha level of 0.20. The data team understood that the alpha level is a measure of tolerance for false positives due to sampling error. Based on the low cost of the program and its supposedly immeasurable benefits, the data team decided that to select a looser standard than the customary scientific standard of tolerance.

Furthermore, the school committee reasonably infers that the data team found no evidence against the program. The data team reported only evidence in support of the program. Presumably, if the data team had found evidence against the program, then the data team would have reported that evidence as well. Actually, however, the data team did not report the finding that, on average, students who participated in the program were statistically significantly less likely to pursue elective courses in music, drama, and the visual arts.

Finally, the school committee reasonably infers that the finding is "significant" and so warrants program implementation. The school committee is making the common mistake of confusing statistical significance with policy significance. From an analytical

perspective, the school committee's inference is not reasonable, because it is based on a common error of interpretation. What makes the inference reasonable from the perspective of the school committee, however, is that the data team has implicitly endorsed this misconception. The school committee is treating the data team as data *educators*, not just as researchers. That is, the school committee is trusting that the data team will disabuse the school committee of common misconceptions. Because the data team did not disabuse the school committee of its misconception, the committee assumes that it does not have any misconceptions.

At this point, the reader may be contemplating whether the data team in this instance is more likely to be incompetent or insincere. For this project, I focus on the problem of incompetence. The fields of statistics and philosophy already have established literatures on the insincere use of true statements for the purpose of misleading. With his *How To Lie with Statistics*, Huff (1954) kicked off a series of works demonstrating that true statistics can be used in purposefully misleading ways. Dynel (2011) reviewed a line of philosophical literature starting with Grice's (1989) pragmatic framework of conversational implicature, which explains how true statements can mislead by falsely implicating. There is a debate in the philosophical literature over whether the definition of lying should encompass falsely implicating. From the Gricean framework, Adler (1997) argued that lying is distinct from (and morally worse than) falsely implicating, but from the same framework Meibauer (2005) contended that falsely implicating should be included within a general definition of lying. Both Adler and Meibauer agreed, nevertheless, that falsely implicating for the purpose of misleading is generally immoral.

In Chapter 5 of *Truth and Truthfulness*, Bernard Williams (2002) also used a Gricean framework to criticize the insincere use of true statements for the purpose of misleading. I agree with the philosophers, especially Williams (on whose arguments I will build in the next section), that purposefully misleading is bad, but my focus is on accidentally misleading, which is also bad though surely not in the same way. I find the philosophers' analyses very helpful to my analysis, because they focus on the deviousness of twisting language and context to arrive at false implicatures. I focus on the difficulty of unraveling language and context to get true implicatures.

An Introduction to Implicature

Although the context in which data teams perform their work can be tricky, true implicatures are often easy to draw, and we regularly infer them with little to no conscious effort. A conversational implicature is a pragmatic inference from what is said by a speaker in conjunction with the shared assumptions of the speaker and audience. Following is an example of conversational implicature, adapted from Grice (1989, p. 32). Consider this exchange between a tourist and a passerby:

The tourist says, "My car is out of gas."

The passerby replies, "There is a garage around the corner."

The tourist pragmatically infers that the garage can solve his gas problem. According to Grice (1989, p. 31), the tourist's chain of inference is as follows:

1. The passerby said that there is a garage around the corner.
2. The tourist has no reason to believe that the passerby is not being helpful.
3. The passerby would not have said there is a garage around the corner unless he thought that the garage can solve the tourist's gas problem.

4. The passerby knows that the tourist will suppose that the passerby thinks that the garage can solve the tourist's gas problem. Furthermore, the passerby knows that the tourist knows that he knows. In other words, the passerby knows that, with mutual awareness, they are on the same page.
5. The passerby has done nothing to stop the tourist from thinking that the garage can solve his gas problem.
6. The passerby intends the tourist to think, or is at least willing to allow the tourist to think, that the garage can solve his gas problem.
7. Therefore, the passerby has implicated that the garage can solve the tourist's gas problem.

A conversational implicature depends on the shared beliefs of the speaker and the audience; therefore, it is not purely a deductive inference from what has been stated explicitly. "There is a garage around the corner" deductively entails "if the garage can solve his gas problem, then the solution to his gas problem is just around the corner." The deductive entailment is only an "if-then" statement with nothing more about the "if." The conversational implicature imports contextual information so as to permit a bolder conclusion than the deductive entailment. Through conversational implicature, the tourist concludes that the garage must be able to solve his gas problem, and that therefore the solution to his gas problem is just around the corner.

Conversational implicature allows us to communicate more than we actually say. It is incalculably useful for every human endeavor that requires information sharing. Once in a while, we must make our otherwise implicit assumptions explicit. Those instances usually begin with a deep breath and the idiom, "Let me spell it out for you."

Those instances, however, are the exception, not the rule. Even in those instances, we spell out only some of the assumptions. It would be impossible to get anything done if everything had to be spelled out completely. Thus conversational implicature is a necessary and powerful tool for communication, but it is also susceptible to abuse. Moreover, and more directly relevant to my point, since conversational implicature depends on the context of shared assumptions, it can be tricky (but no less powerful and necessary) in contexts such as that of a data team report, where the base of shared assumptions is still under construction.

The Gricean framework of conversational implicature begins with the *Cooperative Principle*: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (1989, p. 26). In short, the participants in a conversation should be mutually helpful when contributing to the conversation. Grice uses the term "conversation" broadly to encompass much more than face-to-face, back-and-forth discussion; it also refers to written reports, short answers, and classroom instruction. In general, for Grice and for this thesis, a conversation is any use of language for the sake of conveying information. Thus, the *Cooperative Principle* tells us to be helpful when using language to convey information.

Because a helpful contribution at one stage of a conversation may not be so helpful at another stage, the *Cooperative Principle* explicitly anchors the contribution to the stage at which the contribution occurs. This is important, because the context can change from one stage of a conversation to another, and context is essential for conversational implicature. Consider a conversation between two statistically savvy data

team members, Jen and Joe. Jen says to Joe, "The difference in average GPAs is statistically significant at the 0.05 level." Walter, a statistical novice, joins the conversation, so Jen rephrases her statement in terms that Walter will not misinterpret: "We used scientifically rigorous methods to determine that our observed achievement gap in GPAs is *not* merely an accident of sampling such that one group of students just happened to earn higher grades on average this year." As Walter learns the lingo, perhaps over the course of the year, Jen moves back to her earlier locution, "The difference in average GPAs is statistically significant at the 0.05 level."

When there is a mutually agreed-upon purpose or direction for a conversation, that agreement also provides essential context. In general, whenever we are working together for a common goal, we have reasonable expectations of one another. This is true for all types of work. Grice illustrates this point with scenarios in which a mechanic or baker asks a helper for a tool. Similarly (because I prefer classroom illustrations), consider a scenario in which a teacher and her helper are decorating a bulletin board and some construction paper requires trimming. While holding the paper in place, the teacher asks the helper for scissors. When the helper returns from rummaging through the teacher's drawer, the teacher has reasonable expectations about what the helper will deliver to the teacher's outstretched free hand. The teacher reasonably expects that the helper will deliver the paper scissors rather than the pinking shears or pruning clippers. The teacher also reasonably expects to receive functional scissors, not broken scissors. Furthermore, the teacher reasonably expects that the helper will deliver one whole pair of scissors, not two pairs or half a pair. Finally, the teacher reasonably expects that the helper will deliver them gently and securely, not stab the teacher while making the

handoff. These are all reasonable expectations of a competent helper, and Grice formalizes these expectations with his maxims.

Before I discuss Grice's conversational maxims, I want to foreshadow my argument about conversational competence. If the teacher's helper is a young student, competence may not be a reasonable expectation. The child, familiar with child scissors, may be confused by the adult scissors in the drawer—paper and pinking and pruning, oh my! The child may not realize that some scissors may be broken or may be unable to distinguish functional from broken scissors. The child might hedge her bets by delivering multiple scissors, not understanding that the teacher cannot handle multiple items with only one free hand. The child may not know how to deliver scissors safely, or (more likely, considering the heavy emphasis on scissors safety in early education) the child may not have the impulse control to deliver the scissors safely. It is, after all, very exciting to be the teacher's big helper! Helping the teacher can provide many teachable moments. Such competence requires development. Such competence can be hard. If it is not clear already, it will soon be clear enough that the teacher's helper is an apt metaphor, I think, for data teams and their consultants alike. Lest this metaphor be interpreted as condescending, I appeal to the reader's deep respect for lifelong learning. We were learners in elementary school, and we are learners now.

The data team that presents correlational findings when causal findings are demanded is like the child who presents pinking shears instead of paper scissors. The data team that does not check the quality of its findings by using tests for statistical significance is like the child who does not check to see if the scissors are broken. The data team that overwhelms its audience with information is like the child who hands over

a large handful of scissors. The data team that uses hot jargon without concern for the dangers of misinterpretation is like the child who runs excitedly and dangerously with scissors in hand.

The Gricean Maxims

Data teams can accidentally mislead with true statistics by unintentionally violating the Gricean maxims. Grice enumerates four categories of conversational maxims that follow from the *Cooperative Principle*: the *Maxim of Relevance*, the *Maxims of Quality*, the *Maxims of Quantity*, and the *Maxims of Manner*. The Maxim of Relevance states simply, "Be relevant" (Grice, 1989, p. 27). The Maxims of Quality state, "Try to make your contribution one that is true," and therefore, "1. Do not say what you believe to be false. 2. Do not say that for which you lack adequate evidence" (Grice, 1989, p. 26). The Maxims of Quantity state, "1. Make your contribution as informative as is required (for the current purposes of the exchange). 2. Do not make your information more informative than is required" (Grice, 1989, p. 26). Finally, the Maxims of Manner state, "Be perspicuous," with the following implications: "1. Avoid obscurity of expression. 2. Avoid ambiguity. 3. Be brief (avoid unnecessary prolixity). 4. Be orderly" (Grice, 1989, p. 26). Grice does not contend that his list of conversational maxims is exhaustive, but he proposes it as a good start. An exhaustive list of conversational maxims would be such that, if all participants in a conversation were to follow them, then the participants could generate many true conversational implicatures but would generate no false conversational implicatures.

A speaker in a conversation can either adhere to the maxims or not. Grice discusses four types of non-adherence: *violating*, *opting out*, *clashing*, and *flouting*. The

first type of non-adherence, violating, is liable to mislead because, by definition, the audience is unaware of the non-adherence. The violating of conversational maxims generally stems from incompetence and insincerity, so I will come back to it later after I have discussed competent and sincere non-adherence. In contrast to violating, the other types of non-adherence—opting out, clashing, and flouting—depend on the audience being aware of the non-adherence, and they often appear in competent and sincere conversation. A speaker who engages in opting out makes it clear that she is not adhering to a maxim. A data team, for example, might opt out of the Maxim of Relevance by saying, "We know that an observed correlation falls short of direct relevance to your causal question for these reasons, but it is the best we can do for now." A team might opt out of a Maxim of Quality by saying, "Standards of adequacy for statistical evidence do vary. The standard we adopted is looser than the conventional scientific standard. By 'standards of statistical adequacy,' we mean … . We settled on our standard for these reasons … . We understand that your standards are tighter than ours, but we think that, if you hear us out, you will come to see it our way." A team might opt out of a Maxim of Quantity by saying, "We are omitting an important piece of evidence until we work out some methodological issues and confidentiality issues." Finally, a team might opt out of a Maxim of Manner by saying, " 'Statistical significance' is an obscure expression. Most people, perhaps most researchers, deeply misunderstand the concept. Nevertheless, we will, with all necessary care, use the expression. The right way to understand 'statistical significance' is as follows."

Clashing, the second type of obvious non-adherence to the maxims of conversation, occurs when two maxims clash in such a way that the speaker cannot

adhere to both, but—and this is crucial—the audience recognizes the clash. If a data team has observed only correlations but is asked to discuss causal relationships, then it cannot adhere to both the Maxim of Relevance, which demands direct relevance, and the Maxims of Quality, which demand adequate evidence. In order to make a directly relevant report, the data team would have to make a causal statement for which it does not have adequate evidence. Now, consider a methodologically sophisticated audience of people who know that an estimated correlation cannot directly address a causal question. Furthermore, this audience knows the standards of adequate evidence for causal inference. When the data team reports a sample correlation in answer to a causal question, the methodologically sophisticated audience reasonably infers that the data team does not have adequate evidence to support a directly relevant statement of causal impact. This audience immediately grasps the clash between the two maxims, relevance and quality. The data team can thus rely on that understanding to communicate without fear of misleading. On the other hand, a methodologically naïve audience will not understand the clash and, consequently, will likely be misled into believing that the report was directly relevant to the question and was supported by adequate evidence. Thus, in the sophisticated context, the non-adherence to the conversational maxims is a *clash*, but in the naïve context it constitutes a *violation*. The difference lies in whether or not the non-adherence is obvious to the audience. I will have more to say about clashing in a later section of this chapter where I discuss useful fictions, which I characterize as mediating the clash between the Maxim of Quality, "Try to make your contribution one that is true," and the Maxims of Manner, such as "Be perspicuous."

Flouting is the third type of obvious non-adherence to the maxims of conversation. If the conversation becomes too tense, a completely irrelevant joke may be in order: “Did you hear the one about the statistician? Probably.” Sometimes, we make patently untrue statements to get a reaction: “Low academic achievement is associated with eligibility for free lunch, so we should charge more for lunch.” Tautologies may be totally uninformative, but they can bring flights of fancy back down to the ground: “An average is an average.” I have found that most audiences take delight in small doses of wildly obscure jargon: “That pattern you just described is called ‘heteroscedasticity.’ ” Flouting can provide a mini-vacation from one or more conversational maxims, but it works only if everyone is on board. Grice characterizes metaphors as flouting a Maxim of Quality, because they are literally false, but I think that metaphors are adherent to the Maxim of Quality when they are figuratively true (and when there is adequate evidence for their figurative truth). For example, a useful pedagogical strategy for teaching the arithmetic mean is to describe the arithmetic mean as the fulcrum point of a balanced see-saw. Of course, the audience must be attuned to the figurative truth of the metaphor, but to achieve that goal one must simply avoid obscure and ambiguous metaphors, a requirement that can already be deduced from the Maxims of Manner. To use the language of that category of maxims, metaphors can shine in their brevity and orderliness.

Aside from instances of opting out, clashing, flouting, or perhaps some other form of obvious non-adherence that Grice overlooked, we have reasonable expectations that a sincere, competent speaker is adhering to the maxims of conversation. Therefore, we can make reasonable inferences from the speaker’s statements in conjunction with the context

by relying on the maxims of conversation. Such reasonable inferences are called *conversational implicatures*. An insincere speaker, however, can exploit the logical chain that links statements, context, and maxims and leads to implicatures. The insincere speaker can be surreptitiously uncooperative. The insincere speaker can lie and thereby violate the Maxim of Quality. Perhaps more deviously, the insincere speaker can mislead by telling truths but violating other maxims. I think that Huff's (1954) *How To Lie with Statistics* could be effectively reorganized in terms of violated maxims. His chapter "Much Ado about Practically Nothing," could be filed under violations of the Maxim of Relevance, since it consists primarily of attacks on differences that prove to be irrelevant due to measurement error and sampling error. Similarly, the chapter called "The Sample with the Built-in Bias" describes violations of the Maxims of Quality, as it exposes the use of biased samples that do not supply adequate evidence. The chapter "The Little Figures That Are Not There" discusses the omission of important information, which constitutes a violation of the Maxims of Quantity. And several chapters—"The Gee-Whiz Graph," "The One-Dimensional Picture" and "The Semi-Attached Figure"—are related to violations of the Maxims of Manner, as Huff attacks data-analytic graphs for being disorganized, obfuscating, distracting, and generally imperspicuous.

Huff calls statistics "lies" when they are purposefully misleading (via false implicature from violated maxims). Unlike Huff, Adler (1997) distinguishes between lying and falsely implicating, but he also condemns the practice of purposeful misleading through false implicature. Alder eloquently describes the victim's frustration at feeling complicit in one's own deception. After all, in cases of false implicature, it is the victim's own flawed reasoning that leads him or her to the wrong conclusion. I would add that, in

addition to this tendency to cause victims to blame themselves for being misled, falsely implicating (as opposed to lying) closes certain avenues—such as legal recourse—for holding the perpetrator accountable, since he or she did not make any false statements *per se*. I suppose that this sense of self-infliction without suitable resource makes the use of true but misleading statistics worse than damnable. I am reminded of the quip that Mark Twain attributed to Disraeli: "There are three kinds of lies: lies, damned lies, and statistics." As a statistical consultant to data teams, I regularly hear this and similar comments from skeptical teachers. Indeed, those expressions of skepticism have formed the inspiration for this essay.

True statistics can also mislead unintentionally. This phenomenon is my primary interest here. After all, I believe that, in general, data teams sincerely intend to lead their audiences to true understandings. In this essay—and in my professional work—I proceed under this assumption. I would be remiss, however, if I neglected to mention that data teams can fall under pressure, external and internal, to justify educational programs and systems. It is not farfetched that an administrator or official would ask the data team for numbers that "make the program/department/school/district look good" for the purposes of securing financial funding or good will. Or the data team may have its own reasons to make the program, department, school, or district look good. This is what Murnane, Sharkey, and Boudette (2005) had in mind when they stated that data teams sometimes use educational data *symbolically*—that is, when they mobilize data not to determine whether an aspect of the educational system is working but to promote educational agendas to which they are already committed. As noted earlier, Murnane et al. called for using data *conceptually*, to gain insight into the teaching and learning process for the sake

of school improvement. It is my contention that, to use data conceptually, data teams must strive for accuracy, sincerity, and conversational competence.

The Value of Not Misleading in the Context of Data Teams

In *Truth and Truthfulness*, Williams (2002) analyzed two "virtues of truth," Accuracy and Sincerity. I add a third, Conversational Competence. "Virtues of truth" are, according to Williams (2002), "qualities that are displayed in wanting to know the truth, in finding it out, and in telling it to other people" (p. 7). Because data teams want to find out the truth and communicate it to others, they seek to acquire virtues of truth. Williams capitalized *Accuracy* and *Sincerity* to signify that he is using them as technical terms in ways that may not conform exactly to ordinary usage. I do not think that *Conversational Competence* has an ordinary usage, but I adopt Williams's convention of capitalization when integrating this term into his analysis.

Williams analyzed Accuracy and Sincerity first as instrumental goods and then as intrinsic goods. He argued that Accuracy and Sincerity are not sustainable as instrumental goods, but, since they must be sustained in order to sustain our way of life, they must be elevated from instrumental to intrinsic goods. I argue that Conversational Competence follows the same pattern: that is, it suffers the same unsustainability as an instrumental good, so it can be and should be (and actually is) elevated to an intrinsic good. In the final analysis, data teams strive for the goods of Accuracy, Sincerity, and Conversational Competence because they are virtuous, even in instances when some other good (e.g., financial funding or good will) might be attainable through Inaccuracy, Insincerity or Conversational Incompetence.

Accuracy, Sincerity, and Conversational Competence as Instrumental Goods

To initially define Accuracy and Sincerity, Williams sketches a fictional story that provides a minimalist context in which to clearly observe the two concepts in action. Williams later shows that, in the end, the story is too minimal; actually, as I will discuss below, I think it is too minimal from the beginning. Nevertheless, the most important element of his story is a group of people, each with their own beliefs, who need to pool their beliefs. To keep things simple, Williams assumes a low-tech society in which everyone is a competent speaker of a common language. These people need to find food and avoid danger. The pooling of beliefs is essential to these efforts because each person is limited by his or her particular spatial-temporal point of view. For example, a saber-toothed tiger approaching the village from the north may be discernible only by villagers on the north side of the village. Accuracy is the disposition to form true beliefs; Sincerity is the disposition to contribute one's beliefs to the pool. When the Accurate and Sincere villager sees the saber-toothed tiger approaching from the north, that villager forms the true belief that a saber-toothed tiger is approaching, and the villager contributes his belief to the pool, perhaps by shouting "Tiger!"

I think that Williams's story gives us a solid introduction to Accuracy and Sincerity, but I cannot help wondering if the Accurate and Sincere villager might be dumbstruck by the approaching saber-toothed tiger. From humbling personal experience with dangerous wild animals, I have reason to believe that my reaction would be to retreat while pointing and grunting. I can all too easily imagine that an Accurate and Sincere villager, due to (temporary) Conversational Incompetence, would contribute his true belief to the pool so ineffectively that his contribution would engender false beliefs.

When I tried to communicate, "One-and-half steps away, there is a rattlesnake in our path," my hiking partner mistook my articulations and gesticulations for, "I am having a stroke," and she came to my aid instead of watching out for the rattlesnake. (Luckily, she was not bitten, and only my pride was hurt.) Of course, if saber-toothed tigers are a frequent threat to the villagers, the villagers may train (perhaps through role-playing and drilling) to be Conversationally Competent during times of such crisis.[1]

Conversational Competence is the disposition to *effectively* contribute beliefs to the pool. How does Conversational Competence differ from Sincerity, which is the disposition to contribute one's beliefs to the pool? It might seem as though we could wrap Conversational Competence into Sincerity by taking "effectively" as understood and defining Sincerity as the disposition to (effectively) contribute one's belief to the pool. However, this departure from the common usage of sincerity would be too radical, because sincere people are not always effective. In the village, a person who did not speak the language (e.g., a foreigner or a mute) would necessarily be Insincere, no matter how willing he was to contribute his beliefs to the pool. By my reckoning, therefore, Conversational Competence and Sincerity are distinct. A person can be Conversational Competent and either Sincere or Insincere; conversely, a person can be Sincere and Conversationally Competent or Incompetent. I suppose that the most dangerous combination can be found in the proficient deceiver who is both silver-tongued (i.e., Conversationally Competent) and fork-tongued (i.e., Insincere).

Williams' minimalist story, as I observed earlier, is too minimal from the start. By assuming a low-tech society in which everyone is a competent speaker of a common

[1] Soldiers are trained to yell "grenade." Football players are trained to yell "fumble." Interestingly, students in Rape Aggression Defense (RAD) classes are trained to yell "fire" instead of "help" or "rape."

language, Williams masks the challenges involved in telling the truth to other people, and consequently he masks a third virtue of truth, namely Conversational Competence. For now, I am willing with Williams to stipulate that the villagers are competent in the syntax (i.e., grammar) and semantics (i.e., meanings) of their common language. Of course, we can and will add to the minimalist picture by relaxing the stipulation, because for high-tech societies such as the ones in which data teams operate, specialized syntax and semantics are present and thus it would be absurd to assume general competence in these areas. I am not, however, willing to stipulate with Williams that the villagers are competent in the pragmatics (i.e., contextualized use) of their common language. I fully expect that the villagers will grasp the pragmatics most if not all of the time, but I do not feel that the assumption can be granted automatically. In my version of the primitive village, the villagers consider the context before offering up their beliefs to others. A villager may believe that there are saber-toothed tigers somewhere in the world, but, lest he (or she) give grandpa a heart attack for no reason, he will think twice before he blurts out, "Tigers!" A villager may see the rustling of a bush, and form the uncertain belief that a saber-toothed tiger may be present, but the villager does not share his belief until he gets more confirmation, because the village is not on high alert. A villager believes that there are many saber-toothed tigers fast approaching, and therefore he also believes that there is one saber-toothed tiger fast approaching, but the villager judiciously shares the first belief, not the second, by shouting "Tigers!" rather than "Tiger!" Another villager, upon seeing a saber-toothed tiger, forces himself to shout loudly and clearly, "Tiger!" The point of these various illustrations is that, in my village and in Williams's village, the villagers are *generally* Accurate and Sincere, but those virtues cannot be assumed;

similarly, in my village, the villagers are Conversationally Competent, but this virtue is not a given either. In my village, Accuracy, Sincerity, and Conversational Competence are goods instrumental to the villagers' way of life. Insofar as the villagers value the pooling of beliefs, because this pooling advances their self-interests (e.g., for food and security), they value Accuracy, Sincerity, and Conversational Competence. Consequently, the villagers teach and reward these three traits.

Williams might argue that he is going further back to the beginnings of language, perhaps to something like the chimpanzee-human last common ancestor (CHLCA). As far back as Williams can go with Accuracy and Sincerity, I can go with Conversational Competence. To be clear, Williams is not making an evolutionary argument about how language actually came to be. He is not appealing to fossil records, DNA evidence, or atavistic organs. Rather, Williams is telling a fictional story about how the justification for a relatively simple epistemology of truthfulness can lead to the justification for a relatively complex epistemology of truthfulness, since his ultimate goal is to understand the justification of our current, complex epistemology of truthfulness. To that end, he starts with a simple scenario, where the justification is survival (i.e., self-interest) in a primitive world with a primitive language. If the CHLCA possessed the disposition to form true beliefs (Accuracy) and the disposition to contribute one's beliefs to the pool (Sincerity), those dispositions could not amount to a survival-promoting pool of beliefs unless the CHLCA also possessed the disposition to *effectively* contribute beliefs to the pool (Conversational Competence). Consider the most primitive of warning cries. What would justify use of the warning cry if it were not immediately relevant to survival, based on adequate evidence, sufficiently informative, and perspicuous? In brief, what would

justify the warning cry if it were not Conversationally Competent? Nothing, I would claim.

Free-Rider Problems of Accuracy, Sincerity, and Conversational Competence

As instrumental goods (e.g., goods useful for survival), Accuracy, Sincerity and Conversational Competence suffer from free-rider problems. If the sole purpose of Accuracy, Sincerity and Conversational Competence were to advance self-interest (e.g., to obtain food and security), then Inaccuracy, Insincerity, and Conversational Incompetence would be preferred in those instances where they better advance self-interest. In a village where everybody else is Accurate, Sincere, and Conversationally Competent, an individual may very well advance his self-interest by being Inaccurate, Insincere, or Conversationally Incompetent. Perhaps most obviously, the Insincere individual can fleece the suckers. Less obviously, but nonetheless significantly, the Inaccurate or Conversational Incompetent individual can repurpose the energy that he would otherwise spend on Accuracy and Conversational Competence. The key consideration here is that attaining Accuracy and Conversational Competence require effort. Specifically, Accuracy requires an effort to overcome external and internal obstacles to learning the truth about the target system. Undetectability is an external obstacle to Accuracy, and wishful thinking is an internal obstacle to Accuracy. Saber-toothed tigers are stealthy, so even if they can be detected from a safe vantage point, such detection demands sustained vigilance, and sometimes it involves putting oneself at risk as well. Villagers with a particularly strong desire for safety may be inclined to false negatives when detecting saber-toothed tigers (because maybe, if they ignore the rustling in the bushes, it will go away) or false positives (because the villagers huddle together for

security when the alarm is called). If everybody else in the village is Accurate (and Sincere and Conversationally Competent), the free rider can rely on the efforts of others and save his own effort for occasions of immediate personal threat.

Conversational Competence also requires an effort to overcome external and internal obstacles. These obstacles, however, are not about difficult-to-detect truths of the target system or wishful thinking about the target system. Rather, the obstacles to Conversational Competence are about the conversational context. From the perspective of pragmatics, the shared commitments of the speaker and audience are the most important part of the conversational context (Stalnaker, 1999). Those shared commitments can be difficult to detect. What is relevant for my audience? What counts as adequate evidence for my audience? What is too much (or too little) information for my audience? What is perspicuous for my audience? Furthermore, the commitments are not only difficult to detect in one's audience, but can even be difficult to detect in oneself (e.g., what is relevant, or what counts as adequate evidence, for me?). It takes effort to find common ground, and it takes even more effort to build common ground when there is insufficient common ground to start with. I think that, when it comes to Conversational Competence, the most dangerous piece of wishful thinking is that if I say only true things, my audience will form only true beliefs.

Free riders partake in the benefits of a common good but do not pay their share of the costs for the common good (Hardin, 2013). Some goods are *joint in supply* and *nonexcludable* (or nearly so). A *joint-in-supply* good is one that, once acquired by a sufficient number of group members, can be consumed by others at no marginal cost. A *nonexcludable* good is one that, once acquired by a sufficient number of group members,

cannot be kept from others. As standard examples of joint-in-supply and nonexcludable goods, Hardin (2013) cites radio broadcasts, national defense, and clean air. Once a sufficient number of group members have paid the costs to acquire those goods, those paying group members cannot prevent free riders from equally partaking in those goods.

The free-rider problem arises when too many people choose to ride free and there are too few group members to pay the required costs of the potentially joint-in-supply or nonexcludable common good, with the result that the common good is either a nonstarter or unsustainable. When enough people choose to be free riders, their self-interest in essence prevents the advancement of self-interest. Attempting to distinguish between unenlightened and enlightened self-interest does not provide a practical solution to this problem, but it does allow us to reframe the problem as one of overcoming unenlightened self-interest in order to advance enlightened self-interest.

There are three possible practical solutions (Hardin, 2013). First, one can manipulate the incentives to align unenlightened self-interest with enlightened self-interest. Governments can punish free riding with monetary penalties and jail time, vigilante groups can punish it with physical or psychological force, and social groups can punish it by ostracizing moochers. Conversely, these entities can offer special rewards to contributors. A second practical solution is to rely on and perhaps foster the foolishness of potential free riders, causing them to mistake their enlightened self-interest for unenlightened self-interest. This seems to be the strategy of "You Make the Difference" campaigns, which try to convince potential free riders that each of them personally represents the tipping point between insufficient and sufficient group support. A third option is to appeal to motives other than self-interest, such as duty, love, or virtue.

Williams argues there is an insidious free-rider problem in pooling beliefs, and he further argues that appealing to virtue is the solution.[2]

A pool of true beliefs is to some extent a joint-in-supply, nonexcludable common good. The pool is joint in supply insofar as, once a contributor pays the costs of forming the true belief with Accuracy and contributing the true belief to the pool with Conversational Competence, others can partake in the benefits of that true belief at no marginal cost. The pool is nonexcludable insofar as it is impossible to keep the true beliefs secret among only the contributing group members. In Williams's village, it would be in everybody's enlightened self-interest if all villagers shouted "Tiger!" whenever a tiger was in their vicinity. However, because free riders can reap the rewards while depending on others to expend their energy in detecting and announcing tigers, Williams's village has an insidious free-rider problem.

Of the three possible solutions to the free-rider problem and its threat to the pooling of true beliefs, Williams argues for the third by appealing to the virtues of truth as intrinsic goods. Williams does not deny that societies can punish Inaccuracy, Insincerity, and Conversational Incompetence and reward Accuracy, Sincerity, and Conversational Competence in order to align unenlightened with enlightened self-interest. Indeed, societies do try to do this, and they partially succeed. It is impossible, however, for a society can succeed completely in such an endeavor, because it is impossible to perfectly police Accuracy, Sincerity, and Conversational Competence.

[2] I think that Williams does not spend much time considering duty or love as a solution because we are (and ought to be) trustworthy beyond the bounds of duty and love. Consider the passerby in the Gricean vignette of the out-of-gas tourist. Even if the passerby has no duty or love toward the tourist, it would nonetheless be wrong for the passerby to intentionally mislead the tourist. (Having grown up in a region infested with tourists, I do not think it would be wrong for the passerby to ignore the tourist.)

Also, Williams does not deny that, through propaganda, societies can brainwash members into falsely believing that it is in their own unenlightened self-interest to be Accurate, Sincere, and Conversationally Competent. Nonetheless, it is a perverse strategy to instill a respect for truthfulness via untruthfulness, and I doubt that it is justifiable. In any case, insofar as society cannot justifiably align unenlightened with enlightened self-interest, it must appeal to some justification other than self-interest to bridge the gap.

Accuracy, Sincerity, and Conversational Competence as Intrinsic Goods

Williams has argued that Accuracy and Sincerity contain intrinsic value, and I further submit Conversational Competence as a third intrinsically valuable good. If these three are established as intrinsic goods, this solves the free-rider problem of pooling true beliefs. Intrinsic goods are valuable in and of themselves, as opposed to instrumental goods, which are valuable for the sake of some other purpose such as advancing self-interest. Even if a potential free rider does a cost-benefit analysis of Accuracy, Sincerity, and Conversational Competence and finds that the costs outweigh the benefits, he may nevertheless rationally choose Accuracy, Sincerity, and Conversational Competence for their own sake—that is, trustworthiness for the sake of trustworthiness.

Williams (2002, p. 91) suggests that "it is in fact a sufficient condition for something (for instance, trustworthiness) to have an intrinsic value that, first, it is necessary (or nearly necessary) for basic human purposes and needs that human beings should treat it as an intrinsic good; and, second, they can coherently treat it as an intrinsic good." With his fictional story of the villagers, Williams makes perspicuous the necessity of Accuracy and Sincerity for basic human purposes. I have attempted to do the same for Conversational Competence. In order to solve the insidious free-rider problem, Williams

argues that human beings need to treat Accuracy and Sincerity as intrinsic goods. Again, I tag along with Williams, stating that human beings need to treat Conversational Competence as an intrinsic good. Williams spends the remaining two-thirds of his book discussing how, in our culture, we can coherently treat Accuracy and Sincerity as intrinsic goods. Williams's project in this regard is not exhaustive, nor can it be, nor does he intend it to be so. Williams argues that intrinsic values are and must be woven into the fabric of our value systems (and our way of life). It is not okay, he says, to casually construct a new "intrinsic value" every time a new free-rider problem needs to be solved. As per the philosophical method of reflective equilibrium, our value statements (among other statements) must mesh, and some statements, especially those that have stood the test of time and have proven crucial to our way of life, are more deeply enmeshed than others. Different cultures weave them differently, so whereas in the first one-third of his book (including the fictional story of the village) Williams deals with human universals, in the remaining two-thirds he must delve into particulars. There are many particulars, even within a single culture, and Williams chooses a few. I choose to delve into particulars of the school culture of data-analytic inquiry within which data teams strive to operate. How can data teams weave the intrinsic value of Conversational Competence into their use and understanding of statistics to answer questions about teaching and learning in their schools, so as to benefit the schools?

My goal is to frame the project for data teams. Conversational Competence is a challenging aspiration. For data teams, the challenge is enormous. Audiences place considerable trust in data teams. Audiences assume instinctively that data teams are being helpful, as per Grice's Cooperative Principle. Consequently, as per the Gricean maxims

of conversation, audiences assume (unless given reason to believe otherwise) that data teams' contributions are relevant, adequate (in terms of evidence), thorough, streamlined, and perspicuous. These assumptions are necessary for efficient communication. If data teams had to spell out everything to their audiences, it might take forever to present even the simplest finding. Based on what the data team says, in conjunction with the assumptions, audiences will naturally draw implicatures that round out the data team's contribution, thus allowing the data team to say a lot with a little. This invaluable conversational logic, however, comes at a price. Data teams must invest diligent effort to make sure that their contributions are indeed relevant, true, adequate (in terms of evidence), thorough, streamlined, and perspicuous. Data teams must resist the temptation to free-ride on their audiences' natural (and healthy) dispositions to draw implicatures. In general, nobody can police data teams to make sure that their contributions are Conversationally Competent. As data teams push out into new frontiers, they exist in a sort of conversational Wild West. Data teams are embedded in a larger culture that values Accuracy, Sincerity, and Conversational Competence, but there are special features of data teams that prevent them from importing wholesale the values from that larger culture or any single subculture. First, data teams straddle (and strive to bridge) school culture and scientific culture; second, they are charged with spreading their values, in part, by exemplifying those values.

Because data teams straddle two subcultures, school culture and scientific culture, they have to mesh the values of the two. There is no reason to believe that the two subcultures value Accuracy, Sincerity, and Conversational Competence in the same way. Granted, these values are intrinsic in both subcultures, but they are integrated differently

into the two value systems. Scientific culture, through prescribed methods, exclusive societies, and peer-review processes, sets relatively tight standards for Accuracy, Sincerity, and Conversational Competence. On the other hand, school culture cannot admit such tight standards. Scientists often have the luxury of months or years to uncover a fact, integrate that fact into a theoretical framework, and situate the framework within a larger literature (not necessarily in that order), and then they can leave the practical implications to others. Educators are often flooded with facts and have little or no time for reflection before they must act.

The Massachusetts Department of Education (MDOE) distributes a *District Data Team Toolkit* for the purpose of "helping districts establish, grow, and maintain a culture of inquiry and data use" (MDOE, n.d., p. 1). In the toolkit, the MDOE outlines a data team process and explicitly recommends that data teams "model the process." Its descriptions of how to do so include "Lead by example, not by edict" and "Publicly demonstrate how the District Data Team is moving toward the vision" (MDOE, n.d., p. 51). In principle, to exemplify the values of Accuracy, Sincerity and Conversational Competence, data teams need only to possess the values and illustrate by example that they possess the values. Not all examples, however, are useful examples. Useful examples make manifest the exemplified properties. Often we use fictions (e.g., average students, frictionless planes, Williams's village) as examples because they allow us to abstract away the background noise in order to highlight the properties to be featured. For instance, in his field guides, John Audubon used bird drawings instead of bird photographs. A sketch of an idealized bird can be better than the real thing if the purpose is to highlight the properties important for identification.

Data teams, of course, are not fictional exemplars; they are the real thing. As living exemplars, they cannot magically abstract away the background noise; instead, they must amplify the foreground in order to cut through the noise. I think that the required amplification depends on the expertise of the audience. One aspect of expertise is the ability to cut through the noise when presented with a particular situation. Imagine a master carpenter demonstrating the proper use of an obscure woodworking tool, a chain mortiser. The master carpenter gives the demonstration to a journeyman carpenter and to an apprentice. After the demonstration, the master carpenter asks, "See?" The journeyman carpenter understands what to do, but the apprentice carpenter needs further explanation. Data team audiences are usually apprentices, so data teams must be loud and clear about their values.

My concern is that, in amplifying the values, the data team will (perhaps necessarily) distort the values. For instance, consider a data team that draws attention to its Conversational Competence in this way: "Notice that our findings were relevant. Note that we used standard scientific methods to check the adequacy of our evidence. Know that we streamlined our presentation to include all the vital results and only the vital results. Notice how concise our presentation was … at least until we doubled the total time required in order to tell you how concise it was." The problem with separating the foreground from the background is that we lose the often crucial interaction between foreground and background. Williams insightfully uses his fictional village to introduce the virtues of truth, but just as insightfully he eventually abandons the village for a more contextualized approach, calling his stepwise analytic method "genealogical." Williams writes, "A genealogy is a narrative that tries to explain a cultural phenomenon by

describing a way in which it came about, or could have come about, or might be imagined to have come about." It is obvious that ontology need not recapitulate genealogy, because genealogies do not purport to describe the only ways in which cultural phenomena *must* come about. Nevertheless, perhaps, in order to "establish, grow, and maintain a culture of inquiry and data use" (MDOE, n.d., p. 1), data teams can draw lessons from Williams's genealogy. While establishing such a culture, data teams might be bold in asserting their adherence to the intrinsic values of Accuracy, Sincerity and Conversational Competence, but as the culture grows, data teams must integrate those values with the myriad of other school values. In the following section, I hope to contribute to the early stages of a (necessarily ongoing) project by data teams to integrate Conversational Competence into their value systems.

Recommendations

In this final section, I offer recommendations to data teams striving for Conversational Competence in their statistical reporting. I structure my recommendations around the Gricean maxims regarding relevance, truth, evidential adequacy, informational moderation, and perspicuity. I suggest that data teams should become acutely aware of their conversational context and how it interacts with what they say, because that interaction inevitably yields conversational implicatures. Data team audiences will draw implicatures—that cannot be avoided; thus the challenge facing data teams is to ensure, insofar as possible, that those implicatures are true. For my analysis, I use the Gricean framework of conversational implicature, not because it is the only framework or a perfect one, but because it can serve as an excellent first approximation.

Relevance

The Gricean Maxim of Relevance tells data teams to make their conversational contributions relevant. Audiences will assume that data teams are presenting relevant statistics unless the data teams ostentatiously non-adhere to the Maxim of Relevance by opting out, clashing, or flouting. I began this chapter with a vignette about a data team called on by a school committee to give evidence for the effectiveness of a program. The data team reports, "On average, students who participated in the program achieved statistically significantly higher GPAs than students who did not participate in the program." The school committee, based on the Maxim of Relevance, draws the false implicature that the program *caused* the difference in scores, because the testimony was supposed to be assessing the program's effectiveness.

Rarely will data teams have adequate statistics or other evidence to be perfectly relevant. They may not have experimental data to make relevant causal claims (as in the current vignette). They may not have longitudinal data to make relevant developmental claims. Imperfect relevance is a fact of life for data teams. Nevertheless, to paraphrase Tukey (1962), an imperfectly relevant answer to a perfectly relevant question is better than a perfectly relevant answer to a perfectly irrelevant question.[3] When data teams do not have adequate evidence to be perfectly relevant, they experience a clash between the Maxim of Relevance and the Maxim of Quality, "Do not say that for which you lack adequate evidence." Some audiences truly understand that correlation does not imply causation, but most data team audiences will not. If the audience does understand, then maybe it is acceptable to report a correlational answer to a causal question, because the

[3] "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" (Tukey, 1962, p. 13).

sophisticated audience will draw the true conversational implicature (based on a clash of maxims) that the data team is giving imperfectly relevant information because it does not have adequate evidence to be perfectly relevant. If, however, the audience does not understand, then the data team had better teach the audience, or the audience will draw false implicatures.

In their effort toward Conversational Competence, teaching is a critical tool for data teams. The conversational context is not static. In "Scorekeeping in a Language Game," Lewis (1979) sketches rules for the shaping of conversational presuppositions that constitute the conversational context. Lewis emphasizes the subtlety with which the conversational context can shift. Data teams, however, need not be subtle. At the beginning of a conversation, data teams can probe their audience to determine what counts as relevant to them. At the back end of a conversation, data teams can probe their audiences to determine what implicatures they drew. At either point, the data team can take corrective actions if necessary, teaching the audience to make different assumptions. This works at the back end as well as the front end, because, as Grice (1989) argues, conversational implicatures are cancelable by their very nature. It is not incoherent or self-contradictory to add more information to the conversation so as to negate an identified false implicature. Here is a revision of a previous example:

The tourist says, "My car is out of gas."

The passerby replies, "There is a garage around the corner."

The tourist pragmatically infers that the garage can solve his gas problem.

The passerby remembers, "But, darn it all, the garage is closed today."

The implicature is effectively canceled.

Cancelability distinguishes conversational implicature from logical entailment.

The tourist says, “Socrates is a man.”

The passerby replies, “All men are mortal.”

The tourist logically infers that Socrates is mortal.

There is nothing that the tourist can say in this case, aside from going back on what was said, to block the logically unassailable deductive inference. Pragmatic inferences are different. Conversational implicatures can be blocked without going back on what was said, and a data team should block the implicatures when they prove to be wrong. Of course, data teams should not hesitate to go back on what they said either if what they said turns out to be wrong, even if correcting a previous statement means losing a little face. My point is that correcting a misunderstanding due to conversational implicature need not involve the loss of face that accompanies going back on what one said. Data teams should plan, as a matter of routine, to correct misunderstandings that occur due to conversational implicature. Teachers have a bag of tricks to test for misunderstandings in the classroom; they must bring that bag of tricks to their data team work.

Truth

The Gricean Maxims of Quality state, "Try to make your contribution one that is true” (Grice, 1989, p. 26), followed by two specific examples of that general statement: “1. Do not say what you believe to be false. 2. Do not say that for which you lack adequate evidence” (Grice, 1989, p. 26). I begin with the maxims involving truth and falsity. In this essay, I presume that data teams are striving to report true statistics. My core argument is that true statistics can nevertheless be misleading. Although there may truly be a sample-mean difference in GPAs between two subgroups, the difference may

nonetheless be misleadingly irrelevant, inadequate, underinformative, or imperspicuous to a general audience of educators. The maxims involving truth and falsity are only a few maxims among many. Even if the data team adheres to the maxims involving truth and falsity, it may violate other maxims, thus leading to false implicatures. Therefore, the truth is not a *sufficient* condition for non-misleadingness. In this subsection, I want to make the further point that the truth is not a *necessary* condition for non-misleadingness.

Sometimes two maxims clash, and, consequently one of them must give way—even, in some cases, the maxims involving truth and falsity. Take, for example, the rounding of numbers. Although Samuel Johnson (Boswell & Hill, 1921, p. 396) may have been right when he said that rounded numbers are always false, rounded numbers are often times less misleading than unrounded numbers. Suppose that a data team finds an average difference of 0.34348235 in GPAs between students ineligible for free lunch and those eligible for free lunch. According to Cohen's (1990) recommendation, "less is more," the data team should probably round the average difference to 0.3. Cohen, implicitly appealing to the Maxim of Relevance, argues that data analysts should round in order avoid the reporting of insignificant digits that are essentially just random numbers due to sampling error and measurement error. Based on the Maxim of Relevance, the audience may infer that the insignificant digits are relevant, because, presumably, the data team would report only relevant numbers. Cohen also implicitly appeals to the Maxim of Manner, "Be brief," when he argues for rounding because it avoids unnecessary clutter. I would appeal further to the Maxims of Quantity, which demand that data teams provide neither too much nor too little information for the purposes of the current exchange. Even when digits are potentially informative, data teams should not

report them unless they are *actually* informative (in the current context). Perhaps the hundredths digit of 0.34348235 is potentially informative because the average difference is precise to the hundredths digit. Nevertheless, if the tenths digit provides all the necessary information in the current context, then, by the Maxims of Quantity, the hundredth digit should be rounded off to 0.3 in this situation. Note that the Maxims of Quantity not only encourage data teams to round but also discourage them from rounding too much, because they call for providing sufficient information for the current purposes of the exchange. If data teams round too much, their audiences will draw the false implicature that the rounded-away information is irrelevant.

If Samuel Johnson is right, and rounded numbers are indeed false, then rounding is a non-adherence to the Maxim of Quality that states, "Do not say what you believe to be false" (Grice, 1989, p. 26). I think that a simple use of the word "about," e.g., "the average difference is about 0.3," blocks the allegation of falsehood. Even without this block, a non-adherence to the Maxim of Quality would be justified by the adherence to the Maxims of Relevance, Manner and Quantity. If the non-adherence is ostentatious such that the audience recognizes it, then the non-adherence is a clash. For example, the data team might report to a savvy audience that the average difference is 0.30, and the audience, being savvy, would know that this use of a zero in the hundredths place communicates that the calculated average difference is between 0.295 and 0.305. If the audience does not recognize the non-adherence, then it is a violation but probably a non-misleading violation. As per the Cooperative Principle, the data team's duty is to avoid misleading rather than to avoid falsehood.

Evidential Adequacy

A Gricean Maxim of Quality states, "Do not say that for which you lack adequate evidence" (Grice, 1989, p. 26). Standards of evidential adequacy vary from one context to another. In the data team context, averages and differences in averages often serve as evidence. The relationship between a statement and evidence is exemplificational: evidence supports a statement by exemplifying properties relevant to the truth of the statement (Goodman & Elgin, 1988, p. 20). Data teams talk about average GPAs and average students because such statements exemplify properties relevant to the truth of their statements about achievement gaps and program effects.

In my experience, however, some educators find talk of averages repugnant. A disgusted educator might say something like "No student is average." Data teams must teach the evidential adequacy of averages. By my account, the data team can look that disgusted educator in the eye and respond, "You're right. The average student is a fiction, but a scientifically useful fiction." I think that, when data teams talk about average students, they are talking about fictional students. Talk of fictional students, however, clashes with the Maxim of Quality, "Try to make your contribution one that is true." After all, fictions are not true nor do they purport to be true. So now we have a clash between one Maxim of Quality (involving evidential adequacy) and another Maxim of Quality (involving the truth). Data teams use averages to meet scientific standards of evidential adequacy.

Data teams must understand (perhaps through training) that an average group difference is inadequate evidence, in any context, for a stereotype. Data teams must make sure that their audiences know this fact as well. This is easier said than done. People tend

to mistake average group differences for individual differences, thus turning a sociotype into a stereotype. In the literature on cultural assimilation, a sociotype is a group-level generalization and a stereotype is an individual-level overgeneralization; moreover, the sociotype sometimes provides the grain of truth for the stereotype (Wiest, 2003). This trouble haunts the interpretation of averages, no matter how one construes them. McGarty et al. (2002) have argued cogently that stereotype formation involves social, cultural, and cognitive factors. Misinterpreting averages is a symptom of a complex problem. To alleviate the symptom, I think that my construal of averages as fictive may be helpful. But even more forcefully and conclusively, one great strategy for demonstrating that averages are inadequate evidence for stereotypes is to graphically display the raw data alongside the averages, enabling the audience to see the within-group variation. In other words, the raw data allow the audience to see that generally nobody is average, and to perceive clearly that, even if the average score in group A is higher than that in group B, some individuals in group B outperformed many or most of the members of group A.

Statistical bias and statistical imprecision are other sources of evidential inadequacy. A statistic is a biased estimator if, across random samples, it tends to either overestimate or underestimate the population parameter; it is an imprecise estimator insofar as its estimates of the population parameter vary across random samples. Imprecision is generally unavoidable, but there are methods for reducing it. Some of those methods trade off a small increase in bias for a large decrease in imprecision. The Maxim of Quality states, “Do not say that for which you lack adequate evidence.” In keeping with the maxim, data teams should not report results that are too biased or too imprecise. Unless the data team opts out of the maxim, the audience will draw the

conversational implicature that the results are not too biased and not too imprecise. But, what exactly is "too biased" and "too imprecise"? I focus on the baffling difficulties of defining "too imprecise," but defining "too biased" involves analogous difficulties. In an ideal world, the data team's audience would have (1) an understanding of statistical imprecision, (2) a set level of tolerance for statistical imprecision (e.g., an alpha level), and (3) a strategy for policymaking in light of statistical imprecision. In the ideal world, the data team could adhere to the Maxim of Quality by adopting the audience's alpha level, and the data team could verify that true implicatures have been drawn by observing whether the proposed educational policies driven by the results appropriately reflect the statistical imprecision of the results. In the real world, however, I do not think that data teams can rely on implicature, as their audiences will generally have little to no understanding of statistical imprecision, no set level of tolerance for statistical imprecision, and no ability to take imprecision into account in their policymaking strategy. Therefore, data teams must teach their audiences about evidential adequacy, tolerance levels, and educational policy in light of statistical imprecision. Otherwise, data team audiences will blithely draw the implicature that the evidence is "adequate" when in fact nobody, neither the data team nor the audience, clearly understands evidential adequacy in context.

I think that, with respect to statistical imprecision, data teams must opt out of the Maxim of Quality, "Do not say that for which you lack adequate evidence." Instead of relying on implicit rules of conversation in context, data teams must make explicit their standards of evidential adequacy. Then, the data team audiences can take the information for what it is worth to them. Perhaps even better, data teams can present multiple

confidence intervals for each of their key findings (e.g., 66%, 90%, 95%, and 99% confidence intervals). If taught carefully, confidence intervals convey not only the fact of imprecision but also the magnitude of imprecision. Multiple confidence intervals allow different audiences to find their own tolerance levels. The lower and upper bounds of confidence intervals speak to policymaking strategies that prefer to err on the side of either action or caution. Sometimes, both the lower and upper bounds will support the same policy, and the confidence interval will thus provide powerful proof for the adequacy of the evidence despite statistical imprecision. In my experience as a data team consultant, different audiences do indeed choose different confidence levels, and this justifies their use of differing alpha levels, because a confidence level is simply 1 minus the alpha level.

Informational Moderation

The Maxims of Quantity state, "1. Make your contribution as informative as is required (for the current purposes of the exchange). 2. Do not make your information more informative than is required" (Grice, 1989, p. 26). The data team must walk the fine line between providing too little and too much information. In general, this is a major challenge. I suggest that data teams create multiple versions of their reports with differing levels of detail. Some audiences need only to know the gist of the results; other audiences need to know the details; still others need to know not just the results but the methodology as well. With proper formatting (e.g., headings, abstracts, emphases, and appendices), a single document or presentation may serve the differing informational needs of multiple audiences.

Perspicuity

The Maxims of Manner state, "Be perspicuous. … 1. Avoid obscurity of expression. 2. Avoid ambiguity. 3. Be brief (avoid unnecessary prolixity). 4. Be orderly" (Grice, 1989, p. 26). Statistics can be obscure, but, once clarified, they are generally unambiguous, brief, and orderly. Statistics are functions of the data, and, as such they are unambiguously defined. They can certainly be brief: consider statistical summaries and graphs that describe, with a few numbers or a simple graph, a potentially huge amount of data. Finally, statistical summaries and graphs impose an order on the data and exploit that order for informational purposes. Imagine by contrast the imperspicuity, obscurity, ambiguity, prolixity, and disorder of data analysis in a world without summary statistics and graphs. To answer a question about the SES achievement gap in GPA for a live audience, the data team could read off the GPA for every student eligible for free lunch (perhaps in random order, or in alphabetical order by last name, or in some other irrelevant order), followed by the GPA for every student ineligible for free lunch. Or the data team could display two unordered lists. What could an audience make of that display? Virtually nothing.

I hope that the deeply contextual nature of statistical reporting has become increasingly evident throughout my discussion of implicature. At least one matter of interpretation, however, is the same across policymaking contexts. No matter what the policymaking context is, if statistical imprecision is the sole explanation for a difference in subsample means, then that difference is policy-irrelevant. The term *statistically significant* is often abused, but its proper use is in fact useful. Properly used, it denotes a statistical result for which statistical imprecision has been ruled out (albeit fallibly) as the

sole explanation. By setting the alpha level or confidence level, data analysts control the degree of fallibility, which should be context-sensitive as discussed above. The concept of statistical significance must also be treated as context-sensitive, because the term is more or less obscure depending on the context.

In accordance with the Maxim of Manner, data teams should avoid obscurity. The trouble with referring to a finding as "statistically significant" is that many audiences interpret this term as clear to them when they actually misunderstand it. Castro Sotos, Vanhoof, van den Noortgate, and Onghena (2007) reviewed the literature on misconceptions of statistical inference. Unfortunately, the delusion of clarity is often supported by conversational implicature. I break down the conversational logic involved by closely following Grice's (1989, p. 31) analysis:

1. The data team said that the result is statistically significant.
2. The audience has no reason to believe that the data team is not being helpful.
3. The data team would not have said that the result was statistically significant unless the data team thought that "statistically significant" was clear to the audience, i.e., avoiding obscurity of expression as per the Maxim of Manner.
4. The data team knows that the audience will suppose that the data team thinks that "statistically significant" is clear to the audience. Furthermore, the data team knows that the audience knows that the data team knows. In other words, the data team knows that, with mutual awareness, they are on the same page.
5. The data team has done nothing to stop the audience from thinking that "statistically significant" is a clear expression.

6. The data team intends the audience to think, or is at least willing to allow the audience to think, that "statistically significant" is a clear expression.
7. Therefore, the data team has implicated that "statistically significant" is clear to the audience.

Unless data teams know that "statistically significant" is clear to their audiences, data teams must break this chain of reasoning that would lead their audiences to mistakenly conclude that the term is indeed generally understood accurately. Data teams can break this chain at the first link by avoiding the expression completely. Or they can break the chain at the third link by stating explicitly that the expression is not very helpful due to its misleading obscurity, thus opting out of the Maxim of Manner. Or they can break it at the fifth link by adding to the conversational context in order to block the implicature—e.g., by clarifying the meaning of the expression. I cannot recommend which strategy is best without a specific context; my purpose here is to make clear why some strategy is necessary. I think that the pivotal link is actually the fourth link: "The data team knows that the audience will suppose that the data team thinks that 'statistically significant' is clear to the audience." Generally, I believe, data teams do not know what their audiences will suppose. My point of emphasis is that, in order to prevent misleading implicatures, data teams must learn what audiences will suppose.

Conclusion

I want to conclude by posing a monumental problem for data teams, and I can offer only hints at a solution. If data teams are presenting to diverse audiences (perhaps simultaneously), and if the diverse audiences have diverse suppositions that lead to diverse implicatures, how do data teams ensure that the implicatures are true

implicatures? I do not know the solution, but I worry that the solution for one audience may raise problems for another audience. One idealistic solution would require data teams to segregate their audiences by suppositions, interpret their results accordingly, and place their audiences under gag orders to prevent the sharing of results between audiences. As a slightly more realistic alternative, data teams should find and foster, as much as possible, a common ground of suppositions as part of their mission of "helping districts establish, grow, and maintain a culture of inquiry and data use" (MDOE, p. 1). For the sake of Conversational Competence, data teams must be attuned to the suppositions that make up the conversational context, but they themselves can also tune the suppositions by teaching data literacy. Suppositions change, and data teams can drive some of the change in order to make the challenge of Conversational Competence more manageable. I think that, realistically, data teams will always be herding cats, playing whack-a-mole, and putting out fires. Organizational chaos is a fact of life, as evidenced by all the colorful clichés used to describe it. To interpret results effectively, perhaps the greatest strength of a data team is its diversity. Because the data team is diverse, it contains representatives from multiple audiences. If art teachers have different suppositions from science teachers about relevance, evidential adequacy, informational moderation, and perspicuity, then it can helpful to have an art teacher and a science teacher on the team. The diversity of a data team may support achievement of a broad level of trustworthiness, because a diverse data team may have the resources to be Conversationally Competent for many different audiences.

Chapter 2: The Sample Mean and Sample Median as Exemplars

As a teacher of applied statistics, I take my students' objections seriously. As a philosopher, I try to understand the logic underlying each objection, which generally involves at least a grain truth. Therefore, I cannot help but wonder when a student makes the following objection: averages are useless for understanding people in the real world, because, in the real world, there is no such thing as an average person. Perhaps the average person is a fiction, but fictions can nonetheless be useful for understanding the real world. Before I taught applied statistics to college students and data teams, I taught literature to high schoolers. I thought then and I think now that we have a lot to learn from Odysseus and Penelope, Romeo and Juliet, and Atticus and Scout. Likewise, I think that we have a lot to learn from averages. In this essay, I explore how and what we learn from averages, fictive though they may be.

In an article entitled "What Does the Mean Mean?" Watier, Lamontagne, and Chartier (2011) discussed various methods for calculating the sample mean, and for each method they offer insights that a teacher of applied statistics might incorporate into a lesson on the topic. As a teacher, I find their effort laudable; as a philosopher, I find their question still inadequately answered. What does the sample mean mean? From a philosophical point of view, *meaning* is a muddled concept, perhaps hopelessly muddled. Therefore, I take the liberty of rephrasing the question: To what does the sample mean refer, and how does it refer? My philosophical answer is: The sample mean refers to the property of the *non-outlying value of the sample distribution*, and it does so by exemplification. As I will discuss in detail, the sample mean refers by a particular sort of exemplification—attenuated, moderate, fictive exemplification. As an attenuated

exemplar, the mean is easy to interpret. As a moderate exemplar, the mean avoids any vagueness of the label "non-outlier." As both an attenuated and a moderate exemplar, the mean provides a scientifically objective basis for between-group comparison. As a fictive exemplar, the mean is not restricted to being a value in the sample distribution, so it can be as attenuated and as moderate as science demands. After a general discussion of exemplification, I will discuss attenuated exemplification, moderate exemplification and fictive exemplification in turn.

This schema provides insight into the epistemic functioning of not only the sample mean but also the sample median. Therefore, I can answer a second research question: To what does the sample median refer, and how does it refer? My philosophical answer is: The sample median also refers to the property of the *non-outlying value of the sample distribution*, and it does so by attenuated, moderate exemplification, but it refers by fictive exemplification only when the sample has an even number of values and the two middlemost values differ; otherwise the sample median refers by non-fictive exemplification.

Because non-fictive exemplification is easier to understand than fictive exemplification, in the early stages of my analysis I focus on sample medians from samples with an odd number of values, since these cases yield medians that are non-fictive exemplars. I rationally reconstruct the median in order to gain insight into the epistemic functioning of attenuated, moderate exemplars as a basis for between-group comparison. Between-group comparison is scientifically important because, for instance, it allows experimental researchers to compare treatment groups to control groups in order

to detect causal effects; it also allows educational researchers to compare minority groups to majority groups in order to detect gaps in academic achievement.

This essay will proceed in accordance with the following outline:

- Exemplification
- Attenuated vs. Replete Exemplification
- Moderate vs. Extreme Exemplification
- The Sample Median Rationally Reconstructed as an Attenuated, Moderate Exemplar
- Fictive vs. Non-Fictive Exemplification

Exemplification

What is exemplification? According to Goodman and Elgin, an exemplar exemplifies by meeting three criteria (Elgin, 2012; Goodman, 1976; Suárez, 2009). First, it possesses the property that it exemplifies. Second, it refers to the property that it exemplifies by way of possessing the property. Third, it elucidates the property that it exemplifies. This third condition of elucidation is a success condition that strictly does not need to be met, because exemplars can conceivably fail to support understanding. A good exemplar, however, supports understanding (more or less selectively) about that which it exemplifies (Goodman & Elgin, 1988, p. 69). I will discuss each criterion in turn, using as my illustrations a paint swatch from a hardware store and the median grade from a sample of test scores.

First, consider a Sherman-Williams paint swatch exemplifying seven shades of grey, labeled “Quicksilver,” “North Star,” “Krypton,” “Jubilee,” “Storm Cloud,” “Granite Peak,” and “Outerspace” from light to dark, respectively, as one moves across the paint

swatch from left to right. Each of the seven paint samples exemplifies the shade, hue, and gloss of its respective color. Second, consider a sample of test scores (where $n = 7$): 98, 97, 97, 95, 90, 88 and 80. The sample median of 95, I argue, exemplifies non-outlying values of the sample distribution.

The Possession Criterion of Exemplification

According to the first criterion of exemplification, an exemplar must possess the property that it exemplifies. On the far left of the paint swatch, the square patch of grey is labeled "Quicksilver." As an exemplar of the color property *Quicksilver*, the paint swatch must possess the color property *Quicksilver*. Indeed, it does. Its color is Quicksilver. In particular, the paint swatch possesses the hue, shade, and gloss of Quicksilver. (*Hue*, *shade*, and *gloss* are themselves properties, more fine-grained than the general property of *color*. In other words, the properties *Quicksilver hue*, *Quicksilver shade,* and *Quicksilver gloss* are more fine-grained than the property *Quicksilver color*. As a convention, when I am describing a property but also need to note finer-grained aspects in my discussion of the property, I will call those finer-grained properties "features" in order to avoid a pileup of "properties.") The paint swatch possesses the color features of Quicksilver: *Quicksilver hue*, *Quicksilver shade*, and *Quicksilver gloss*. It also possesses many other features such as size, shape, weight, price, age, flammability, conductivity, hardness, digestibility, and flavor, but only hue, shade, and gloss figure in its exemplification of *Quicksilver*. Some of these other features may be non-essential but convenient, such as the item's size and shape; most of them are utterly irrelevant, such as digestibility and flavor. (Paint samples are not for eating.) Anything with the hue, shade, and gloss of *Quicksilver* has the potential to serve as an exemplar of *Quicksilver*

(provided that the other necessary condition of exemplification, reference, is met), but paint samples of *Quicksilver* are particularly convenient exemplars.

Now let us turn to the sample median as an exemplar. What feature must the sample median possess in order to exemplify the property of a *non-outlying value of the sample distribution*? Simply, the sample median must itself be a non-outlying value of the sample distribution. This property is easy to observe for sample medians from sample distributions with an odd number of values. To calculate the sample median value, one ranks all the sample values from lowest to highest. The middle-ranked value is then the sample median, if there is indeed a middle-ranked value—which there is whenever the sample distribution contains an odd number of values. The sample median of 95 possesses the property of *non-outlying value of the sample distribution*, and therefore 95 can exemplify the property of *non-outlying value of the sample distribution*. Of course, other values in the sample possess and can exemplify this property too, but the median is a particularly elucidating exemplar of *non-outlying value of the sample distribution*.

Note that I have not explicitly defined *non-outlying value of the sample distribution*. The label "non-outlying" is generally vague, and I treat it as such. Take the aforementioned distribution of sample values (n = 7): 98, 97, 97, 95, 90, 88 and 80. The label "non-outlying" clearly applies to the values of 98, 97, 97, 95, 90 and 88. It is not so clear, however, whether the label appropriately applies to the value of 80. A vague label is a label the applicability of which is indeterminate in at least some actual or possible cases. Generally, in a unimodal distribution of sample values, there is a middle range of possible cases where the label "non-outlying value of the sample distribution" clearly applies, and there are also an upper extreme range and a lower extreme range of possible

cases where the label clearly does not apply, but there are also borderline ranges where the applicability of the label is problematic. Vagueness presents labeling problems. The median systematically avoids those problems, and this avoidance contributes to the median's particular value in elucidation.

The Reference-by-Possession Criterion of Exemplification

According to the second criterion of exemplification, an exemplar must refer, by way of possessing the property, to the property that it exemplifies. It is not enough for an exemplar of a property to possess the property; the exemplar must also refer to the property. This reference is fundamentally an act of symbolization, and it must be interpreted as such. Paint swatches and sample medians are symbols. Paint companies fabricate and distribute paint swatches to refer to the variety of colors that they offer. Similarly, data analysts calculate and report sample medians to refer to the central tendencies of sample distributions.

The Elucidation Criterion of Exemplification

An exemplar of a property, then, requires both possession of the property and reference to the property by way of its possession of the property. The possession condition and the reference-by-possession condition are jointly sufficient conditions for exemplification. In other words, anything that possesses the property and refers to the property by way of possessing the property is an exemplar of the property. Nevertheless, not every exemplar is a good exemplar. Consider one not-so-good exemplar. An English-language learner (perhaps a child) is first learning a basic schema of six colors: red, orange, yellow, green, blue, and violet. Whereas a red fire hydrant might be a good exemplar of red, a chartreuse fire hydrant would be a bad exemplar of green. At least for

me, “green” is a vague label for chartreuse fire hydrants. When I force myself to label such a fire hydrant as either “green” or “yellow,” I am torn. Even if I finally decide (perhaps with the help of a color wheel) that the chartreuse fire hydrant possesses the color property *green*, it would nonetheless be a terrible example for an initial lesson on the difference between yellow and green, because a good exemplar elucidates the label. (The chartreuse fire hydrant might be a perfect example for an advanced lesson, but a raw beginner needs clear-cut exemplars, not borderline exemplars. Borderline exemplars may be useful for teaching nuances, but the raw beginner first needs to be taught the essential basics.

Exemplars elucidate by highlighting, underscoring, displaying, conveying, and thereby “honing one’s ability to recognize, synthesize, reorganize and so on” (Elgin, 1996, p. 183). It is beyond the scope of this essay to present a learning theory that provides a general account of the fit between exemplars and learners such that the exemplar facilitates the learner’s recognizing, synthesizing, reorganizing, and so on. Undoubtedly there are many ways in which an exemplar can facilitate or fail to facilitate these tasks, but I will focus on a limited few. My dimensions of particular focus will become clearer in the coming sections as I discuss the distinctions between attenuated and replete exemplars, moderate and extreme exemplars, and non-fictive and fictive exemplars.

Attenuated vs. Replete Exemplification

Every exemplar possesses features irrelevant to its exemplification. The exemplar paint swatch is hardly digestible and blandly flavored, but that is irrelevant. The relevant features are hue, shade, and gloss. The sample median was calculated on a Tuesday, but

that is irrelevant to its exemplification of a *non-outlying value of the sample distribution*. The relevant feature is simply that it is a non-outlying value of the sample distribution. An *attenuated exemplar* is one for which it is easy to parse the relevant features from the irrelevant features. An attenuated exemplar exemplifies along relatively few dimensions, and one can recognize relatively easily what is exemplified. A *replete exemplar*, on the other hand, is one for which parsing is difficult. In standard contexts, the paint swatch is an attenuated exemplar of *Quicksilver*, and the sample median is an attenuated example of the concept of *non-outlying value of the sample distribution*. Most hardware store customers can readily parse a paint sample, and any data-literate person can readily parse a sample median. By way of contrast, I will introduce two replete exemplars.

Socrates' method is a replete exemplar of the property *Socratic method*, and Gauguin's *D'où Venons Nous / Que Sommes Nous / Où Allons Nous* is a replete exemplar of the property *expansive*. First, consider Socrates' method. Three Hellenic schools of philosophy, Stoicism, Cynicism, and Skepticism, each adopted Socrates' method as an exemplar of philosophical and pedagogical method, but each school interpreted different features of Socrates' method as relevant. In broad strokes, the Stoics took Socrates' emphasis on reason over emotion as essential, the Cynics focused on his abandonment of worldly things, and the Skeptics prized Socrates' radical doubt. Proficient interpreters may all agree that Socrates' method exemplifies the concept of *Socratic method*, but they can and do disagree about the features relevant to the exemplification. Socrates' method is, therefore, a *replete* exemplar because the relevant features are difficult to discern.

Gauguin's *D'où Venons Nous / Que Sommes Nous / Où Allons Nous* is a replete exemplar of the property *expansive*. It is both literally and metaphorically expansive, and

it begs a lifetime of study. Reasonable interpretations are various and shifting. Whereas hardware store customers can instantly agree on the meaning of a paint swatch, art critics can argue forever about the meaning of Gauguin's painting. This is the difference between attenuation and repleteness. In *Languages of Art*, Goodman (1975, p. 230) introduced the distinction between attenuated and replete symbols to help in explaining how art functions cognitively. Art often uses replete exemplars. Goodman acknowledged that this distinction also has implications for explaining how science functions cognitively, and Elgin (2012) fleshed out some of those implications, arguing in particular that science favors attenuated exemplars.

Attenuated Exemplars in Science

The relationship between theory and evidence is exemplificational: evidence supports a theory by exemplifying properties relevant to the truth of the theory (Goodman & Elgin, 1988, p. 20). Consider Galileo's purported experiment in which he dropped a cannonball and a musket ball simultaneously from the Leaning Tower of Pisa. It is important that Galileo set the stage in order to make obvious which features mattered and which did not (see Cartwright, 1983 on stage-setting in science). It was relevant that he dropped the balls from the same height at the same time and that they hit the ground at the same time. That is, the features *same height* and *same time* were relevant. It was also important that the balls were different masses. It was not important, however, that Galileo dropped the balls from the Leaning Tower of Pisa, or precisely which cannonball or which musket ball he selected. It is fairly easy for scientists to agree on the labels exemplified by Galileo's simultaneous dropping of a cannonball and a musket ball from the Leaning Tower of Pisa. Thus, Galileo's experimental exemplar is attenuated.

If we look more closely, however, we see that the exemplar is not *perfectly* attenuated. The cannonball and musket ball share the feature of *same specific gravity*. For Galileo, this was relevant to the exemplification, but for modern scientists it is irrelevant. (Specific gravity is an early measure of density that does not require the modern concept of mass, because it is a unitless ratio of an object's weight per volume divided by liquid water's weight per volume.) Galileo's theory of falling bodies is explicitly conditioned on the sameness of specific gravity: "Large and small bodies [fall] with the same speed provided they have the same specific gravity" (Galilei, 1914, p. 110). Modern theories, however, reject the relevance of density and, consequently, specific gravity. Modern scientists take the experiment as exemplifying the fact that bodies with different masses fall with the same acceleration, provided they are dropped in the same gravitational field, and provided they are slowed by friction equally. Thus, modern science has a different interpretation of Galileo's experiment from that given by Galileo himself. Nevertheless, the interpretations are not contradictory. Roughly speaking, the relevant features for Galileo and for the modern scientist belong to the same family. Attenuation and repleteness are two ends of a continuum, and Galileo's experiment is near the attenuated end of the continuum, which helps explain why it is good science. An essential part of scientific training involves learning which features are relevant and which are irrelevant to a datum's exemplification (see Elgin, 1999, who discussed Kuhnian paradigms in this regard).

Attenuated Exemplars at the Intersection of Science and Policymaking

Now let us consider (1) the sample median as a mathematical function of the sample data and (2) the sample median as fodder for a policy discussion. I argue that the

median is an attenuated exemplar, but the median is a more attenuated exemplar in a mathematical context than in a policy context. Take a sample median test score just after it has been calculated, just as it is transitioning from a mathematical function to a political fact, and let us say that, in this transitional moment, the sample median is at "the mathematical/political borderline." The sample median score belongs to Marie, whose score of 95 exemplifies non-outlying values of the sample distribution of test scores. At the mathematical/political borderline, Marie's score has extremely few relevant features, and they are all akin to each other: *non-outlying value of the sample distribution*, *value of the sample distribution*, *typical test score*, etc. At the mathematical/political borderline, the only features relevant about Marie's score are its value of 95 and the fact that one can arrange the other scores into two groups of equal size, one of which contains only values equal to or greater than 95 whereas the other contains only values equal to or less than 95. At the mathematical/political borderline, the sample median is an extremely attenuated exemplar. With a minimum of training in data analysis, the reference of the median is clear at the mathematical/political borderline.

But when we consider the policy implications of the median—having now crossed the mathematical/political borderline—the reference of the median may not be so clear. A teacher worried about grade inflation may ascribe the property *too high* to 95. A teacher who knows that Marie can do better than average may ascribe the property *too low* to 95. Reasonable interpreters may disagree about the practical or policy implications of the sample median. However, because the sample median is such an attenuated exemplar at the mathematical/political borderline, they will always agree that the sample median represents a non-outlying value of the sample of test scores. At the

mathematical/political borderline, the sample mean is neither too high nor too low (assuming that it was calculated correctly), and it refers to the property of *non-outlying value of the sample of test scores* or *test-score value* or *typical test score* or some other property in the same family. Once we begin to consider policy implications, the features relevant to the median as an exemplar can explode, but the center of the explosion is the point where the statistic transitioned from containing purely mathematical to political content, and at that point the exemplar was attenuated. All legitimate policy implications from the statistic should be traceable back to that point. Thus, the median marks common ground (or ground zero, as the case may be).

Replete Exemplars in Science

Hesse's (1966) argument about analogies in science may also apply to exemplars in science, suggesting that some repleteness in exemplars can be important for science. In a nutshell, Hesse argues that analogies in science are partly positive, partly neutral, and partly negative. An analogy, "A is like B," is positive insofar as A and B clearly share the same properties. The analogy is negative insofar as the properties of A and B clearly differ. An analogy is neutral insofar as there are some properties which are neither clearly shared nor clearly differing.

Neutral analogies can be very interesting scientifically. Hesse contemplates the analogy between light waves and sound waves. Some properties clearly apply to both light waves and sound waves: they are *diffractive* and *reflective*. Other properties clearly apply to either light waves or sound waves, but not both: one is *visible* and the other is *audible*. With regard to still other labels, we are undecided, and this neutral analogy becomes a clarion call for scientific investigation. For example, consider the question of

whether both types of waves *move through a medium*. Sound waves move through a medium (e.g., air), but do light waves move through a medium (e.g., the "ether")? Michelson and Morley designed an experiment to exemplify the relative motion of ether wind, but their experiment failed to exemplify that property. Modern physicists now believe that their experiment was actually one of the first to exemplify properties relevant to the truth of the following theories: "ether wind does not exist" and "the speed of light is constant."

Adapting Hesse's reasoning to our own discussion, we can say that an exemplar has positive, negative, and neutral aspects of exemplification. Some features are clearly relevant to an exemplar's exemplification, other features are clearly not relevant, and still others may or may not be relevant. Sometimes we have a lot to learn from those features of uncertain relevance, because they suggest replications of the experiment that vary those features while holding others constant so as to determine if the uncertain features indeed matter. To refine and update Galileo's famous experiment, we would take as neutrally exemplified the feature of *equal specific gravity* shared by Galileo's cannonball and musket ball. We would replicate the experiment, keeping the positively exemplified features, e.g., *different sizes* and *dropped at the same time*, but changing the neutrally exemplified feature of *equal specific gravity*, and see if the two balls still share the experimentally crucial feature of *landing at the same time*.

Replete Exemplars in Policymaking

At the mathematical/political borderline, the sample median is so attenuated that there is only positive and negative exemplification but no neutral exemplification. Not until we move deeper into political territory, by embedding the sample median in policy

arguments, do we get neutral labels, and they are generally abundant, perhaps more so than science would favor. The test score of 95 possesses the property of *non-outlying value of the sample distribution*, but now it also possesses other properties such as *purportedly indicative of subject mastery* and *belonging to Marie who is an underachiever*. Are these other properties relevant to the median's exemplification? The answer depends on how the median is framed in a policy debate. The former property may be relevant in a policy debate about grade inflation. The latter property may be relevant in a parent-teacher meeting about Marie's performance. For these debates, it is important for the median to refer to more than the property of *non-outlying value of the sample distribution,* but once we have moved beyond that undisputed property there may be a lot of room for disagreement. Attenuation is neither good nor bad, but the context can make it either good or bad. Insofar as data analysts seek a starting point for discussion on which everyone can agree, then the median's attenuation at the mathematical/political borderline makes it a good exemplar. If, however, the data analysts seek neutral exemplification to guide deeper inquiry, then they must cross the mathematical/political borderline into applied territory and bring outside information to bear. That additional information may include qualitative analysis and policy analysis.

Moderate vs. Extreme Exemplification

In this section, I argue that that the sample median, as a perfectly moderate exemplar, avoids the problems of vagueness associated with the label of *outlier*. I draw heavily from Scheffler's (1979) analysis of vagueness. First, I discuss the labeling of features that belong to ordinal schemes, and then I discuss some advantages and

limitations of using numeric labels. Finally, I argue that moderate exemplars can be particularly elucidating for quantified ordinal schemes.

Ordinal Schemes and Labeling

Some features belong to ordinal schemes. Consider the features exemplified by paint swatches. The features *light* and *dark* belong to the same ordinal scheme, and we might label that scheme *shade*. *Matte* and *glossy* belong to the ordinal scheme of *gloss*. The features *red*, *orange*, *yellow*, *green*, *blue*, *indigo*, and *violet* belong to the scheme o *hue*, and the scheme is ordinal if we take the color spectrum as an ordering guide. Not all ordinal schemes are continuous, but the schemes mentioned so far lend themselves to a continuous construal, such that between any two ordered features we can conceive and label an intermediate feature. We might generate intermediate labels by modifying existing labels with adverbs, adjectives, prefixes, suffixes, or compound constructions (e.g., "very dark," "midnight blue," "semi-gloss," "purplish" and "yellow-green," respectively) or by inventing entirely new labels such as "Quicksilver," "North Star," and "Krypton." Another way to generate new labels is to correlate the scheme with the set of real numbers line by mapping colors onto the number line based on their shade, gloss, or hue. Then we can use numbers as labels, and those numbers can have as much decimal precision as we require for our purposes.

Practically speaking, even if we correlate color labels with real numbers so that any level of decimal precision is conceivable, only a limited decimal precision is reasonable in any given context. Suppose that you are choosing paint for a bathroom wall. No consumer can perceive indefinitely minute differences in gloss, shade, and hue, and no producer has the complexity of paint engineering or financial justification to

supply indefinitely many differences in gloss, shade, and hue. Color labels tend to be vague labels. “Quicksilver,” “North Star,” “Krypton,” “Jubilee,” “Storm Cloud,” “Granite Peak” and “Outerspace” share (approximately) the same gloss and hue, but they differ with respect to shade, with “Quicksilver” being the lightest and “Outerspace” the darkest. In terms of shade, where does “Quicksilver” end and “North Star” begin? Where does “North Star” end and “Krypton” begin? The labels are vague insofar as their applicability is uncertain. Vagueness is context-dependent. We can decrease the extent of vagueness by more precisely defining our labels and more precisely measuring our colors. Of course, there are practical limits to that precision. In the end, we work with ranges that meld imperceptibly into one another.

Ordinal Schemes and Exemplification

The “Quicksilver” label applies to a range of colors, but a paint swatch exemplifying “Quicksilver” might contain only one color from that range. Recall the jointly necessary and sufficient conditions of exemplification: possession of the property and reference to the property by possession of the property. The “Quicksilver” label that applies to each color in the exemplified range of colors also applies to the exemplifying paint swatch. Furthermore, the exemplifying paint swatch refers to the exemplified range of colors. By these two criteria for exemplification, any color from the range of “Quicksilver” colors can exemplify “Quicksilver,” but that is not to say that any color from the range should be used as an exemplar. An extremely dark “Quicksilver” will be imperceptibly different from an extremely light “North Star.” In turn, an extremely dark “North Star” will be imperceptibly different from an extremely light “Krypton.” If the main purpose of the paint swatch is to differentiate the seven shades of grey, then, for the

purposes of exemplification, instances in the middle of each color range will serve better than extreme instances. Thus, in the context of differentiating ranges within a continuum, we want moderate exemplars as opposed to extreme exemplars. In other contexts, we may want extreme exemplars, for instance, when the two poles of a continuum would be easy to label (e.g., "dark" and "light") and the middle portion would pose the difficulties. In still other contexts, the moderation or extremity may not matter.

Marie's score of 95, the sample median, is a perfectly moderate exemplar of a *non-outlying value of the sample distribution*. Marie's score is certainly not an outlier with respect to quantity (assuming that the sample distribution is unimodal). As a perfectly moderate exemplar, the median avoids any vagueness of the label "non-outlying value of the sample distribution." For some distributions, "outlier" is not a vague label; every value falls neatly under the label or outside the label. For other distributions, "outlier" is a vague label. To understand the difference, consider the common practice in data analysis of using the interquartile range (IQR) to label outliers. The IQR is the difference between the first quartile and the third quartile, essentially a range encompassing the middle 50% of values. Data analysts use the IQR as a yardstick for outlier labeling. Data analysts have agreed to label as an "outlier" a value that falls two IQRs below the first quartile or above the third quartile. Likewise, they agree that a value that falls only one IQR below the first quartile or above the third quartile is not an outlier. There is, however, room for disagreement in the area between these two conventions. John Tukey, the greatest advocate of this method of outlier labeling, used 1.5 IQRs as his breakoff point. When a student asked him to explain the rationale for using 1.5, Tukey answered, "Because 1 is too small, and 2 is too large" (Paul Velleman, personal

communication). A data analyst's choice can depend on custom, intuition, and preference, but there are objective considerations as well: sample size, tolerance for false positives, tolerance for false negatives, and theoretical probability distributions (Banerjee & Iglewicz, 2007). This latter group of considerations can be called objective in the sense that they are open to reasoned criticism from the relevant community (e.g., fellow researchers or members of the school community). Nevertheless, aside from sample size, the objective considerations are difficult to pin down at best. Rough approximations are the order of the day. Therefore, an observation that is 1.6 IQRs above the third quartile may prove extremely difficult to label objectively as either an outlier or a non-outlier. In this way, "outlier" is a vague label. It is outside the scope of this dissertation to make "outlier" any less vague than it is. A great virtue of the sample median is that it steers clear of the vagueness associated with outlier labeling, because the sample median is a perfectly moderate exemplar.

The Median, Rationally Reconstructed as an Attenuated, Moderate Exemplar

As an attenuated, moderate exemplar, the sample median supports the objective comparison of values within and between groups. This claim warrants an extended discussion. I will begin with a rational reconstruction of the median, incorporating my prior arguments that (1) the median as an attenuated exemplar is easy to interpret and (2) the median as a moderate exemplar avoids any vagueness of the "outlier" label.

What is a Rational Reconstruction?

A rational reconstruction is a story that takes facts, principles, standards, methods, categories, goals, and/or values and shows how their interplay can reasonably lead to a conclusion. My rational reconstruction of the median is a story about two data analysts

with diametrically opposed theories. Each one selects a particular exemplar from a given sample so as to support their respective theories. I first consider a biased method of selecting exemplars, which I call the cherrypicking method. Then I consider an objective method of selecting exemplars, which I call the jury-selecting method. The objective method ultimately selects the median as the exemplar. Thus, the story contends, the median is an objective basis for comparison. Rational reconstructions may be schematic, and they do not purport to reflect the actual course of events. Their goal is to be just concrete enough to concretize an abstraction. They omit details that are, for current purposes, irrelevant. The purpose of my generic story is to show how the sample median can provide an objective basis of comparison.

Objectively Comparing Values within a Sample

Let us consider the minimum score in a sample. The lowest test score in a sample may be of particular worry to an educator, because it represents the poorest performance in the group. After all, if any score is low, the lowest score is low. But how low is "low"? To answer this question, there are standard-setting procedures (e.g., the Angoff method and the bookmark method) that rely on information extrinsic to the sample distribution of test scores. Nevertheless, let us restrict ourselves only to the sample distribution of test scores. We can answer our question by comparing the score to others in the sample distribution. What makes one score an elucidating basis of comparison for another score?

Imagine two teachers with axes to grind. One teacher wants to show that the lowest score is very low indeed, so she points to the difference between the lowest score and the highest score. Another teacher wants to show that "very low indeed" is an exaggeration, so he points to the difference between the lowest score and the next-to-

lowest score. Neither teacher will be happy with the other's basis of comparison, nor should they be. The problem with either basis of comparison is that, in the deliberation, the will gets more weight than the world. That is, the wishes of each proponent determine, to a large extent, each person's findings. Each one is cherrypicking the data. Cherrypicking is a biased sampling method born of wishful thinking. The cherrypicker systematically neglects exemplars that suggest the falsity of a preferred theory in favor of those that suggest the truth of the preferred theory.

I propose an alternative to the cherrypicking method: a *jury-selection method*. Suppose we allow the two opponents to take turns, each ruling out a score as a basis of comparison until only one score remains. That remaining score will be close to the sample median. If we seek a general basis of comparison (a basis of comparison for any score, not just the lowest score), we can include the lowest score as a candidate for the general basis of comparison. Of course, as a candidate, the lowest score will be eliminated early along with the highest score, and we will end up with the median score as the compromise value.

I call this method *jury selection* in loose reference to the legal process of peremptory challenges in the process of selecting trial juries. Generally, a trial will have a prosecuting attorney and a defense attorney, each with a preferred theory: guilty and not guilty, respectively. For the sake of fairness, it is essential that both sides be represented. Each side is expected to prefer a biased sample when choosing jurors. If each side gets a fair chance, however, the sampling biases will be offset, or so the rationale goes. In the United States, each side is allowed a certain number of peremptory challenges whereby some of the randomly selected jury candidates can be eliminated based on sampling bias

alone. A justification for the process of peremptory challenge is that random selection alone is not sufficient to ensure a jury representative of common citizens. When a jury system selects citizens randomly to serve as jurors, the underlying presumption is that the group of citizens chosen is generally typical in the appropriate respects. Nevertheless, there are likely exceptions to the general presumption, so, according to the process of peremptory challenge, the prosecution and defense should each have an opportunity to eliminate those potential jurors with extreme views. A statistical analogy is the "trimmed" (or "Winsorized") sample mean that is computed in a sample from which extreme values have been eliminated systematically. The sample median is the ultimate trimmed sample mean.

Objectively Comparing Values between Subsamples

The median student is not only a reasonable basis of comparison within a group, but also a reasonable basis of comparison between groups. Suppose that we are comparing scores for female and male students, in a situation where the males generally outperform the females. Again, imagine two teachers with theoretical axes to grind. One teacher wants to show that the gap in achievement by sex is huge, so she points to the difference between the lowest score from the subsample of female students and the highest score from the subsample of male students. Another teacher wants to show that that there is no such gap, so he points to the difference between the highest female score and the lowest male score. Both teachers are cherrypicking. Their choices of exemplars are biased in favor of their preferred theories. The jury-selection method, however, offsets their biases, because each side takes turns disallowing a member of the group as

an exemplar for comparison. In the end, the two teachers will settle on the median student in each subgroup as exemplars for comparison.

Perhaps one may wonder if the fairness of comparing one subgroup median to another subgroup median has nothing to do with the median *per se* but rather everything to do with the equality of moderation or extremity between the two subgroups. In other words, perhaps comparing the highest score in each subgroup (or the lowest) would be just as fair. Indeed, in athletic competitions, the two teams generally pit their best athletes against one another. What is unfair about that? In the case of athletic competitions, nothing is unfair as long as the competitors play by the rules. Scientific inquiry, however, is fundamentally different from athletic competition in this respect. In scientific inquiry, there is no set of hard and fast rules such that, if they are followed and the conditions that they set forth are met, victory (i.e., the expansion of knowledge) is certainly attained (see Elgin, 1999, for an analysis of pure, perfect, and imperfect procedural epistemologies). Accordingly, for scientists, there are guidelines instead of hard and fast rules. One such guideline is that outliers are not a reliable basis for comparing groups. The maximum and the minimum may be outliers, but the median is never an outlier (at least in unimodal distributions). What are outliers, then? And why are they not a good basis of comparison?

Outliers vs. Medians as Bases of Comparison

A sample outlier is an especially extreme value in a sample distribution. Not all sample distributions have outliers, and some sample distributions have borderline outliers, making "outlier" is a vague label in context. Because outliers are especially extreme, we cannot rely on two sample distributions having equally extreme outliers. Scientists generally want exemplars for between-group comparisons to be equal in

moderation or extremity. Therefore, they tend to avoid outliers as bases of between-group comparison. That is the gist of my argument, and it may seem fairly elementary, but two questions merit further consideration. First, why do scientists generally want exemplars for between-group comparisons to be equal in moderation or extremity? Second, given that outliers from different groups can differ in their extremity, can medians from different groups also differ in their extremity?

Scientists generally want exemplars for between-group comparison to be equal in moderation or extremity because this feature gives them a way to account for individual-level differences within groups. Between-group comparisons are group-level comparisons that answer questions about group differences, whereas within-group comparisons are individual-level comparisons that answer questions about individual differences. Scientists do not want to confound between-group differences and within-group differences, because science favors attenuation. Generally, when scientists ask questions about differences, they want to understand either group differences or individual differences. Unless the data analyst can account for within-group differences separately, it will be unclear whether the difference being exemplified by the two students selected is a between-group difference exemplifying group-level differences or a within-group difference exemplifying individual-level differences. In order to get clear answers to their questions about group-level differences, scientists want to account for within-group differences.

One unsatisfactory method for ruling out the exemplification of within-group differences is to stipulate that the two students are drawn from different groups. In this case, one might argue, the two students cannot exemplify within-group differences,

because they are not within the same group. This strategy is unscientific, however, because it assumes a fact that should be evidenced (or not evidenced) by the data. The strategy assumes that the two groups are objectively different in terms of test scores. A scientifically objective method should be agnostic as to whether there are between-group differences in test scores. Such a scientifically objective method involves choosing exemplars of equal moderation or extremity. Differing in moderation or extremity is a hallmark of within-group differences. If the data analyst compares two equally moderate or two equally extreme exemplars, then the analyst is effectively accounting for within-group differences, and consequently any difference in test scores of the equally moderate or extreme exemplars exemplifies in an attenuated fashion group-level differences in values.

My argument against outliers hinges on the point that scientists want exemplars for between-group comparison to be equal in moderation or extremity. But that brings us to the second question: can medians from different groups differ in their moderation? Can the sample median score of female students be more (or less) moderate than the sample median score of male students? If so, then my argument against outliers would apply to medians as well. But this is not the case, because all medians are equally moderate. Once we have defined *moderate*, moderate is moderate, but extreme is more or less extreme depending on its distance from moderate. We measure extremity in terms of distance from a moderate point. Consider the rationale behind the construction of box plots, which use IQRs from the first or third quartile to identify potential outliers. Or consider *z*-scores, which use the standard deviation of a sample as a yardstick to measure the distance of a typical observation from the sample mean, with the sample standard

deviation itself being a sort of "average distance from the sample mean" among all observations in the sample. Our practice of measuring extremity in terms of distance from a moderate point suggests that defining *moderation* must necessarily precede measuring *extremity*. Thus, as long as we use the same definition, such as the median, to define *moderate* for each group, then the medians of two groups are equal in extremity and, consequently, moderation.

Why make so much of measuring extremity in terms of distance from a moderate point? After all, if we measure moderation in terms of distance from the extremities, would defining *extreme* not become conceptually prior to measuring *moderation*, thus turning my argument on its head? No, my argument stands because, even if we measure moderation in terms of distance from the extremities, we must nevertheless define *moderation* along with *extremity* in order to do so. Whether we measure moderation or extremity from the middle point or an end point, the middle point is a pivotal point and needs to be defined.

Consider two points, A and B, which differ in their distance from the middle point. The point further from the middle point is more extreme than the one closer to the middle point. The middle point alone is a sufficient guide to relative extremity. Now consider two points, C and D, which differ in their distance from an endpoint. The point further from the endpoint may or may not be more moderate than the one closer to the end point. It all depends on the middle point. If points C and D are between the endpoint and middle point, then we get one answer. If the middle point is between points C and D and the end point, then we get another answer. As we observe points increasingly distant from the extreme, we may note that the points become increasingly moderate, but only

until we observe the middle point, after which the points again become increasingly extreme. The role of the middle point as a pivotal point gives rise to the childhood riddle: "How far can the rabbit run into the woods?" The answer: "Halfway, because if it runs any farther in the same direction, then it is running *out of* the woods." Defining moderation is both necessary and sufficient to give us a guide to relative extremity or moderation, so it makes sense to treat a definition of moderation as conceptually prior to measuring extremity. Therefore, if we define moderation in terms of the median, then all medians are equally moderate and consequently good bases of between-group comparison.

Fictive vs. Non-Fictive Exemplification

I have argued that sample medians (at least for samples with an odd number of values) are attenuated, moderate exemplars, and now I will argue that sample means are also attenuated, moderate exemplars. As an exemplar of a non-outlying value of the sample distribution, the sample mean must be a non-outlying value of the sample distribution. That is the possession condition of exemplification. The possession condition, however, does not restrict possession only to non-fictive possession. Fictions possess properties. Achilles fictively possesses the properties *Greek* and *wrathful*. Hamlet fictively possesses the properties *Danish* and *melancholic*. Don Quixote fictively possesses the properties *La Manchan* and *quixotic*. (If Don Quixote does not possess the property *quixotic*, who does?) Similarly, the mean test score among students in a class fictively possesses the property of *non-outlying test score among students in the class*. Articulating and defending this point is the work of the present section.

First, note that, for sample sizes with an even number of values, the sample median is the mean of the two middlemost values. Unless those two middlemost values are identical, such a sample median is itself not a value of the sample distribution. Nevertheless, I think that medians from even-numbered distributions function epistemically in the same way as medians from odd-numbered distributions. Data analysts do not treat the two types differently, because they fulfill identical functions. The median value retains its nature even when that value belongs only fictively to the sample distribution. This point merits further discussion, but I hope that my insight about medians of even-numbered distributions will encourage the reader to suspend disbelief regarding the usefulness of fictive exemplification.

In this section, I make the following argument that the mean of a sample distribution indirectly elucidates the world by this referential chain:

1. The mean is a *fictive representation* of a non-outlying value of the distribution.
    a. The mean belongs to the *denotative* symbol system of mathematics.
    b. The mean itself *neither denotes nor purports to denote*.
2. The mean *exemplifies* the property of *representation of a non-outlying value of the distribution*.
    a. The mean *possesses* the property.
    b. The mean *refers* to the property by possessing the property.
3. As an *attenuated* and *moderate* exemplar, the mean *elucidates* representations of non-outlying values of the distribution.

a. As an *attenuated* exemplar, the mean is easily interpretable as representing a non-outlying value of the distribution.
b. As *a moderate* exemplar, the mean represents an unambiguously non-outlying value of the distribution.
c. As an *attenuated*, *moderate* exemplar, the mean is a good basis for comparison.

4. Because the reference of the non-fictive representations of non-outlying values of the distribution is denotative, *elucidating* the representations is tantamount to *elucidating* the world.

Since (as stated in item 4 above) elucidating sample values, which are denotative representations, is tantamount to elucidating the world, and since (as indicated in item 3) means elucidate sample values, therefore means, although they are fictive, elucidate the world. I will now devote a subsection to each of the four points in the argument.

The Mean as a Fictive Representation

A fiction is a type of representation. Some representations denote, but others do not. A fictive representation neither denotes nor purports to denote. Consider three maps: a map representing the location of Brookline's Green Hill, a map misrepresenting the location of Brookline's Green Hill, and a map representing the location of Narnia's Green Hill. The first (correct) map both denotes and purports to denote the location of Brookline's Green Hill. The second (erroneous) map does not denote the location of Brookline's Green Hill, but it *purports* to denote the location of Brookline's Green Hill. The third (fictive) map neither denotes nor purports to denote, but it nonetheless represents the location of Green Hill in a fictive world, namely C. S. Lewis's magical

world of Narnia. In this section, I argue that the mean is like the third map: it is a representation that neither denotes nor purports to denote.

What is denotation? Denotation is a type of reference. Hitherto, I have focused on another type of reference—exemplification. The primary difference between exemplification and denotation is that an exemplifying symbol refers by possession (i.e., exemplifying a property requires possessing the property) but a denoting symbol does not refer by possession. Whereas exemplification is reference by possession, denotation is reference irrespective of possession. Sometimes a denoting symbol happens to possess the denoted property, but, insofar as the reference is denotational, the possession of the property is incidental to the reference to the property. The map that accurately represents the location of Brookline's Green Hill refers to *the location of Brookline's Green Hill*, but the map does not literally possess *the location of Brookline's Green Hill*, because, presumably, the map itself is not sitting on Brookline's Green Hill. Even if the map happened to be located in Brookline's Green Hill, its location would be irrelevant to its usefulness in representing the location of Brookline's Green Hill (unless the map contains a you-are-here marker, in which case the location of the map does matter). Thus, the accurate map denotes *the location of Brookline's Green Hill*. Again, denotation is reference, not by possession.

The map is a symbol, and it belongs to a symbol system of geographical maps. Some knowledge and skill are required to interpret, for the purpose of geographical orientation, a piece of paper inscribed with two-dimensional symbols. However, because geographical maps belong to a system of symbols, the knowledge and skill acquired to interpret one geographical map largely carry over to interpreting other geographical

maps. In general, much of learning, especially book learning, is the acquisition of knowledge and skill in and through various symbol systems. Some symbol systems teach us about the world by denoting properties, objects, relations, and so forth in the world. This fact does not imply either that such a symbol system always denotes or that such a symbol system can teach us about the world only by denotation. With these important caveats in mind, I describe as *denotational* any symbol system that can teach us about the world by denoting properties, objects, relations, or anything else that exists in the world. This will allow me also to succinctly define *representation* for my purposes: a representation is a symbol belonging to a denotational symbol system.

A representation can fail to denote. Such is the case with the erroneous map that misrepresents the location of Brookline's Green Hill. Suppose that, by typographical error, Fisher Hill was mislabeled "Green Hill," and Green Hill was not labeled at all. In this case, the map would erroneously put Green Hill on the wrong side of Route 9. Thus, the map would purport to represent the location of Brookline's Green Hill but would fail to denote the location of Brookline's Green Hill, because it fails to refer to the location of Brookline's Green Hill. Such things can happen in mapmaking. But my point here is that an erroneous map is still a map; a misrepresentation is still a representation. Even though it fails to denote, the erroneous map still belongs to the denotational symbol system of geographical maps.

Some representations succeed at denoting (e.g., the correct map), and other representations fail to denote (e.g., the erroneous map), but still other representations do not even try to denote. The map of Narnia's Green Hill falls into this category, because Narnia does not exist, nor does Narnia's Green Hill, nor does the location of Narnia's

Green Hill. The map contains a symbol for the location of Narnia's Green Hill, but that symbol refers to nothing that exists. The map thus denotes nothing. Nevertheless, since it belongs to the denotational symbol system of geographical maps, the map is a representation. More precisely, it is a fictive representation. A fiction is simply a type of representation that neither denotes nor purports to denote.

The sample mean is a fictive representation. It is a representation because it belongs to the denotational symbol system of mathematics with its numbers, operators, variables, and logic. Mathematical symbols can refer to properties, objects, relations, and so forth in the concrete world. A number can denote a student's test score and, by extension, the academic achievement of that student. A theoretical mathematician, however, may not give one hoot whether her instance of the symbol system denotes anything beyond the purely mathematical realm. Mathematical symbols need not refer to properties, objects, relations, or anything else in the concrete world. The mean test score, for instance, does not denote the test score of a student, nor does it denote or purport to denote anything else beyond the purely mathematical realm. If the mean test score happens to equal the test score of one or more students, the equality is coincidental. The mean does not refer by denotation but, rather, by exemplification.

The Mean as an Exemplar Representation

The mean test score is a fictive representation of a non-outlying value of the sample distribution of test scores, so it possesses the property of *representation of a non-outlying value of the sample distribution*. Therefore, according to the possession condition of exemplification, the mean test score can exemplify a representation of a non-outlying value of the sample distribution. According to the reference condition of

exemplification, the mean score does exemplify a representation of a non-outlying value of the sample distribution when we interpret it as referring to representations of non-outlying values of the sample distribution. I think that we do interpret the mean as referring to non-outlying values of the sample distribution. This interpretation is evidenced when we describe the mean value as a "typical value," even when it is not an actual value in the sample distribution.

In the final subsection, I will discuss the difference between, and bridge the gap between, *exemplifying a representation of a thing* and *exemplifying the thing itself*. In the meantime, suffice it to say that the difference boils down to practically nothing in the present case. We ultimately want to talk in terms of exemplifying things themselves, but fictions are essentially representations, so any thorough discussion of fictive exemplification must take us through the exemplification of representations first. We can, nevertheless, do some terminological housekeeping to cut through the clutter of "representation" talk. Let us say that a fictive representation of a property *fictively possesses* the property. Thus, Don Quixote, a fictive representation, fictively possesses the property *quixotic*. Likewise, let us say that a fictive representation of a property *fictively exemplifies* the property itself when the fictive representation exemplifies a representation of the property. Thus, Don Quixote fictively exemplifies *quixotic*. With this locution, instead of talking about the mean (non-fictively) exemplifying a representation of a non-outlying value of the sample distribution, we can describe the mean as *fictively* exemplifying a non-outlying value of the sample distribution.

The Mean as an Elucidating Exemplar

For a sample distribution, the mean can be arithmetically calculated by summing all the values in the distribution and dividing by the sample size. To arithmetically calculate the mean test score for a class, we can sum the test scores from every student and divide by the number of students. A mathematical equivalent to the arithmetic method for calculating the mean is the method of ordinary least squares (OLS). The mean of a distribution is the value that minimizes the sum of squared differences between the value and every value in the distribution. By the OLS method, calculating the mean is a minimization problem solvable through differential calculus. The OLS method can be visualized as striking a balance between values above and below the mean. In my rational reconstruction of the median, I proposed an alternative to the cherrypicking method, which I called the jury-selection method. In this subsection, I discuss the features of the jury-selection method that make it conducive to scientific objectivity, and I argue that (under some conditions, at least) the OLS method shares those features even if it is otherwise disanalogous to the jury-selection method.

In the jury-selection method, two data analysts with diametrically opposed theories are choosing exemplars. The theories are diametrically opposed in the sense that an exemplar supporting the truth of one theory equally supports the falsity of the other, and vice versa. From each subgroup, the data analysts are allowed to take turns eliminating observations for candidacy as exemplars until only one observation remains, and that observation becomes the methodically chosen exemplar for the subgroup. In my introduction of the jury-selection method, I assumed an odd sample size, and consequently each data analyst had used an equal number of peremptory challenges when

only the last observation remained. Suppose now that the sample size is even. When the last observation remains, one data analyst has had one fewer peremptory challenge than the other. Some legal systems actually give the defense attorney more peremptory challenges than the prosecuting attorney because the presumption of innocence favors the defendant. To achieve scientific objectivity, however, neither data analyst should have any presumption in her favor. Therefore, unless the last two candidate exemplars are identical (in which case either will be equally acceptable to the data analysts), we have a problem. There is an easy solution, however, because the exemplar is attenuated and quantified. Because there is only one relevant feature (i.e., the outcome construct), and because that feature is quantified, the two remaining candidates are quantities. Therefore, we can identify a third quantity halfway between those two quantities, the mean between them, which, by compromise, can serve as the exemplar (albeit fictive exemplar).

What does the fictiveness add? Fictions can be as attenuated and as moderate as scientific objectivity requires. Taking the mean of the last two remaining values gives each side the same number of peremptory challenges. Moreover, the fictiveness yields more attenuation than the amount in the case with an odd sample size. When the sample size is odd, the median student has not only a perfectly moderate test score but also a last name, first name, birthdate, astrological sign, and a vast collection of other biological, social and biographical properties that may or may not be germane to the inquiry at hand. When the sample size is even, the median student has only a perfectly moderate test score—and no distractions. At the beginning of this section, I noted that data analysts treat both types of medians in the same way, and I conjectured that they do so because the two types function in the same way. Here is the basis for my conjecture: I think data

analysts treat both non-fictive and fictive medians as if they were as fully attenuated as the fictive median. Even though the non-fictive median student has a name, quantitative data analysts generally ignore it (at least until they cross over to qualitative data analysis).

What are the features of the jury-selection method that make it scientifically objective? First, the jury-selection method has no *a priori* presumption in favor of any theory; it is theory-neutral. Second, as a result of this theory-neutrality, every subgroup is treated in the same way, and within every subgroup, every observation is treated identically. Third, the jury-selection method yields moderate exemplars. I will discuss in turn each of these features of the jury-selection method, explaining how the feature lends itself to scientific objectivity and how the OLS method shares the feature.

The jury-selection method has no *a priori* presumption in favor of any theory, nor does the OLS method. An *a priori* presumption of a data-analytic method is a presumption, before any observations are made, about those observations and how they will be counted. For instance, the jury-selection method has an *a priori* presumption that the observations are rank-ordered. The OLS method has an *a priori* presumption that the observations are intervally scaled. The jury-selection method also has an *a priori* presumption that the two theories of the cherrypicking data analysts are diametrically opposed, and it further presumes that the cherrypickers will systematically eliminate observations according to those theories, but that because the two theories are diametrically opposed and absolutely extreme, they will effectively negate each other, causing the jury-selection method to end up being theory-neutral. The OLS method, in contrast, makes no presumptions about any theories involved. The OLS method does not even presume that any theories are involved at all; one can employ the OLS method with

or without a theory. The theory-neutrality of the method of exemplar selection lends itself to scientific objectivity because, as much as possible, scientifically objective exemplars should be chosen independently of the wishful thinking of any theory's proponents. Data do not speak for themselves, but that does not imply that the analyst (who may or may not have a theoretical axe to grind) should speak for the data. Rather, the analyst should, as much as possible, apply theory-neutral methods to choose spokespersons—i.e., representatives or exemplars—for the data.

Due to their theory-neutrality, the jury-selection method and the OLS method treat every subgroup and every observation within subgroups in the same way. This statement demands qualification, however. If the jury-selection method treated every observation in absolutely the same way, and if it treated one observation as the median, then it would treat every observation as the median. This and other absurdities follow unless we qualify the statement. To serve as a point of contrast, consider the differential treatment of observations by the cherrypicking method, which treats observations differently depending on whether the observations support or undermine the cherrypicker's theory. That differential treatment may be systematic by subgroup; the cherrypicker may favor high-valued observations in one subgroup and low-valued observations in another. It may be the case that, across groups, the cherrypicker may favor maximums or minimums or even medians, *if* they support the cherrypicker's theory. In fact, once a cherrypicker has found exemplars that support her theory, she may not bother to look through the rest of the data. On the other hand, the jury-selection method and the OLS method take into account the value of each observation, not relative to any one theory, but only relative to the other values in the distribution. The jury-

selection method gives each value identical weight, and it balances the values above and below the median. The OLS method gives each value a weight in accordance with its value, and it balances the values above and below the mean. Therefore, both the jury-selection method and the OLS methods seek to balance weighted values, and, moreover, each method weights each value based on a consistent, theory-neutral criterion.

It is not enough, however, to assure scientific objectivity that the method be neutral to theories and fair to observations. If that were enough, then simple random selection of one observation from each subgroup to be that subgroup's exemplar would be a scientifically objective approach. Such a method may be objective, but it is not *scientifically* objective. Scientists would reject this method because the sample mean, as an estimator of the population mean, would be wildly imprecise given the sample size of one. In this essay, however, I am focusing on samples only, not populations (although my conclusions about sample means generally apply to population means). In comparisons of subsamples, the random selection of exemplars is not scientifically objective, because scientists have a particular interest in exemplars that are equal in moderation or extremity. As I argued in my rational reconstruction of the median, by comparing exemplars of equal moderation or extremity, the data analysts are accounting for within-group differences so that the difference in exemplars exemplifies between-group differences in attenuated fashion. Moreover, science generally prefers moderate exemplars because they avoid the vagueness problems associated with the labeling of outliers, which are especially extreme values. Means, because they are fictive, can be exemplars as attenuated and as moderate as scientific objectivity requires, which is the thesis of this essay.

Elucidating Sample Values Is Elucidating the World

How do we learn from fictive representations such as the mean? Fictive representations can teach us about non-fictive representations. As a fictive exemplar of non-outlying values of a sample distribution, the mean teaches us about the sample values, which are themselves representations. Representations directly elucidate the world by perspicuously correlating in the right way with the world. Representations indirectly elucidate the world by elucidating representations that *do* directly elucidate the world. With regard to the correct map, which is a representation of *the location of Brookline's Green Hill*, I briefly wave my hand at "elucidate" and "perspicuously correlating in the right way." For sample means, I promise to cash out "elucidate" and "perspicuously correlating in the right way," after I do a little more foundational work. As a catchall, I will simply note that "elucidate" and "perspicuously correlating in the right way" can vary from context to context.

The correct map representing *the location of Brookline's Green Hill* elucidates the world by perspicuously correlating in the right way with the world. One way in which a geographical map correlates with the world is that the relative spatial relations between symbols on the map correlate with the relative spatial relations between locations in the world. In the world, Brookline's Fisher Hill is north of Route 9 and Green Hill is south of Route 9. Even a quickly sketched map of Fisher Hill, Route 9, and Green Hill should have the symbols for Fisher Hill and for Green Hill on opposite sides of the symbol for Route 9. On a high-quality map, the distances between each hill symbol and Route 9 will be proportional to the distances in the real world. These correlations are perspicuous to anybody familiar with the symbol system of geographic maps. Because the map

perspicuously correlates with the world, it elucidates the world. The map teaches us the relative spatial relations among locations in the world.

Now, let us explore the role of reference in my rough-and-ready analysis of how the map teaches us the relative spatial relations among locations in the world. In this section I introduced denotation, which is a symbol's reference to a property, object, or relation without the requirement that the symbol necessarily possesses the property, object, or relation. Most of this essay, however, has been about exemplification, which is a symbol's reference to a property, object, or relation by means of possessing it. The map denotes, but it also exemplifies. The map denotes locations, because it need not possess the locations to which it refers. The map, however, exemplifies the relative spatial relations among the locations, not because it possesses the locations themselves, but because it does possess the relative spatial relations to which it refers. The map's exemplification is attenuated, because it is easy to interpret, and the attenuation contributes to the presence of perspicuity, consistent with the phrase "perspicuously correlating in the right way with the world." Thus, representations can teach us about the world by denotation and exemplification.

Even though it does not denote, a fictional representation may nonetheless teach us about the world through exemplification. The fictive map representing the location of Narnia's Green Hill has little to teach us about the world, but it has served my purpose of exemplifying representations that neither denote nor purport to denote. Thus, fictive representations can teach us about representations via exemplification. Representations exist in the world, so if the fictive map representing the location of Narnia's Green Hill teaches us about representations, then it teaches us about the world. Perhaps fictive

representations can teach us about the world in other ways, but for the present purpose I need to establish only that fictive representations can teach us about non-fictive representations, because data are non-fictive representations of the world, and data analysts are trying to understand their world by understanding their data.

I will now proceed to discuss how fictive representations can teach us not only about representations but also about the greater world. Like a Russian nesting doll, the *Chronicles of Narnia*, containing C. S. Lewis's stories and maps, belong to a symbol system of Narnia narratives, which is a subsystem of the symbol system of fictional narratives, which is in turn a subsystem of the symbol system of narratives. Lewis's stories of Narnia, as opposed to his maps of Narnia, have much to teach us about the greater world, albeit indirectly. The *Chronicles of Narnia* exemplify representations of courage, betrayal, adventure, deception, nobility, faith, forgiveness, and self-sacrifice.

There is a difference (having to do with existence proof) between exemplifying a representation of self-sacrifice and exemplifying self-sacrifice. Aslan's self-sacrifice in *The Lion, the Witch and the Wardrobe (LWW)* does the former, not the latter. Recall that exemplification requires reference by possession. *LWW* possesses the property of *representation of self-sacrifice*. In other words, Aslan's self-sacrifice is a representation of self-sacrifice. In order to exemplify self-sacrifice, the exemplar would need to be an actual self-sacrifice, such as that performed by the four chaplains of various faiths on board the sinking USAT *Dorchester*, who stayed with the ship to hand out every last life preserver including the ones on their backs. An exemplar of self-sacrifice refers directly to self-sacrifice in the greater world because the exemplar is itself a self-sacrifice in the greater world. Thus, the four chaplains can serve as existence proofs that there is self-

sacrifice in the greater world, whereas Aslan cannot. In general, a key difference between exemplifying a representation of something and exemplifying the thing itself is that the former cannot serve as an existence proof, whereas the latter can.

In terms of direct reference to the greater world, there is a difference between exemplifying a representation of self-sacrifice and exemplifying self-sacrifice, but the difference evaporates insofar as the representation is denotative. Consider two interpretations of the biblical representation of Jesus' self-sacrifice. The first interpretation is that the biblical representation is an accurate historical representation and thus denotes actual historical events. The second interpretation is that the biblical representation is purely fictional and denotes nothing. The first interpreter views the denotation in this instance as providing a tight correlation between the biblical representation and the historical events, and so this interpreter takes understanding the biblical representation as equivalent to understanding the historical events of Jesus' self-sacrifice. Therefore, if C. S. Lewis's fictional representation of Aslan's self-sacrifice helps the first interpreter to understand the biblical representation of Jesus's self-sacrifice, then the first interpreter is indebted to Lewis for a lesson about the greater world. Non-fictive representations teach us directly about the world by denoting the greater world, where the denotation is a direct reference to the greater world. Fictive representations, on the other hand, by exemplifying representations, can teach us about the greater world indirectly through teaching us about non-fictive representations.

The second interpreter agrees that, if the biblical representation denoted the historical events, then there would be no difference, in terms of direct reference to the greater world, between understanding the biblical representation and understanding the

historical events. The second interpreter disagrees, however, with the premise. The second interpreter takes the biblical representation as fictive and thus as denoting neither historical events nor anything else for that matter. Thus, for the second interpreter, Jesus' self-sacrifice, just like Aslan's, is a fictive exemplar of sacrifice, denoting nothing but exemplifying representations of self-sacrifice. The second interpreter might nevertheless respect the Bible as a work of fiction that provides interesting or useful fictive exemplars. Although fictive exemplars cannot serve as existence proofs for the represented properties, objects, relations, and so forth, they can be powerful tools for indirectly understanding the greater world when they exemplify representations (such as representations of self-sacrifice) that at least sometimes do denote the greater world.

As a fictive representation, the mean teaches us about the world indirectly through teaching us about non-fictive representations. Those non-fictive representations are, for instance, the numbers that represent students' test scores and, by extension, their levels of academic achievement. Now, with a lot of help from my friends the psychometricians, I can fulfill my promise to thoroughly substantiate "elucidating" and "perspicuously correlating in the right way" for the mean, in its role as a (fictive) representation of *non-outlying value of the sample distribution*. The numbers that represent academic achievement are perspicuously correlated with the world in the right way for elucidating the world, at least according to the psychometricians (professional, amateur, or unwitting) who validated the measurements for the data analyst's inferences.

A set of measurements is *construct-valid* relative to an inference. Data analysts make inferences about properties that are not directly observable, such as academic achievement. A construct is a property that is not directly observable. A set of

measurements is construct-valid for an inference about a construct if a relevant inference about the set of measurements is tantamount to a relevant inference about the construct. The data analyst is asking about construct validity when she asks, "When we talk about differences in test scores, are we talking about differences in academic achievement?" She is asking about the correlation between the representations (i.e., the numbers) and the represented properties (i.e., the levels of the construct such as *academic achievement*). In the introduction to his *Mathematical Logic*, Quine (1981) commented that the symbol system of mathematics is a tremendously useful tool for reasoning about the world … *if* we can correlate objects and properties of interest with the series of real numbers. The conscientious psychometrician knows that Quine's *if* is a big one.

The sample mean elucidates the world only insofar as the sample values denote relevant objects, properties, relations, and so forth in the world. This fact gives rise to the data-analytic adage, "garbage in, garbage out." If the sample values denote relevant properties in the world, then the mean can serve as an attenuated, moderate exemplar of those properties, and the attenuation and moderation bring with them the scientific virtues that I have discussed throughout this essay. As an attenuated exemplar, the mean is easy to interpret. As a moderate exemplar, the mean avoids any vagueness of the label "non-outlier." As an attenuated, moderate exemplar, the mean provides a scientifically objective basis for between-group comparison.

Chapter 3: Why Infinite Populations for Statistical Inference?

As a teacher of applied statistics and consultant to data teams, I seem to face the same question whenever I describe the notion of statistical inference. In those special cases where we conduct a census of the entire population rather than drawing a random sample from the population (e.g., when we gather data from every student in the school, and the school population is the target population), why do we nevertheless use confidence intervals and null-hypothesis tests to make statistical inferences that are intended to provide generalizations from the sample to the population? In the case of a census, isn't the sample the population? My short answer is that, in order to rule out sampling error as the sole explanation, we conceive of the population as infinitely large, including not only actual subjects but also hypothetical subjects, and we can never sample such a population entirely. So, no, the sample is not the population, if our purpose is explanatory. The goal of this essay is to unpack that short answer. In the process, I address the component questions: What is sampling error, what are explanations, and what are populations? I devote a section to answering each of these important sub-questions.

In order to answer these questions, I delve into the philosophy of *counterfactuals* (i.e., counterfactual-conditional statements). In general, a counterfactual is a statement that can be framed in a standard form, as follows: If it were the case (which presumably it is not) that p, then it would be the case that q. In English grammar, we use the subjunctive mood to mark counterfactuals. Consider the following counterfactual, which I emphasize by the use of boldface text, embedded in Jacob Cohen's introduction to

statistical inference from his textbook *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*:

> In most circumstances in which regression and correlation coefficients are determined, the intention of the investigator is to provide valid inferences from the sample data at hand to some larger universe of potential data—from the statistics obtained for a samples to the parameters of the population from which it was drawn. Because random samples from a population cannot be expected to yield sample values that exactly equal the population values, statistical methods have been developed to determine the confidence with which such inferences can be drawn. … **A *sampling distribution* is a distribution of the values of a sample *statistic* that would occur in repeated random sampling of a given size, *n*, drawn from what is conceived as an infinite population.** Statistical theory makes possible the estimation of the shape and variability of such sampling distributions. We estimate the population value (*parameter*) of the sample statistic we obtained by placing it within a *confidence interval* (*CI*) to provide an estimate of the margin of error (*me*), based on these distributions. (Cohen et al., 2003, pp. 41-42)

I can restate the counterfactual in standard form: If it were the case (which presumably it is not) that an infinite number of repeated random samples of size *n* are taken from the population, then it would be the case that the distribution of values of the sample statistic obtained in each random sample is the sampling distribution of the sample statistic. In this essay, I analyze Cohen's introduction to the notion of statistical inference not because it is perfect but because it has proven useful to applied statisticians. In general, I intend to answer the question why it is useful for applied statisticians to reason counterfactually about repeated random samples from hypothetical populations that are infinitely large.

I argue that the concepts of *sampling error*, *explanation* and *population* involve counterfactuals. For *sampling error*, I make a philosophical distinction between *method counterfactuals* (i.e., counterfactuals about the method of inquiry, such as sampling from a census) and *target counterfactuals* (i.e., counterfactuals about the target of inquiry, such

as the students in a school). With regard to *explanation*, I argue that policy-relevant explanations involve target counterfactuals. As for *population*, I argue that infinitely large populations including both actual and hypothetical subjects can provide the target counterfactuals necessary to rule out sampling error as the sole explanation for the findings.

Thus far, I have used the terms *sample*, *census* and *population*, and I will develop the significance of these terms throughout this paper. In this introduction, I can offer preliminary definitions. A *sample* includes all the observed subjects that are the target of statistical inquiry. A *census* includes all the actual subjects (observed and unobserved) that are the target of statistical inquiry. A *population* includes all the subjects (actual and counterfactual) that are the target of statistical inquiry. Counterfactual subjects are possible-but-not-actual subjects, and I will have much more to say about them. Figure 1 is a two-by-two table that allows us to categorize all possible subjects into three types, based on two dichotomies: observed and unobserved, and actual and counterfactual. A *sample* includes all subjects of Type I. A *census* includes all subjects of Types I and II. A *population* includes all subjects of Types I, II and III.

	Actual Subjects	Counterfactual Subjects
Observed Subjects	I	--
Unobserved Subjects	II	III

Figure 1. A table for categorizing possible subjects into three types.

Some researchers (see for example Abo-Zena's 2010 entry "Sample Size Planning" in *The Encyclopedia of Research Design*) make a distinction between

"theoretical populations" (e.g., high-school principals in the United States) and "accessible populations" (e.g., high-school principals within a two-hour drive from the research university), but according to my definitions, their distinction is between censuses, not populations. That is because they are distinguishing between two sets of actual subjects, depending on the convenience with which they can be observed. Their distinction does not involve counterfactual subjects (e.g., possible-but-not-actual high-school principals in the United States). The thesis of this essay is that, for explanatory purposes, researchers must consider not only actual subjects but also counterfactual subjects.

As I proceed, I further define *counterfactual subjects* as possible-but-not-actual subjects. I argue that counterfactual subjects should be treated as fictional subjects—scientifically useful fictions—and *not* as future subjects. Researchers use random-sampling methodology to draw statistical inferences, so in this essay I focus exclusively on random sampling. Researchers cannot randomly sample from the future. Therefore, scientists cannot draw statistical inferences about subjects from the future. Until time travel is invented, subjects from the future cannot be the targets of statistical inquiry. Researchers can draw statistical inferences about past and present subjects, and they can also make the non-statistical bridging argument that the future will resemble the past and present in certain respects, such that statistical inferences about past and present subjects can be extended to future subjects. Nevertheless, the direct target of statistical inquiry must be limited in scope to the past and present, because these are the time periods from which researchers can randomly sample subjects in order to draw statistical inferences about past and present populations.

I argue that my use of counterfactuals to define *population* is consistent with theoretical statistics, but I must note that statisticians need not think of statistical inference as inference from a sample to a population. Instead, statisticians can think of statistical inference as inference from a statistic to a parameter. A statistician who takes the latter approach can eschew the concept of *population* altogether. To demonstrate, I offer two interpretations of the same model. The first, which I call a *population-minded interpretation*, invokes the concept of *population*, but the second (a *process-minded interpretation*) does not. After offering the two interpretations, I will demonstrate that, although they are different, they are not contradictory.

Consider the following statistical model, which a data team can specify in order to investigate the relationship between academic achievement and socioeconomic status (SES) among students at Anonymous Middle School (AMS):

$$GPA_i = \beta_0 + \beta_1 FREELUNCH_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$$

In particular, by fitting this model to appropriate data, the data team can estimate and investigate gaps in average student achievement by SES, using eligibility for free lunch as an indicator of low SES. In the model, the component variables, outcome GPA_i and question predictor $FREELUNCH_i$, represent the grade-point average (GPA) and free-lunch status (1 if eligible for free lunch, 0 otherwise) for the *i*th student. Although the model ostensibly describes the GPA of the *i*th student, its purpose, according to the population-minded interpretation, is to represent the underlying relationship between *GPA* and *FREELUNCH* in the population of students from which the *i*th student was randomly drawn. Thus, in the model, parameter β_0, represents the mean GPA for the subpopulation of students ineligible for free lunch. Model parameter β_1, represents the

difference in subpopulation means between students ineligible and eligible for free lunch. The error term, ε_i, represents the difference in the population due to all unobserved causes of the outcome, including measurement error, the omission of unspecified predictors, and individual variations in the outcome, between the *i*th student's GPA and the mean GPA of the subpopulation to which the *i*th student belongs. As an essential part of the model, along with the formal statement of the model, the data team hypothesizes that, in the population, these latter errors are distributed identically and independently, with a normal (Gaussian) distribution that has a mean parameter of 0 and a variance parameter of σ^2 which is homogeneous (homoscedastic) across subpopulations. During data analysis, the data team can fit this hypothesized linear statistical model to data from a sample that the data team has drawn randomly from the underlying population, using the traditional method of ordinary least squares. The fitted model that the data team obtains then provides an estimate for each of the unknown model parameters, β_0, β_1, and σ^2. As per convention, I indicate an estimated parameter with a hat (^). For example, $\hat{\beta}_1$ denotes an estimate of β_1. Of particular interest in practical research is the estimate $\hat{\beta}_1$, because it is an estimate of the difference in mean GPAs between the two subpopulations of students, those eligible for free lunch (i.e., low-SES students) and those ineligible for free lunch (i.e., mid- to high-SES students). Therefore, $\hat{\beta}_1$ can provide researchers and makers of educational policy with information about a gap in academic achievement according to level of SES.

By way of contrast, I can interpret the model not as a population-minded statistician but as a process-minded statistician. Process-minded statisticians reason about the data-generating process. For them, the hypothesized process that generated any GPA_i

datum has both a probabilistic component and a deterministic component. Process-minded statisticians model the probabilistic component of the data-generating process as a random variable, ε_i, the probability-density function of which is assumed to be normal with a mean of 0 and an unknown constant variance σ^2. They model the deterministic component as the sum of an unknown constant, β_0, and the product of an unknown constant and the value of a (nonrandom) variable (β_1 and $FREELUNCH_i$, respectively). In essence, process-minded statisticians treat the GPA of the ith student as a function of unknown constants, the value of a (nonrandom) variable, and a random variable. There is no need for a process-minded statistician to refer to any population. For the process-minded statistician, a parameter is a property of a model, not a property of a population. A parameter is just an unknown constant in a model, so a process-minded statistician would read Cohen's phrase "population parameter" as "model parameter." Process-minded statisticians use statistics (which necessarily derive from samples) to estimate model parameters.

Consider the physical process of radioactive decay. The process is essentially probabilistic. The half-life of a carbon-10 isotope is 19.26 seconds. This means that we can model the time until decay of a carbon-10 isotope as an exponentially distributed random variable with a median of 19.26 seconds (Stanford & Vardeman, 1994, p. 55). If, for the sake of simplicity, we focus on the half-life, we can imagine that, for each carbon-10 isotope at time t, Mother Nature flips a fair coin, and if the coin turns up heads, the isotope decays before t + 19.26 seconds; if tails, it does not. (I personify the model's probabilistic component as the result of Mother Nature's coin flip, in the hope of aiding the understanding of readers who, like me, have trouble with the abstractness of random

variables.) There is no need to invoke the concept of *population*. If there were only one carbon-10 isotope in all of existence at time *t*, then Mother Nature would flip a fair coin for that isotope to determine whether or not the isotope decayed before *t* + 19.26 seconds. For a process-minded statistician, the 19.26 is just a parameter (formerly unknown, presently known) of a model with a random variable (which I personify as the result of Mother Nature's coin flip) that describes the time until decay of a single radioactive isotope.[4] On the other hand, for a population-minded statistician, 19.26 is the number of seconds it takes for half the population to decay, where the population is conceptualized as an infinitely large number of carbon-10 isotopes at time *t*, including both actual and hypothetical ones.

By making this distinction, I am not implying that process-mindedness and population-mindedness are contradictory. A process-minded statistician would readily

[4] It may seem that the concept of population is implicit in the process-minded interpretation, in two ways. First, it may seem that the concept of population is implicit in the interpretation of probability, e.g., the 0.50 of Mother Nature's coin. Second, it may seem that the concept of population is implicit in the discovery or justification of model parameters, e.g., the half-life of 19.26. Neither is necessarily the case.

Interpretations of probability differ (Hájek, 2012). Indeed, for frequentists, the concept of population may be implicit in the interpretation of probability, but frequency interpretations are not the only interpretations. Frequentists interpret probabilities as frequencies in populations. Subjectivists, however, interpret probabilities as degrees of belief, and propensity theorists interpret probabilities as natural tendencies. The concept of population is not implicit in all interpretations of probability.

The logic of discovery or justification of a parameter does not fix the reference of the parameter, so even if we discover or justify a parameter by invoking the concept of population, we need not invoke the concept of population to interpret the parameter. We might discover a natural tendency by observing a frequency in a random sample that we take as representative of a population. Nevertheless, the natural tendency is not the frequency in the population; rather, the natural tendency *causes* the frequency in the population. Similarly, we might justify our degree of belief by observing a frequency in a random sample that we take as representative of a population. Nevertheless, the degree of belief is not the frequency in the population; on the contrary, the frequency in the population causes our degree of belief. The decoupling of discovery/justification and reference becomes especially clear when we use alternative methods to discover natural tendencies and justify degrees of belief. For example, we might conclude that a flipped coin has a 0.50 probability of turning up heads by conducting a sufficiently large number of experiments, but we might also reach the same conclusion by carefully inspecting the coin for symmetry. Likewise, there might be telltale signs in a radioactive isotope that indicate its half-life. If so, we could use those signs to determine the parameters of our model for the radioactive decay of the single isotope.

concede that, if the modeled process generated an infinitely large amount of data, and if we called that dataset "the population," then certain properties of "the population" would be equal to certain parameters of the model. For example, the difference in mean GPAs between the subpopulations of students ineligible and eligible for free lunch, respectively, would be equal to β_1. In my exposition of the population-minded perspective, I rely on this relationship between the process-minded perspective and the population-minded perspective to develop my preliminary definitions. Nothing in this essay should be construed as an argument against process-mindedness. The only perspective against which I do argue here is census-mindedness, which gives no consideration to important counterfactuals necessary for explanation. Insofar as their work is explanatory, applied statisticians cannot do their work from a census-minded perspective. It may be the case that applied statisticians could do all their explanatory work from a process-minded perspective, but, as a matter of fact, many applied statisticians do their explanatory work from a population-minded perspective.[5]

I conjecture that applied statisticians are population-minded for two related reasons, both of which lean toward concrete thinking. First, population-mindedness allows applied statisticians to avoid the highly abstract concept of *random variable*. Kachapova and Kachapov (2012) discussed random variables and misconceptions thereof. The concept of *random variable* is tricky. From the population-minded perspective, applied statisticians can replace the highly abstract concept of *random*

[5] A process-minded statistician researching the students of AMS might view the students' GPAs as the product of an incredibly complex process that encompasses everything from nature and nurture, to aptitude and attitude, to politics and pedagogy. At the same time, the process-minded statistician might specify a simple model for the complex GPA-generating process such as our model with its one variable (*FREELUNCH*), two parameters, and an error term. The model, of course, is inevitably a gross oversimplification of the complex process. Nonetheless, the model may prove to be a useful oversimplification.

variable with the relatively concrete concept of *random sampling*. Second, data themselves are more concrete than the concept of a *data-generating process*, and population-mindedness allows applied statisticians to think in terms of data, albeit sometimes hypothetical data. Understanding the mean difference in a given sample does not require a high level of abstract thinking. If the sample falls short of a complete census, then it hardly requires a leap in abstraction to understand that the complete census also has a mean difference, and that the census mean's difference is probably different from the sample mean difference. It does require a great leap of abstraction, however, to go from understanding the actual sample, which is finite, to the hypothetical population, which is infinite. Nevertheless, the great leap from sample to population is scaffolded by the small leap from sample to census. In this essay, I discuss the concepts of *sampling error*, *explanation* and *population* in terms of counterfactual reasoning, and, in so doing, I provide an empirically grounded justification for the leap in abstraction from sample to population.

This is a philosophical essay for non-philosophers. As such, it contains many subtle but nonetheless crucial distinctions. I expect non-philosophers to have trouble keeping track of all the distinctions. Therefore, I conclude this introduction with a summary (Figure 2) of crucial distinctions that can serve the reader as a sort of dramatis personae. I invite the reader to refer back to this figure early and often.

Introduction	
A census includes all *actual* subjects who are the target of inquiry (regardless of the convenience or inconvenience with which they may be observed).	A population includes all *possible* subjects who are the target of inquiry including possible-but-not-actual subjects (i.e., counterfactual subjects).
Census-mindedness	Population-mindedness, with its like-minded ally in abstraction, process-mindedness
What Is Sampling Error?	
Sampling error is the difference between a sample statistic and a *census statistic*.	Sampling error is the difference between a sample statistic and a *population superstatistic*.
Method counterfactuals	*Target* counterfactuals
Sampling distributions involve counterfactual samples of *actual subjects.*	Sampling distributions involve counterfactual samples of *counterfactual subjects.*
What Are Explanations?	
Descriptive studies	Explanatory studies
Descriptions tell us what *is*.	Explanations tell us what *would be if …* .
Descriptions *do not* involve target counterfactuals.	Explanations *do* involve target counterfactuals.
Estimate census means	Estimate population means
What *is* the census-mean difference between two subgroups?	Does random variation explain the census-mean difference between two subgroups? In other words, *would there be* no census-mean difference *if* there were no sampling error?
What Are Populations?	
Censuses are finite (in the social sciences, at least).	We conceive of populations as infinite.
A census includes all and only *real* subjects.	A population includes all and only *realistic* subjects, including both real subjects and realistically fictional subjects.
Use so-called *"finite-population corrections"* to estimate imprecision due to sampling error.	Use *traditional methods* to estimate imprecision due to sampling error.
Use the actual random sample to allude to counterfactual random samples as "*actual* subjects that are the target of inquiry."	Use the actual random sample to allude to counterfactual random samples as "*realistically possible* subjects that are the target of inquiry."
Use the actual random sample as a random sample *from the census*.	Use the actual random sample as a random sample *from the population*.

Figure 2. A table of distinctions in this essay.

What Is Sampling Error?

In this section, I answer the question, "What is sampling error?" I answer it question from both a process-minded perspective and a population-minded perspective, and I discuss how the two perspectives relate to one another. In short, from a process-minded perspective, sampling error is the difference between a statistic (which belongs to a sample) and a parameter (which belongs to a model). From a population-minded perspective, sampling error is the difference between a statistic (which belongs to a sample) and a superstatistic (which belongs to a population). After defining my terms and clarifying my answers, I discuss the role of counterfactuals in reasoning about sampling error from either perspective.

A Process-Minded Perspective on Sampling Error

From a process-minded perspective, sampling error is the difference between a statistic (which belongs to a sample) and a parameter (which belongs to a model). Generally, we specify models with *unknown* parameters, and we estimate those *unknown* parameters by fitting the model to appropriate sample data. For the sake of exposition, however, consider a model with *known* parameters:

$$GPA_i = \beta_0 + \beta_1 FREELUNCH_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$$

Assume that the parameters are known: $\beta_0 = 3.02,\ \beta_1 = -0.56,\ \sigma^2 = 0.69.$ Therefore:

$$GPA_i = 3.02 - 0.56 FREELUNCH_i + \varepsilon_i,\ \varepsilon_i \sim N(0.00, 0.69)$$

The model represents the GPA of the *i*th student as though it were generated by a process in which Mother Nature considers the free-lunch eligibility of the *i*th student. If

the *i*th student is ineligible for free lunch, then Mother Nature determines the student's GPA by adding the value of a random variable, ε_i, to 3.02:

$$(GPA_i \mid FREELUNCH_i = 0) = 3.02 - 0.56(0) + \varepsilon_i = 3.02 + \varepsilon_i$$

The random variable, ε_i, is distributed normally with a mean of 0.00 and a variance of 0.69. In modeling the process of a carbon-10 isotope's radioactive decay before *t* + 19.26 seconds, we can imagine Mother Nature flipping a fair coin for the carbon-10 isotope at time *t*, because the random variable is distributed binomially with a trial parameter of 1 and a probability parameter of 0.50. In our model for the *i*th student's GPA, the random variable is distributed not binomially but normally; however, to personify the probabilistic component of the process, we can take advantage of the relationship between binomial and normal distributions. For the *i*th student, Mother Nature flips a fair coin a million or a billion or (better yet) a trillion times and counts the number of flips that turn up heads. The number of heads closely approximates a normally distributed random variable with a mean equal to half the number of flips and a variance equal to a quarter the number of flips. Mother Nature can then rescale (linearly, thus preserving the normality) the random variable to have a probability distribution with a mean of 0.00 and a variance of 0.69, as per the parameters of the model. Now Mother Nature can flip a coin (a very large number of times) to determine the value of ε for the *i*th student. If the *i*th student is eligible for free lunch, then Mother Nature determines that student's GPA by adding the value of the random variable, ε_i, to 3.02 minus 0.56:[6]

[6] Note that the interpretation is causally agnostic, because Mother Nature may consider the *i*th student's free-lunch eligibility as a direct cause of lower GPA or merely as a signal of lower GPA.

$$(GPA_i \mid FREELUNCH_i = 1) = 3.02 - 0.56(1) + \varepsilon_i = 3.02 - 0.56 + \varepsilon_i$$

In all probability, ε_i does not equal zero, so, in all probability, the *i*th student's GPA differs from the GPA "suggested" by the deterministic component of the process.

Now, suppose that the parameters of the model ($\beta_0 = 3.02,\ \beta_1 = -0.56$ and $\sigma^2 = 0.69$) are unknown to us. Suppose further that we cannot observe the modeled process in sufficient detail to directly determine the model parameters. We can nonetheless observe products of the modeled process. We can observe the GPA of the *i*th student, treat the *i*th student's GPA as a sample product, and use the sample product to provide information about the modeled process. If the *i*th student is ineligible for free lunch, then we can use the student's GPA as an estimate of β_0. If the *i*th student is eligible for free lunch, then we can use the student's GPA as an estimate of $\beta_0 + \beta_1$.[7] Unfortunately, because the modeled process has a probabilistic component, ε_i, a single datum is an extremely imprecise estimator. If the *i*th student is ineligible for free lunch, then the student's GPA is $\beta_0 + \varepsilon_i$. If the *i*th student is eligible for free lunch, then the student's GPA is $\beta_0 + \beta_1 + \varepsilon_i$. The difference, ε_i, is sampling error. In general, sampling error is the difference between an estimate of a parameter and the parameter itself. The imprecision due to sampling error of the *i*th student's GPA as an estimator is directly related to the variance of the random variable's probability distribution.

[7] By treating $\beta_0 + \beta_1$ as a single parameter to be estimated, I can skirt the irrelevant issue that, because a single observation lacks the necessary degrees of freedom, a single observation is insufficient to estimate more than one parameter.

Statistics from larger samples can serve as more precise estimators of model parameters. The mean GPA for students ineligible for free lunch can provide an estimate of β_0, or $\hat{\beta}_0$. The difference in mean GPAs between students ineligible for free lunch and students eligible for free lunch can provide an estimate of β_1, or $\hat{\beta}_1$. Thus, statistics from a sample can serve as estimators for parameters of a model. In this case, sampling error is the difference between a statistic as an estimator of a parameter and the parameter itself. We can estimate the imprecision of our estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, by reasoning counterfactually about the modeled process. We can reason counterfactually about what would happen if the modeled process generated not only the factual sample but also an infinite number of equal-sized counterfactual samples. A factual sample is an actual sample; a counterfactual sample is a possible-but-not-actual sample. We can reason counterfactually about the mean GPAs from each counterfactual sample, and, consequently, we can reason counterfactually about $\hat{\beta}_0$ and $\hat{\beta}_1$ from each counterfactual sample. Statistical theory guides our reasoning about the value of $\hat{\beta}_0$ and $\hat{\beta}_1$ in each counterfactual sample. We can reason about the distribution of the many estimates of β_0 or about the many estimates of β_1. In other words, we can reason about the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. In particular, statistical theory provides estimates for the shape and variability of each sampling distribution, and the variance of an estimator's sampling distribution can serve as a measure of the estimator's imprecision due to sampling error.

From a process-minded perspective, it is easy to see why we must continue to worry about imprecision due to sampling error even when we have census data (e.g., when we have gathered data from every student in the school). We are not modeling the census data. Rather, we are modeling the probabilistic process that generated the census data. Similarly, it is not the ultimate goal of random sampling to gather a sample that is representative of the census. Rather, the ultimate goal of random sampling is to gather a sample that is a representative product of the process under investigation. Presumably, the census is a representative product of the process under investigation, so by gathering a sample representative of the census, we thereby also gather a sample that is a representative product of the process under investigation. Even if we do manage to gather data for the entire census, the census is still just a representative product of the process under investigation. Because the process under investigation is probabilistic, any representative product of the process, including the census, will provide imprecise estimates of the process.[8] This is true whether the gathered census data are composed of millions of observations, hundreds of observations, a few observations, or even just a single observation. Of course, the fewer the observations we have, the less the precision is.

A Population-Minded Perspective on Sampling Error

From both a process-minded perspective and a population-minded perspective, we worry about the imprecision of our estimators resulting from sampling error. From either

[8] The methods that I am discussing apply not only to random samples but also to cherrypicked samples. The problem with cherrypicked samples, from a process-minded perspective, is that they are a product of not only the process under investigation but also the cherrypicking process. The cherrypicking process involves the researcher's conscious and unconscious thought process (including biases) in selecting some potential subjects over others for observation. Therefore, in modeling the process that generated the cherrypicked sample data, we are modeling a combination of the two processes.

perspective, we consider sampling distributions in order to quantify the imprecision. The two perspectives differ, however, in the objects of their statistical estimation and the composition of their sampling distributions. Whereas model parameters are the objects of statistical estimation for process-minded statisticians, for population-minded statisticians the objects of statistical estimation are population superstatistics. I will define the term *population superstatistics*, but first I discuss the composition of sampling distributions from both perspectives. This preliminary discussion provides a working definition of *population*, which I formalize in the fourth section of this essay.

For a process-minded statistician, a sampling distribution is composed of statistics, each from an equal-sized sample that was counterfactually generated on demand by the process under investigation. When a process-minded statistician needs to consider a sampling distribution, it is as though she orders up an infinitely large number of samples, and the process starts generating the samples (with infinitely great celerity). On the other hand, for a population-minded statistician, a sampling distribution is composed of statistics, each taken from an equal-sized sample that was randomly drawn from an infinitely large population. When a population-minded statistician needs to consider a sampling distribution, it is as though she draws a random sample from some infinitely large population, and then she draws another random sample, and then another, and then another and so on (with infinitely great celerity). What exactly is this infinitely large population? In the fourth section of this essay, I answer this question from a population-minded perspective. In the meantime, I can answer it from a process-minded perspective.

From a process-minded perspective, the infinitely large *population* is just an infinitely large *sample* that would be generated by the process under investigation if the process were continued long enough to generate an infinitely large sample. The infinitely large sample is not an actual sample, but a counterfactual sample. We calculate statistics from samples, so we can imagine calculating statistics from the counterfactual sample that is the population. To distinguish such imagined statistics from actual statistics, I call them *superstatistics*. Population means and variances are superstatistics. Likewise, subpopulation means and variances are also superstatistics, as are differences in subpopulation means. If we were to use a population superstatistic as an estimator for a model parameter, that population superstatistic would possess infinitely great precision due to the infinitely great sample size of the population. (Statistical theory tells us that, all other things being equal, the greater the sample size, the greater the precision of our statistical estimators.) Thus, a population-minded statistician can define the model parameter β_1 as the difference in subpopulation means between students ineligible for free lunch and students eligible for free lunch. For population-minded statisticians, model parameters are equal to population superstatistics. Based on this equivalence, population-minded statisticians use statistical models to understand populations.

For population-minded statisticians, sampling error is the difference between a sample statistic and the analogous population superstatistic. To estimate the imprecision (due to sampling error) of statistical estimators, population-minded statisticians use the same sampling distributions as process-minded statisticians. Population-minded statisticians, however, conceive of sampling distributions as coming about via infinitely repeated random sampling from a population, whereas process-minded statisticians

conceive of sampling distributions as occurring by means of the same process generating an infinite number of samples. What are the practical implications of this slight difference in conceptualization?

According to my preliminary definition of *population* as an infinitely large sample generated by the process under inquiry, there are no practical implications. It is perfectly superfluous to conceive of the population as an intermediary between the data-generating process and the sampling distribution. If, however, we can define *population* independently of the data-generating process, then we can conceive of the population not as an intermediary but as a source. In the fourth section of this essay, I will define *population* independently of the data-generating process.

There are practical implications to conceiving of the population, as opposed to the process, as the source of sampling distributions. For concrete thinkers, random processes are generally difficult to conceptualize, but random processes with discrete, equally probable outcomes are exceptionally easy to understand, especially when we can experiment with the processes (Kay, 2006). Flipping a fair coin is one such conveniently available random process with a discrete set of equally probable outcomes, which is why, as much as possible, I have personified random processes in terms of Mother Nature flipping fair coins. A lottery is another such random process. One practical implication of conceiving of the population as the source of sampling distributions is that concrete thinkers can conceptualize the random process as a lottery. The members of a population are discrete, and each member has an equal chance of being sampled, just as if the sampling were conducted by lottery. Furthermore, concrete thinkers have experience with sampling from a census by way of lottery, so they can analogize from that experience to

sampling from a population by the same means. Population-minded statisticians need their abstract-thinking skills only to conceptualize random processes as lotteries, whereas process-minded statisticians need them to conceptualize random processes as random variables that rarely have discrete and uniform probability distributions.

Target Counterfactuals and Method Counterfactuals

In considering sampling distributions, both population-minded statisticians and process-minded statisticians reason counterfactually. That is, they reason about possible-but-not-actual observations. When a data team at AMS specifies and estimates a statistical model in order to understand SES achievement gaps at the school, the data team reasons about possible-but-not-actual AMS students. If the data team is population-minded, then it imagines a population composed of possible-but-not-actual AMS students in addition to the actual AMS students. If the data team is process-minded, then it imagines that the process under investigation generated not only the actual sample of actual AMS students but also countless possible-but-not-actual samples of possible-but-not-actual AMS students. I call these types of counterfactuals *target counterfactuals*, as distinguished from *method counterfactuals*. As we observed earlier, target counterfactuals are counterfactuals about the target of inquiry. For population-minded statisticians, the target of inquiry is the population, and the population includes counterfactual subjects. For process-minded statisticians, the target of inquiry is the process, and the process counterfactually produces more samples. In the following section, I argue that target counterfactuals are fundamentally involved in explanation.

Whereas a target counterfactual is about the target of inquiry, a method counterfactual is about the method of inquiry. As examples of the latter category, a data

team's method of inquiry may involve observing a sample of students from the census of students. In such cases, there are observed actual students and unobserved actual students. The data team can counterfactually reason about what the team would have discovered if it had observed a different sample of students from the census of students. It can also counterfactually reason about what the team would observe if the team had sampled the entire census completely. In neither of these cases does the counterfactual reasoning involve counterfactual students. Rather, in both cases, it involves the counterfactual observation of actual students who were not actually observed. These can be called method counterfactuals because they are counterfactuals about the method of inquiry, which is observational.

In order to clarify further the distinction between target counterfactuals and method counterfactuals, I can take advantage of our strong intuitions about the random outcomes of the repeated flipping of a fair coin. Imagine that, after a certain number of flips, the coin is destroyed so that no more flips can occur. The outcomes from the actual flips constitute the census data.[9] Note that, if there is an odd number of actual flips, it is impossible that exactly half the actual flips turned up heads. If there is an even number of actual flips, it is improbable that precisely half the actual flips turned up heads (and the

[9] Here and throughout the essay, by *actual* I mean "actual at one time or another." The census includes all actual flips of the coin—past, present, and future. In general, censuses include all actual members of the target population—past, present, future, north, south, east, west, up and down. If the universe is infinitely large, then perhaps a census can be infinitely large. For example, a physicist might make a census-level generalization about all carbon-10 isotopes that ever did exist, do exist, and will exist (but not regarding all carbon-10 atoms that *could* exist), such as that all carbon-10 isotopes happen to decay in less than a million years. My arguments do not hinge on censuses being finite, but I focus on finite censuses because social scientists study finite censuses (e.g., a definable set of people in a particular, limited space and time). The essential features of a census are that all its members are actual and that all actual members of the target population (past, present, and future) are included. In practice, it is impossible to sample randomly from the future, so random-sampling methodology does not warrant inferences into the future, at least without the bridging assumption that the future will resemble the past in relevant respects. Therefore, social scientists generally limit the target populations of their studies to the past and/or present, and consequently, actual people from the past and/or present (but not the future) generally constitute censuses for social scientists.

greater the number of flips, the more improbable this result becomes). Because the coin is fair, *ex hypothesi*, we know that the outcomes can be modeled as a random variable distributed binomially with a probability parameter of 0.5. Now, imagine that we estimate the probability parameter ($p = 0.5$) from an observed sample of size n taken from the census of actual coin flips. The estimate will be imprecise due to sampling error. In order to estimate the imprecision, we can conceive of a sampling distribution. The sampling distribution involves counterfactual reasoning, but does it involve method-counterfactual reasoning or target-counterfactual reasoning?

Suppose that we conceive of a sampling distribution but limit ourselves to method-counterfactual reasoning, so we imagine infinitely repeated random samples of size n from the census. Because the census is composed only of actual outcomes, we do not consider counterfactual outcomes when we imagine repeated random samples from the census. We do not consider counterfactual outcomes, but we do consider counterfactual samples. Our actual method involved taking an actual sample, but we can reason method-counterfactually about taking other samples from the census. What does this method-counterfactual reasoning buy us? If we were to take the proportion of heads from each of the counterfactual samples, the proportions would be distributed with a mean equal to the proportion of heads in the census and a variance that could be used as a measure of precision for the proportion of heads in a sample of size n as an estimator of the proportion of heads in the census. But are we interested in estimating the proportion of heads in the census, or are we interested in estimating the probability parameter, p? These are surely two different things if the census consists of an odd number of coin flips, and they are most probably two different things if the census consists of an even

number of coin flips. If we are interested in estimating *p*, then we must not limit ourselves to method-counterfactual reasoning.

If we reason about possible-but-not-actual outcomes, we reason target-counterfactually. A process-minded statistician would reason about an infinitely large number of new samples of size *n* generated by the coin. A population-minded statistician would reason about an infinitely large number of random samples of size *n* from taken an infinitely large population of coin flips, which would include not only all the actual coin flips from the census but also an infinitely large number of counterfactual coin flips. Whether it is process-minded or population-minded, the target-counterfactual reasoning buys us a sampling distribution with a mean equal to the probability parameter, *p*, and with a variance that can be used as a measure of precision for the proportion of heads in a sample of size *n* as an estimator of *p*. In short, we need target counterfactuals to reason about the fairness of a purportedly fair coin. Method counterfactuals are not enough.

Target-counterfactual reasoning and method-counterfactual reasoning support different inferences. We can see this from the differing sampling distributions that arise from the different reasoning. To make this crucial point more concrete, first I consider a fair coin that has been flipped only a few times (and will never be flipped again), and then I consider a small subset of a group of students. The smallness of a small census makes it triply useful for concretizing my point. First, a small census permits the easy enumeration of all the possible samples of actual outcomes, and that enumeration provides intimate acquaintance with the sampling distribution from the perspective of method-counterfactual reasoning. Second, for a given sample size, the smaller the census, the greater proportion of the census included in the sample, and the greater the proportion

of the census included in the sample, the smaller the variance of the sampling distribution using method-counterfactual reasoning, but, on the other hand, the proportion of the census sampled has no effect on the variance of the sampling distribution using target-counterfactual reasoning, so the contrast is extreme. Third, a small census, which yields only small samples, capitalizes on our strong intuitions about the inadequacy of small samples for some statistical inferences but not for other statistical inferences. I focus on small samples to concretize my point that the different forms of counterfactual reasoning yield different sampling distributions and, ultimately, different inferences. My point, however, is not restricted to small censuses; it applies to any census.

Consider a fair coin that has been flipped a total of five times and will never be flipped again. Suppose that the coin happened to land on heads three times, for a proportion of 0.60. Suppose further that we observe only a random sample of three coin flips ($n = 3$) from the census of five coin flips ($N = 5$).[10] We can use the sample proportion of heads as an estimator of either the census proportion (0.60), the probability parameter (0.50), or the population proportion (0.50). The proportion of heads in the sample ($n = 3$) is an imprecise estimator of either 0.60 or 0.50. In fact, the proportion of heads is a necessarily inaccurate estimate of 0.60 or 0.50, because the proportion of heads in the sample ($n = 3$) must be 1.00, 0.66 or 0.33. (Note that the proportion of heads in the sample cannot be 0.00 because there are only two instance of tails in the census, so at least one instance of heads must be sampled.) In order to quantify the imprecision of the estimator, we can reason counterfactually about the sampling distribution. The type of counterfactual reasoning (method counterfactual vs. target counterfactual) differs

[10] As per convention, I use lowercase *n* to denote sample size and uppercase *N* to denote census size.

depending on whether we are estimating the census proportion (0.60) or estimating either the probability parameter (0.50) or the population proportion (0.50).

If we are estimating the census proportion (0.60), we reason about all possible samples ($n = 3$) of actual coin flips ($N = 5$). If we are estimating the probability parameter (0.50) or the population proportion (0.50), then we reason about possible-but-not-actual coin flips in addition to actual flips. The sampling distribution (Figure 3) of possible samples ($n = 3$) from the actual set of coin flips ($N = 5$) is different from the sampling distribution (Figure 4) of possible samples ($n = 3$) of possible coin flips. For starters, a sample ($n = 3$) with no heads (i.e., three tails) is not a possible sample from the actual coin flips, but it is a possible sample of possible coin flips. If we name each of the five actual coin flips (H1, H2, H3, T1 and T2, respectively), then we can list the ten possible and equally likely types of samples ($n = 3$) of those five actual coin flips: {H1, H2, H3}, {H1, H2, T1}, {H1, H2, T2}, {H1, H3, T1}, {H1, H3, T2}, {H1, T1, T2}, {H2, H3, T1}, {H2, H3, T2}, {H2, T1, T2} and {H3, T1, T2}. We can see that there is no TTT pattern among them. However, if we derive our sample from all possible coin flips, a headless sample (TTT) is one of eight equally probable types of samples: HHH, HHT, HTH, HTT, THH, THT, TTH and TTT.

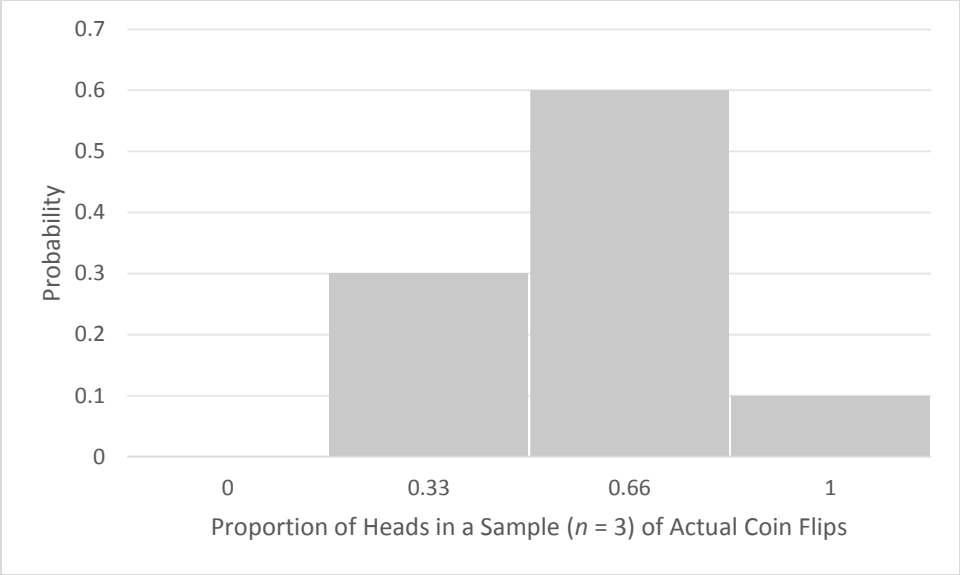


Figure 3. A sampling distribution of proportions from all possible samples (n = 3) of actual coin flips (N = 5), with a mean of 0.60, variance of 0.04, and standard deviation of 0.20.

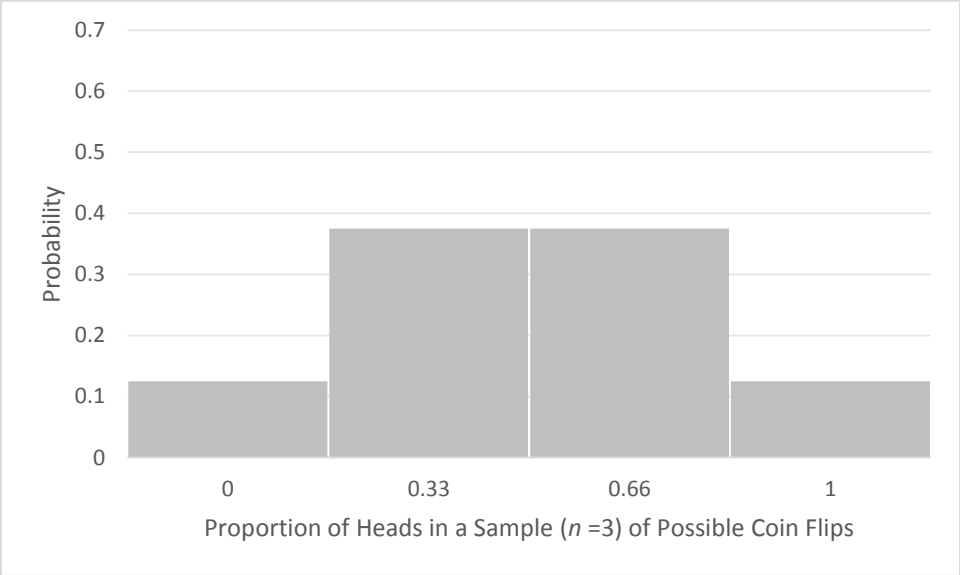


Figure 4. A sampling distribution of proportions from all possible samples (n = 3) of possible coin flips, with a mean of 0.50, variance of 0.08, and standard deviation of 0.29.

In general, the two sampling distributions have different means and variances. The mean of a sampling distribution (of an unbiased statistic) is equal to the value for which the statistic is an estimator. The mean of the method-counterfactual sampling

distribution is 0.60, which is the census proportion. The mean of the target-counterfactual sampling distribution is 0.50, which is the probability parameter or the population proportion. Thus, the proportion of heads in the sample is an estimator for two different values (0.60 or 0.50) depending on whether we are reasoning method-counterfactually or target-counterfactually. The sample proportion is a more precise estimator for the census proportion (0.60) than for either the probability parameter (0.50) or the population proportion (0.50), as evidenced by the smaller variance of the sampling distribution (0.04 compared with 0.08). The question remains, however: when do we want to estimate the probability parameter and/or population superstatistic rather than the census statistic?

Let us turn our attention from flips of a coin to students in a school. Consider a school interested in creating an inclusive educational environment for students with Asperger's syndrome (AS). The data team is tasked with providing insight into current gaps in academic achievement associated with AS. However, only one AS student, named Christopher, is currently attending the school. If census statistics are sufficient for the purposes of inquiry, then the data team's job is easy. The sampling distribution for the census-mean GPA has no variance, so Christopher's GPA is a perfectly precise estimator of the census mean; in fact, Christopher's GPA is the census-mean GPA. If Christopher's GPA is below average, it is true that, on average, students with AS are performing academically worse than students without AS. But is this information valuable to the data team? Hardly. In such cases, data teams want more than census statistics. Rather, they want to make true generalizations about not only actual AS students currently in the school but also possible-but-not-actual AS students currently in the school. What if, instead of Christopher, a fictional child named Albert were the only AS student in the

school? Or Vernon, Phillipa, or William? Would these possible-but-not-actual AS students tend to perform below average? I think that data teams want information about target-counterfactual, possible-but-not-actual students, and model parameters and population means represent that information.

The distinction between census statistics on one hand and model parameters and population superstatistics on the other hand is most evident in small censuses, such as my examples of the five coin flips and the student with Asperger's syndrome, but it is not restricted to small sample sizes. Consider a coin that was flipped a million and one times before it was destroyed. No matter how fair the coin, it cannot possibly have yielded an equal number of heads and tails, because it was flipped an odd number of times. (Even if we discounted the final flip, the likelihood of having precisely 500,000 heads and 500,000 tails would be approximately 0.000798.) Now consider a nationwide census of students diagnosed with Asperger's syndrome. There is some probability that a particular boy will be diagnosed with AS, and there is some probability that a particular girl will be diagnosed with AS, and it appears that those two probabilities are very different. Boys are perhaps three times more likely than girls to be diagnosed with AS. Whatever the probabilities and their ratio, it is extremely unlikely that they will manifest themselves exactly in the census proportions, especially if the census is large. The paradox of larger sample sizes is that our estimates of model parameters and population pseudostatistics are more likely to be close but are generally no more likely, and often less likely, to be perfect.

When do we care about model parameters and population superstatistics rather than census statistics? Sometimes our goal is description, but at other times our goal is

explanation. When our goal is to describe the census, then we care about census statistics. If our goal is to explain the census, however, then we care about model parameters and population superstatistics, because they support explanatory reasoning. I devote the next section to the concept of *explanation*. I discuss the relationship between counterfactuals and explanations through their mutual relationship with lawlike generalizations. I argue that lawlike generalizations can support both counterfactuals and explanations. I further argue that both model parameters and population superstatistics can support the lawlike generalizations that in turn support counterfactuals and explanations.

What Are Explanations?

Explanation is a philosophically thorny concept, but Hempel's (1962) nomological-deductive (ND) and inductive-statistical (IS) accounts of explanation are a standard starting point for any introductory discussion of *explanation*. I cannot offer a fully mature account of *explanation* (I doubt that anyone can), but I think that, for my purposes, Hempel's account will suffice as a first approximation. Any refinement of Hempel's account or any alternative account must, however, accommodate Hempel's key insight: explanations involve lawlike generalizations. After I introduce Hempel's account of explanation, I define lawlike generalizations, contrasting them with accidental generalizations. Finally, I discuss sampling error as an explanation, because, as per the thesis of this essay, we conceive of populations as infinitely large in order to rule out sampling error as the sole explanation for our statistical results.

Hempel's Account of Explanation

Both of Hempel's accounts (the ND account and the IS account) give a central role to lawlike generalizations (Hempel & Oppenheim, 1948). The ND account invokes

deterministic lawlike generalizations, and the IS account invokes probabilistic lawlike generalizations. According to Hempel, we explain a statement by noting relevant background statements and demonstrating how the statement to be explained follows (either logically or probabilistically) from the relevant background statements in conjunction with one or more statements of lawlike generalization. Since data analysts deal with probabilistic lawlike generalizations, I will illustrate Hempel's IS account. As my illustration of Hempel's IS account, I present a result from an actual randomized controlled trial (RCT), which found that, on average, students in the sample who were randomly assigned to a program ($M = 3.22$, $SD = 0.34$) earned higher GPAs than students randomly assigned to the control condition ($M = 3.01$, $SD = 0.33$). What explains this result? According to Hempel's IS account, the result is explained by relevant background statements and probabilistic lawlike generalizations. One relevant background statement (among others) is that the expectations for students assigned to the program and to the control condition were equal at the outset, due to the randomness of the assignment. One probabilistic lawlike generalization (among others) is that the program tends to improve GPAs for students like the ones in our sample. In short, the program's effectiveness explains the results.

We can rationally reconstruct the background statements and lawlike generalizations as premises of an argument in which the result to be explained follows (probabilistically) from the premises:

Premise 1: The students assigned to the program and the control condition were equal in expectation at the outset due to the randomness of the assignment (among other conditions).

Premise 2: The program tends to improve GPAs for students like the ones in our sample (among other lawlike generalizations).

These premises make the following conclusion highly likely:

Conclusion: On average, in the sample, students randomly assigned to the program ($M = 3.22$, $SD = 0.34$) earned higher GPAs than students randomly assigned to the control condition ($M = 3.01$, $SD = 0.33$).

In theory, the premises (including the omitted premises) constitute a complete explanation of the result. In practice, a relevant subset of the premises functions as an adequate explanation of the result. According to Hempel's IS account, a complete explanation always invokes lawlike generalizations. My project is to show how model parameters and population superstatistics contribute to lawlike generalizations (and consequently explanations) in a way that census statistics do not.

Lawlike vs. Accidental Generalizations

Whereas accidental generalizations tell us what happens to be the case, lawlike generalizations tell us what must be or what tends to be the case. Therefore, lawlike generalizations support target counterfactuals. I exemplify the support of counterfactuals with two parallel pairs of generalizations. Each pair includes a lawlike generalization and, for purposes of contrast, an accidental generalization. I adapt the first pair of generalizations from van Fraassen (1989, p. 27), who attributes them to Reichenbach and Hempel:

All solid chunks of enriched uranium (U235) are less than one mile thick.

All solid chunks of gold (Au) are less than one mile thick.

As noted by van Fraassen, both generalizations are plausible, and their form is identical. The difference is in the substance. Enriched uranium has a critical mass that would cause an explosive nuclear chain reaction before such a large chunk could ever form. Both generalizations may be true, but the first is necessarily true (given certain background statements and lawlike generalizations that pertain according to contemporary physics).

Suppose that both generalizations are true. Both generalizations support method counterfactuals, but only the generalization about U235 supports target counterfactuals. Both generalizations tell us that, if we were to comb the universe and observe every solid chunk of U235 or gold, we would find that all are less than one mile thick. Both elements are so rare that such an aggregation has never happened and will never happen. Both generalizations tell us what would happen if we counterfactually observed actual samples. Only one generalization, however, tells us what would happen if we counterfactually observed counterfactual samples. Consider a possible-but-not-actual planet that is superabundant with U235 and gold. The generalization about uranium holds for this counterfactual planet, but the generalization about gold may not. This is because the generalization about uranium tells us not only what happens to be the case but also what must be the case. It is a lawlike generalization.

The second parallel pair of generalizations is a bit more relevant to school-based data teams. Here the generalizations involve an RCT of a type that would be rare in educational settings but that provides clear-cut examples. In the next section, I will delve into a more common example. For now, suppose that the data team assigns all students in the school to a program condition or a control condition depending on the flip of a fair coin. All students whose coin flip turns up heads, and only those students, are assigned to

the program. Thus, each student has a value for a variable that we can label $HEADS_i$, and this value is 1 if the *i*th student's coin flip turns up heads or 0 if it does not. Likewise, each student has a value for a variable $PROGRAM_i$, which takes on the value of 1 if the *i*th student is assigned to the program and 0 otherwise. Note that the values of the two variables, $HEADS_i$ and $PROGRAM_i$, are identical. Now consider two more variables, $GPA\_BASELINE_i$ and GPA_FINAL_i. $GPA\_BASELINE_i$ is equal to the *i*th student's GPA when the coin was flipped. GPA_FINAL_i is equal to the *i*th student's GPA after the program has taken place. Consider the following two models fitted to the census data, each model providing evidence for a generalization (stated immediately after the model):

$$\hat{GPA\_FINAL}_i = 3.01 + 0.21 PROGRAM_i$$

On average, current students who were randomly assigned to the program achieved higher final GPAs than current students who were randomly assigned to the control group.

$$\hat{GPA\_BASELINE}_i = 3.02 - 0.02 HEADS_i$$

On average, current students whose fair coin flip turned up heads achieved lower baseline GPAs than current students whose fair coin flip turned up tails.

I contend that the first generalization is lawlike and that the second generalization is merely accidental. Both generalizations support method counterfactuals, but only the generalization about the program effect supports target counterfactuals. Target counterfactuals involve possible-but-not-actual students from Anonymous Middle School. This is tricky, because the possible-but-not-actual students are different from the actual students, but they cannot be so different that they are no longer the type of students who currently attend AMS. After all, the scope of the generalization is about current

AMS students. To grasp this point, first consider slightly different students, so slightly different that they are obviously still AMS-type students. Then consider entirely different students who are nonetheless AMS-type students. Consider the possibility that the outcomes of the coin flips were reversed. This would change the data for the AMS students for the variables $HEADS_i$, $PROGRAM_i$ and GPA_FINAL_i, but not for $GPA\_BASELINE_i$. Thus, these possible AMS students would be slightly different from the actual AMS students. Nevertheless, the generalization about the program effect would probably still hold. No matter what AMS students were randomly assigned to the program condition, the program would tend to improve their final GPAs. The generalization about the baseline differences, however, would not hold. In fact, it would be reversed. The actual students whose coin flips turned up heads happened to have a lower baseline GPA, on average, but they would still have had a lower baseline GPA, on average, if the coin flips had been reversed. Whereas the program appears to have a direct causal impact on final GPA, the coin flip has no causal relationship with baseline GPA.

Consider an entirely different census of AMS-type students. Because the program is effective, on average, for AMS-type students, the generalization about the average program effect will probably still hold. Because the result of the coin flip has no causal relationship with baseline GPA, the average baseline difference will tend to be zero. It may be negative in the actual census, but it has an equal chance of being positive in a possible census. If we average over all possible censuses, the average baseline difference will be zero. In other words, the average baseline difference in the population is zero. An accidental generalization is a generalization about the census, which in the present example includes only actual AMS students. A lawlike generalization is a generalization

about the population, which in the present example includes not only actual AMS students but also possible AMS students.

I distinguish lawlike generalizations from accidental generalizations based upon the support of target counterfactuals. Not every philosopher, however, is so prone to rely on counterfactuals for definitional purposes. Hempel, for example, is completely committed to lawlike generalizations as essential to explanation, but he is ambivalent about counterfactuals. As Woodward (2011) notes, Hempel is skeptical about counterfactuals because, according to Hempel (1965, p. 339), counterfactuals "present notorious philosophical difficulties"; nevertheless, Hempel agrees that lawlike generalizations, as opposed to accidental generalizations, support counterfactuals. I do not share Hempel's skepticism about counterfactuals, because I think that Goodman (1983), by construing counterfactual possibilities as fictions, has gone a long way toward clearing up the murky metaphysics surrounding counterfactuals.

Once we construe counterfactual possibilities as fictions, a counterfactual-based definition of lawlike generalizations fits neatly within Hempel's theory of explanation. Although Hempel is reluctant to embrace the close relationship between explanation and counterfactual reasoning, he fully embraces the close relationship between explanation and predictive reasoning. Hempel and Oppenheim (1948) argue that explanation and prediction have the same formal structure and, thus, that explanation and prediction differ only pragmatically. Both explanation and prediction are based on the logical or probabilistic relationship between (1) background statements and lawlike generalizations and (2) their consequences. Whereas explanation involves determining the background statements and lawlike generalizations that lead to the given consequence to be

explained, prediction involves determining the consequence that follows from the given background statements and lawlike generalizations. This argument by Hempel applies, however, not only to prediction in particular but also to inference in general, where predictive reasoning is one type of inferring and counterfactual reasoning is another type of inferring. Whether we are talking about a future world or a fictional world, if you give me the relevant background statements and relevant lawlike generalizations, I can infer what is likely to happen in that world. In the fourth section, I will discuss the scientific evaluation of counterfactuals as fictions that are supported by lawlike generalizations.

Sampling Error as an Explanation

One purpose of a traditionally trained data analyst is to always consider sampling error as a possible explanation (or perhaps a partial explanation) of statistical results calculated from a sample. Data analysts ask, "Are the results statistically insignificant?" In other words, are the results explainable by sampling error alone? Sampling error is a complete explanation for at least some statistical results. Take the aforementioned result of an average baseline difference in GPA between students whose fair coin flip turned up heads and students whose fair coin flip turned up tails. The coin flip is causally independent of every baseline variable including *GPA_BASELINE*. Any correlation between *HEADS* and a baseline variable such as *GPA_BASELINE* is a spurious correlation, which means that it is a correlation due solely to sampling error.[11] Sampling error is a label for the influence of stochastic processes on statistical results. By *stochastic process* I mean a process that is random (i.e., unpredictable) according to the

[11] Under the rubric of "spurious," some data analysts include correlations due to a third variable in addition to correlations due to chance. For the purposes of this essay, I include only correlations due solely to chance.

statistical model. The result of a coin flip may not be random to a Laplacean demon with the knowledge and computing power to predict the results of a coin flip from the initial conditions and physical laws. Nevertheless, the results of a coin flip are random according to a statistical model in which the result is modeled as a binomial random variable with a probability parameter of 0.50. That randomness, sampling error, explains the baseline difference.

If sampling error is an explanation, then it involves lawlike generalizations that support target counterfactuals. What are those lawlike generalizations and target counterfactuals? The lawlike generalizations are about the tendency of estimates across repeated random samples. From a process-minded perspective, the repeated random samples are generated by the same process. From a population-minded perspective, the repeated random samples are drawn from the same population. To explain a baseline difference as an artifact of sampling error is to assert that, on average, across repeated random samples (generated by the same process or drawn from the same population), the baseline difference is zero. These lawlike generalizations support target-counterfactual claims about what would happen if an infinite number of possible-but-not-actual samples of possible-but-not-actual subjects were generated by the same process or drawn from the same population.

Data teams are interested in explanation for the sake of *control* and *prediction* in addition to *understanding*. Data teams seek to identify programs and practices that improve educational outcomes, so that educational leaders can take *control* of the educational system, instead of leaving it to chance. Data teams also seek to *predict* failure so that educational leaders can prevent it through proactive programs and practices.

Descriptions alone do not help with control and prediction. Descriptions tell us *what is*, but explanations suggest *what can be* and *what will be*. Sampling error as an explanation, however, is deflationary with respect to control and prediction. For example, in an RCT designed to yield an unbiased estimate of the causal impact of a program, if the estimate is truly explained by sampling error—in other words, if the estimate is truly explained by random chance rather than the causal impact—then offering the program affords no control over educational outcomes. In a correlational study, if the sample correlation is truly explained by sampling error, then the correlation is of no help in predicting educational outcomes. For these reasons, sampling error as an explanation should be a constant consideration for data teams and their consultants.

Epidemiologists make a distinction between *descriptive* epidemiology and *analytic* epidemiology. Descriptive epidemiology describes the frequency and distribution of disease, and analytic epidemiology draws conclusions about the causes of disease. This distinction mirrors the one that I want to make between descriptive science and explanatory science. Descriptive science aims at describing the world, and explanatory science aims at explaining the world. What distinguishes a description from an explanation? One distinguishing factor is that explanations imply counterfactuals about their target systems. My distinction between description and explanation mirrors Deming's (1953) distinction between enumeration and analysis. My contribution is to draw the distinction in terms of counterfactual reasoning.

What Are Populations?

For the purposes of statistical inference in the context of an explanatory inquiry, the population is composed of not only all the actual subjects but also all the possible-

but-not-actual subjects that are the target of inquiry. In short, the population is composed of all the possible subjects that are the target of inquiry. If a data team wants to explain the current SES achievement gap at AMS, then the data team draws an inference to the population of current AMS students, which includes not only all the actual current AMS students but also the possible-but-not-actual current AMS students. Of course, the data team cannot hope to provide a full explanation of the SES achievement gap, but the data team can perhaps rule out sampling error as the explanation. In this section, I elaborate on what is referred to by "population," or "all the possible subjects that are the target of inquiry." Then I delve into qualifications of the term "possible," because data teams are interested in the realistically possible as opposed to the fantastically possible. Finally, I conclude this essay by summarizing the reasons for population-mindedness.

Reference and Populations

Samples exemplify. Exemplification is one type of reference. Denotation is another type of reference. For this essay, I use Goodman's (1976) theory of reference. First I will define *denotation*, and then I will define *exemplification* in terms of denotation.

Denotation is reference to a property, object, or relation by a label that applies to that property, object, or relation. For instance, the label "green" applies to the color of (some) grass, the color of (some) money, and the color of (some) frogs. When we use *green* to refer to the color of grass, money, or frogs, we use the word denotationally. Because the label "current AMS student" applies to the current students of AMS, we can use "current AMS student" to denote a current AMS student. Whereas denotation is reference to a property, object, or relation *by a label*, exemplification is reference *to a*

label by a property, object, or relation denoted by the label. We can use a sample of green grass, green money, or green frogs to exemplify the concept of green. Likewise, we can use a sample of current AMS students to exemplify "current AMS students."

When we use a sample to exemplify the *census* of current AMS students, we are using actual current students of AMS to exemplify "actual current students of AMS." When we use a sample to exemplify the *population* of current AMS students, we are using possible current students of AMS to exemplify "possible current students of AMS," because the population of students includes both actual (i.e., possible and actual) and counterfactual (i.e., possible-but-not-actual) students. The same property, object, or relation can exemplify multiple labels. A sample of green frogs can exemplify "frogs" as well as "green," because the labels "frogs" and "green" both apply to the sample. Likewise, a sample of current AMS students can exemplify "actual current AMS students" and "possible current AMS students," because the current AMS students are both actual and possible.

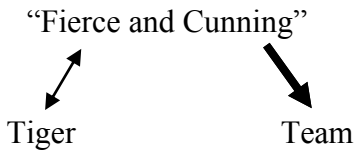
If the same sample can exemplify different properties, objects, or relations, how do we know exactly what a sample in a given context is intended to exemplify? Is the sample of green frogs exemplifying "green" or "frogs"? Is the sample of current AMS students exemplifying "actual current AMS students" or "possible current AMS students"? The answers depend on how we use the samples. We can use the sample of green frogs to exemplify "green" or to exemplify "frogs." We can use the sample of current AMS students to exemplify "actual current AMS students," or we can use the same sample to exemplify "possible current AMS students." Often the use of a sample is evident from the context. The sample of green frogs may be used in the context of a

lesson on color or of a lesson on animals. The sample of current AMS students may be used in a descriptive context or an explanatory context. The context, however, is not always sufficiently clear to make the sample's use evident. In fact, the difference between a descriptive context and an explanatory context is a bit esoteric.

When the use of a sample is not evident from the context, what can we do to determine the sample's reference? We can clarify the context. With the sample of green frogs, we can ask whether the lesson is about color or animals. With the sample of current AMS students, we can ask whether the inquiry is descriptive or explanatory, but, because the distinction between descriptive and explanatory inquiry is esoteric, I do not think the answer will be helpful for most applied statisticians. Happily, there is another method of determining a sample's reference: we can ask about acceptable alternative samples. To determine whether the sample of green frogs exemplifies "green" or "frogs," we can ask which is an acceptable alternative sample: a sample of green grass or a sample of red frogs. To determine whether the sample of current AMS students exemplifies "actual current AMS students" or "possible current AMS students," we can ask which is an acceptable alternative sample: a random sample from the census or a random sample from the population. In other words, we can ask to what acceptable alternative samples the sample alludes.

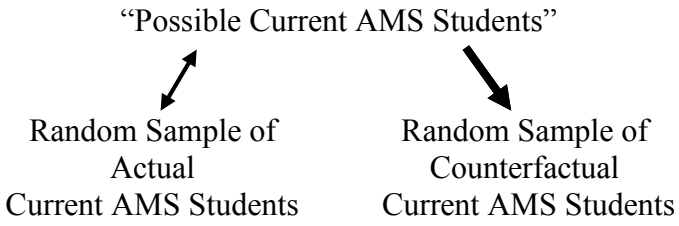
Allusion is a form of indirect reference. In short, allusion is the use of an exemplar to refer to other possible exemplars. Through allusion, a sample of actual students of AMS can indirectly refer to samples of counterfactual students of AMS. The actual students allude to the counterfactual students. What, exactly, is allusion? To answer this question, I draw on Elgin's (1983, p. 143) analysis of allusion. Elgin used

diagrams in which single-headed arrows represent denotation and double-headed arrows represent exemplification. To begin with Elgin's simplest diagram, consider a football team that adopts the tiger as its mascot because the tiger alludes to the team's fierceness and cunningness:

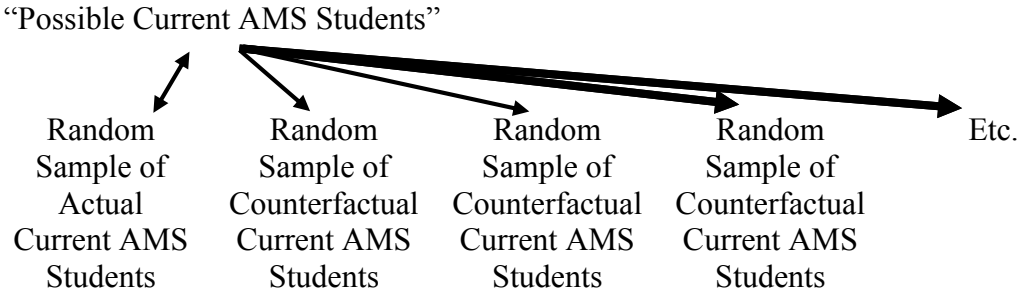


A single-headed arrow runs from the label "fierce and cunning" to the team, because the label applies to the team and, consequently, the label denotes the team. Because the label denotes the team, the team is a possible exemplar of "fierce and cunning." The label "fierce and cunning" also applies to tigers, but we give the arrow a second head because not only does the label "fierce and cunning" refer to the tiger via denotation, but also the tiger refers to the label "fierce and cunning" via exemplification. In our culture, the tiger is an exemplar of "fierce and cunning." Because the tiger exemplifies "fierce and cunning," the team can use this symbol to allude to its own fierceness and cunningness.

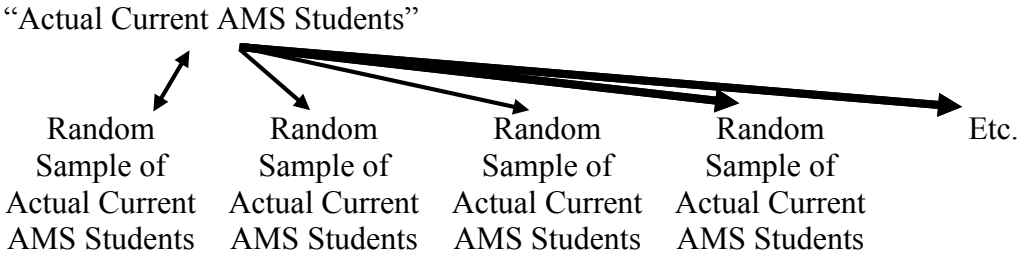
The label, "possible current AMS students," applies to the actual current students of AMS, who are both possible and actual. A random sample of actual current AMS students can exemplify "possible current AMS students." Just as the tiger can allude to the team as fierce and cunning, a random sample of actual current students can allude to a random sample of counterfactual current AMS students as possible current AMS students:



Indeed, a random sample of actual students can allude to an infinite number of random samples of counterfactual students:



I am not claiming that random samples *must* allude to target-counterfactual random samples; I am claiming only that they *can*. Random samples can just as easily allude to method-counterfactual random samples. After all, any sample can exemplify in many ways. A random sample of actual current AMS students can allude to other random samples of actual students:



What a sample exemplifies and, consequently, what it alludes to are a matter of usage. Researchers can use random samples to estimate census statistics, model parameters, and population superstatistics. In order to assess the imprecision of their

estimates, researchers consider the variance of sampling distributions that involve the conception of infinitely repeated random sampling. First, consider the case in which researchers want to estimate census statistics. If the infinitely repeated random sampling is from a finite census, then researchers use what is called a "finite-population correction" (though by my definitions it should be called a "finite-census correction"). A researcher's use of a finite-census correction declares that the researcher is using the random sample of actual census members to allude to other random samples of actual census members. Finite-population corrections are a special case in applied statistics. Now consider the typical case in which researchers want to estimate model parameters or population superstatistics. Typically, applied statisticians do not use finite-population corrections because the applied statisticians are conceiving of infinitely repeated random sampling from a data-generating process or from an infinite population. This makes sense, because, typically, applied statisticians are doing explanatory work that requires the consideration of counterfactual subjects produced by the data-generating process or belonging to the infinite population. A data team at AMS would typically use a random sample of actual current AMS students to allude to infinitely repeated random samples of counterfactual current AMS students. In so doing, the data team would be supporting an inference about model parameters or population superstatistics rather than about census statistics, all in the service of its explanatory work.

Evaluating Claims about Largely Fictive Populations

Not all samples are equally useful for scientific investigation. Scientific objectivity requires that a sample be chosen in a theory-neutral way. Random sampling is theory-neutral, and it allows the application of important results from theoretical

statistics. When we draw a sample by fair lottery from a census, it is obvious that the sample is a random sample from the census. It is not so obvious, however, that the sample is a random sample from the population. After all, we did not draw the sample by fair lottery from the population. Nonetheless, if the census were a random sample from the population, then a random sample from the census would also be a random sample from the population. But is the census a random sample from the population? The answer depends on how we define the population. In fact, for the purposes of statistical inference, we define the population exactly in such a way as to make the answer to this question "yes."[12]

First, consider a definition of the population that is ill-suited for the purposes of statistical inference. Thus far, I have argued that the population of current AMS students includes not only actual current AMS students but also counterfactual current AMS students. According to Goodman's analysis of counterfactuals as fictions, counterfactual current AMS students are fictional current AMS students. Fictions can be unrealistic. Of course, actual AMS students are ordinary kids, but presumably a data team could whimsically fantasize that the current AMS students also include some covert undead Martian magician werewolf mutant bionic ninja gladiators. The data team would not only have the seed for the next blockbuster series of young-adult novels, but it would also have students of a hypothetical AMS population, albeit one ill-suited for the purposes of

[12] The trick is to explicitly define our populations by carefully defining the censuses from which we randomly draw our samples. If a data team draws a random sample from the census of current AMS students, the data team must be careful not to mistake the census of current AMS students for the census of all AMS students (past, present, and future) and consequently draw a mistaken inference to the population of all AMS students, past, present, and future. If an educational researcher draws a random sample of principals from within a two-hour drive of the research university, then the researcher must be careful not to mistake the census of nearby principals for the census of U.S. principals.

statistical inference. Fictions can be unrealistic, yet fictions can also be realistic. This latter type of fiction is what data teams need to make useful statistical inferences.

For the purposes of statistical inference, let us consider only realistically fictional current students of AMS in addition to non-fictional current students of AMS. The label "realistic current AMS students" applies to realistically fictional current students of AMS and non-fictional current students of AMS.[13] If we were relying on storytelling rather than statistical modeling for our counterfactual insight, the story would have a non-fictional setting, AMS, but with realistic fictional characters, the AMS students. The characters would be as just as typical of AMS students as actual AMS students are, but they would also be as quirkily individual as the actual AMS students. I doubt that anyone has the imagination to write such a story without cheating by altering or creating composites of actual AMS students. The fictional students cannot be mere alterations or composites of actual AMS students, however, because actual AMS students are not mere alterations or composites of other actual AMS students. (Statistically, this would be a violation of independence assumptions.) Happily, we do not rely on storytelling, but on statistical modeling. Statistical modeling simplifies the world picture. The deterministic component of the model describes the AMS pattern, and the probabilistic component of the model describes the individual variation from the AMS pattern. In the end, we seek to

[13] Two small philosophical points arise from this statement. First, in his nominalist theory of mention-selection, Scheffler (1996) gives a philosophical account of how labels apply to fictions. For my purposes, I hope it will suffice to assert that the label "quixotic" applies to Don Quixote, a fictional character. In the same way that "quixotic" applies to Don Quixote, the label "realistic current AMS students" applies to realistically fictional current students of AMS. Second, it is admittedly awkward to label actual students "realistic." In general, it is awkward to talk about real things being realistic. The awkwardness comes from the pragmatics of language. Pragmatically, to say that something is realistic is to imply that the thing is not real, because it would be uninformative (or under-informative) to knowingly say of a real thing that it is realistic. Such an implicature, however, is cancelable, and I am canceling this implicature for the purposes of my exposition. Although it generally goes without saying that real things are realistic, in this case the logic of my argument requires that I state the obvious.

parse the AMS patterns from the individual variation. Inferring to the population, with its actual and counterfactual students, helps us do that parsing. Theoretical statistics guides our statistical inferences because it gives us information about how the estimates of our fitted model's deterministic components vary over counterfactual samples from the infinitely large population.

Sampling error as an explanation implies counterfactuals, and considering sampling error as the explanation for a result involves evaluating the truth of those counterfactuals. Evaluating the truth of counterfactual conditional statements is a challenging philosophical problem. In 1946, Goodman (1946/1983) introduced the problem of counterfactual conditionals, and philosophers have yet to agree on a solution to this problem. Goodman divided the problem into two parts. The first part of the problem involves specifying the background statements supporting the counterfactual, and the second part involves specifying the lawlike generalizations supporting the counterfactual. In this essay, I focus on the first part of the problem.[14] The crux of the first part is that the background statements of a counterfactual world must differ from those in the actual world, because the counterfactual, by definition, makes assumptions contrary to fact. But exactly how does the counterfactual world differ from the actual world? After all, if counterfactual worlds are fictional worlds, what is to keep a counterfactual world from being a psychedelic fantasy where anything goes?

[14] The second part of the problem is the problem of projectable predicates, also known as the grue paradox. In order to set aside the grue paradox, I will take for granted the projectability of the predicates in the following lawlike generalization: If the observed mean difference in a random sample is explained solely by sampling error, then, on average, across infinitely repeated random samples (generated by the same process or drawn from the same population), the difference would be zero.

I think that members of an AMS data team considering the observed baseline correlation between coin-flip results and baseline GPAs would be fairly comfortable reasoning about the correlations in counterfactual worlds in which there are different AMS students with different coin-flip results and baseline GPAs. I think the data team would readily agree that the correlation would not hold up consistently across counterfactual versions of AMS: in some worlds the correlation would be negative (as it was in our actual case), but in other worlds the correlation would be positive, and in the average world there would be no correlation at all. Such counterfactual reasoning supports the intuition that the sample correlation between coin-flip results and baseline GPAs is accidental rather than lawlike. Suppose, however, that the data team has a mischievous member who runs wild with the notion that counterfactual worlds are fictional worlds. The member says, "In my counterfactual world, the smartest students, who happen to have the highest baseline GPAs, had supernatural divinations about the program and wanted to get into it, so they supernaturally rigged their coin flips to guarantee program assignment. In that counterfactual world, and in counterfactual worlds like it, the correlations between coin-flip results and baseline GPAs would tend to be positive." I think that even the mischievous member herself would agree that her counterfactual world, with all its supernaturalism, is somehow irrelevant, presumably because the counterfactual claim is supposed to be about AMS students, not Hogwarts students. Nevertheless, it is a tricky matter to define precisely the relevancy criteria for background statements in counterfactual worlds.

As of yet, there are no universally objective criteria for counterfactual relevancy that neatly sort background statements (or counterfactual worlds) into the categories of

admissible and inadmissible. I doubt that such criteria will ever exist. In the meantime, the best that we can do is to strive for intersubjective agreement within self-critical communities. Kim and Maslen (2006) argue that counterfactuals are short stories and should be evaluated as such. Their account leaves the details vague, and the devil is in the details. When it comes to evaluating stories, communities can demonstrate a remarkable ability to build critical consensus about the internal logic of even the most fantastical fictions. Fans of the Superman character willingly suspend disbelief about a man with x-ray vision and heat vision, but they protest when Superman is depicted as exercising *ad hoc* powers such as rebuild-the-Great-Wall-of-China vision. In his *Poetics*, Aristotle (2006, p. 46) criticized the insertion of the unnecessary and improbable into the Greek drama, even though the intervention of the gods was perfectly acceptable: "It is therefore evident that the unraveling of the plot, no less than the complication, must arise out of the plot itself, it must not be brought about by the *Deus ex Machina*." In my work as a teacher of ninth-grade English literature, I found that even unsophisticated readers had visceral reactions against plot contrivances. In principle, people understand that not everything is permissible in fiction, even fictions that contains superheroes and gods, and moreover people can at least sometimes reach agreement on what is and is not permissible.

Data teams have two great advantages in following the counterfactual logic implied by sampling error as an explanation. First, quantitative data are so bare that very little is left to the imagination. Quantitative data are relatively easy to interpret. As analysts of quantitative data, we are not talking about the factual world in all its richness, or about counterfactual worlds in all their richness; we are talking about distilled

representations—quantifications. In the dataset, the data team has a distilled representation of the factual world, and from the dataset, the data team can consider distilled representations of counterfactual worlds. It is the work of psychometricians to validate the quantifications such that inferences about the quantities warrant inferences about the world. Psychometrics is outside the scope of this project, but the work of data teams depends critically on psychometrics. The complex work of psychometrics makes the otherwise complex work of statistical inference fairly simple.

Results from theoretical statistics provide the second great advantage in following the counterfactual logic implied by sampling error as an explanation. Theoretical statistics yields critical insight into the behavior of sample means across indefinitely repeated random sampling from an indefinitely large population. Theoretical statistics provides practical information about the sampling distributions of means, mean differences, and correlations, among other statistics. The measures contained in the primary example of this essay (i.e., the program effect, baseline GPA difference, and SES gap) are quantified as mean differences, so for the sake of concreteness, I focus on unbiasedly estimated mean differences, denoted $\hat{\beta}_1$. The sampling distribution of $\hat{\beta}_1$ is a distribution of an infinite number of sample-mean differences from an infinite number of random samples of equal size. From a process-minded perspective, the infinite number of samples are all generated by the same random process. From a population-minded perspective, the infinite number of samples are all randomly drawn from the same population. Theoretical statistics gives data teams the necessary insight into the counterfactuals implied by sampling error as an explanation. If sampling error were the explanation, then theoretical statistics would tell the data team three things. First, the

mean of the $\hat{\beta}_1$ distribution would be zero. Second, if the factual sample is sufficiently large, the shape of the $\hat{\beta}_1$ distribution would be approximately normal. Third, if the factual sample is sufficiently large, the variance of the $\hat{\beta}_1$ distribution would be estimable (with good precision and little bias) from sample statistics.

In Figures 5 through 8, I visually depict the counterfactual logic implied by sampling error as an explanation. In Figure 5, I depict the real-number line that I will use to compare the sample-mean differences in baseline GPAs ($\hat{\beta}_1$) from the factual sample and from many counterfactual samples. Much of the logical labor is accomplished by psychometricians who map the target system to the real-number line. In Figure 6, I depict the sample mean difference in baseline GPAs from the factual sample alone ($\hat{\beta}_1$ = -0.07). I mark the point on the real-number line with a grey dot above the point. The grey dot encloses a globe that represents the world in a distilled and quantified form. The world is the factual world. In Figure 7, I depict the results from 136 counterfactual samples in addition to the factual sample. For each counterfactual sample, I mark its $\hat{\beta}_1$ on the real number line with a white dot enclosing a globe. Each white dot represents a counterfactual world in distilled and quantified form. Theoretical statistics tells us that the counterfactual worlds will tend to fall out in this way if sampling error does indeed explain the difference in sample means that we observe in the actual sample. We can see that the actual sample would not be improbable among the counterfactual samples if sampling error were the explanation. Of course, this is not definitive proof that sampling error is the explanation, but it does suggest that we cannot reject sampling error as the explanation. Contrast the results depicted in Figure 7 with those in Figure 8, which

presents the estimated program effect size from the factual sample as 0.21. The factual sample is highly improbable among the counterfactual samples if sampling error is the explanation. This fact does not provide absolute proof that sampling error is *not* the explanation, because the tails of the normal distribution asymptotically approach infinity such that no value of $\hat{\beta}_1$ has zero probability. Nevertheless, the conventions of null-hypothesis significance testing (NHST) would have us reject sampling error as the sole explanation. The grounds for this rejection are that the probability of drawing a random sample with $\hat{\beta}_1$ as extreme as or more extreme than 0.21 is less than .05, if sampling error is the sole explanation for $\hat{\beta}_1$.

Figure 5. The real-number line upon which sample-mean differences in baseline GPA by program assignment are placed.

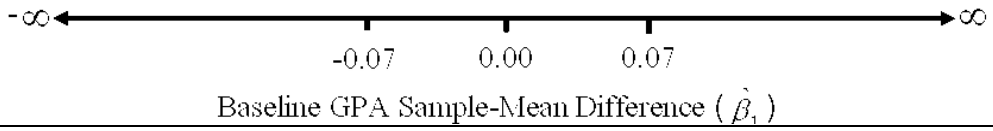


Figure 6. For baseline GPA by program assignment, the sample-mean difference from the factual sample (grey).

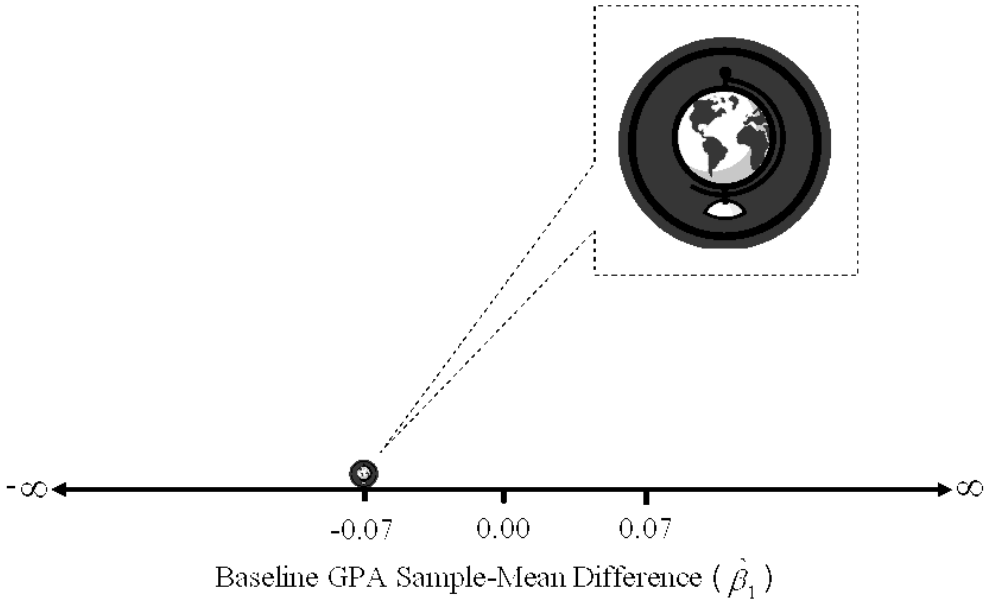


Figure 7. For baseline GPA by program assignment, sample-mean differences from 136 counterfactual samples (white) and the factual sample (grey).

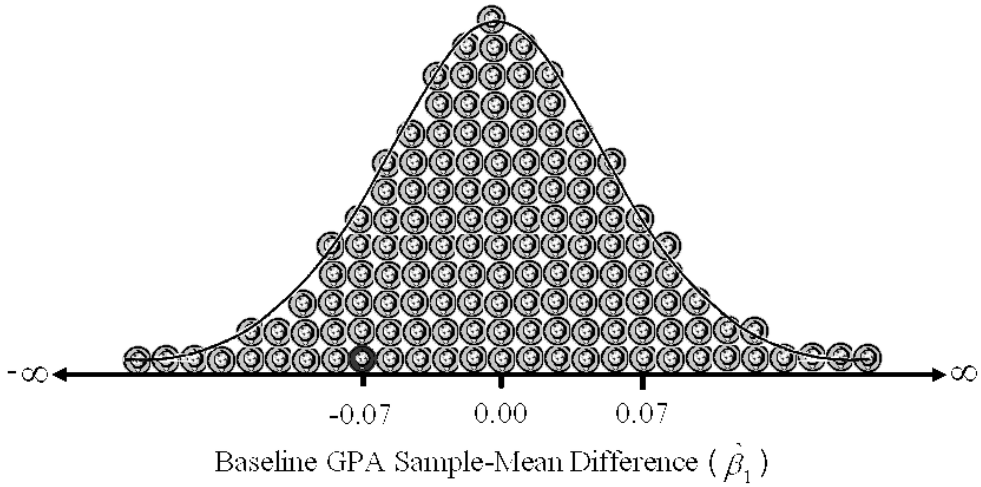
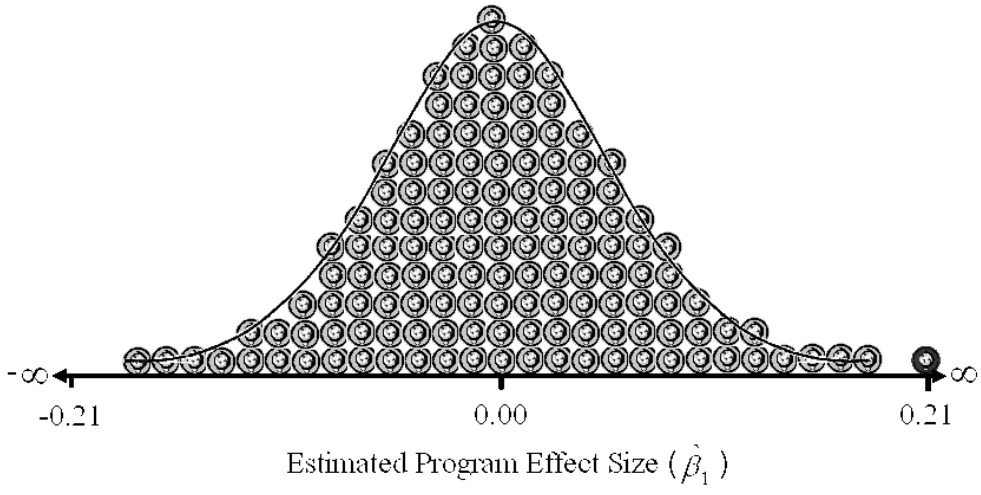


Figure 8. Estimated program effect sizes from 137 counterfactual samples (white) and the factual sample (grey).



A discussion of the merits of NHST would not be relevant to my purpose. My purpose is to highlight the counterfactual reasoning involved in NHST, which depends on the results from theoretical statistics. Similar reasoning supports the construction of confidence intervals. In order for the results of theoretical statistics to be applicable, the factual sample must be a random sample from an indefinitely large population. It is easy to conceptualize the drawing of a random sample from a demographic population such as

a school population: hold some sort of lottery in which every member of the demographic population has an equal selection probability of 1 divided by the size of the demographic population. It is not so easy to conceptualize the drawing of a random sample from a statistical population. After all, the statistical population is indefinitely large, and it includes a great number of counterfactual members. Nevertheless, I think there is a short but subtle argument for treating the census as a random sample taken from the population.

First, consider the census as a sample from the population, so that we can ask whether the census is a non-random sample or a random sample. As a sample from a population, the census can always exemplify the population. Recall, however, that a single sample can exemplify many labels. A sample of green frogs can exemplify "frogs" or "green." Similarly, the census as a sample can exemplify many populations. The census can exemplify a population from which the census was non-randomly drawn or a population from which the census was randomly drawn. What a sample exemplifies depends on usage. By using statistical theory that assumes random sampling, applied statisticians declare that they are using the census to exemplify the population from which the census was randomly drawn.

At the start of this subsection, I noted that the sampling mechanism must be theory-neutral for the sake of scientific objectivity. Random sampling not only undergirds the applicability of theoretical statistics but is also theory-neutral, at least for a while. I have just argued that the census can be treated as a random sample, and since a random sample from a random sample is a random sample, a random sample from the census can always be treated as a random sample from the population.

There is, however, a caveat that accompanies any random sample. That caveat is against *post hoc* theorizing. The theory-neutrality of a random sample is fragile. For the purpose of testing a theory, researchers can strip the theory-neutrality from a random sample by developing the theory based on the random sample. In this case, the theory and sample are no longer independent, and the sample no longer provides an independent test of the theory. If a data team observes a surprisingly strong correlation in a random sample, theorizes that the correlation is not due solely to sampling error, and tests the theory using the same random sample, then the data team is making the mistake of *post hoc* theorizing. According to NHST methodology and a customary alpha level of 0.05, on average, statistical tests on the strongest 5% of spurious correlations in a random sample will yield false positives, falsely suggesting that the spurious correlations are not explainable by sampling error alone. If a data team thoroughly explores a dataset, the data team will find the strongest 5% of spurious correlations. Of course, those spurious correlations do not come labeled as such. They will appear to be interesting findings, but they are properly explained as mere artifacts of sampling error. The problem of *post hoc* theorizing looms even when we sample the entire census. The entire census may be a random sample from the population, but, once we use it to do *post hoc* theorizing, it is no longer a theory-neutral sample from the population, and since we sampled the entire census, we cannot take another random sample to test our *post hoc* theories independently. When we do *post hoc* theorizing from the census, we paint ourselves into an inescapable corner. This is an extension of my overall point that the status of the census as a sample (as a random sample or as a theory-neutral sample) depends very much on how we treat the census.

Why Be Population-Minded?

I hope to have answered two questions in this essay: (1) Why should we be population-minded (or process-minded) rather than census-minded? (2) Why should we be population-minded rather than process-minded? I consider each question in turn.

Why not be census-minded? Inferences to the census support descriptions of the census, but they do not support explanations of the census. The census includes all actual subjects that are the target of inquiry, and only these subjects. A focus on the census is perfectly appropriate in the context of a descriptive inquiry. If the data team wants to know how many current AMS students are eligible for free lunch, then the data teams wants to infer to the census of current AMS students. Likewise, if the data team wants to know whether the students eligible for free lunch just *happen* to earn lower GPAs, on average, than their ineligible counterparts, then the data team wants to infer to the census. If, however, the data team wants to know *why* there is an SES achievement gap, then the data team wants to establish a *tendency* (not just a one-time occurrence) according to which students eligible for free lunch earn lower GPAs, on average, than their ineligible counterparts. In other words, the data team wants to rule out that the correlation is spurious, or that the correlation would not hold up over repeated counterfactual samples of counterfactual students randomly drawn from the population from which the census was randomly drawn. The data team needs to establish a lawlike generalization for the purposes of explanation. In order to establish the requisite lawlike generalization about the target of inquiry, current AMS students, the data must infer to a population that includes not only actual current AMS students but also counterfactual current AMS students.

Second, why not be process-minded? As far as I can tell, there is no good reason to be population-minded rather than process-minded, if everybody is comfortable with random variables that have probability-density functions that are not uniform and discrete. Process-mindedness supports the target-counterfactual reasoning necessary for explanatory inquiry. The problem with process-mindedness is that many applied statisticians (not to mention the audiences of applied statisticians) do not really understand random variables. If an applied statistician is population-minded, then the applied statistician needs to understand only the randomness involved in lotteries. Whereas the process-minded statistician conceives of samples as generated by a random process modeled with a random variable, the population-minded statistician conceives of samples as randomly drawn by lottery from a target population. For this simplification, the population-minded statistician must pay a price. The population-minded statistician must define the target population in such a way that it is consistent with being a product of the modeled process. In this essay, I have shown how we can define target populations independently of, but consistent with, the modeled process.

We can trace the definition of the target population from the actual sample. Thus, the actual sample provides empirical grounding for the theoretical population. The actual sample is a sample, and as such it can exemplify many things. What exactly the sample exemplifies depends on how the sample is used. When the data team uses the actual sample along with traditional methods of statistical inference, then it declares that it is using the actual sample to allude to other samples from the same population. Specifically, the data team is using the actual random sample to allude to counterfactual random samples as "possible subjects that are the target of inquiry." The population, for the

purposes of statistical inference in the context of an explanatory inquiry, is simply all the possible subjects that are the target of inquiry.

It is a mouthful to say "all the possible subjects that are the target of inquiry," so I propose a shorthand that is consistent with the parlance prevalent among applied statisticians. If the AMS students are the target of inquiry, applied statisticians sometimes make a distinction between the actual AMS students and *AMS-type* students. In practice, they are making a distinction between the census and the population. I theorize that applied statisticians are using the hyphenated suffix "-type" to denote "all the possible subjects that are the target of inquiry." To test the theory, we can inspect whether the labels have the same extension. Let us focus on current AMS students as the targets of inquiry. Consider two labels: "possible current AMS students" and "current AMS-type students." Both labels apply to actual current AMS students. Both labels apply to realistically counterfactual current AMS students. Furthermore, if rightly understood, neither label applies to fantastically counterfactual current AMS students. As I discussed in the previous section on evaluating claims about largely fictive populations, the range of "possibility" is restricted by statistical theory. The sampling distribution that statistical theory leads us to consider is a distribution of possible statistics from possible samples that have been randomly drawn from the same population from which the census was randomly drawn.

In conclusion, let us revisit an earlier example: Consider a school interested in creating an inclusive educational environment for students with Asperger's syndrome. The data team is tasked with providing insight into current gaps in academic achievement associated with AS. There is, however, only one AS student currently in the school. The

one student provides sufficient information to draw an inference to the census of current students with AS. That one student, however, does not provide sufficient information to draw an inference to the *population* of current students with AS. In other words, that one student does not provide sufficient information to draw an inference to possible current students with AS, or to AS-type students in the current educational environment. If the current AS student happens to be performing below average, the data team cannot rule out sampling error as an explanation for the statistical finding that, in the sample, students with AS earn lower GPAs, on average, than their non-AS counterparts. Based on NHST methods of statistical inference, it is plausible that, if the team had more data, the team would find no achievement gap, so the statistical findings are inconclusive. In order to conduct an explanatory inquiry, or at least an inquiry in which sampling error can be ruled out as the explanation for an observed achievement gap, data teams must consider not only actual students but also counterfactual students. That is, data teams must consider the population.

References

Abo-Zena, M. M. (2010). Sample size planning. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1301-1309). Thousand Oaks, CA: Sage.

Adler, J. E. (1997). Lying, deceiving, or falsely implicating. *Journal of Philosophy, 94*(9), 435-452.

Aristotle. (2006). *Poetics*, trans. by S. H. Butcher. New South Wales, Australia: Objective Systems.

Banerjee, S., & Iglewicz, B. (2007). A simple univariate outlier identification procedure designed for large samples. *Communications in Statistics—Simulation and Computation, 36*(2), 249-263.

Boswell, J. & Hill, G. B. N., eds. (1921). *Life of Johnson: Including Boswell's journal of a tour to the Hebrides and Johnson's diary of a journey into North Wales*. New York, NY: Bigelow, Brown & Co..

Boudett, K. P., City, E. A., & Murnane, R. J. (2005). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning.* Cambridge, MA: Harvard Education Press.

Boudett, K. P., & Steele, J. L. (2007). *Data wise in action: Stories of schools using data to improve teaching and learning.* Cambridge, MA: Harvard Education Press.

Cartwright, N. (1983). *How the laws of physics lie.* Oxford, UK: Clarendon Press.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98-113. doi:http://dx.doi.org/10.1016/j.edurev.2007.04.001

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 4*5(12), 1304.

Cohen, J., Cohen, C., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Daniels, N. (2011). Reflective equilibrium. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring ed., pp. 1–31). Retrieved from http://plato.stanford.edu/archives /spr2011/entries/reflective-equilibrium/

Deming, W. E. (1953). On the distinction between enumerative and analytic surveys. *Journal of the American Statistical Association, 48*(262), 244-255.

Deming, W. E. (2000). *Out of the crisis.* Cambridge, MA: MIT Press.

Descartes, R. (1907). *The method, meditations, and selections from the principles of Descartes* (14th ed.), translated by J. Vietch. Edinburgh and London: W. Blackwood and Sons.

Dynel, M. (2011). A web of deceit: A neo-Gricean view on types of verbal deception. *International Review of Pragmatics,* 3(2), 139-167.

Elgin, C. Z. (1983). *With reference to reference*. Indianapolis, IN: Hackett Publishing.

Elgin, C. Z. (1999). *Considered judgment.* New Haven, CT: Princeton University Press.

Elgin, C. Z. (2001). Reflective equilibrium. In R. Audi (Ed.), *The Cambridge dictionary of philosophy* (p. 782). Cambridge, MA: Cambridge University Press.

Elgin, C. Z. (2012). Making manifest: The role of exemplification in the sciences and the arts. *Principia, 15*(3), 399-413.

Galilei, G. (1914). *Dialogues concerning two new sciences*, translated by H. Crew & A. de Salvio. New York, NY: Macmillan.

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols.* Indianapolis, IN: Hackett.

Goodman, N. (1983). *Fact, fiction, and forecast.* Cambridge, MA: Harvard University Press.

Goodman, N., & Elgin, C. Z. (1988). *Reconceptions in philosophy and other arts and sciences*. London, UK: Routledge.

Grice, P. (1989). *Studies in the way of word*s. Cambridge, MA: Harvard University Press.

Hájek, A. (2012). Interpretations of probability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter ed., p. 1-71). Retrieved from http://plato.stanford.edu/entries/probability-interpret/

Hardin, R. (2013). The free rider problem. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring ed., pp. 1–10). Retrieved from http://plato.stanford.edu/entries/free-rider/

Hempel, C. (1962) Explanation in science and history. In R. C. Colodny (Ed.), *Frontiers of science and philosophy* (pp. 9-19). Pittsburgh, PA: University of Pittsburgh Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135-175.

Hesse, M. B. (1966). *Models and analogies in science.* Notre Dame, IN: University of Notre Dame Press.

Kachapova, F., & Kachapov, I. (2012). Students' misconceptions about random variables. *International Journal of Mathematical Education in Science and Technology*, *47*(7), 963-971.

Kay, S. (2006). *Intuitive probability and random processes using MATLAB*. New York, NY: Springer.

Kerr, K. A. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4), 496–520.

Kim, S., & Maslen, C. (2006). Counterfactuals as short stories. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 129*(1), 81-117.

Kuhn, T. S. (1996). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic, 8*(1), 339–359.

McGarty, C., Yzerbyt, V. Y., & Spears, R. (2002). Social, cultural and cognitive factors in stereotype formation. In C. McGarty, V. Y. Yzerbyt, & R. Spears, (Eds.) *Stereotypes as explanations: The formation of meaningful beliefs about social groups* (pp. 1-15). Cambridge, UK: Cambridge University Press.

Meibauer, J. (2005). Lying and falsely implicating. *Journal of Pragmatics, 37*(9), 1373-1399.

Murnane, R. J., Sharkey, N. S., & Boudett, K. P. (2005). Using student-assessment results to improve instruction: Lessons from a workshop. *Journal of Education for Students Placed at Risk, 10*(3), 269–280.

MDOE. *District Data Team Toolkit*. Retrieved from http://www.doe.mass.edu/apa/ucd/ddtt/toolkit.pdf

Newton Public Schools (2008). *Systemwide goals*. Retrieved from http://www3.newton.k12.ma. us /sites/default/files/Systemwide_Goals_2008-2010_10_14_08.pdf

Quine, W. V. (1951). Two dogmas of empiricism. *Philosophical Review, 60,* 20–43.

Quine, W. (1981). *Mathematical logic*. Cambridge, MA: Harvard University Press.

Ronka, D., Lachat, M. A., Slaughter, R., & Meltzer, J. (2008). Answering the questions that count. *Educational Leadership, 66*(4), 18–24.

Scheffler, I. (1979). *Beyond the letter: A philosophical inquiry into ambiguity, vagueness, and metaphor in language*. Boston, MA: Routledge & Kegan Paul.

Scheffler, I. (1996). Denotation and mention-selection. In *Symbolic worlds, art, science, language, ritual* (pp. 11-22). Cambridge, UK: Cambridge University Press. doi: http://dx.doi.org/10.1017/CBO9780511663864.002

Scherer, M. (2008). Driven dumb by data? *Educational Leadership, 66*(4), 5.

Stalnaker, R. (1999). *Context and content: Essays on intentionality in speech and thought.* Oxford, UK: Oxford University Press.

Stanford, J. L., & Vardeman, S. B. (1994). *Statistical methods for physical scienc*e. San Diego, CA: Academic Press.

Suárez, M. (2009). *Fictions in science: Philosophical essays on modeling and idealization*. New York, NY: Routledge.

Tukey, J. (1962). The future of data analysis. *Annals of Mathematical Statistics 33*(1), 1-67.

Van Fraassen, B. C. (1989). *Laws and symmetry*. New York, NY: Oxford University Press.

Watier, N. N., Lamontagne, C., & Chartier, S. (2011). What does the mean mean? *Journal of Statistics Education, 19*(2), 1-20.

Wiest, L. R. (2003). Twelve ways to have students analyze culture. *Clearing House, 7*6(3), 136-138.

Williams, B. (2002). *Truth and truthfulness: an essay in genealogy*. Princeton, NJ: Princeton University Press.

Woodward, J. (2011) Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter ed., p. 1-78). Retrieved from http://plato.stanford.edu/archives/win2011/entries/scientific-explanation

VITA

Sean Parker

1994-1998	**Clark University, Worcester, MA** **B.A., May 1998**
2001-2004	**English Teacher** **Dennis-Yarmouth Regional High School** **South Yarmouth, MA**
2004-2005	**St. John's College, Santa Fe, NM** **M.A., August 2005**
2005-present	**Doctor of Education Candidate** **Graduate School of Education** **Harvard University** **Cambridge, MA**
2008-2013	**Lecturer** **Tufts University** **Medford, MA**
2008-present	**Data Team Consultant**