DENVER, COLORADO
JSM 2019
JULY 27–AUGUST 1

MAKING AN
IMPACT
WITH STATISTICS
FREEBOOK
EXCLUSIVELY FOR ATTENDEES OF 2019 JOINT
STATISTICAL MEETINGS
TABLE OF CONTENTS

Introduction

1 • Data Science in Sports
   • The Birth of the Pool
   • Improving the Big East Conference Basketball Tournament

2 • Statistics and Healthcare
   • Discovery of New Fraud Patterns
   • On the Use of the Concentration Function in Medical Fraud Assessment

3 • Statistics and Social Issues
   • Thinking Statistically About Crime
   • Fighting Slavery through Statistics: A Discussion of Five Promising Methods to Estimate Prevalence in the United States
   • Jobs
   • Is the Gini Index of Inequality Overly Sensitive to Changes in the Middle of the Income Distribution?
   • Measuring Aspects of Sexuality and Gender: A Sexual Human Rights Challenge for Science and Official Statistics

4 • Bonus Journal Content

5 • Call for Authors ASA-CRC Series of Statistical Reasoning in Science and Society

6 • ASA Portal Information
READ THE LATEST ON DATA SCIENCE AND STATISTICS WITH THESE KEY TITLES AND JOURNALS

VISIT WWW.CRCPRESS.COM/STATISTICS TO BROWSE FULL RANGE OF STATISTICS BOOKS

SAVE 30% AND FREE STANDARD SHIPPING WITH DISCOUNT CODE ASA18
Introduction

With a focus on the Joint Statistical Meetings 2019 theme, *Statistics: Making an Impact* the JSM Freebook is brought to you by CRC Press/Taylor and Francis Group and the American Statistical Association and is an exclusive resource for JSM attendees. Our editors have handpicked chapters from books in the popular *ASA-CRC Series on Statistical Reasoning in Science and Society* and trending ASA journal articles. This book features content surrounding JSM key topics including:

- Data Science in Sports
- Statistics and Healthcare
- Statistics and Social Issues
- Bonus content that includes articles from the special issue: *Statistical Inference in the 21st Century: A World Beyond p < 0.05*

The chapters are from:

**Books from the ASA-CRC Series on Statistical Reasoning in Science and Society**

**Improving your NCAA® Bracket with Statistics** by Tom Adams

This book is both an easy-to-use tip sheet to improve your winning odds and an intellectual history of how statistical reasoning has been applied to the bracket pool using standard and innovative methods. It covers bracket improvement methods ranging from those that require only the information in the seeded bracket to sophisticated estimation techniques available via online simulations.

**Statistics and Healthcare Fraud: How to Save Billions** by Tahir Ekin

This book helps the public to become more informed citizens through discussions of real-world health care examples and fraud assessment applications. The author presents statistical and analytical methods used in health care fraud audits without requiring any mathematical background. The public suffers from health care overpayments either directly as patients or indirectly as taxpayers, and fraud analytics provides ways to handle the large size and complexity of these claims.

**Measuring Crime: Behind the Statistics** by Sharon Lohr

Crime statistics are everywhere, but how do you know when they’re valid? If a newspaper report says "the rate of overall violent crime decreased by 0.9 percent," how can you tell where that statistic came from, what it measures, and how accurate it is? Is it worth repeating or sharing? *Measuring Crime: Behind the Statistics* gives you the tools to interpret and evaluate crime statistics’ quality and usefulness.

**Measuring Society** by Chaitra H. Nagaraja

This book is a short, accessible guide to six topics: jobs, house prices, inequality, prices for goods and services, poverty, and deprivation. Each relates to concepts we use on a personal level to form an understanding of the society in which we live: We need a job, a place to live, and food to eat. Using data from the United States, the authors answer three basic questions—why, how and for whom these statistics have been constructed. They add some context and flavor by discussing the historical background. The intention is to provide the reader with a good grasp of these measures.
Introduction

Publications from the American Statistical Association:

*The American Statistician* is a quarterly, peer-reviewed scientific journal covering statistics published by Taylor & Francis on behalf of the American Statistical Association. It was established in 1947. The journal contains timely articles organized into the following sections: Statistical Practice, General, Teacher’s Corner, History Corner, Interdisciplinary, Statistical Computing and Graphics, Reviews of Books and Teaching Materials, and Letters to the Editor.

*CHANCE* is published quarterly by Taylor & Francis on behalf of the American Statistical Association. The magazine, created in 1988, is designed for anyone who has an interest in using data to advance science, education, and society. *CHANCE* is a non-technical magazine highlighting applications that demonstrate sound statistical practice. *CHANCE* represents a cultural record of an evolving field, intended to entertain as well as inform.

*Statistics and Public Policy* (open access) applies statistical methodology to problems in the realm of public policy and/or relevant political science. Established in 2013, it publishes open access articles that address international, national, or local policy questions, with an emphasis on application rather than methodological novelty.

*Note to readers:* References from the original chapters have not been included in this text. For a fully-referenced version of each chapter, including footnotes, bibliographies, references and endnotes, please see the published title. As you read through this FreeBook, you will notice that some excerpts reference subsequent chapters. Please note that these are references to the original text and not the FreeBook.
CHAPTER 1
Data Science in Sports
A Note From the Editor

Most major sports franchises and university athletic departments have now fully embraced data science and analytics as a way to win more games and to improve player performance. This chapter includes an article and a book excerpt about NCAA college basketball. The first discusses how individuals can improve their odds during “March Madness” tournament pools, while the second discusses how the Big East can improve its conference tournament.

**The Birth of the Pool** excerpted from *Improving your NCAA® Brackets with Statistics* by Tom Adams, published by the ASA and CRC Press, 2018.

Twenty-four million people wager nearly $3 billion on college basketball pools each year, but few are aware that winning strategies have been developed by researchers at Harvard, Yale, and other universities over the past two decades. This chapter looks at the history of the bracket pool.


Through data-based modeling, the authors show that the current Big East tournament format is not very effective in determining the true best team. Specifically, by considering a variety of alternate formats, the authors find that certain formats that exclude all but a handful of teams substantially outperform the current format in determining the true best team.
CHAPTER 1

THE BIRTH OF THE POOL

This article is excerpted from
Improving Your NCAA Bracket with Statistics
by Tom Adams
©2018 Taylor & Francis Group. All rights reserved.

Learn more
The Birth of the Pool

It all began at Jody’s Club Forest. Jody’s bills itself as “A Taste of Irish in Staten Island.” The family-friendly North Shore West Brighton neighborhood eatery and bar organized the first known bracket pool in 1977. The entry fee was $10 and 88 patrons filled out brackets (Rushin 2009). The bracket had only 32 teams that year. This first bracket pool involved picking only the Final Four®, the champion, and the number of points scored in the championship game, not the full bracket. Marquette beat North Carolina in the championship game that year.

A year later, Bob Stinson, a U.S. Postal Service employee, started a bracket pool that may stand as the first using the typical modern format and rules in Louisville, Kentucky. According to Stinson, before that people would just pick random team names out of a hat and the newspaper did not print a bracket. He wanted a contest that rewarded knowledge of the game of basketball. The pool had 15 participants in its first year (Hill 1997). Tim Trowbridge of Kent, Ohio independently started a pool that involved picking a winner for every game in the bracket back in 1981 (Allard 2017). The NCAA tournament had 48 teams back then. The National Collegiate Basketball Association, or NCAA, is the governing body for much of U.S. college basketball. Trowbridge and his
friend Jeff Hunt managed the pool. They worked out the format over a few beers in a local pub. Their goal was to “once and for all determine who knew the most about college basketball.” This pool also had 15 entries the first year and that grew to 200 entries by the fourth or fifth year. The filled-out bracket entry sheets were collected from local bars. An engraved plaque memorializing the names of all the winners of the “Trowbridge Hunt & Trowbridge Annual NCAA Tournament” used to hang in the Rusty Nail bar. It now hangs near the door of Trowbridge’s office. The first description of an “office pool” involving the NCAA tournament bracket was in 1984 among books and sources indexed by Google (Wayne 1984).

THE TOURNAMENT

In the first intercollegiate basketball game, played in 1896, the Chicago Maroons beat the Iowa Hawkeyes. Yale was a powerhouse team in the early years. The Ivy League was the first conference. A conference is an organized group of teams in a region that mostly play each other. The Southern Conference held the first college tournament at the end of its regular season in 1921 (ESPN 2009).

Tournaments caught on. Tournaments are interesting and exciting for a number of reasons. These are single elimination tournaments—so each game is a sudden death match that (in the case of the NCAA tournament) ends the season for the losing team. Since the seniors on the team will graduate or lose eligibility, it marks the last game that a losing team would ever play together. This can be a very emotional moment. Only one team emerges from the tournament victorious in all games. Another appeal of a tournament is that it tends to be an equalizer. The best team in a conference may be favored to win all the conference games it plays all season. But playing in a tournament is like running a gauntlet. It’s often the case that even the best team is more likely than not to lose the tournament. A tournament is a challenge for even the best teams, and it gives every team a chance to shock the rest and emerge victorious. This level of unpredictability is riveting for the fans.
The first NCAA tournament, consisting of eight teams, was held in 1939. It expanded to 16 teams in 1951. In 1953, the NCAA tournament added six more teams; this was the first unbalanced bracket requiring six play-in games. The tournament expanded to 32 teams in 1975, and 64 teams in 1985. In 2011, the field expanded to 68 teams, requiring four play-in games, but most bracket pools ignore the play-in games and use a balanced format involving the 64 teams that include the winners of the play-in games. (The NCAA has started calling the play-in round of four games the “first round.” But this book will stick to the earlier convention of calling the round of 64 teams the “first round.”)

THE BRACKET POOL EMERGES
A number of events happened around 1977 that might explain why the bracket pool caught on. UCLA dominated the tournament for 12 years before 1976, winning 10 of 12 championships, making the championship game somewhat anti-climactic. UCLA's coach for that period, John Wooden, retired in 1975, and the UCLA dynasty was over. The championship game became more competitive. Fans started anticipating a close, exciting game. Television ratings soared when Larry Bird dueled Magic Johnson in the 1979 championship game. The NCAA tournament became a national institution in the United States. The phrase “Final Four” was coined in 1975. The Xerox machine was becoming more common in offices, allowing easy duplication of the bracket printed in newspapers just before the tournament. The stage was set for the office bracket pool to emerge.

Why did the bracket pool begin in Staten Island? Why not Las Vegas, the center of the legal sports betting universe in the United States? Las Vegas is a tourist town, and tourists rarely stay for the four weeks between placing bets and seeing the outcome of the tournament. Bracket pools made sense in neighborhoods or offices where the participants live and work together. Betting on the win/lose outcomes of sport competitions is illegal in most of the United States, including Staten Island, but some state penal codes carve
out exceptions for “social gambling” when the pool organizer does not profit by charging a fee or taking a cut of the pot. The fairly expansive social gambling exception in the New York State penal code led to the toleration of a public bracket pool, open to all, at a business establishment. Even in states where social gambling is illegal or when federal laws apply, bracket pools are rarely prosecuted (Edelman 2017). Edelman recommends limiting entrants to close friends and paying out all entry fees as prizes. After 29 years of continuous operation, Jody’s pool was discontinued in 2007 after Jody Haggerty was investigated by the IRS and fined by the New York State Liquor Authority for promoting gambling in an establishment licensed to serve liquor (Daily News 2007). Press reports indicate that the 2006 pool winner had reported his winnings as income to the IRS, leading to investigations. The pool pot had grown to $1.5 million, and Jody was not reporting the payout of the winnings to the IRS as required by regulations.

The tournament is played after the regular college basketball season, which runs each year from October to March. There are about 350 teams in the NCAA Division I. Division I is composed of the larger colleges, and these generally have the better teams. These teams are organized into conferences consisting of 8 to 16 teams each. During the regular season, each team plays about 30 games, mostly with other teams in their conference. Just after the regular season, each conference plays a conference tournament. The winners of all these tournaments get an automatic invitation (called an “automatic bid”) to the NCAA tournament. The rest of the tournament field is fleshed out with teams chosen by an NCAA selection committee. These “at-large bids” are awarded to teams that had a good season with many wins but were eliminated during the conference tournaments.

In the course of the NCAA tournament, all but one team is eliminated by losing a game. The tournament starts with 64 teams (not counting the play-in games) and ends after 63 are eliminated in 63 games. The NCAA tournament is organized into four regional tournaments of 16 teams each. The teams are
seeded 1 to 16 in each regional tournament. The teams are seeded by the selection committee based on their season performance and a few other considerations. The lower numbered seed is actually the higher ranked seed, typically referred to as the “higher seed.” So, the number 1 seeds tend to be the favorites. Figure 1.1 shows how the brackets for each regional tournament are structured. The victors in the four regions, the Final Four, meet in the final two rounds of the NCAA tournament to decide the national champion.

FIGURE 1.1  The bracket structure for an NCAA tournament region.
The selection committee meets in March on “Selection Sunday©” to determine the at-large bids, to assign teams to regions, and to seed each team in their region. The bracket, with all teams assigned to a specific slot, is revealed around 6:00 p.m. on Selection Sunday. The first game of the first round of the tournaments begins around noon on the following Thursday. During those 90 hours (between the revelation of the bracket and the tip-off of the first game) is when the bracket pool game is played, because you can’t fill out a bracket until you know how the tournament is seeded and you are not allowed to enter a bracket in a pool after the games have begun. Your goal is to rack up the most points. You get points for each game when you correctly predict the winner of that game. The specific scoring rules vary from one bracket pool to another, but more points are awarded for games in the later tournament rounds in most pools. Your bracket along with your entry fee must

FIGURE 1.2 President Obama’s 2015 Men’s bracket (Wall 2015). Licensed under CC BY 3.0.
be delivered to your office pool’s manager before the games begin. Figure 1.2 shows President Barack Obama’s 2015 bracket.

Now comes the waiting and watching to see how your bracket fares against your opponents’ brackets. The 32 games of the first round of the four regional tournaments are played over Thursday and Friday, 16 games per day, with two or three teams defeated and eliminated per hour at some points. The second round of 16 games among the 32 victors in the first round is played on Saturday and Sunday. Then there is a breather till the next Thursday, then four more days of tournament play that winnow the field down to just four teams. Then another breather till the next Saturday night when two games are played and only two teams are left standing. Then, after a rest day, the final game of the tournament, the championship game, is played the following Monday night. The last two teams learn their fate, as do all the hopeful bracket pool players who still have a bracket sheet in the running. In some years, all fates are determined by a ball flying toward a hoop in the last split-second of the last game. The tournament champion is crowned, and all the pooled money is distributed to the bracket pool winners who have the best scores on their bracket sheets.

All gambling is illegal in some states. In 1992, Robert S. Plain of East Greenwich High School in Rhode Island was arrested for possession of gambling paraphernalia. He had in his possession bracket pool entry sheets that he was handing out during homeroom. Mary McNulty, his math teacher of all people, reported him to the police. District Court Judge Robert K. Pirraglia ordered Plain to pay $84.50 in court costs (AP 1992). Fortunately for the future of the bracket pool, this conviction did not create a trend in the United States. Some students at Greenwich protested that betting pools involving teachers were common at the school. The bracket pool was on its way to becoming something of a national institution.

The NCAA created the tournament, but it has no love for the peculiar institution of the bracket pool that the tournament spawned. In the NCAA’s own words: “Does the NCAA really oppose the harmless small-dollar bracket office pool for the Men’s
Final Four? Yes! Office pools of this nature are illegal in most states. The NCAA is aware of pools involving $100,000 or more in revenue. Worse yet, the NCAA has learned these types of pools are often the entry point for youth to begin gambling. Fans should enjoy following the tournament and filling out a bracket just for the fun of it, not on the amount of money they could possibly win” (NCAA 2010). The NCAA is OK with some bracket contests that give prizes but do not require a wager. The NCAA sponsors its own bracket contest at the website bracketchallenge.ncaa.com.

A NATURAL EXPERIMENT IN ECONOMICS

About 15 years after the first bracket pool at Jody’s, Andrew Metrick somehow convinced his thesis advisor, Eric Maskin (winner of the Nobel Prize for Economics in 2007), to allow him to write a PhD thesis at Harvard on topics that included the Jeopardy!® game show and the NCAA bracket pool. Metrick was an avid chess player in high school and developed an interest in games and decision making. Games that involve decisions and monetary prizes can constitute a natural experiment in economics. The constraints and rules of such a game make it somewhat like a well-designed experiment, one where the results are available without the cost of carrying it out in a laboratory. Metrick used data from 24 bracket pools conducted in 1993 for his analysis of strategic behavior. He also wrote a paper on the bracket pool (Metrick 1996).

Simplify the Problem

In his quest to analyze the strategies of pool players, Andrew Metrick had a problem: there were a lot of strategies. In the parlance of game theory, each possible distinct filled-out bracket represents one pure strategy (a pure strategy is one that is completely defined with no uncertainty). The number of pure strategies equals the number of possible tournament outcomes. A basketball game has two possible outcomes, and each additional game doubles the number of possible outcomes for the set of 63 games that
are played. It’s like the legend of the wheat and the chessboard. In one telling of the story, Sessa, the inventor of chess, had so pleased his master that the master offers Sessa his heart’s desire. Sessa asked that one grain of wheat be placed on the first square of the chessboard, two on the second, four on the third, and so on in a doubling progression until the board was filled. His ruler laughed at such a meager prize, only to find that the 64th square alone must hold over 18.4 quintillion grains of wheat. So it goes, with the possible outcomes when 63 basketball games are played, only half as many, over 9.2 quintillion outcomes. Each pool player can choose from over 9.2 quintillion possible brackets. There are over 9.2 quintillion pure strategies.

Metrick proceeded as if he was following the advice from Polya’s *How to Solve It*: “If you cannot solve the proposed problem, then do not let this failure afflict you too much but try to find consolation in some easier success, *try to solve some related problem* …” (Polya 1973). Metrick simplified the rules of the bracket pools that he was analyzing. There were just two rules:

**Rule 1:** Pick one winner at random from all players who correctly chose the tournament champion. This amounts to taking all the brackets that picked the champion correctly, throwing them in a hat, and blindly picking one winner.

**Rule 2:** If nobody qualifies under Rule 1, then pick one winner at random from all the entrants in the pool. That is, throw all the brackets in the hat and pick one winner.

The rules only use the player’s pick for champion. The player’s down-bracket picks have no bearing on who wins the pool. But these simplified rules retain one important characteristic of actual pools: one player who correctly picks the champion usually wins the pool. Typically, more points are awarded for picking the champion. The most common pool scoring system has the points awarded for each game doubling with each round, awarding 1, 2, 4, 8, 16,
and 32 points for tournament rounds 1, 2, 3, 4, 5, and 6, respectively. Under this scoring system and other common pool scoring systems, the winner of the pool is almost always someone who picks the champion (assuming anyone picks the champion). If the champion pick is not decisive in determining the pool winner, then the winner of the pool will be decided before the championship game, making the championship game anticlimactic as far as the pool is concerned. Making it anticlimactic is not necessarily bad, but pool managers tend to set up rules that make this an uncommon occurrence.

Under the simplified rules, there are only 64 pure strategies, corresponding to the 64 teams that a pool player could pick as champion.

A Team’s Chances of Winning
Metrick wanted to know if there were any profit opportunities available to players in bracket pools. He had all the 1993 data from 24 bracket pools. So, he had information on the proportion of players that bet each team for champion. He also needed information on the probability that any given team would win the tournament. He got this information from the Las Vegas odds. The bookmakers take bets on the champion and they adjust the odds so that they will have to pay out the same amount of money whether the team wins or loses. Say a bookmaker gets a bet from Larry for $100 that Duke will win and a bet from Moe for $70 that Duke will lose, at even odds. The bookie will charge a fee (or vigorish as it is called, or vig for short) from each bet, let’s say 10%. So, the bookie will have $100 + $70 + $17 = $187. But if Duke wins, then he will have to pay out $200 to Larry! The bookie will lose money. Ironically, bookies are not gamblers, at least not in their role as bookies. They are functionaries in a business that will not be viable without a reliable profit. So, the bookie will go to Curly (or anyone else who is interested in betting) to see if he can find someone to bet $30 against Duke. And, he may have to incentivize Curly by giving better than even odds on Curly’s bet.
The odds will shift to equalize the amount bet on each side of the proposition.

The betting market futures provide estimates of the win probabilities of the possible champ picks.

So, the bettors cause the odds to shift. In effect, the determination of the odds is crowdsourced. The odds are determined by a market mechanism. On average, sports gambling is a losing proposition because of the vigorish. The bookmakers make a profit and the profit equals the gambler’s average losses. But presumably the gamblers make a rational attempt to avoid bad odds and seek good odds, and, in this push and pull, the market ends up producing an estimate of the probability that each team will win the tournament.

Are There Favorable Strategies?

Table 1.1 presents the eight teams from Metrick’s study with the highest probability of winning the tournament. The bulk of pool players, 78.6%, chose a 1 seed for champion. Yet, the probability of a 1 seed winning the pool (according to the Vegas odds) was only 37.3%. Among the next four teams, Seton Hall looks to be an outlier, picked much more than the other 2 or 3 seeds listed. This is

<table>
<thead>
<tr>
<th>Team (Seed)</th>
<th>Pool Player Behavior: Share of Champion Picks (%)</th>
<th>Win Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNC (1)</td>
<td>22.6</td>
<td>10</td>
</tr>
<tr>
<td>Indiana (1)</td>
<td>15.8</td>
<td>9.1</td>
</tr>
<tr>
<td>Kentucky (1)</td>
<td>18.7</td>
<td>9.1</td>
</tr>
<tr>
<td>Michigan (1)</td>
<td>21.2</td>
<td>9.1</td>
</tr>
<tr>
<td>Arizona (2)</td>
<td>1.5</td>
<td>7.8</td>
</tr>
<tr>
<td>Duke (3)</td>
<td>5.4</td>
<td>7.8</td>
</tr>
<tr>
<td>Seton Hall (2)</td>
<td>10.3</td>
<td>6.9</td>
</tr>
<tr>
<td>Cincinnati (2)</td>
<td>0.7</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Source: Data reprinted from Metrick (1996), pages 163 and 166, Copyright 1996, with permission from Elsevier.
because a good proportion of Metrick’s data came from New York City where Seton Hall is a hometown team. Almost 16% of players in New York City pools picked Seton Hall for champion, whereas less than 7% picked Seton Hall in the other pools.

Consider picking UNC versus picking Arizona in a pool with 100 players (each playing 1 bracket) under Metrick’s simplified rules. If you pick UNC, then you have a 10% chance of ending up in the hat under Rule 1. But around 22 other players will be in that hat, so your chance of being picked at random from the hat is less than 1 in 20, less than 5%. Multiply that by the 10% chance of being in the hat in the first place and you have only a 0.5% chance of winning, less than 1 chance in 200. If the entry fee for the pool is $1, then there will be $100 in the pot. If you win once in every 200 years, then you only average 50 cents per year. Your average return is about 1/2 of your entry fee.

Now consider picking Arizona. You have a 7.8% chance of ending up in the hat. But there will typically only be one or two names in the hat, including yours. Even with two opponent names in the hat, you have a 1/3 chance of winning. Overall, you have a $7.8/3 = 2.6$ or a 2.6% chance of winning the pool. You have a 2.6% chance of winning $100. So, your expected return from the pool will be $2.60 for the dollar you invested in the pool entry fee.

Picking a contrarian champ reduces the competition from opponents that have the same champ.

This is not the more detailed analysis that Metrick performed, but it’s a good first cut at estimating the return on investment. One difference is that this example used the average pick share for all pools. Metrick found that the players in larger pools were prone to pick the favorites, like UNC, less than the average player. Hence, those who picked UNC in a pool of size 100 got a return of 80 cents on the dollar, better than the 50 cents or so, but still a losing proposition.
Rather than just using the raw pick share numbers (like 22 for UNC) to directly represent the number of names in the hat, Metrick treated them as point estimates of pick share probabilities. If the raw data showed that 22% picked UNC in a pool, then he estimated that each player had a 22% likelihood of picking UNC. For a pool of size 100, the number of UNC picks could, in principle, vary from 0 to 100, but those extreme values are unlikely. This is like 100 flips of a biased coin with a 22% likelihood of landing on heads. This is called a binomial probability distribution. The outcome will center around 22. Twenty-two will be the average number of UNC picks. The number of opponent names in the hat, if UNC wins, will be between 17 and 27 about 80% of the time. If only 1.5% picked Arizona, then the pick distribution for a pool of size 100 would center around 1.5. The number of opponent names in the hat if Arizona wins would be less than three about 80% of the time. Metrick estimated a $4.39 return for Arizona in a pool of size 100, quite a bit more than my seat-of-the-pants estimate.

Pooled Betting
The sort of pooled betting used in bracket pools is called pari-mutuel betting. *Pari mutuel* is just the French term for pooled betting that found its way into English. Pooled betting has interesting implications for your choice of a champion. The likelihood that you will win the pool depends on how many of the other pool players choose that same champion. If you are the only one to correctly pick the champion, then you win the pool; it’s a 100% sure thing under the simplified rules. But if 10 other players choose that same champion, then there are 10 brackets in the hat and you have a 10% chance of winning or a 0.10 probability of winning. However, this has to be balanced against the fact that the popular picks for champion tend to be the most likely winners. This is the bracket pool player’s dilemma. The pool player has to balance these two factors: (1) the probability that their champion pick will win the tournament, (2) the probability that the pool player will score enough points in the earlier rounds to beat any opponents.
who picked the same champion. Or (under Metrick’s simplified rules) the probability that their bracket will be picked from the hat under Rule 2.

Horse racing and some lotteries use pari-mutuel betting. Economists had analyzed the strategic behavior of bettors in those other games before Metrick analyzed bracket pools. In horse racing, bettors evidence a long-shot bias; that is, they over-bet the long shots. Lottery players over-bet certain lucky numbers (such as birthdays). So, a smart bettor might be able to gain an edge by betting the favorites in horse racing or avoiding the lucky numbers in the lottery. But in practice, the edge is typically not there because a fraction of the total bets is taken by the house. (Office bracket pool organizers almost never take a cut of the pot because this removes the real or de facto social gambling exclusion and could lead to prosecution.)

Metrick found the opposite of a long-shot bias in bracket pool betting. There is a bias in the direction of the favorites. Why this difference? Metrick speculated that this was due to the fact that the odds offered are explicit in horse racing. The odds are right there on the racetrack tote board, visible to all. In the bracket pool, the odds offered are more obscure. Lottery picks show something similar to a “favorites bias” where birthdays and “lucky” numbers are over-bet. The lottery odds offered for specific numbers are also obscure. It takes some investigation to estimate the odds offered in a bracket pool, some awareness and analysis of past patterns of opponent play. Few pool players evidence any awareness of the odds offered. Most bracket pool players act like someone who goes to the racetrack and plays the ponies without looking at the odds on the tote board.

Some Pool Players Already Knew

Metrick found that there was less over-betting of the 1 seeds in larger pools. The fact that players in larger pools were significantly less prone to pick the favorites indicates something interesting. As Metrick puts it, “a small number of players are
changing their behavior to reflect the changing strategic situation.” Some pool players had already discovered the gist of Metrick’s findings and were playing the pool for a profit. But in spite of the few players that deployed an effective profit-oriented strategy, there were still profits to be made. Metrick estimated that in pools of size 200 betting, Arizona still returned $5.23 on each dollar invested.

What a Competitive Pool Would Look Like
Metrick calculated the overall optimal betting strategy assuming all players had the goal of maximizing their wins/profits. He did this using a “what if” process of progressively shifting player’s champion picks to whatever profit opportunities existed until there were no more profit opportunities. This produced an “equilibrium” distribution of champion picks. The equilibrium is where no single player could improve his expected payoff. This is called the Nash equilibrium. John Nash, whose life story was presented in the movie *A Beautiful Mind*, won the Nobel Prize for development of this equilibrium analysis for non-cooperative games.

At equilibrium, the pick percentage of each team is approximately the same as the win probability expressed in a percentage. The deviations from this rule occur in the smaller pools. The smallest pools that Metrick analyzed had 25 players. The deviations were partly due to the fact that there is no profit in betting a team with a win probability of less than 1/25 even if you are the only player who bets them, because your expected payout will be less than 1/25th of the sum of the 25 entry fees in the pot; less than 25*1/25 = 25/25 = 1 entry fee. So, only nine teams (those with a win probability greater than 1/25) could be profitably bet in a pool with 25 players.

Don’t go too contrarian. Don’t pick a champ that has a win probability of less than 1/N of winning the tournament, where N is the total number of brackets in your pool, if you can avoid it.
As the pool gets larger, a larger fraction of teams is profitable to bet. In pools with 200 entries, more than 12 teams were profitable to bet and 16% of the bets went to 4 seeds or lower. (Metrick did not report the details beyond the 3 seeds.) As the number of pool players grows to infinity, the fraction of picks for each team approaches the win probability for that team.

In calculating the Nash equilibrium, Metrick used the pick distributions from the pools. But this detailed information about opponent play would typically not be available when the brackets were being filled out back in 1996. It would have been difficult to arrive at a detailed quantitative estimate of the payoff for betting a specific team. But the fact that 1 seeds would be over-bet could be determined from the pattern of play in past pools. By 2002, the large online bracket contests sponsored by Yahoo and ESPN started posting the pick distributions for all teams in those contests days before the submission deadline for brackets. This data provides a useful estimate of opponent play in office pools except for the hometown effect. If there is a hometown team in the tournament, then it tends to be over-bet relative to the large nationwide bracket contests.

Are Most Pool Players Irrational?
The actual pick distributions were far away from the rational profit-oriented equilibrium. Are pool players just irrational? One viewpoint in economics is that everyone is rational and that some find value in something other than monetary (or pecuniary) profit. Another viewpoint, going under the name “behavioral economics,” posits that we are indeed sometimes irrational in our decision making.

Much of standard economic theory is based on the notion that people are a collection of rational agents seeking to maximize profit. Sometimes the notion is just a simplifying assumption that allows progress in an abstract analysis. But it can also function as an invalid assumption in real-world decision making. Metrick’s paper on the bracket pool was part of an effort by economist and
psychologist to better understand the non-rational component of people's economic behavior, an effort that came into prominence in the latter half of the twentieth century. This effort is important because it can help improve the effectiveness of organizations and decision making and perhaps even help us find better ways to prevent the instability that on occasion plagues the whole economy, as in the severe 2008 recession.

Metrick also calculated a different kind of Nash equilibrium. For this calculation he assumed (for the purpose of the analysis) that all play was rational but that some pool players were getting some kind of non-pecuniary payoff, a payoff in something other than money. A player might simply enjoy correctly picking the correct champ even if they do not win the pool. He considered the case of a player betting for 1 seed Michigan instead of 2 seed Arizona in a 50-player pool with an entry fee of $5. This is one of the hardest decisions to justify because Michigan's win probability exceeded Arizona's only 1.3%, yet Arizona's expected payoff exceeded Michigan's by $11.55. He found that, under this assumption, the player must be getting almost $890 worth of non-pecuniary utility simply from picking the winner. (The $890 figure seems large since the pool pot is only $250. The large figure is due to the fact that the player is forgoing a significant amount of expected pecuniary payoff for a very small 1.3% chance of actually getting the additional non-pecuniary payoff of successfully picking the champion.) Of course, the enjoyment from picking the ultimate winner might not be the only benefit. But the $890 figure gives an idea of the magnitude of the seemingly non-optimal pool betting behavior that needs to be explained.

Figure 1.3 represents a diagram of Metrick's paper. The betting market futures and the competitor's brackets provide sufficient information for two different financial analyses, two different Nash equilibriums. One analysis yields a what-if picture of what rational financial behavior based on a pure dollars-and-cents profit motive would be. The other analysis yields a set of estimates for the value that each competitor is apparently assigning to their
There is a gap between the two. This gap is typically given the pro forma name “non-pecuniary payoff.” But the term is theory-laden. The term assumes an explanation based on the bettor’s value system and avoids the assumption that the bettor’s behavior is, at least to some extent, irrational. The gap is really just an unexplained gap. The dotted line boxes indicate some possible influences on the competitor’s behavior.

At one point in his paper, Metrick referred to players who push the pools toward the profit-oriented equilibrium as “smart,” seeming to imply that their opponents are dumb. That’s one theory that could help explain the gap. Three years earlier economist Richard Thaler referred to a large proportion of stock market investors as “blockheads” (Thaler 1993). Tendencies in economic behavior that appear to have no rational basis are referred to as behavioral anomalies. Some behavioral anomalies may exist due to the fact that our brains evolved in an environment where solving problems like filling out the best bracket or solving modern financial puzzles was not important.

Metrick provides a number of explanations of how seemingly suboptimal player behavior might be rational because it made

---

**FIGURE 1.3** Key concepts in Metrick’s paper.
playing the pool more enjoyable. The hometown effect (where Seton Hall gets a higher fraction of picks in New York City, and 6 seed Berkeley was picked by seven players in San Francisco and none elsewhere) might be due to players getting more enjoyment from betting for their favorite team. Also, players might pick objective favorites because they get enjoyment from remaining in contention as long as possible.

Don’t pick the hometown favorite for champ.

Racetracks provide explicit payoff odds. But the payoff odds are more obscure in the bracket pool. Even the raw data that could be used to calculate the payoff odds are unknown since each player’s bracket is not revealed to other players until after the deadline for making bets. The odds would have to be inferred from patterns observed from past years’ pools. It takes time and effort to do that. The cost of time and effort can be viewed as diminishing the potential profit.

Metrick pointed out that the profits come with the risk of loss. Arizona had only a 7.8% chance of winning. The player who routinely plays a strategy like that will get less than one win per decade on average even in a small pool with 25 players. Much of this financial risk could be diversified away if it were possible to play multiple entries in a pool. Metrick did not consider this, perhaps because it was against the pool rules or otherwise not allowed. Metrick did consider playing different pools, which provides even more effective diversification. Different champions could be bet in different pools. But Metrick pointed out that it was not possible to “poach” office pools all over the nation. Most office pools were closed to outsiders. A few players could not aggregate and capture all the potential profits. It was not feasible to create a bracket pool hedge fund in 1996. This is an important feature that keeps office bracket pool play from being forced to the conventional financial equilibrium by a few profit-oriented rational players.
Was the Simplified Problem Good Enough?

This is all premised on the idea that Metrick’s simplified rules are close enough to reality. The simplified rules assume that the winner of the pool always picks the correct champion. Metrick found that all 24 pools he analyzed would have been won by a UNC picker if UNC had won and the rules were the steep points-per-round progression of 1, 2, 4, 8, 16, 32 for rounds 1, 2, 3, 4, 5, 6, respectively. Even with the flatter, more gradual, points-per-round progression of 1, 2, 3, 4, 5, 6, 18 of 24 pools would have been won by a UNC picker. Among the 24 pools, eight of the pools use the steep progression, at least one used the flatter progression, and the median progression was 1, 2, 3, 5, 9, 15. So the assumption that the winner of the pool picks the champ often holds true, particularly for pools that award a relatively large number of points for correctly picking the champion.

Down-BracketChalk

To further defend the idea that picking the correct champion is usually key to winning, Metrick suggests a particular down-bracket strategy: he suggests that a player who correctly picks a contrarian champion could defend himself from getting overtaken by a player who scored more points in the earlier round games by simply picking objective favorites in all other games that were not constrained by his championship pick. Picking all favorites would maximize expected points in the earlier rounds. This makes it less likely that any player who picked a losing champion could overtake a player who picked the winning champion by winning enough points in the earlier rounds to compensate for the championship round points.

“Chalk” is a slang term for picking the favorites that originated in the days when bookies would write the horse racing odds in chalk on a blackboard. The odds on the favorites tended to have to be updated more often, so the spot where they were erased and rewritten showed more chalk dust. Hence, a “chalk horse” was a favorite. The first use of the term “chalk bracket” turned up
in 2008 when the outcome of the NCAA tournament was the closest it has ever been to a chalk bracket, with all four 1 seeds making it to the Final Four that year.

The goal of Metrick’s paper was not to give bracket advice, but the profit-oriented player can derive a likely favorable strategy from Metrick’s analysis: Pick a 2 seed that is not a hometown team for champion and go with the favorites everywhere else. Fill out a bracket that is 100% chalk except for advancing one team to the championship that is likely to be under-backed in your pool.

A down-bracket with no upsets provides you with at least an average chance of coming out ahead of opponents that pick the same champ as you picked.

Metrick proposed the all-chalk down-bracket merely as a simplifying assumption. He assumed it was average versus opposing down-brackets. More recent analysis has shown that an all-chalk down-bracket is typically above average. Hence, in situations where Metrick’s analysis recommends a 2 seed for champ, more recent methods may find that a less contrarian 1 seed provides the higher return on investment.

OBAMA’S BRACKETS

Andrew Metrick served on President Obama’s Council of Economic Advisors from 2009 through 2010. But it’s clear that Obama did not use advice from Metrick on how to fill out a bracket. Obama’s champion picks were too chalky and his down-bracket picks were not chalky enough. When it comes to filling out a bracket, Obama is more of an everyman than a strategic player who is optimizing his chances of winning.

Obama is an avid basketball player and fan. He filled out a bracket in his first March in office, and he assented to ESPN’s request to make this a public event. Predictably, Obama got political flack from a number of quarters for filling out a bracket. Duke’s coach, Mike Krzyzewski, had this to say about
Obama’s 2009 bracket picks: “Somebody said that we’re not in President Obama’s Final Four, and as much as I respect what he’s doing, really, the economy is something that he should focus on, probably more than the brackets” (Fox News 2009). Obama had Duke, a 2 seed, playing to seed expectations, but Obama also picked Duke’s archrival, UNC, to win the tournament. Perhaps Krzyzewsky, known to be a shrewd coach, was trying to fire up his team. Obama was criticized for ignoring the women’s bracket in 2009. There were complaints about the male-dominated work environment at the White House and the all-male pick-up games on the White House basketball court. Obama filled out a women’s bracket every year after 2009 (Wolff 2016). Figure 1.4 shows Obama’s 2015 bracket for the women’s tournament.

According to the manager of the White House bracket pool, the pool had hundreds of entries (Tracy 2017). Every year, Obama played one of the top two most popular 1 seeds for champion.

FIGURE 1.4 President Obama’s 2015 Women’s bracket (Wall 2015). Licensed under CC BY 3.0.
Even if one of his champion picks won (as UNC did in 2009) he would have faced lots of competition. The buzz in the media was that his down-bracket picks were relatively chalky, but not chalky enough by Metrick’s standards. He picked 52 seed upsets in the first round and 47 seed upsets in the later rounds over eight years of men’s tournament brackets. He had good results on his first-round upset picks, going 28–24. But in the later rounds, where correct picks typically are rewarded more points, his upset picks hurt his brackets; He went 13 for 47 (Mather 2017). Overall, he would have improved his chances of winning if he had just picked the higher seed in all of his men’s brackets. In the women’s brackets, Obama had a fondness for Princeton in the later years; his niece Leslie Robertson was a starter on the team. He advanced 8 seed Princeton to the Final Four in 2015, but they played to seed expectation and lost to 1 seed Maryland in the second round.

However, the White House “pool” did not require an entrance fee, probably as a nod to federal regulations against gambling in the workplace. It was a contest for bragging rights, not really a betting pool at all. So, all payoffs were non-pecuniary. And Obama surely got some enjoyment from some of his upset picks. One of his riskiest picks was Hawaii to win the first round of the 2016 men’s tournament. Obama grew up in Hawaii and he got to go to one of their games with his grandfather when he was age 10, so they were a sentimental favorite of his. And his pick panned out: 13 seed Hawaii beat 4 seed California 77–66. Also, even if Obama had optimized his brackets, it’s still more likely than not that he would have never won any of the eight pools he played as president. In a pool of size 200, a middling player will win once every 200 years on average. An optimized bracket that is five times better than average will still only average one win every 40 years.

METRICK’S IMPACT
There is every indication that word did not get around quickly about Metrick’s analysis. The pundits in the media kept advising
pool players to pick lots of upsets in the early rounds and to pick a champion with no thought about their opponent’s picks.

In 2005, the *New York Times* published a piece that referenced Metrick’s findings that the No. 1 seeds are over-bet (Leonhardt 2005). Since 2005, the notion of betting a contrarian champion while picking few or no upsets in the early rounds has gotten somewhat more attention.

In 2015, Metrick gave a rare interview on the bracket pool to the *Wall Street Journal* saying, “It often pays to be contrarian, you want to take good bets that others don’t want: stocks that are out of style for behavioral reasons and basketball teams that are good but may not have the most fans” (Cohen 2015).

The *Journal* even managed to get a quote on the bracket pool from John Nash. In an email, Nash wrote “I guess I am myself pretty far from being the sort of person who would ‘fill out an NCAA tournament bracket’.”
IMPROVING THE BIG EAST CONFERENCE BASKETBALL TOURNAMENT

This article is excerpted from *The American Statistician* by Christopher Abdul-Chani and Jesse Frey. © 2016 Taylor & Francis Group. All rights reserved.

Learn more
ABSTRACT
The Big East Conference basketball tournament is a four-day, 10-team, knockout tournament that is used to decide which team receives the conference's automatic bid to the NCAA basketball tournament. Through data-based modeling, we show that the current tournament format is not very effective in determining the true best team. Specifically, by considering a variety of alternate formats, we find that certain formats that exclude all but a handful of teams substantially outperform the current format in determining the true best team. We also find that among formats that involve all ten teams, a format in which the top two seeds each receive two byes is relatively effective. We show that our conclusions are robust to several key modeling choices. We also investigate the effectiveness of the tie-breaking scheme used by the Big East Conference, finding that it is little better than random and may even favor weaker teams.

1. Introduction
The NCAA men's basketball tournament, a 68-team knockout tournament to determine a national champion, is one of the most anticipated events on the American sports calendar. It takes place over a 3-week span in late March and early April each year. Before the NCAA tournament, however, most of the individual NCAA basketball conferences play conference tournaments. Each conference tournament champion receives an automatic bid to the NCAA tournament, and those teams are joined in the field of 68 by a group of at-large teams, nominally the best teams that did not receive automatic bids, that are chosen by a committee. The one conference that does not send a tournament champion to the NCAA tournament is the Ivy League, which does not play a tournament and instead awards an automatic bid to its regular season champion.

The Big East Conference, though much newer than conferences such as the Big Ten, the Atlantic Coast Conference, and the Ivy League, has been among the most successful college basketball conferences over the past 35 years, winning seven national championships over that time period. In its current incarnation, the conference has ten teams, and the conference regular season consists of each team playing each other team twice, once at home and once on the road, for a total of 18 in-conference games. After the regular season, the teams are seeded from 1 (best) to 10 (worst) according to their number of in-conference wins. The tournament takes place over four days, with the first day featuring one game between seeds 7 and 10 and another between seeds 8 and 9. The second day then includes four games such that, if all goes according to seed on the first day, teams with seeds that sum to nine play each other, and so on. In other words, the tournament is played according to virtually the same format that is used for a 16-team NCAA regional, the one difference being that seeds 1–6 receive byes to the second round.

There is a substantial literature on tournament design, with early statistical contributions including Glenn (1960) and Sears (1963). A variety of types of tournaments have been studied, and a variety of goals have been considered. One might be interested in designing a tournament to maximize the entertainment value. One might also be interested in satisfying certain notions of fairness as in Hwang (1982) and Schwenk (2000). We focus here on knockout tournaments, and we judge the quality of a particular tournament format by looking at the degree to which the tournament succeeds in determining the true best team. Annis and Wu (2006) pursued a similar goal in the context of college football, as did Glickman (2008) for college basketball. Other recent studies of tournament design in sports include McGarry and Schutz (1997), Scarf, Yusof, and Bilbao (2009), and Baumann, Mathe- son, and Howe (2010). In this article, we compare the current format of the Big East Conference tournament to other possible formats. We use data from recent seasons to create a model that produces results that are consistent with past results, and we then evaluate different potential tournament formats by simulating seasons under the model and estimating the chance that the true best team emerges as the tournament champion.

One assumption that we make in what follows is that a team's strength does not change within a season. Obviously, this assumption cannot hold exactly, as team strength may decline if a key player is injured or improve as young players gain experience. Still, given that the conference season is less than three months long, the assumption may be a good approximation.

We develop our model for team strengths and the home court advantage in Section 2. In Section 3, we compare the performance of two types of tournaments, namely those that include a fixed number of teams and those that include all teams with conference win totals within a certain distance of the highest win
total. In Section 4, we compare potential tournament formats that involve all ten teams. In Section 5, we assess the robustness of our conclusions in Sections 3 and 4 by examining how changing the model setup in various ways would affect our results. In Section 6, we use our model from Section 2 to study the effectiveness of the tie-breaking scheme currently used in the Big East Conference, and in Section 7, we conclude with a discussion.

2. Modeling Team Strengths and the Home-Court Advantage

To make meaningful comparisons between potential formats for the Big East Conference tournament, we need a good model from which to simulate regular season and tournament results in the conference. Several elements are clearly essential for success of the model. We need a model that accurately captures the extent of the home court advantage within the conference. We also need a model such that the distribution of wins closely matches the empirical distribution of wins in recent conference seasons. In particular, the model distribution for wins should match the empirical distribution closely in terms of shape and spread.

To get a sense of the size of the home court advantage and the distribution of wins within the conference, we obtained data on the past eight Big East Conference seasons from Sports Reference, LLC (2015). Data from earlier seasons are also available, but the number of conference games played by each team was smaller in those earlier years. The number of teams in the conference has varied over the eight seasons, but the number of conference regular-season games for each team was fixed at 18 throughout the period. Thus, the average number of wins per team was nine for each of the eight seasons. The standard deviation has varied over the eight seasons, but the number of conference games played by each team was consistent over the eight years. Thus, it is reasonable to seek a model where the standard deviation of team wins is close to the overall empirical standard deviation of 3.925 wins and where the percentage of games won by the home team is close to the overall empirical percentage of 60.8%.

We obtained $\sigma$ and $H$ in stages. When team 1 hosts team 2, the chance that the home team wins is $\Phi(\theta_1 + H - \theta_2)$. We also have that team strengths are independently distributed $N(0, \sigma^2)$ random variables. Thus, for fixed $\sigma$ and $H$, the proportion of games won by the home team is $E[\Phi(\theta_1 + H - \theta_2)]$, which, since $\theta_1 + H - \theta_2 \sim N(H, 2\sigma^2)$, can be written as $F(H|\sigma) \equiv E[\Phi(H + \sqrt{2}\sigma Z)]$, where $Z \sim N(0, 1)$. For a given $H$, we can compute $F(H|\sigma)$ to any desired accuracy via numerical integration. Then, using the fact that $F(H|\sigma)$ is an increasing function of $H$, we can use the bisection root-finding algorithm to find an $H$ such that $F(H|\sigma) \approx 0.608$. This observation reduces our task to that of choosing a value for $\sigma$.

To choose $\sigma$, we used simulation. For a given value of $\sigma$, we found an $H$ such that $F(H|\sigma) \approx 0.608$. We then, for each choice of $\sigma$, simulated many different independent sets of season win totals by (i) generating independent team strengths $\theta_1, \ldots, \theta_{10} \sim N(0, \sigma^2)$, (ii) playing a season in which each team plays each other team once at home and once on the road, and then (iii) computing the number of wins for each team. We then computed the standard deviation of wins over all teams and all seasons that we simulated using a given $\sigma$. Some results are shown in Figure 2.

Figure 2 plots the simulated standard deviation of yearly team wins against $\sigma$ both for the case where the home court advantage parameter $H$ is included (the solid curve) and for the case where $H$ is not included (the dashed curve). Values were simulated for $\sigma$ ranging from 0.55 to 0.75 in steps of 0.02, with

![Figure 1. Histogram of yearly team win totals in conference games for all Big East teams during the years 2008–2015.](image)

![Figure 2. Distribution of wins within the conference, we obtained data from earlier seasons are also available, but the number of conference games played by each team was smaller in those earlier years. The number of teams in the conference has varied over the eight seasons, but the number of conference regular-season games for each team was fixed at 18 throughout the period. Thus, the average number of wins per team was nine for each of the eight seasons. The standard deviation has varied over the eight seasons, but the number of conference games played by each team was consistent over the eight years. Thus, it is reasonable to seek a model where the standard deviation of team wins is close to the overall empirical standard deviation of 3.925 wins and where the percentage of games won by the home team is close to the overall empirical percentage of 60.8%.

We obtained $\sigma$ and $H$ in stages. When team 1 hosts team 2, the chance that the home team wins is $\Phi(\theta_1 + H - \theta_2)$. We also have that team strengths are independently distributed $N(0, \sigma^2)$ random variables. Thus, for fixed $\sigma$ and $H$, the proportion of games won by the home team is $E[\Phi(\theta_1 + H - \theta_2)]$, which, since $\theta_1 + H - \theta_2 \sim N(H, 2\sigma^2)$, can be written as $F(H|\sigma) \equiv E[\Phi(H + \sqrt{2}\sigma Z)]$, where $Z \sim N(0, 1)$. For a given $H$, we can compute $F(H|\sigma)$ to any desired accuracy via numerical integration. Then, using the fact that $F(H|\sigma)$ is an increasing function of $H$, we can use the bisection root-finding algorithm to find an $H$ such that $F(H|\sigma) \approx 0.608$. This observation reduces our task to that of choosing a value for $\sigma$.

To choose $\sigma$, we used simulation. For a given value of $\sigma$, we found an $H$ such that $F(H|\sigma) \approx 0.608$. We then, for each choice of $\sigma$, simulated many different independent sets of season win totals by (i) generating independent team strengths $\theta_1, \ldots, \theta_{10} \sim N(0, \sigma^2)$, (ii) playing a season in which each team plays each other team once at home and once on the road, and then (iii) computing the number of wins for each team. We then computed the standard deviation of wins over all teams and all seasons that we simulated using a given $\sigma$. Some results are shown in Figure 2.

Figure 2 plots the simulated standard deviation of yearly team wins against $\sigma$ both for the case where the home court advantage parameter $H$ is included (the solid curve) and for the case where $H$ is not included (the dashed curve). Values were simulated for $\sigma$ ranging from 0.55 to 0.75 in steps of 0.02, with

![Figure 1. Histogram of yearly team win totals in conference games for all Big East teams during the years 2008–2015.](image)
Figure 2. Plot of the standard deviation of team wins against $\sigma$ both with (solid curve) and without (dashed curve) a home court advantage. The home court advantage is chosen so that home teams win 60.8% of games on average. The horizontal reference line is at 3.925 wins.

![Plot of the standard deviation of team wins against $\sigma$.](image)

Figure 3. Histogram of team win totals for 100,000 10-team seasons simulated from the normal model.

![Histogram of team win totals.](image)

Figure 4. Empirical and model-based average numbers of home conference wins as a function of the total number of conference wins. The empirical values are plotted as individual points, and the model-based values are shown as a curve.

![Empirical and model-based average numbers of home conference wins.](image)

Each value being based on 100,000 simulated ten-team seasons. We see from the figure that including the home court advantage parameter $H$ makes the win totals less variable. We also see that to get a standard deviation for wins that matches the empirical value of 3.925 when including the home court advantage, we need $\sigma \approx 0.65$. This leads to $H \approx 0.37$.

As a check on our model fitting, we generated many ten-team seasons using the fitted model and then compared the distribution of team wins to the empirical distribution for the years 2008–2015. Figure 3 gives a histogram of the season win totals for 100,000 10-team seasons simulated using $\sigma = 0.65$ and $H = 0.37$. We see from the figure that, as intended, the distribution of wins is symmetric about the mean of nine. We also see that the shape and spread are both comparable to what we see in Figure 1 for the years 2008–2015.

Our model uses a single home court advantage parameter for all teams. In Section 5, we assess how sensitive our conclusions are to this assumption by examining what happens when we assign home court advantage parameters at random, thus making the home court advantage bigger for some teams than for others. Another potentially problematic possibility is that the home court advantage might be bigger for strong teams than for weak teams (or vice versa). To show that this is not in fact the case, we compared two sets of averages. First, for each possible number of conference wins from 0 to 18, we computed the average number of home wins for Big East teams with that number of conference wins during the years 2008–2015. Second, using the same 100,000 seasons that we used to obtain Figure 3, we computed the simulated average number of home wins for teams with each possible number of conference wins. The results are shown in Figure 4, where both the empirical average numbers of home wins (the points) and the average numbers of home wins from the model (the solid curve) are plotted against the total number of conference wins. Figure 4 shows that there is no systematic difference between the two sets of averages.

3. A Comparison of Tournament Formats

A variety of possible formats are available, most of which require that the teams be seeded from 1 to 10. We assume that the 10 teams are seeded from 1 (best) to 10 (worst) according to their number of regular season wins (out of 18), with ties broken at random. The Big East Conference actually has an elaborate tie-breaking procedure (Big East Conference 2015) that starts with looking at head-to-head records in the mini-conference created by considering only games between teams from a given set of tied teams. However, we find in Section 6 that this tie-breaking scheme is little different from breaking ties at random. Thus, we assume in the next few sections that ties were broken at random.

The first type of tournament format that we considered is a tournament with a fixed number of teams $T$, where $T$ is an integer between 1 and 10 inclusive. Regardless of how many teams are included, the games actually played are exactly those that would be played in one of the 16-team NCAA tournament regionals, the one difference being that games that would have involved teams with seeds higher than $T$ are not played. For example, if $T = 7$, then the first day of play features games between seeds 2 and 7, 3 and 6, and 4 and 5. On the second day, seed 1 plays the winner of seeds 4 and 5, and the winner of seeds 2 and 7 plays the winner of seeds 3 and 6. The winners of those two games then play on the third day to determine the tournament champion. Thus, $T = 10$ corresponds to the current tournament format, and $T = 1$ corresponds to playing no games and simply declaring the top-seeded team the tournament champion.

A second type of tournament format that we considered is a tournament in which the number of teams included is random. Specifically, the tournament includes all teams with a number of wins that is no more than $M$ wins smaller than the maximum win total, and the tournament structure, conditional on
The left side of each table shows results when reseeding is not used, and the right side shows results when reseeding is used.

Each tournament format was used on one million independent seasons, and the probabilities in the table are the empirical probabilities that the tournament champion was the true best team or among the true top 2, 3, 4, or 5 teams. In running each tournament, we assumed that all games were played at neutral sites. Thus, the parameter \( H \) was used only when simulating regular season games to determine tournament seeding.

Table 2 shows that the best fixed-size tournament format is the \( T = 1 \) format in which the top seed is simply declared the champion. It gives nearly a 63% chance that the true best team wins the tournament. The next best fixed-size formats are \( T = 2 \) and \( T = 3 \), and the main factor driving how the formats compare seems to be the number of wins needed for the top seed to become the champion. Using \( T = 2 \) rather than \( T = 1 \) means that more wins are required from the top seed, and the chance that the best team wins goes down. Similar drops appear when \( T \) goes from 3 to 4 and from 7 to 8. For the fixed-size tournament formats with \( T \geq 6 \), reseeding seems to have a small positive effect on the chance that the true best team wins. For the more effective \( T \leq 5 \) designs, however, reseeding does not come up, and the left and right side formats are exactly equivalent. The current \( T = 10 \) tournament is among the worst of the choices shown in Table 2, giving the best team only a 47% chance of winning the tournament.

Table 3 shows that among the random-size tournament formats, the best formats are \( M = 0 \) and \( M = 1 \), each of which gives the true best team a roughly 65% chance of winning the

<table>
<thead>
<tr>
<th>( M )</th>
<th>( w/o ) reseeding</th>
<th>( w/ ) reseeding</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>0.649, 0.857, 0.941, 0.976, 0.991</td>
<td>0.649, 0.857, 0.941, 0.976, 0.991</td>
</tr>
<tr>
<td>( 1 )</td>
<td>0.649, 0.859, 0.943, 0.977, 0.992</td>
<td>0.650, 0.859, 0.942, 0.977, 0.992</td>
</tr>
<tr>
<td>( 2 )</td>
<td>0.626, 0.844, 0.934, 0.973, 0.990</td>
<td>0.626, 0.844, 0.934, 0.973, 0.990</td>
</tr>
<tr>
<td>( 3 )</td>
<td>0.590, 0.818, 0.919, 0.966, 0.986</td>
<td>0.590, 0.818, 0.919, 0.966, 0.987</td>
</tr>
<tr>
<td>( 4 )</td>
<td>0.555, 0.791, 0.902, 0.956, 0.982</td>
<td>0.556, 0.791, 0.902, 0.956, 0.982</td>
</tr>
<tr>
<td>( 5 )</td>
<td>0.528, 0.766, 0.884, 0.946, 0.977</td>
<td>0.528, 0.767, 0.885, 0.946, 0.977</td>
</tr>
<tr>
<td>( 6 )</td>
<td>0.508, 0.747, 0.869, 0.936, 0.970</td>
<td>0.509, 0.747, 0.870, 0.937, 0.971</td>
</tr>
<tr>
<td>( 7 )</td>
<td>0.496, 0.731, 0.856, 0.926, 0.964</td>
<td>0.498, 0.734, 0.858, 0.928, 0.965</td>
</tr>
<tr>
<td>( 8 )</td>
<td>0.488, 0.721, 0.846, 0.918, 0.958</td>
<td>0.493, 0.726, 0.851, 0.922, 0.960</td>
</tr>
<tr>
<td>( 9 )</td>
<td>0.483, 0.713, 0.838, 0.912, 0.954</td>
<td>0.488, 0.719, 0.843, 0.916, 0.957</td>
</tr>
</tbody>
</table>
tournament. The chance that the best team wins the tournament gets progressively smaller with bigger \( M \). Reseeding seems to improve the performance of the formats where \( M \geq 6 \), but offers no noticeable gain for the more effective \( M = 0 \) and \( M = 1 \) formats.

Overall, the best choices from among the formats we have discussed seem to be the \( M = 0 \) and \( M = 1 \) formats. The \( T = 1 \) format is also effective, but it presumably loses out to the \( M = 0 \) format because it involves breaking ties at random rather than by playing additional games.

4. Tournaments that Involve All Teams

Our results suggest that having the tournament involve only a small fraction of the teams in the conference gives the best chance of having the true best team win the tournament. However, having the tournament exclude most teams is unlikely to be an attractive prospect for fans, players, coaches, and broadcasters. Thus, we also consider which format might be best if the tournament must include all ten teams. There are several available alternatives to the current format.

One possibility, due to Marchand (2002), is to have all teams participate, but to assign the seeds completely at random. A second possibility, due to Schwenk (2000), is to use cohort seeding. Teams are split into cohorts of prespecified sizes using the current seeding method, and teams in the same cohort are seeded at random. With a 10-team conference, it makes sense to have four cohorts, one consisting of the top two teams, one of the next two teams, one of the next two teams, and one of the bottom four. The first cohort contains the top two teams, who are randomly assigned to seeds one and two and thus ensured of not facing each other until the final game. The second cohort contains the next two teams, who are randomly assigned to seeds three and four and thus ensured of not facing any other top-four teams until the semi-finals, and so on.

Two other possibilities involve modifying the current bye structure. If the tournament is to be completed in four days, with no team playing more than once per day, then there are only a handful of possible bye structures. The current format gives the top six seeds byes to the second day, with the remaining four teams playing for two quarterfinal spots. One could also give the top seed a bye to the third day and the next three seeds a bye to the second day. Finally, one could give the top two seeds byes to the third day, with the other eight teams playing on the first day. We write these tournament structures as \( B(6, 0) \), \( B(3, 1) \), and \( B(0, 2) \), where the two entries indicate the number of seeds receiving one and two byes, respectively. Figure 5 shows the bracket structure for the \( B(3, 1) \) and \( B(0, 2) \) formats.

We compared the current format, the random format of Marchand (2002), the cohort seeding format, the current format with reseeding, and the \( B(3, 1) \) and \( B(0, 2) \) formats using the same method and the same number of simulated seasons (one million) that we used in Section 3. The results are given in Table 4. We see from the table that while the current format is more effective than the random format and comparable in effectiveness to the cohort seeding format, it is less effective than either the \( B(3, 1) \) or the \( B(0, 2) \) format in determining the best team. The most effective of the six formats is the \( B(0, 2) \) format, which would thus be the recommended choice.

5. A Sensitivity Analysis

The specific results that we obtained in Sections 3 and 4 depend on the choices that we made in fitting the model. These include the choice of the win probability formula \( \Phi(\theta_i - \theta_j) \), the choice of \( \sigma \), and the choice of the distribution for the team strengths. There are also certain home court advantage issues that we have not yet addressed. First, we assumed that the home court advantage is the same for all teams, but in fact some teams may have bigger advantages than others. Second, the tournament is played in Madison Square Garden, which also serves as the home court for selected regular-season games played by conference member St. John’s University. Thus, St. John’s should arguably be given the home court advantage when we simulate the tournament.

To examine whether our results are sensitive to our choice to use the win probability formula \( \Phi(\theta_i - \theta_j) \) for simulating game results, we reid our work from Sections 2 and 3 using the alternate win probability formula \( 1/(1 + \exp(\theta_j - \theta_i)) \), which matches the formula used in the Bradley and Terry (1952) model for paired comparisons. We continued to take the team strengths \( \theta_i \) and \( \theta_j \) to be \( N(0, \sigma^2) \), but the change in win probability formula required that we modify our choices for \( \sigma \) and \( H \). Using exactly the strategy that we used in Section 2 for the earlier probability formula, we obtained estimates \( \sigma \approx 1.11 \) and \( H \approx 0.63 \) for the new formula. We then compared the various fixed-size and random-size tournament formats as in Section 3.

To assess the impact of changing \( \sigma \), we considered letting \( \sigma \) be either 0.55 or 0.75 rather than 0.65. After changing \( \sigma \), we used either \( H = 0.35 \) (with \( \sigma = 0.55 \)) or \( H = 0.40 \) (with \( \sigma = 0.75 \)) to ensure that home teams would continue to win 60.8% of games on average.

Table 4. Estimated chance that the winning team is the true best team or among the true top \( N \) teams for \( N = 2, \ldots, 5 \), given as a function of the tournament design. Win probabilities \( \Phi(\theta_i - \theta_j) \) are used. Simulation standard errors: 0.0005 or less.

<table>
<thead>
<tr>
<th>Method</th>
<th>Best</th>
<th>Top 2</th>
<th>Top 3</th>
<th>Top 4</th>
<th>Top 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>0.473</td>
<td>0.703</td>
<td>0.831</td>
<td>0.907</td>
<td>0.951</td>
</tr>
<tr>
<td>Random</td>
<td>0.443</td>
<td>0.660</td>
<td>0.789</td>
<td>0.872</td>
<td>0.925</td>
</tr>
<tr>
<td>Cohort</td>
<td>0.472</td>
<td>0.703</td>
<td>0.830</td>
<td>0.907</td>
<td>0.951</td>
</tr>
<tr>
<td>Reseeding</td>
<td>0.476</td>
<td>0.705</td>
<td>0.832</td>
<td>0.907</td>
<td>0.951</td>
</tr>
<tr>
<td>( B(3, 1) )</td>
<td>0.501</td>
<td>0.731</td>
<td>0.855</td>
<td>0.924</td>
<td>0.962</td>
</tr>
<tr>
<td>( B(0, 2) )</td>
<td>0.509</td>
<td>0.748</td>
<td>0.869</td>
<td>0.934</td>
<td>0.968</td>
</tr>
</tbody>
</table>
To assess the impact of having the home-court advantage vary from one team to another, we let the home-court advantage for a particular season be a random draw \( H_i \) from the uniform distribution on the interval \([-0.63, 0.63]\) rather than simply a constant \( H = 0.65\). In this case, how teams rank at home may differ from how they rank on a neutral court, but we continued to view the true best team as the team with the best chance of winning on a neutral court. To assess the impact of playing the tournament on the home court for St. John’s University, we simulated tournaments in which all other model features matched the model from Section 2, but one randomly selected team was given the home court advantage for the entire tournament. Simulation results both for the model from Section 2 and for the seven modified models are given in Table 5.

![Histogram of team win totals for 100,000 10-team seasons simulated using the skew-normal model. The shape parameter \( \lambda = -3\) was used.](image)

Table 5. Estimated chance that the true best team wins the tournament, given as a function of the tournament design and the way in which the model differs from the model fit in Section 2. The seven right-most columns correspond to (i) using alternate win probabilities \( 1/(1 + \exp(\theta_i - \theta_j)) \), (ii) using \( \sigma = 0.55 \), (iii) using \( \sigma = 0.75 \), (iv) using a skew-normal distribution for team strengths, (v) using a uniform distribution for team strengths, (vi) letting the home court advantage be bigger for some teams than for others, and (vii) giving one randomly selected team the home court advantage in the tournament. Simulation standard errors: 0.0005 or less.

<table>
<thead>
<tr>
<th>Format</th>
<th>Original</th>
<th>B-T</th>
<th>Smaller ( \sigma )</th>
<th>Larger ( \sigma )</th>
<th>Skew-Normal</th>
<th>Uniform</th>
<th>Random ( H )</th>
<th>Home court</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T = 1 )</td>
<td>0.629</td>
<td>0.630</td>
<td>0.590</td>
<td>0.657</td>
<td>0.556</td>
<td>0.525</td>
<td>0.618</td>
<td>0.629</td>
</tr>
<tr>
<td>( T = 2 )</td>
<td>0.573</td>
<td>0.579</td>
<td>0.539</td>
<td>0.601</td>
<td>0.510</td>
<td>0.485</td>
<td>0.567</td>
<td>0.572</td>
</tr>
<tr>
<td>( T = 3 )</td>
<td>0.572</td>
<td>0.578</td>
<td>0.537</td>
<td>0.600</td>
<td>0.507</td>
<td>0.483</td>
<td>0.565</td>
<td>0.569</td>
</tr>
<tr>
<td>( T = 4 )</td>
<td>0.503</td>
<td>0.510</td>
<td>0.467</td>
<td>0.536</td>
<td>0.441</td>
<td>0.423</td>
<td>0.498</td>
<td>0.498</td>
</tr>
<tr>
<td>( T = 5 )</td>
<td>0.503</td>
<td>0.510</td>
<td>0.467</td>
<td>0.536</td>
<td>0.440</td>
<td>0.422</td>
<td>0.498</td>
<td>0.498</td>
</tr>
<tr>
<td>( M = 0 )</td>
<td>0.649</td>
<td>0.651</td>
<td>0.611</td>
<td>0.679</td>
<td>0.577</td>
<td>0.544</td>
<td>0.641</td>
<td>0.650</td>
</tr>
<tr>
<td>( M = 1 )</td>
<td>0.649</td>
<td>0.653</td>
<td>0.611</td>
<td>0.679</td>
<td>0.577</td>
<td>0.544</td>
<td>0.641</td>
<td>0.650</td>
</tr>
<tr>
<td>( M = 2 )</td>
<td>0.626</td>
<td>0.630</td>
<td>0.589</td>
<td>0.655</td>
<td>0.553</td>
<td>0.521</td>
<td>0.618</td>
<td>0.624</td>
</tr>
<tr>
<td>( M = 3 )</td>
<td>0.590</td>
<td>0.596</td>
<td>0.552</td>
<td>0.619</td>
<td>0.517</td>
<td>0.491</td>
<td>0.583</td>
<td>0.588</td>
</tr>
<tr>
<td>( M = 4 )</td>
<td>0.555</td>
<td>0.562</td>
<td>0.517</td>
<td>0.586</td>
<td>0.482</td>
<td>0.461</td>
<td>0.550</td>
<td>0.552</td>
</tr>
</tbody>
</table>

6. Does the Big East Use an Appropriate Tie-Breaker?

Our model from Section 2 puts us in position to evaluate the effectiveness of the current tie-breaking system used by the Big East Conference. This system, detailed in Big East Conference (2015), involves multiple steps, the first of which is to take each set of teams that are tied, create a mini-conference that includes just those teams, and then to break the tie by using the number of wins inside the mini-conference. Thus, if two teams are tied for first, the first tie-breaker is head-to-head competition. If one team won both games, then that team would get the top seed.

To assess the effectiveness of this first step of the tie-breaker, we simulated 10 million seasons using our model from Section 2. Among the 10 million simulated seasons, there were roughly 1.5 million in which there was a two-way tie for most wins. In roughly one-third of these seasons, the first step of the tie-breaker broke the tie, with the weaker team (as measured by team strength) winning 262,368 of the tie-breakers and the stronger team only 250,280. Thus, the estimated chance that the weaker team is favored is 51.2%, with standard error 0.07%. This finding suggests that the tie-breaker is essentially random and that to the extent that it differs from complete randomness, it actually favors the weaker team.

A similar result holds for three-way ties. Among the 10 million simulated seasons, there were roughly 230,000 that ended with a three-way tie for most conference wins. In just under two thirds of these seasons, the first step of the tie-breaker broke the tie, with the weakest team winning 51,758 of the ties, the middle...
team winning 51,204, and the strongest team winning 49,681. These yield estimated success probabilities 33.9%, 33.5%, and 32.5% for the weakest, middle, and strongest teams, each with standard error 0.12%.

One way to make the failure of the tie-breaker more intuitive is to compare the tie-breaker with its exact opposite. The current tie-breaker system gives the better seed to the team with the better head-to-head record. Doing the exact opposite would amount to giving the better seed to the team with the best record against the teams not involved in the tie, which also makes sense. Thinking about the tie-breaker in this way raises the question of whether the size of the conference matters. The bigger the conference, the more non-head-to-head games each team plays, and this could conceivably affect the performance of the tie-breaker.

To study this question, we repeated our simulation experiment for conferences with even numbers of teams from 4 to 12, each time assuming that each team plays each other team once at home and once on the road. The results are given in the first three columns of Table 6. We see from the table that the probabilities are all near 51%, but that there is a tendency for the chance that the weaker team wins the tie-breaker to be higher when the number of teams is higher.

To test the sensitivity of our conclusions here to the win probabilities that we used, we ran the same experiment using the Bradley–Terry win probabilities that we described in Section 5. The results are given on the right side of Table 6. We see from the table that with the modified win probabilities, the chance that the weaker team wins the tie-breaker is indistinguishable from 50%. Thus, we do not have clear evidence that the tie-breaker favors weaker teams, but it does seem clear that the tie-breaker is little better than random.

### Table 6: Estimated chance that the weaker of two tied teams wins the first step of the tie-breaker, given as a function of the number of teams in the conference and the type of win probabilities used. Each value is based on 10 million simulated seasons. Simulation standard errors are also given.

<table>
<thead>
<tr>
<th>Number of teams</th>
<th>Normal Win probability</th>
<th>Normal SE</th>
<th>Bradley–Terry Win probability</th>
<th>Bradley–Terry SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.5071</td>
<td>0.0009</td>
<td>0.5006</td>
<td>0.0009</td>
</tr>
<tr>
<td>6</td>
<td>0.5082</td>
<td>0.0007</td>
<td>0.4990</td>
<td>0.0007</td>
</tr>
<tr>
<td>8</td>
<td>0.5119</td>
<td>0.0007</td>
<td>0.5013</td>
<td>0.0007</td>
</tr>
<tr>
<td>10</td>
<td>0.5115</td>
<td>0.0007</td>
<td>0.4990</td>
<td>0.0007</td>
</tr>
<tr>
<td>12</td>
<td>0.5129</td>
<td>0.0007</td>
<td>0.4998</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

7. Conclusions

We have shown through modeling and simulation that the current format for the Big East Conference tournament is not very effective in determining the true best team. Switching to either the $M = 0$ format or the $M = 1$ format would considerably increase the chance that the true best team wins, but would perhaps make the tournament less attractive for fans, players, coaches, and broadcasters. If the conference prefers to stay with a 10-team format, then the $B(0, 2)$ format would be a good choice.

We have also shown that the current tie-breaker procedure is ineffective. One possible alternative would be to bring in new information, breaking ties by looking at how teams rank based on their full schedule of in-conference and out-of-conference games. In support of this idea, we present Figure 7, computed based on one million simulated seasons, which shows the distribution of the probability that the stronger team would beat the weaker team on a neutral court when exactly two teams are tied for first under the model from Section 2. We see from the figure that even though the two teams are tied, their strengths are frequently quite different. This difference in strength is likely to be evident from the performance of the two teams in nonconference games.

Our results here, which suggest that keeping the tournament very small maximizes the chance that the true best team will win the tournament, differ from those of Annis and Wu (2006), who found that a tournament with eight teams offered the best chance of identifying the true best team in college football. The difference here can presumably be attributed to a difference in information. In college football, there are many teams, relatively few games per team, and highly unbalanced schedules. In the current Big East Conference, there are few teams, more games, and a perfectly balanced schedule. Thus, regular season win totals offer substantial information about which team is the best, and a smaller tournament is effective in identifying the best team.

Though our results here are specific to the Big East Conference basketball tournament, the general strategy that we have used is applicable more broadly. It can be used to compare tournament formats at other levels and in other sports and nonsport settings. Within NCAA basketball, there are other major conferences. The 10-team Big Twelve Conference uses the same tournament format as the Big East Conference. The 12-team PAC-12 Conference awards first-round byes to four teams, which is analogous to what the Big East Conference does with 10 teams. The 14-team Big Ten, Southeastern, and Atlantic Coast Conferences each award two byes to four teams and one bye to six additional teams. This awarding of multiple byes to the top teams makes their tournament format similar to the recommended $B(0, 2)$ format from Section 4 and thus likely more effective than the 14-team analog of the current Big East format.
CHAPTER 2

Statistics and Healthcare
A Note From the Editor

It is estimated that in the U.S., 3 – 10% of national healthcare spending is lost each year to fraud, waste and abuse. This chapter discusses how statistical and analytical methods can be used in health care fraud detection.

*Discovery of New Fraud Patterns* excerpted from *Statistics and Health Care Fraud* by Tahir Ekin, published by the ASA and CRC Press, 2019.

The global average health care fraud loss is estimated to be 6% of the overall health care spending or approximately $450 billion. This chapter discusses the use of unsupervised methods - outlier detection, clustering and association methods to discover new fraud patterns.


The authors propose a simple, but effective, tool to detect possible anomalies in the services prescribed by a health care provider (HP) compared to his/her colleagues in the same field and environment. Their method is based on the concentration function that is an extension of the Lorenz curve widely used in describing uneven distribution of wealth in a population. The proposed tool provides a graphical illustration of a possible anomalous behavior of the HPs and it can be used as a prescreening device for further investigations of potential medical fraud.
DISCOVERY OF NEW FRAUD PATTERNS
Discovery of New Fraud Patterns

OVERVIEW

Rockwall is a small city in Texas, near Dallas. It was an ordinary small town, until the national news broke about one of its respected physicians. Dr. Jacques Roy was noticed to submit by far the most Medicare claims in the nation for home health services. Further investigation showed that Roy’s office handled more home health care visits from January 2006 to November 2011 than any other physician’s office in the country. As it was put by an Assistant U.S. Attorney during the trial, “A doctor cannot care for 11,000 patients at once.”

The more his activities were investigated, the more dirt came out. Auditors found out falsified visits for un-rendered services that were not necessary to start with. People from homeless shelters were taken to lunch in exchange for allowing their Medicare information being used. Dr. Roy was eventually sentenced to 35 years in federal prison and $373,331 in restitution for health care fraud and false statements. His lawyers continued to suggest he was just a hardworking doctor who did what was necessary for his patients.
In reality, when agents searched his home in 2011, they found books about vanishing without a trace and hiding money in offshore bank accounts. Agents also found two different fake identities with Roy’s photo along with passport documents from Canada, where he was born. According to federal prosecutors, if he had pulled off his escape plan, Dr. Roy might have been living in Canada or France now under the alias of Michel Poulin. The investigators got to Dr. Roy early enough, so he could not manage to run away from prosecution. However, most fraudsters manage to pull out an escape, while all taxpayer dollars vanish and become almost impossible to recover.

Dr. Roy’s extraordinary activities were revealed with the help of an outlier detection tool. How do these methods work? Why did Jacques Roy go unnoticed for at least five years? How can we reveal new fraud patterns and recover the overpayments before it is too late?

Predictive methods rank claims with respect to their likelihood of fraud, using known fraud patterns. But what about new patterns of fraud? If you were a fraudster, would you continue committing fraud the same way or mix and match your ways a little bit every time?

Fraudsters adapt to investigation outcomes as well as to changes in health care policies, and, therefore, the nature of fraud patterns change. Predictive methods are not able to capture changing fraud patterns unless they are frequently updated with labeled data. Such frequent updating can be expensive since retrieving labeled data in the context of health care fraud detection corresponds to the need for an actual audit, which can also be time-consuming. This lack of adaptability and the need for constant tuning have increased the attention on outlier detection, clustering, and association methods. These are also referred to as unsupervised methods—since they do not require labeled data.

These methods generally serve as pre-screen filters that list potentially fraudulent claims before the actual audit. This initial screening can decrease personnel costs as fewer transactions
are reviewed. They are not dependent on a particular labeled data set. That is why they can be used to help detect changing fraud patterns.

Outlier detection methods are used to reveal activities which are abnormal compared to average or expected behavior. Clustering is the most common segmentation method and is based on grouping similar objects into clusters. These findings can be later used along with predictive methods including classification. Association methods, link analysis, and network-based methods study attributes that go together in order to discover hidden relationships in a large data set. This chapter introduces these methods which can enable auditors to detect dynamic fraud patterns.

The questions to be addressed include

- How can we detect excessive or very different claim submissions, using outlier detection methods?
- How can we group claims?
- How can association type algorithms be used to find links among claims?
- What are the aspects to consider while assessing the effectiveness of the analytical methods?
- What should auditors consider during the deployment of the statistical methods?

OUTLIER DETECTION: FINDING EXCESSIVE BILLINGS

Have you ever heard about the importance of dental braces in Texas? Between 2005 and 2010, Texas Medicaid has spent more money on braces than the other 49 states combined. Given that Texas state population is less than 10% of the population of the United States, it is surprising. Either patients take really good care of their teeth in Texas compared to the nation, or providers bill excessively.
Audits revealed that doctors have routinely billed Texas Medicaid for uncovered procedures. These include putting braces on youngsters for purely cosmetic reasons and performing unnecessary root canals on small children.

Some of these unnecessary procedures even resulted in the deaths of pediatric patients. Extra scrutiny is good to ensure patient health is protected and taxpayer money is not wasted. However, extra audits also prompted some dental offices to stop serving Medicaid patients. So, health care fraud eventually hurts legitimate providers and deserving patients, one way or another.

Durable medical equipment, such as wheelchairs, scooters, and walkers, increases quality of life for many senior citizens. However, this equipment is also abused a lot. For instance, there were instances of billing everyone in a particular nursing home, as if all residents needed and received a wheelchair.

Seniors are especially targeted in such schemes. It is illegal for a medical supplier to make unsolicited telephone calls to people with Medicare, other than consented to and follow-up calls. Some suppliers get around this by hiring independent marketing firms. These high utilization rates can be identified by outlier detection methods.

A common aspect of these examples is excessive activity. How do we define excessive? This is the tricky part because of the subjective nature of health care and treatments. One specialist may be legitimately billing for hundreds of thousands of dollars for a patient who is getting a specialized cancer treatment, whereas the activity of a hundred-dollar billing may be excessive if the diagnosis only requires a ten dollar procedure. So, it is important to compare apples to apples, although the complexity of health care makes this challenging.

There are a number of ways to define normal activities. For each set of providers or procedures, there may be benchmarks that define normal ranges and acceptable thresholds. Then, claims that exceed a certain dollar amount or providers that overcharge for a given service can be flagged.
Outliers are defined as out of the norm, abnormal, and, therefore, exceeding the thresholds. The biggest challenge is to determine what to measure and the respective threshold levels. Setting thresholds or deciding the extent of normal require subject domain expertise.

If you set the thresholds too high, then you may miss the fraudulent activities. If you set it too low, then there will be too many claims to investigate. Integrated adaptive outlier detection methods may help to detect excessive billings, even before the payment. The bottom line is that labeled data may not be needed to reveal new or previously unknown patterns of fraud.

Some approaches analyze the data of billings over time and aim to detect spikes and billings outside of the trend line. For instance, a primary care physician may see more patients during flu season, and claims will spike. That would be deemed as normal. Whereas spikes occurring at odd times for that provider can be flagged for further audits. Sudden increases in billings or number of patients also trigger investigations. Even a provider suddenly starting to bill for very old patients can be noteworthy and may warrant assessment.

In order to compare apples to apples, we need to work with homogenous groups, or so-called peer groups. These peer groups are set by the characteristics of the providers, claims, or patients. For instance, the billings of a particular provider and the billings of providers in the same specialty can be compared. Then it can be argued that the providers who bill for more services than 99% of similar providers are suspicious and should lead to further investigations.

An alternative is to examine the overall activity, and aim to come up with homogenous groups. We will present a more detailed discussion on those grouping activities later in this chapter.

Descriptive statistical and basic visualization methods, which are discussed in Chapter 2, can be used as simple outlier detection methods. For instance, providers’ billings for a given procedure
can be compared using a boxplot. These can result in audit lists such as:

- Doctors who treated more patients a day than their peers
- Providers administering far higher rates of tests than their peers
- Providers billing far more costly claims, on a per patient basis, than their peers
- Providers with a higher ratio of distance patients than their peers
- Providers that prescribe certain drugs at a higher rate than their peers

What if we want to compare the provider’s billing behaviors for more than one procedure? One option is to use a composite ranking. You can score the difference from the expected value, or average, for each procedure, and come up with an aggregate measure.

An alternative approach could be to use concentration functions and Lorenz curves. This can help in investigating the billing differences of a provider among a variety of prescribed services. The main focus is to compare each provider’s billing behavior to the average behavior in their peer group. Selecting that peer group becomes even more important.

The peer groups are selected using assumptions such as a group of providers having similar characteristics based on provider specialty, region, and number of years of experience. Providers within peer groups are assumed to be providing similar services to similar patient populations. The distributions of these billings can be compared using a similarity measure. We will use the Lorenz measure in the following example.

Let’s assume we have cardiologists billing for a total of 100 procedure codes in a homogenous region. We are interested if
any of them have very different billing patterns compared to the benchmark—assumed to be average in this case. We compute the cumulative billing percentages for all procedures. The plot of these cumulative billing percentages versus the cumulative percentage of the corresponding procedure is referred to as the Lorenz curve. If a provider exactly mimics the average billing behavior, then his/her line will be on top of the benchmark line. Here is an example that compares two providers (see Figure 2.1). The provider that is denoted by the dashed line is more similar to the benchmark (bold line) compared to the other provider (dotted line).

Once peer groups and benchmarks are determined, statistical measures and visualization tools can be used to display outliers. This is one of the many methods to detect outliers. Please see the end of this chapter for references of more applications.

FIGURE 2.1  Lorenz curves for two providers (dashed and dotted lines) compared to the benchmark (bold line).
CLUSTERING: GROUPING HEALTH CARE CLAIMS

The Office of Inspector General (OIG) often receives calls from whistleblowers about health care fraud. Sometimes, it is a whistleblower suggesting that one hospital has billed Medicare for millions of dollars’ worth of up-coded (overcharged) services. The auditors aim to act on the provided information, but many hospitals bill more than a million claims every year. How do you decide if these claims are up-coded or not?

The audited item can be compared with identical claims or providers, the so-called peer groups. While presenting outlier detection methods, we discussed the importance of finding peer groups. At times, we know what the peer group is, such as the specialty of the provider. For example, Berenson–Eggers Type of Service (BETOS) categories can be used to decide on peer groups for providers. BETOS codes are assigned to every procedure and consist of readily understood clinical categories. They are stable over time.

As an alternative, further data such as patient profiles for each provider can be utilized to construct sub-peer groups. That is where clustering can be beneficial.

Clustering is simply used for grouping, or so-called segmentation. The objective is to find groups, aka clusters, that have similar observations within the group, and are as different as possible from other groups. Let’s say you have 100,000 movie goers who may be fond of different genres of movies. You have access to their characteristics, such as age, time spent on social media, and favorite TV shows. But you do not know their favorite genre(s).

Clustering can help group them based on the type of movie they may like. Ideally, when you group these moviegoers, you would put all action genre lovers into the same group, while drama and comedy moviegoers would be in separate clusters. Keep in mind you do not really know their favorite genre, so you try to group them with respect to other characteristics. Given these attributes, the objective is to find clusters such that data points in one cluster are more similar to one another whereas data points in separate clusters are less similar to one another.
Once the clusters are identified, an option is to apply labels to each cluster to classify each group based on its characteristics. These findings can be later used to build predictive models including classification or decision trees.

How can we use such an approach for health care fraud detection? Let’s say we have all the claims submitted by a hospital. We can aim to find out unusual or outlying clusters that cannot be explained by “normal” claim behavior.

Clusters can also be used as the first step of any statistical analysis to ensure the group of interest is homogenous. For instance, in U.S. Medicare, socio-demographic and location variables are used to cluster geographical regions before analyzing billing behavior. Another way is to form physician practice profiles with respect to number of visits, prescription percentages, and expenditures.

Most of these methods are so-called hard clustering approaches. A movie fan can be assigned to only one group, suggesting she can only like one type of movie. One provider can only correspond to one group at a time. However, the real world does not work that way. Therefore, so-called soft, aka probabilistic clustering methods are proposed. Each provider is assumed to have a certain probability of billing to any given cluster. This can be especially beneficial for providers who work with different sets of patients that require different levels of procedures. One provider may act like a cardiologist 50% of the time while billing like an internal medicine specialist 40% of the time. So-called Bayesian methods can deal with uncertainty by assigning probability distributions to all unknown events and variables.

Some of the relevant methods also help you do both grouping as well as outlier detection, which are referred to as Bayesian hierarchical methods, but we will not get into detail in this book. Hierarchical structure helps with dealing with complex data which can be modeled in layers. This can help you reveal the hidden patterns among providers and medical procedures which would otherwise be missed within a big data set.

Let’s present how this can be applied to health care claims data. Figure 2.2 presents an overview of the data set of interest, which
is a collection of Medicare Part B claims. This is also referred to as a health care procedure cloud. This is very similar to so-called word clouds that are used in text mining. This procedure cloud lists the procedures on a map depending on their relationship. If two procedures are billed by similar doctors, they would be more closely located on the map. The size of each procedure represents the frequency of billings for that procedure. For instance, 99213 is a frequently billed procedure, therefore it is larger in the map. You can use this as a pre-screening mechanism within your pre-payment or post-payment analytics framework.

A Bayesian hierarchical method can be used to understand which procedures are frequently billed together. This is a soft-clustering method, so each procedure is a member of all groups (clusters) with a certain probability. In text mining, these groups are also called topics. Table 2.1 provides an output of this method. It shows the five most frequently billed procedures in a particular group with their descriptions and billing frequencies. What do you recognize when you examine this group?
Even for our untrained eyes, the answer can be relatively simple. You may notice that a collection of eye-related procedures is displayed in Table 2.1. But how can this be used for audits?

You would expect ophthalmologists and optometrists to bill for this group and these five procedures. First, we can check if that is the case. In addition, a follow-up investigation would include specialty providers other than eye doctors billing for this group the most. Or we can investigate the ophthalmologists and optometrists who do not bill for this topic the most.

As you can recognize, in health care, fraud assessment tools do not give you ultimate answers. Most often than not, they result in more questions. It is all about funneling your energy and resources to potential problematic areas. Eventually, an audit would tell you whether the investigated activities are fraudulent or not.

Once you identify a suspicious provider, more statistics of that provider, such as its proportion of billing to each group, can be used. That helps to understand if that physician is an outlier or bills similar to his/her peers. A variety of descriptive statistical methods are already presented in Chapter 2. Such visual evidence helps to see how different that unusual provider is. In Figure 2.3, a boxplot displays the proportions of billing of each doctor in a peer group (referred to as theta in the figure) to four groups (referred to as topics or clusters in the figure). The unusual provider is denoted with a triangle. You can see the differences in the billing pattern of that outlier doctor. For instance, the outlier doctor bills most

<table>
<thead>
<tr>
<th>HCPCS Code</th>
<th>Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>92012</td>
<td>Eye exam established patient</td>
<td>0.324</td>
</tr>
<tr>
<td>92014</td>
<td>Eye exam and treatment</td>
<td>0.156</td>
</tr>
<tr>
<td>92083</td>
<td>Visual field examination(s)</td>
<td>0.075</td>
</tr>
<tr>
<td>S0620</td>
<td>Routine ophthalmic exam including refraction—new patient</td>
<td>0.062</td>
</tr>
<tr>
<td>92134</td>
<td>Computerized ophthalmic imaging retina</td>
<td>0.051</td>
</tr>
</tbody>
</table>
frequently for the procedures in Group (Topic) 8. Whereas all his/her peers most often bill for Group (Topic) 4.

Similarity computation tools used with clustering output can help identify potential teamwork as well as outliers. For instance, you expect a certain kind of cardiologist to have similar billing behavior, and therefore to have similar scores. A cardiologist with a high dissimilarity score compared to other cardiologists can be red-flagged. You may also have an anesthesiologist who bills similarly to these cardiologists. Such billing patterns can be a sign of collaboration or fraud. However, we cannot know that for sure before an actual investigation.
So far, we have discussed methods of grouping with respect to one aspect or one variable. Relatively new approaches are proposed to deal with dyadic data using so-called co-clusters. These aim to link two group types. For instance, the frequency of visits of groups of beneficiaries to groups of providers can be modeled with a co-clustering algorithm. Cardiologists may be found to serve older patients, forming a physician-patient co-cluster. If a particular cardiologist has been serving patients outside of his/her co-cluster, that may signal unusual activity. These methods can potentially reveal an emerging type of fraud called “conspiracy fraud” that involves attributes of more than one party of the health care system.

ASSOCIATION: FINDING LINKS AMONG CLAIMS

Which procedures are billed in a similar fashion? Which prescriptions are written by what kind of doctors? Can we find any potential associations and links among them? Association type algorithms are designed to help with these questions. This class of descriptive methods studies attributes that go together to discover relationships hidden in large data sets.

One of the early and popular applications has been to investigate customer behavior in retail stores. Therefore, this is also referred to as market basket analysis.

Social media companies use such algorithms extensively to understand user behaviors in their networks. In the auto insurance industry, so-called crash-for-cash schemes are uncovered using network algorithms.

The number and quality of relationships within a provider network can be analyzed in a similar setup. These algorithms may help reveal relationships, links and hidden patterns of information sharing and interactions within potentially fraudulent groups of providers and patients. This may be through links between businesses such as contact information, locations, service providers, assets, associates, or the relationships with a group of beneficiaries who are known to be within the fraudulent circle. Financial
statements, private commercial records, property tax records, banking records, and voter records can all help reveal further connections. You can even consider text data and social media data to reveal relationships.

Association algorithms can reveal the connection between pharmacies that were owned by the same people and providers who share a large percentage of the same patients. These algorithms can also compute the similarity between providers based on activity and the likelihood of any given patient to get services from a provider. For example, you can expect any two providers with a shared specialty to bill for relatively similar procedures. If their billing records are too different, that may signal suspicious activity.

In July 2017, the Justice Department announced charges for 412 people, including many doctors, in a $1.3 billion health care fraud. Some of the fraud schemes were committed within a network and were very sophisticated. For instance, one of those consisted of a guy who brought in patients through staged accidents, a lawyer that represented the patients against insurance companies, a few doctors who billed for the services, and the leader who ran the clinic and took care of the financials. It takes a lot of time and effort to get to the leader of such a network scam.

Let’s present a simple demonstration of suspicious provider networks based on their information. Drs. Lecter, Decimus, and Lebowski may have totally different billings and may be serving different sets of beneficiaries. But Dr. Decimus and Dr. Lebowski work for Roman Hospital, while Dr. Lecter works for Lecter Associates. On their websites, Lecter Associates and Roman Hospital seem to use the same toll-free number. In addition, both these providers excessively prescribe home health devices that are filled by Durden Durable Equipment Corporation. Paying attention to details can reveal these connections, and coming up with a simple network may reveal sham companies and potential fraudulent links.
In one case, the owner of a drug-treatment center in Delray Beach, Florida, recruited addicts to aid him in his schemes. His team attended Alcoholics Anonymous meetings and visited “crack motels” to persuade people to be “treated” by him. Offered kickbacks include gift cards, plane tickets, trips to casinos and strip clubs as well as drugs. This drug-treatment center was eventually charged with fraudulently billing insurance companies for more than $50 million for false treatment and urine tests over nearly five years.

The output of these link analysis tools can be displayed to show connections that otherwise would not be detected. Such tools can also be used to understand which drugs are prescribed together. Auditors can explore the existence and strength of links of drugs as well as their frequency of being prescribed by the same or similar providers.

These methods are valuable since a significant amount of fraud, waste, and abuse are found to be committed by organized, sophisticated, and collusive networks of providers and patients. It is more likely that fraudulent people are connected to other fraudulent people. A large number of overlapping patients would raise a red flag for further investigation. These may be fraud, but may also signal complementary services. This cannot be known without an actual audit.

Thresholds can be used to decide on the exceeded frequency level required to start investigation. The frequency of connections between certain types of entities, if they are much greater than normally expected, can be red-flagged. This also can help catch ghost patient billing. There are many unsuspecting patients out there whose identities are used to bill for procedures that are never provided.

Fraud patterns change. The networks should be continuously updated with new data to reflect these changes. They should keep evolving. One should be aware that they are very informative, but these tools take time and infrastructure to build in the first place.
EFFECTIVENESS OF THE ANALYTICAL METHODS

So far, we have presented a variety of analytical methods. But how effective are they? Let’s say you have a pulmonologist who specializes in the surgical treatment of lung issues, and he is the only expert who can conduct that expert level procedure in a geographical region of ten million people. All such patients would come to him, and his billings would be very different to the other pulmonologists. Most outlier detection methods may flag him as an outlier and may recommend an audit. After the audit, this may be seen as a case of method going wrong. However, keep in mind this billing could also have been a result of someone overcharging and billing for unnecessary expensive service codes.

That is why integrated approaches that also use other information, such as clinical data or demographics, are very valuable. Integration of field intelligence, policy knowledge, and clinical expertise into the development of algorithms and evaluation of outputs can help retrieve actionable output. This can reduce false positives and boost efficiency by preventing big losses early and potentially in the pre-payment phase.

Unsupervised methods give us the unique chance of understanding changes in billing behaviors, health conditions, or relationships within the provider network. For instance, network analysis can capture the effect of a new hospital to the other providers in the region. However, for the maximum benefit, we need to consider clinical knowledge for the evaluation of the output. The context of the patient, provider, and their joint history are all important. Evaluation of risk profiles and fraud scores of individuals based on an ensemble of models is crucial.

Measuring the success of statistical methods has been a well-documented challenge. There are not any industry-accepted standard methodologies to calculate savings of predictive analytics technologies for health care program integrity. In order to assess the overall impact, one needs to consider:
• Direct fiscal impact
  • Potential savings from denial of claims based on pre-payment reviews and auto-denial edits
  • Avoiding costs by more thorough screening of providers and revocation of billing privileges and payment suspensions
  • Overpayment recoveries by audits and fraud investigations
  • Contractor costs

• Improvements in fraud, waste, and abuse prevention, detection, and deterrence
  • Removal of systemic vulnerabilities
  • Sentinel effects such as providers self-modifying their behavior so as to not commit fraud, waste, and abuse
  • Greater ability to validate and compare claims across different provider types
  • Optimization of existing staff resources through better and more focused data retrieval
  • Communication of the policy changes to close vulnerability gaps and prevent future risks

In their annual report, the CMS came up with a novel measure to assess the performance of health integrity efforts. The CMS measure of adjusted savings only considers the direct savings, but not the avoidance of ineligible payments. It should be noted that the potential savings due to changes in provider billing behavior are difficult to measure.

Measuring the effectiveness of models can also help us to understand when a model becomes outdated and needs updating or replacing. In addition, it helps to assess the return on
investments. The methods may be working fine and may be generating many accurate leads. However, if they do not translate to action such as overpayment recovery that results in savings, their use may not be justified.

This brings us to the next problem. After having justified the use of an algorithm, how do the auditors deploy it?

**DEPLOYMENT VIA RULES**

How do we deploy these methods within an audit framework? Most often, they are embedded within a rule-based setup to filter fraudulent claims and behaviors and to discover emerging new fraud patterns.

Rule-based frameworks have many sources of information such as whistleblower tips. The rules can reveal providers who are not billing for legitimate beneficiaries or can detect geographically impossible billings. For instance, if a beneficiary is billed in Texas earlier that day, another claim that is submitted for him in Florida one hour later has a high potential of being fraudulent.

You can also flag providers who bill for stolen Medicare identification numbers with a simple rule. Such applications are also referred to as identity analytics that address eligibility fraud. These can be performed by standard enrollment qualification processes that include verification and authentication of all provider and beneficiary identities. Identity analytics can help eliminate unqualified and risky providers from the network. It should be noted that such rules can also be used to reveal the existence of known schemes and obvious patterns that are retrieved based on retrospective reviews and descriptive statistical analysis.

Although they can be simple and effective for fraud detection, rule-based methods have a few disadvantages. First, it is time-consuming to construct rules at first. Then, subject matter experts need to update the existing rules with respect to changing fraud patterns. Otherwise, poor maintenance would result in false positives and a diminishing number of recoveries. In addition, new fraudulent schemes would remain uncovered, so-called false
negatives. But once you create the rules, they easily generate leads. Sophisticated analytical skills are not required; anyone can use these after an initial short training.

On the other hand, fraudsters can game these algorithms with information about the rules and their thresholds. For instance, let’s suppose, claims worth under $100 are only investigated if the overall billing activity of that provider is excessive compared to his/her peer doctors. If the fraudster has this information, he/she can get away with billing for fabricated claims each for $99.90 with a total number less than the threshold; so that will not be deemed as excessive.

That is why it is important to use outlier detection in conjunction with other methods. That may result in decreased false leads. In order to customize methods, it is a good idea to involve the investigators in the process. A close cooperation between physicians, statisticians and policymakers would be very beneficial during the stages of defining and tuning the model as well as analyzing and interpreting the results. Use of medical knowledge is expected to improve performance, despite adding additional cost and complexity.

Hybrid approaches can improve the deployment performance of health care fraud detection methods. For instance, the CMS has been continuously refining their existing models based on the feedback received through the FPS and insights from field investigators, policy experts, clinicians, and data analysts. Ensemble algorithms combine the strengths of the supervised models for certain aspects of the problem and help find the outliers to improve the predictions. Unsupervised methods can improve the performance of supervised methods. For example, a clustering algorithm can be applied to divide all insurance subscriber profiles into groups. Then a decision tree can be built for each group and converted into a set of rules. Clustering can overcome the deficiencies of decision trees with larger data sets and many categories. When an unsupervised method is followed by a supervised method, the objective is usually to discover knowledge in a hierarchical way.
CURRENT EFFORTS

Implementation of effective data analytics programs offers several advantages, including a positive return on investment that can exceed that of traditional methods and strengthen program integrity safeguards. Pre-payment analytics help with early detection of improper payments. Recognized fraud patterns can be used to improve the initial review of claims. The potential areas of further losses can be identified, and the necessary actions such as closing policy loopholes can be taken.

Identification of predictors of improper billing can also lead to the development of new and more effective models for post-payment audits and recoveries. Predictive analytics coupled with unsupervised methods can help with comparing providers of the same type and identifying long-term trends. Then these trends can be used to build novel pre-payment rules to check for future claim submissions.

The CMS reported total savings of more than $39 billion of improper payments in Medicare during 2013 and 2014 (CMS, 2016b). The related efforts include training programs, avoidance of incorrect submissions via improved billing systems and better pre-payment, identity controls and actual overpayment recovery.

A recent news release (CMS, 2016b) lists recovery tools as:

- Enforcement of the False Claims Act by the Department of Justice (DOJ)
- Efforts of the Medicare Fraud Strike Force, which consists of OIG and DOJ members, against organized crime
- Use of advanced fraud detection technologies by OIG and CMS
- Enhanced provider screening and enrollment requirements set by the CMS
- Increased collaboration among the DOJ, OIG, FBI, and CMS via the Health Care Fraud Prevention Partnership
• Senior Medicare patrols by groups of volunteers who educate and empower their peers to identify, prevent, and report health care fraud

• Training and education programs for users such as the “Help Prevent Fraud” campaign

When it comes to application of statistical methods, each method has its own particular advantages and limitations. Relatively more complex methods such as network analysis may be more time-consuming and are complicated to construct at first. However, they generally provide a comprehensive understanding of health care billing patterns, which may result in more accurate leads. Table 2.2 summarizes the methods that are discussed in Chapters 4 and 5 with major application areas as well as their main drawbacks.

Fraud detection schemes will keep evolving. Technological developments also help fraudsters to continually become more inventive and resourceful. Fraudsters are flexible enough to change their tactics based on their knowledge of the government’s focus. Therefore, there will not be a bulletproof fraud or abuse detection technique.

Simultaneous and concurrent use of multiple methods offers the best chance for detecting both opportunistic and organized fraud. This can help in adapting to ever-changing fraud and abuse schemes, successful pre-payment detection, and the ability to deal with billings that are in large amounts and variety. Having

<table>
<thead>
<tr>
<th>Method</th>
<th>Application</th>
<th>Drawback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlier Detection</td>
<td>Detection of different behavior</td>
<td>Need for well-defined peer groups</td>
</tr>
<tr>
<td>Regression</td>
<td>Overpayment prediction</td>
<td>Need for labeled data</td>
</tr>
<tr>
<td>Classification</td>
<td>Grouping with respect to fraud type</td>
<td>Need for labeled data</td>
</tr>
<tr>
<td>Network</td>
<td>Revealing organized crime rings</td>
<td>High initial investment</td>
</tr>
<tr>
<td>Clustering</td>
<td>Grouping claims</td>
<td>Need for clinical insights</td>
</tr>
</tbody>
</table>
a data management and analytics plan in advance can help with the objective of revealing high-risk relationships using all patient, provider, and claims data.

The utilization of better identity checks and pre-payment analytics are expected to help identify claims that are not eligible. Quick processing of post-payment analytics tools can identify leads for early intervention by Medicare administrative contractors. Ongoing projects of the CMS include the integration of the FPS with the claims-processing system to enable claims denials or rejections directly through the FPS. This was successfully piloted for the rejection of certain physician claims. The CMS also launched a pilot project to explore opportunities to leverage other interventions for resolving leads in the FPS, such as provider education or medical review.

Statistical methods have enhanced the data analysis capabilities of health care fraud investigation teams. The interfaces and visual aids facilitate data summary and analysis, especially in the cases of comparisons. Such integrated capabilities allow investigators to capture all findings that are relevant to an investigation, including claims data, network diagrams, case notes, surveillance video, and any other external structured or unstructured data. Combining the auditor expertise, data analysts, as well as the traditional investigation skills, is crucial for the battle against fraud.

For the ideal implementation of health care fraud analytics, an integrated system which can combine a variety of data sources and expert knowledge such as medical and clinical insights, financial analysis, death/birth, and prison records is necessary. Such data resources help algorithms learn and improve themselves. This is crucial especially for proactive detection against newly evolving fraud patterns. For instance, clinical feedback may be the main difference maker for the success of an anomaly detection algorithm. The algorithm may return unusual activities as statistical outliers and flag many specialists such as oncologists and cardiothoracic surgeons. However, medical experts can help prioritize leads, and argue that sicker patients may visit with that provider because of
the expertise of the physicians and hospital. Clinical insights and medical logic can separate the medically needed and legitimate billings. In order to address this, there are ongoing efforts to use artificial intelligence-based tools within the health care domain.

In addition, there are a number of promising methodological advancements in recent years. More advanced time-trend analysis-based tools have the potential to become more widely used. These can show the billing development of networks or regions over a time horizon. They can also help tune other models and update thresholds in rule-based methods. Analysts may also forecast the change in health conditions of beneficiaries using such time-trend analysis tools.

Another relatively new set of tools includes the use of topological data analysis to identify new patterns of aberrant behavior. These have the potential to result in more accurate predictions. The main challenges of operating successful big data analytics programs may remain as the limited ability to integrate and manage unstructured data such as demographics or text data, data quality issues, and the potential shortage of analytics skills. These challenges can only be addressed in the long run through more training and investment.

**KEY TAKEAWAYS**

1. Unsupervised methods are used to explore claims data using unlabeled data and to reveal emerging fraud schemes.

2. Anomaly-based methods can be used for comparisons with peer groups to identify unusual activities.

3. Clustering methods can be used to group claims.

4. Bayesian hierarchical methods can help with both grouping as well as outlier detection.

5. Association type algorithms can reveal links among claims, providers, and patients.
6. Descriptive statistical methods can complement unsupervised methods while visualizing the output.

7. Direct fiscal impact, as well as potential improvements in fraud, waste, and abuse prevention, detection, and deterrence, should be considered while measuring the effectiveness of a statistical tool.

8. The deployment of rule-based methods can be improved by incorporating investigators and health care providers into the process.

9. Use of complementary data sources, consideration of changing legislation and health patterns, and frequent updating of rules can help decrease the ratio of false leads.

10. The choice of appropriate fraud analytics tool depends on the particular case and data.

11. Ease of use for data analysis and visualizations as well as data management are important considerations when choosing a statistical fraud detection tool.
ON THE USE OF THE CONCENTRATION FUNCTION IN MEDICAL FRAUD ASSESSMENT

This article is excerpted from The American Statistician by Tahir Ekin, Francesca Ieva, Fabrizio Ruggeri and Refik Soyer.

© 2017 Taylor & Francis Group. All rights reserved.
On the Use of the Concentration Function in Medical Fraud Assessment

Tahir Ekin\textsuperscript{a}, Francesca leva\textsuperscript{b}, Fabrizio Ruggeri\textsuperscript{c}, and Refik Soyer\textsuperscript{d}

\textsuperscript{a}Department of Computer Information Systems and Quantitative Methods, Texas State University, San Marcos, TX; \textsuperscript{b}MOX-Modeling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Milan, Italy; \textsuperscript{c}Istituto di Matematica Applicata e Tecnologie Informatiche, Consiglio Nazionale delle Ricerche, Milan, Italy; \textsuperscript{d}Department of Decision Sciences, George Washington University, Washington, DC

**ABSTRACT**

We propose a simple, but effective, tool to detect possible anomalies in the services prescribed by a health care provider (HP) compared to his/her colleagues in the same field and environment. Our method is based on the concentration function that is an extension of the Lorenz curve widely used in describing uneven distribution of wealth in a population. The proposed tool provides a graphical illustration of a possible anomalous behavior of the HPs and it can be used as a prescreening device for further investigations of potential medical fraud.

1. Introduction

Health care expenditures have increased significantly over recent decades in developed countries. In addition to the aging population and increasing diversity in health services, fraud as well as abuse and waste dramatically contribute to the increase in health care costs. For example, in the United States, the National Health Care Anti-Fraud Association (\url{www.nhcaa.org}) estimated conservatively that at least 3\%, or more than 60 billion dollars, of annual health care expenditures was due to fraud, waste, and abuse in 2010. Total health care related spending in the United States was almost 3 trillion dollars corresponding to 9523 dollars per person in 2013 (\url{www.cms.gov}). Therefore, efforts for assessment and reduction of expenses due to medical fraud are crucial in the health care industry.

The systematic use of statistical approaches in medical fraud assessment in the United States gained momentum with the Health Care Fraud and Abuse Control Program in 1996. Sampling and overpayment estimation methods help medical auditors to use sample data and make extrapolations for the population. Woodard (2015) demonstrated the use of sampling by US governmental medical insurance programs. The formal governmental guidelines (\url{www.cms.gov}) recommend the use of the lower limit of a one sided 90\% confidence interval of the total overpayments as the recovery amount from the provider under investigation. This protects the provider and results in fair recovery with 95\% confidence. However, application of the central limit theorem to compute the lower bound is based on the assumption that the overpayment population either follows the Normal distribution or that the sample size of overpayments is reasonably large. It is well known that medical claims data mostly exhibit skewness and nonnormal behavior, requiring large sample sizes for a valid application. As alternatives, Edwards et al. (2003) obtained the lower bounds using a non-parametric inferential method whereas Gilliland and Edwards (2011) constructed randomized lower bounds. Ignatova, Deutsch, and Edwards (2012) proposed a sequential sampling framework that aims to make inference on the proportion of claims with overpayments. Ekin, Musal, and Fulton (2015) presented a zero-one inflated mixture model for estimation of overpayments. Musal and Ekin (2017) provided a recent overview of such references before proposing a Bayesian mixture model.

These studies are mostly interested in the percentages of claims to be reviewed to make a reliable assessment on the possible fraudulent behavior of a health care provider (HP). In contrast, our article studies how billing by an HP, split among different prescribed services, differs from the average behavior in a population of HPs.

There are relevant data mining studies that provide sophisticated methods to detect possible fraud. Supervised approaches, including but not limited to neural networks, decision trees, and Bayesian networks, are proposed; see Li et al. (2008) for a survey. These methods require labeled data, which correspond to already audited claims. The results of these methods are dependent on a particular dataset; therefore they cannot adapt to the dynamic nature of fraud patterns. As a potential remedy, unsupervised methods can be used to extract information about the relationships within medical data. Particularly, these may serve as prescreening tools to identify a set of potentially fraudulent claims before domain experts are brought during the investigation phase. Onderwater (2010) provided an overview of such outlier detection methods for fraud assessment. The anomaly detection framework for Australia Medicare spatio-temporal data (Ng et al. 2010), and the use of Benford’s law distributions to detect anomalies in claim reimbursements (Lu and Boritz 2005) are some examples. To group medical claims data, clustering algorithms can also be used. Musal (2010) provided an illustration of clustering of geographical regions as input to his regression model. The Bayesian co-clustering model of Ekin et al. (2013) investigates the dyadic patterns among providers and beneficiaries. Overall, these methods can decrease
personnel costs as fewer transactions are reviewed (Laleh and Azgomi 2009).

Our article presents the use of the concentration function as a prescreening tool to aid in medical fraud assessment. It does not suffer from the issues of supervised methods; in particular, it does not require labeled data and can easily adapt to changes if it is run with a different dataset. This simple unsupervised tool lets the auditor analyze the billing patterns of a particular doctor, which can reveal potential unusual behaviors. Compared to existing data mining approaches, our approach provides a simple tool, both in implementation and understanding, which could be used to detect an anomalous behavior and might hinder potential fraud.

The idea behind the proposed approach is quite simple. We assume that a group of health care providers (HPs) with similar characteristics (age, specialty, years in the area, etc.) are providing similar services to patient populations that are similar in terms of distribution of age, income, gender, etc. We are aware that this assumption might not fully reflect reality, but our goal is to provide a tool that can detect unusual behavior by an HP in terms of deviating from a population of HPs who are expected to display similar prescribed service patterns. Further analyses will be needed to prove if such an anomaly is due to fraud or due to heterogeneity in population of patients. It is possible that an HP might prescribe particular services, and charge more frequently than other HPs with no fraudulent intent. For example, a provider may prescribe a significantly larger number of prostate exams in 1 year due to a larger number of elderly patients in need in his/her area. Assessment of potential causes of such different behavior is the next step to be undertaken through careful review of the claims by the HP. In Section 2, we will introduce the concentration function, an extension of the Lorenz curve, which will be used in Section 3 to analyze real data. Final remarks will be presented in Section 4.

2. Concentration Function and Lorenz Curve

The proposed tool is based on the concentration function, which is a generalization of the Lorenz curve (see, e.g., Marshall and Olkin 1979, p. 5) and is well known in the statistical literature. The Lorenz curve is a graphical tool used to describe the discrepancy between a discrete probability measure \( \Pi \) and a discrete uniform measure \( \Pi_0 \). Its typical application is about the comparison of the actual income distribution in a population \( \Pi \) with an income that is evenly distributed across the population \( \Pi_0 \). The Lorenz curve is obtained by plotting the cumulative wealth of the poorest individuals in the population. In particular, we consider a population of \( n \) individuals with wealth (income) \( x_i, \) \( i = 1, \ldots, n \), assuming no ties for simplicity. We order their incomes in ascending order and obtain the ordered wealths \( x_{(1)}, \ldots, x_{(n)} \), from the poorest to the richest individual. We define \( S_0 = 0 \) and \( S_k = \sum_{i=1}^{k} x_{(i)} \). Therefore, \( S_n \) is the total income of the population and \( S_k/S_n \) is the fraction of wealth owned by the \( k \) poorest individuals. We plot the curve connecting the points \( (k/n, S_k/S_n), k = 0, \ldots, n \). For a given \( k \), the plot displays the fraction \( S_k/S_n \) of the total income owned by the \( k/n \cdot 100\% \) of the poorest part of the population. We obtain a convex, increasing function connecting the points \( (0,0) \) and \( (1,1) \). It is worth mentioning that evenly distributed wealth implies a straight line since \( S_k = k/n \cdot S_n \) for all \( k \). When \( \frac{S_k}{S_n} \) is the concentration function. A prescreening tool to aid in medical fraud assessment. It does not suffer from the issues of supervised methods; in particular, it does not require labeled data and can easily adapt to changes if it is run with a different dataset. This simple unsupervised tool lets the auditor analyze the billing patterns of a particular doctor, which can reveal potential unusual behaviors. Compared to existing data mining approaches, our approach provides a simple tool, both in implementation and understanding, which could be used to detect an anomalous behavior and might hinder potential fraud.

We can compare the Lorenz curves and related indices for two or more populations. The population with the lowest Lorenz curve, if that exists, is the one with the largest disparity in income distribution. Sometimes, Lorenz curves might intersect and this requires the use of Gini's area of concentration and Pietra's index for comparison of the income distributions. The first case is well represented by two populations, \( A \) and \( B \), of three individuals each, where the wealth is distributed according to the following, already ordered, percentages of (10%, 30%, 60%) in \( A \) and (20%, 30%, 50%) in \( B \).

The Lorenz curves for the two populations are presented in Figure 1, where the lowest dashed curve corresponds to the wealth distribution in \( A \), the middle dotted one to the distribution in \( B \), and the highest (straight) solid curve corresponds to the case in which wealth is evenly distributed among the three individuals. In this case, it is evident that both populations have an unequal distribution of wealth, and it is more uneven in population \( A \).

An example of intersecting Lorenz curves is obtained when the population \( A \) and \( B \) have percentages of (15%, 40%, 45%) and (20%, 25%, 55%), respectively. The corresponding curves are presented in Figure 2. In this case, there is no ordering of Lorenz curves. It is true that the poorest individual in \( A \) has...
less income than the corresponding one in B (15% vs. 20%) but the reverse holds when considering the two poorest individuals (55% vs. 45%). Gini’s area is 0.1 for A and 0.117 for B, whereas Pietra’s index is 0.183 for A and 0.217 for B. The indices show that the income distribution is, in general, more uneven in B, where the largest values are obtained.

The Lorenz curve can be extended to compare any pair of probability measures on the same measurable space, as described in Cifarelli and Regazzini (1987). The same authors showed that Pietra’s index is equal to the total variation norm distance between the two probability measures. We do not refer to their elegant, but mathematically sophisticated, definition but we rather prefer to illustrate its use with a very simple example related to HPs’ prescribed services.

Suppose that HPs in a homogenous region are prescribing only three tests (blood, urine, and ECG) for their patients. We are looking at the percentage of the billing for each test with respect to (w.r.t.) the total. We are interested in discovering if an HP has a different pattern w.r.t. the group and, therefore, if further investigation of the individual is worthwhile to detect possible fraud. We suppose the billings in the HPs group for blood test, urine test, and ECG account for 20%, 40%, and 40% of the total amount, respectively.

We consider two HPs and we expect that they behave in a similar way w.r.t. the group. The billing of the first HP (called A) is split into 20% for blood tests, 70% for urine tests, and 10% for ECG, whereas the percentages for the second HP (called B) are 30%, 50%, and 20%, respectively. In probabilistic terms, we are interested in comparing two probability measures: a reference one, \( \Pi_0 \), related to the whole group of HPs with probabilities \( (0.2, 0.4, 0.4) \) (for blood, urine, and ECG, respectively) w.r.t. \( \Pi \) for the selected HP, given by (0.2, 0.7, 0.1) for A and (0.3, 0.5, 0.2) for B.

Therefore, we consider two probability measures, \( \Pi \) and \( \Pi_0 \), assigning probabilities \( p = (p_1, \ldots, p_n) \) and \( q = (q_1, \ldots, q_n) \), respectively, to the same outcomes \( (x_1, \ldots, x_n) \) of a statistical experiment (here the billing for different services, that is, blood, urine, and ECG). We suppose \( q \) represents the distribution for the group and we are interested in measuring how far \( p \) is from it. Earlier, the Lorenz curve was constructed summing the income of the individuals \( x_i \), starting from the poorest. The concentration function is constructed summing the probabilities of the outcomes \( x_i \), which are more unlikely under \( \Pi \) than under \( \Pi_0 \) (i.e., the values where \( \Pi \) is less concentrated than \( \Pi_0 \)).

For each \( i, i = 1, \ldots, n \), we compute the (likelihood) ratios \( r_i = p_i/q_i \) and order the \( x_i \)'s according to ascending values of \( r_i \). We therefore order the outcomes from the ones where \( \Pi \) assigns much less probability than \( \Pi_0 \) toward the ones where \( \Pi \) assigns much more probability than \( \Pi_0 \). The ordered values are denoted as \( x_{(1)}, \ldots, x_{(n)} \) and the corresponding probabilities are \( q_{(1)}, \ldots, q_{(n)} \) and \( p_{(1)}, \ldots, p_{(n)} \). Similar to the Lorenz curve, we plot the curve connecting the points \( (Q_k, P_k), k = 0, \ldots, n \), where \( Q_0 = P_0 = 0, Q_k = \sum_{i=1}^{k} q_{(i)}, \text{ and } P_k = \sum_{i=1}^{k} p_{(i)}. \) As before, we obtain a convex, increasing function connecting the points (0, 0) and (1, 1) and we call it the concentration function of \( \Pi \) w.r.t. \( \Pi_0 \).

Let us focus on a particular value \( k \) and consider \( Q_k \) and \( P_k \) to better understand their meaning. When considering all possible percentages \( Q_i \) of billings for services prescribed by the whole group of HPs, there are some services that have been charged less by the HP under scrutiny. In our case, the 60% of billings by the entire group could be represented either by blood and urine tests or by blood test and ECG. Looking at HP A, the first pair corresponds to 90% of his/her billing for the prescribed services and the second to just 30%. We are interested in the second pair since it provides the smallest possible percentage, \( P_1 \), of billings for A among all those with a given percentage \( Q_i \) for the group of HPs. Such choice corresponds to the largest possible dissimilarity in behavior, and this is what we are after. In probabilistic terms, we are interested in the set of outcomes that assign less probability under \( \Pi \) (here 0.3) among all those with a given probability under \( \Pi_0 \) (here 0.6). Such outcomes are given by \( x_{(1)}, \ldots, x_{(k)} \), whose probabilities sum up to \( P_k \) under \( \Pi \) and \( Q_k \) under \( \Pi_0 \).

Like the Lorenz curve, the concentration function can be used to analyze the distance between the probability measures \( \Pi \) and \( \Pi_0 \). If the distance between the concentration function and the straight line is short, then the distribution of the services prescribed by an individual HP is similar to that of the group of HPs. No warning should be issued in this case since the HP’s behavior is very similar to the group’s. Otherwise, future investigation may be suggested in search of possible causes, potentially including fraud.

In the previous example about two HPs, it can be shown that \( Q_1 \) is obtained in both cases for the ECG billing, accounting for 40% of the group total, whereas the value of \( P_1 \) is given by 0.1 for A and 0.2 for B. Whereas B is not too far from the group, A should be further investigated because of the very small amount charged for ECG and the larger amount for the urine tests that are prescribed.

The concentration curves for the two HPs are presented in Figure 3, where the lowest dashed curve corresponds to A, the middle dotted one to B, and the highest (straight) solid curve corresponds to the case in which the HP charges tests in the same percentages as the group.

The different behavior of HP A is confirmed by looking at Gini’s area of concentration and Pietra’s index. Their previous definitions can be extended to the current situation. Pietra’s index \( \mathcal{I}_{\Pi} \) becomes \( \sup_{1 \leq k \leq n-1} (Q_k - P_k) \) whereas Gini’s area of
concentration becomes \(1/2 - 1/2 \sum_{k \leq n} (P_k + P_{n-k}) (Q_k - Q_{n-k})\). The latter is 0.18 for \(A\) and 0.11 for \(B\), whereas the former is 0.3 for \(A\) and 0.2 for \(B\).

Once a synthetic index exceeds a fixed threshold then a warning should be issued and the individual HP could be subjected to more detailed investigation to detect the causes of the different behavior than the group of HPs.

### 3. Application of the Concentration Function

In this study, we use the public dataset, Provider Utilization and Payment Data Physician and Other Supplier Public Use File, which was prepared by The Centers for Medicare & Medicaid Services (www.cms.gov). It includes information related to payment, number of services, and number of beneficiaries for each provider and prescribed service. We rearrange the data and we consider here a small dataset, with 30 medical doctors (MDs) in Diagnostic Radiology in Vermont and the percentages of their billings within a set of 61 prescribed services. In the Appendix, Table A.1 contains the service descriptions. We chose MDs in the same specialty and in an area with a small number of people (Vermont is the second least populous state in the United States) to make the assumption about uniformity of behavior among HPs more reasonable. The prescribed services include X-rays, Computed Tomography, Magnetic Resonance Imaging for different parts of the human body. We consider just two MDs (named MD1 and MD2, respectively) out of 30 since we would like to show how the concentration function works in the simplest case. The extension to all the MDs is straightforward.

Table 1 presents the percentages of each MD's billing for each 8th prescribed service \(p_i^1\) and \(p_i^2\), respectively, \(i = 1, 2, \ldots, 61\) and the average percentages \(q_i\) among the population. Whereas, Table 2 lists the likelihood ratios for each MD \(r_i^1 = p_i^1/q_i\) and \(r_i^2 = p_i^2/q_i\), respectively.

First of all, the likelihood ratios \(r_i^1\) and \(r_i^2\) provide information on how the charges of each MD differ from the average charges of the population. A value close to 1 denotes a similar behavior in terms of percentage of billing for such service, whereas a smaller (larger) one shows that the MD is charging less (more) than the average. Note that we are not considering the billing in dollars, but are normalizing the figures considering just the percentage w.r.t. the total billing.

There are some services never charged by the MDs but our interest is about those which are overcharged w.r.t. the population: those that might be due to fraud, or, at least, abuse, and waste. In particular, we set a threshold (5 in our case) on the likelihood ratio whose exceedance should trigger further investigations. When the likelihood ratio is more than 5, then the percentage of charges for the related service prescribed by an MD is at least five times larger than the average charge for that service.

In Table 2, we highlight the values of likelihood ratios exceeding the threshold, using boldface for \(r_i^1\) (MD1) and \(r_i^2\) (MD2).

<table>
<thead>
<tr>
<th>Type</th>
<th>(p_i^1)</th>
<th>(p_i^2)</th>
<th>(q_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_1</td>
<td>0.0089</td>
<td>0.0319</td>
<td>0.0000</td>
</tr>
<tr>
<td>p_2</td>
<td>0.0144</td>
<td>0.0000</td>
<td>0.0047</td>
</tr>
<tr>
<td>p_3</td>
<td>0.0098</td>
<td>0.0000</td>
<td>0.0123</td>
</tr>
<tr>
<td>p_4</td>
<td>0.0344</td>
<td>0.0514</td>
<td>0.0000</td>
</tr>
<tr>
<td>p_5</td>
<td>0.0422</td>
<td>0.0028</td>
<td>0.0000</td>
</tr>
<tr>
<td>p_6</td>
<td>0.0238</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>p_7</td>
<td>0.1865</td>
<td>0.0412</td>
<td>0.0000</td>
</tr>
<tr>
<td>p_8</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1155</td>
</tr>
</tbody>
</table>

Table 1. Percentages of prescriptions for MD1 \(p_i^1\), MD2 \(p_i^2\), and population \(q_i\).

<table>
<thead>
<tr>
<th>Type</th>
<th>(r_i^1)</th>
<th>(r_i^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r_1</td>
<td>3.2364</td>
<td>9.2643</td>
</tr>
<tr>
<td>r_2</td>
<td>3.7370</td>
<td>1.3289</td>
</tr>
<tr>
<td>r_3</td>
<td>2.5907</td>
<td>2.0574</td>
</tr>
<tr>
<td>r_4</td>
<td>0.0000</td>
<td>5.2653</td>
</tr>
<tr>
<td>r_5</td>
<td>2.7880</td>
<td>5.8476</td>
</tr>
<tr>
<td>r_6</td>
<td>1.8049</td>
<td>0.0000</td>
</tr>
<tr>
<td>r_7</td>
<td>9.6515</td>
<td>2.1360</td>
</tr>
<tr>
<td>r_8</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 2. Likelihood ratios for MD1 \(r_i^1\) and MD2 \(r_i^2\) w.r.t. population.
First of all, for MD1 there are two ratios exceeding 9: they correspond to Computed Tomography of the abdomen and pelvis (9.2643) and X-ray exam of abdomen (9.6515). The former accounts for more than 3% of the charges by MD1 whereas the latter represents almost 19%. Although both account for charges almost ten times more than average, the major concern is about the X-ray exam of abdomen since it accounts for almost one fifth of the total billing by MD1. Further investigation in this case should be about number of prescriptions and diagnoses, to check for possible overcharges (and fraud, therefore) and waste in prescribing the X-ray. The other three prescriptions with ratios larger than 5 are X-ray exam of thigh (5.0040), a different X-ray exam of abdomen (5.2653), and X-ray exam of hip (5.8476). MD2 is prescribing just an exam, the ultrasound exam of the abdomen back wall, charged well above average. Given that there is just one anomalous exam, accounting only for 2% of the charges, then MD2 can be hardly considered at risk of fraudulent behavior, unlike MD1.

The analysis of the likelihood ratios allows us to identify a different behavior of MD1 w.r.t. the MDs population regarding the charges for five prescribed services, out of 61. The concentration function is a graphical tool that allows for an immediate recognition, at a glance, of the anomalous behavior when considering all the prescribed services.

The concentration functions, plotted in Figure 4, denote clearly how MD1 (dashed line) differs significantly from the population since the corresponding curve is quite far from the straight line, unlike MD2 (dotted line). Such a distance is summarized by Pietra’s index of 0.5504, compared to 0.2198 for MD2. A similar result is obtained when considering Gini’s index: 0.3580 for MD1 and 0.1593 for MD2.

A look at the concentration function in Figure 4 can provide further information on the behavior of the MDs. The flat (dashed) line from 0 to almost 0.5 tells us that MD1 is never giving prescriptions accounting for almost 50% of the billings by the average population of MDs. The sharp increase around 1 is interpreted as an excess of charges by MD1 w.r.t. average (mostly due to X-ray exam of abdomen, as discussed earlier). The dotted curve is almost parallel to the straight line: this is due to very few prescribed services that are not charged by MD2, unlike the MDs population, implying the initial lowering of the curve, whereas all the other services are charged quite similarly to the average, except for the ultrasound exam of the abdomen back wall, as discussed earlier.

In Figure 5, we present the histogram of Pietra’s index values for the 30 MDs in the group. It can be seen that 20 MDs have their index in the first two bins, denoting a substantial concordance among themselves about percentages of charges for the services they prescribe. There are just a few values (7) exceeding 0.5, including the one corresponding to MD1. If a warning limit is set to 0.5, then all the MDs whose Pietra’s index exceeds that value could be subject to further investigation.

As shown in this section, the concentration function, and the likelihood ratios needed to construct it, can be easily computed and can provide insights on the general behavior of an MD and also on the details of his/her prescribed services and related charges.

4. Discussion

In this article, we present a simple tool that could detect anomalous behavior of health care providers w.r.t. a population of providers believed homogenous, with a similar pattern in terms of charges for prescribed services. As discussed in the article, the tool does not provide formal evidence of fraud or abuse but it can be used as a prescreening device to detect possible anomalies in the pattern of charges, which could be further investigated. It should be noted that heterogenous behavior in terms of charges for prescribed services can be totally legitimate, and may be due to specialists having sicker patients and performing necessary operations. However, the tool still provides us information about different doctor billing patterns, which may also be the result of incentives that lead doctors to overcharge for a procedure.

We are working on the assessment of potential fraud in medical practice by different approaches. One approach that will be presented in a forthcoming article considers sophisticated methods based on Bayesian co-clustering to link groups of providers and prescribed services. However, it is important to note that one also needs to provide simple tools that can be used by practitioners and the current work is an attempt to meet this need. At the same time, it is also worth considering critical aspects of data. In fact, data preprocessing takes a fair amount of time before conducting the statistical analysis. Furthermore, another important aspect is medical data security. The data analyst should adhere to proper security, access, and privacy.
controls. Personally identifiable information that can distinguish or trace any identity should be dealt with caution and protected properly.

Those issues should be kept in mind when considering other possible applications. As an example, the proposed tool can be used to analyze deviations within any given category. Therefore, we can conduct the analysis for differently defined peer groups of providers. For example, Berenson-Eggers Type of Service (BETOS) categories have been used to categorize the providers and analyze U.S. Medicare costs. These clinical categories are a collection of Health Care Financing Administration Common Procedure Coding System procedure codes and can serve as peer groups. Another potential extension is to construct sub-peer groups with respect to a co-variate, such as patient profiles. Then, the billings of a given provider can be compared to the peer group to extract potential patterns. These insights can be helpful before bringing domain experts into the investigation. Close cooperation between physicians, statisticians, and people involved in decision making is essential while interpreting the results.

### Appendix: List of Prescribed Services

<table>
<thead>
<tr>
<th>Table A.1. Service descriptions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray exam of humerus</td>
</tr>
<tr>
<td>Ct abdomen w/dye</td>
</tr>
<tr>
<td>Bone imaging whole body</td>
</tr>
<tr>
<td>US exam of head and neck</td>
</tr>
<tr>
<td>X-ray exam of neck spine 2 view</td>
</tr>
<tr>
<td>X-ray exam of ribs/chest</td>
</tr>
<tr>
<td>X-ray exam of thigh</td>
</tr>
<tr>
<td>X-ray exam of thoracic spine</td>
</tr>
<tr>
<td>X-ray exam of lower leg</td>
</tr>
<tr>
<td>Ct angiography chest</td>
</tr>
<tr>
<td>MRI brain w/ &amp; w/dye</td>
</tr>
<tr>
<td>MRI brain w/o dye</td>
</tr>
<tr>
<td>Ct neck spine w/o dye</td>
</tr>
<tr>
<td>X-ray exam of hip</td>
</tr>
<tr>
<td>X-ray exam of lower spine</td>
</tr>
<tr>
<td>MRI lumbar spine w/o dye</td>
</tr>
<tr>
<td>X-ray exam of wrist</td>
</tr>
<tr>
<td>Computer dx mammogram add-on</td>
</tr>
<tr>
<td>X-ray exam of knee 3</td>
</tr>
<tr>
<td>Extracranial study</td>
</tr>
<tr>
<td>X-ray exam of ankle</td>
</tr>
<tr>
<td>Ct abdomen &amp; pelvis</td>
</tr>
<tr>
<td>X-ray exam series abdomen</td>
</tr>
<tr>
<td>DxA bone density axial</td>
</tr>
<tr>
<td>X-ray exam of abdomen</td>
</tr>
<tr>
<td>X-ray exam of foot</td>
</tr>
<tr>
<td>X-ray exam of shoulder</td>
</tr>
<tr>
<td>Ct abdomen &amp; pelvis w/contrast</td>
</tr>
<tr>
<td>Ct head/brain w/o dye</td>
</tr>
<tr>
<td>Total knee arthroplasty</td>
</tr>
<tr>
<td>Inject spine w/cath c/t</td>
</tr>
</tbody>
</table>
CHAPTER 3
Statistics and Social Issues
Collecting and analyzing data on unemployment, inflation, crime, poverty, slavery, sexual identity and inequality helps us describe the complex world in which we live. When published by the government, they are called official statistics. They are reported by the media, used by politicians to lend weight to their arguments and by economic commentators to opine about the state of society.


Where do crime statistics come from and how can you tell if they are accurate? Statisticians apply standard principles and procedures to collect data and to calculate and interpret statistics. This chapter illustrates how those principles apply to data sources and statistics about U.S. crime rates and tells you what to look for when evaluating a statistic.


This chapters describes job statistics. Using data from the United States, this chapter answers three basic questions—why, how and for whom these statistics have been constructed. The intention is to provide the reader with a good grasp of these measures.


Modern slavery prevention efforts will be more effective and targeted if better, more-reproducible slavery prevalence estimates are available. This article discusses five promising methods for getting these results: Multiple Systems Estimation (MSE), Respondent Driven Sampling (RDS), taking online-based surveys of U.S. stakeholders, national aggregation of statewide prevalence estimates, and Gallup public opinion surveys on human trafficking awareness and willingness to report to hotlines and/or law enforcement.


Sexual orientation and gender identity (SOGI) are two demographic and personal dimensions that have received relatively little attention in large-scale surveys and official statistics despite recent advances in social, cultural and legal equality for lesbian, gay, bisexual and transgender (LGBT) people. This article discusses the importance of such reporting and offers advice, best practice and FAQs to ensure that these marginalized populations are a regular part of the data-based
decision-making world in which we live.


The Gini index is the most commonly used measure of income inequality. Like any single summary measure of a set of data, it cannot capture all aspects that are of interest to researchers. One of its widely reported flaws is that it is supposed to be overly sensitive to changes in the middle of the distribution. By studying the effect of small transfers between households or an additional increment in income going to one member of the population on the value of the index, this claim is re-examined.
CHAPTER 3

THINKING STATISTICALLY ABOUT CRIME

This chapter is excerpted from *Measuring Crime: Behind the Statistics* by Sharon L. Lohr. © 2019 Taylor & Francis Group. All rights reserved.

Learn more
Thinking Statistically about Crime

The headline of the February 2018 news story drew me in: “94%: Sexual misconduct in Hollywood is staggering.” The story continued:

The first number you see is 94% — and your eyes pop with incredulity. But it’s true: Almost every one of hundreds of women questioned in an exclusive survey by USA TODAY said they have experienced some form of sexual harassment or assault during their careers in Hollywood.

Almost all of the women who participated in the survey said they experienced sexual harassment or assault. But what about women who did not participate?

We can’t tell from this survey. The USA Today story acknowledged this: “As a self-selected sample, it is not scientifically representative of the entire industry.”

Why don’t the results from this survey generalize to all women who work in Hollywood?

- Email invitations for the survey were sent only to members of two advocacy organizations: The Creative Coalition, and Women in Film and Television. All survey participants have joined an organization that raises awareness about issues such as public funding for the arts and gender parity in the screen industries. Their experiences and perceptions likely differ from those of non-joiners.
• Altogether, 843 women participated in the survey. The story does not say how many were asked to participate, so we do not know what percentage of women responded to the survey invitation. Often, persons who choose to respond to surveys are particularly engaged in the topic or have strong opinions. The women who responded to the survey invitation may have had different experiences than the women who did not elect to participate.

A survey administered to a conveniently chosen sample is often cheaper and easier to conduct than a survey that is statistically representative. The USA Today survey provided timely context about fall 2017 news stories concerning sexual assault in Hollywood by asking a large number of people about their experiences. It demonstrated that the women who had come forward publicly were not the only ones with accounts of harassment and assault, and gave a voice to the survey participants.

However, the statistics from the survey apply only to the women who participated in it. It is correct to say that 94% of the 843 women who responded to this survey reported having experienced sexual harassment or assault—according to the survey’s definitions and questions about those events—during their careers.

The study does not tell us what percentage of women in Hollywood have been sexually harassed or assaulted. That number might be 94%. It might be something else. The statistical procedures used in the survey do not allow us to assess how accurately the statistics describe all women in Hollywood.

**STATISTICAL REASONING**

Where do crime statistics come from, and how can you tell whether they are accurate?

Statisticians employ standard principles and procedures to collect data and to calculate and interpret statistics. This book illustrates how these principles apply to data sources and statistics about US crime rates, and tells you what to look for when evaluating a statistic.

The same principles apply to other statistics you encounter, from any field of study: political polls, unemployment rates, transportation use, size of bald eagle populations, agricultural production, diabetes prevalence, literacy rates, health care expenditures—the list goes on.
All statistics about crime, even the ones that appear to be exact counts such as number of homicides, are estimates. Statistical reasoning methods allow us to quantify uncertainty about estimates and tell how accurate they are likely to be.

WHY DO WE NEED ACCURATE CRIME STATISTICS?

You hear about crime all the time. Almost every edition of a newspaper contains at least one crime report. Stories about crime get high ratings—“If it bleeds, it leads”—and it is natural when you see one account after another to think that crime is everywhere.

Stories are memorable. But statistics tell us whether the stories are isolated events or reflect trends in society. When there are no high-quality statistics, people tend to extrapolate from personal experiences (“I know three people who were robbed last year—crime is really going up”) and opinions.

Accurate crime statistics help answer questions such as:

- How much crime has occurred, and what types of crime are increasing or decreasing?
- Who are the victims and offenders?
- What are the costs of crime to victims and to society?
- What crime-prevention and crime-reduction strategies are effective?
- Where should law enforcement resources be allocated?

WHAT IS ACCURACY?

Any discussion of the accuracy of a statistic has to begin with the question: accurate compared to what? Suppose that there existed an omniscient statistician, who knows about every crime that is committed (even the so-called “perfect crimes” that are undetected), and knows the hearts and intentions of every perpetrator and victim. From a statistical point of view, a crime rate calculated by this omniscient statistician is about as good as one can possibly have. We’ll call this rate the “true value.”

But the true value depends on what is defined to be a crime. The USA Today survey described at the beginning of the chapter included nine types of experiences as harassment or assault, ranging from “having someone make unwelcome sexual comments,
jokes or gestures about you” to “being forced to do a sexual act.” Different definitions would have led to different answers.

A crime rate statistic has a numerator and a denominator—for example, a violent crime rate might be reported as 382 violent crimes per 100,000 inhabitants of the area. The definition used for crime determines the numerator of a crime rate statistic. Crime rates also depend on who is included in the denominator. Are children included? Prison inmates? Nursing home residents? Members of the armed forces? Some sexual assault statistics consider both sexes; others consider only women; others consider only women who are attending a college or university.

The set of persons or entities to whom the statistics are intended to apply is called the population. We would expect statistics about different populations to differ.

**STATISTICAL PROPERTIES**

Even if the same crime definitions and populations are used, crime statistics from different sources or samples are expected to vary. Almost every statistic deviates from the true value it estimates, usually for one or more of the following reasons:

**Missing data.** Almost all data collections fail to obtain some data, and for crime statistics this problem is of particular concern because often the missing data belong to crime victims. Undetected murders (such as deaths mistakenly ascribed to natural causes) will cause homicide statistics to be too low. Robberies not reported to the police will be missing from the law enforcement statistics for that crime.

Some surveys ask people about criminal victimizations they have experienced. If persons who are willing to answer the survey questions are more likely to be crime victims than those who decline to participate, then the victimization rate estimated from the survey data may be too high. The survey estimates may be too low if persons who agree to participate in the survey are less likely to be crime victims than persons who are asked to be in the survey but do not take part.

**Measurement error.** Measurement error occurs when an entry in the data set differs from the true value. Misclassifying a robbery as a purse-snatching is an example of a measurement error in
police records. In surveys, measurement errors can occur because a question is worded confusingly or is misinterpreted, or because one interviewer might elicit a different response than another interviewer, or because the person responding to the survey does not tell the truth or has faulty memory.

Statisticians use the term “error” for anything that causes a statistic to deviate from its true value, but measurement errors should not be interpreted to mean “mistakes.” Rather, they should be viewed as sources of uncertainty about statistics. Some measurement errors are indeed the result of mistakes, as when someone types the wrong value in the database, but others can occur simply because people interpret a question in various ways.

Sampling variability. Some crime statistics come from randomly selected samples of households, persons, businesses, or records. The statistic calculated depends on the particular sample that was drawn. If a different sample had been drawn, a different value of the statistic would have been obtained, and that leads to variability from sample to sample.

If you have taken a statistics class, sampling variability is probably the type of error you learned about—and it is often the only measure of uncertainty that is reported for crime statistics. But effects of missing data and measurement errors should also be considered when interpreting a statistic.

WHAT THIS BOOK IS ABOUT

This book is about the statistical ideas needed to interpret statistics about crime rates: definitions of crime, populations, missing data, measurement error, and variability.

These factors affect all statistics. The two national sources of homicide statistics—one set obtained from death certificates and the other from law enforcement agency reports—show parallel trends over time but have different numbers of homicides. Chapter 2 outlines some of the reasons for these differences, including different definitions, missing data, and classification error.

Law enforcement agency statistics on crimes such as assault and burglary are also estimates. The Federal Bureau of Investigation (FBI) collects and tabulates statistics from US law enforcement agencies on violent and property crimes. The statistic on the back cover about violent crime decreasing by 0.9 percent from 2016
to 2017 comes from the FBI’s Uniform Crime Reporting System discussed in Chapter 3. The chapter describes crime classification errors and some of the statistical methods that can be used to measure and reduce them.

Of course, the FBI statistics include only crimes that are known to and recorded by the police. These statistics cannot tell us about crimes that are not reported to the police.

Surveys, however, can provide information on crimes not known to the police. They ask people about crimes that happened to them, and then ask whether they reported those crimes to the police. The US National Crime Victimization Survey (NCVS), the subject of Chapters 4 through 6, has surveyed US residents age 12 and older every year since 1973.

The NCVS is just one of many surveys that have been taken about crime. Some surveys give more accurate estimates than others. Chapter 5 explains why results from randomly selected samples can be generalized to a population and describes the procedure used to select the NCVS sample.

Chapter 6 describes the weighting methods used to try to compensate for missing data from persons who do not respond to a survey. It also discusses measurement errors in surveys: how do you ask questions and conduct the survey to elicit accurate responses?

Chapter 7 summarizes the statistical principles from the first six chapters, distilling them into eight questions that you can ask to assess the quality of any statistic you encounter. The remaining chapters of the book apply these ideas to two types of crime that are particularly challenging to measure (sexual assault and fraud), and look at how new data sources and procedures might improve crime statistics.

Sexual assault statistics from surveys depend heavily on what is counted as an assault, what questions are asked, and who asks the questions. Chapter 8 describes some of the experiments that have been conducted to study how different survey methods affect statistics about sexual assault.

Fraud and identity theft, the subjects of Chapter 9, were not included in the original list of crimes to be reported in the FBI statistics. In part because of this historical omission, the statistics available about fraud and identity theft are much sketchier than those about violent crime, burglary, and theft. Fraud presents a particular challenge because many victims are unaware they have been defrauded; others, such as persons in nursing homes, are excluded from many data collections.
Chapter 10 talks about the feasibility and challenges of using “big data,” the massive amounts of data now available from financial transactions, social media, and police operations, to learn more about crime. Chapter 11 concludes the book with a historical perspective of US crime statistics along with some ideas for continuing to improve them.

The glossary at the end of the book defines terms and acronyms, and the website http://www.sharonlohr.com contains endnotes and links to data sources.

Although most of the examples are from US crime statistics, the statistical concepts apply to any data collection. The same statistical issues of operational definitions, populations, missing data, measurement error, and sampling variability affect almost any statistic you encounter.

Okay, let’s get started. We explore homicide statistics in the next chapter.

SUMMARY

Crime statistics depend on how crime is defined, who provides data (and who does not provide data), how data are collected, what questions are asked, and how estimates are calculated.

A statistic can be thought of as

\[
\text{Statistic} = \text{“true value”} + \text{“deviation,”}
\]

where the deviation includes anything that causes the statistic to differ from the true value, either in a positive or negative direction. Statistics about crime rates can differ because the true values differ, the deviations differ, or both.

The true values depend on what crimes are measured, how the crimes are defined, and what populations are studied. Missing data, measurement error, and sampling variability cause a statistic to deviate from its true value.

Every statistic should be accompanied by a measure of uncertainty that assesses how close the statistic is likely to be to the true value.
FIGHTING SLAVERY THROUGH STATISTICS: A DISCUSSION OF FIVE PROMISING METHODS TO ESTIMATE PREVELANCE IN THE UNITED STATES

This article is excerpted from *Chance* by Davina P. Durgana and Paul L. Zador © 2017 Taylor & Francis Group. All rights reserved.
Fighting Slavery through Statistics: A Discussion of Five Promising Methods to Estimate Prevalence in the United States

Davina P. Durgana and Paul L. Zador

It is much easier to protect what one can count and identify. Modern slavery prevention and intervention efforts can become more effective and targeted with better information about how many victims exist and where they are likely to be found.

Measuring prevalence in modern slavery has historically presented several challenges. Definitional disagreements, lack of national coordination between all anti-slavery stakeholders and their data collection systems, and limited funding have played roles in the difficulties of obtaining a widely accepted and methodologically rigorous national prevalence estimate for human trafficking in the United States. Some global estimation efforts have included disaggregated national estimates for the U.S. alongside many other countries, but a more-focused and customized effort is required to estimate modern slavery in a country as populous, diverse, and critically nuanced as the U.S.

There are some concerns about the divisive repercussions of multiple, wide-ranging previous estimates in the U.S. and some of their particular foci on specific target populations, such as minors or foreign nationals, and industries such as commercial sexual exploitation of minors or agricultural labor.

As one of the first countries in the world to pass a comprehensive national anti-human trafficking law and one of the only countries in the world to publish its annual Trafficking in Persons Report through the Department of State, the U.S. has the opportunity to maintain its leadership and continue its progress on this critical human rights issue. One possible way forward is through more support for a national human trafficking estimate.

Modern slavery prevalence measurement in developed countries presents a consistent challenge to the existing survey-based methods of estimation, since the ineffective rule of law and generally low levels of extreme vulnerability among the populations in many of these countries make slavery harder to locate. However, other developed countries, such as the UK and the Netherlands, have managed to use alternative methods of obtaining national prevalence estimates. Unfortunately, if we only continue to study modern slavery in developing countries, where it may be traditionally easy to find through survey methods, we will mischaracterize the established and dominant role of stronger market economies and developed countries in this issue.

To better advocate for committed and sustained funding for anti-slavery efforts in the U.S., and to develop benchmarks and baselines for future research and progress measurement, it is critical to come together as a field and engage in collaborative efforts to determine national prevalence estimates. Success in the U.S. context would encourage many other developed countries facing similar constraints and contribute substantially to the information that exists about human trafficking in the United States, both domestically and from other countries.

There are three important reasons why better, and more-reproducible, slavery prevalence estimates are needed—to help:

1. Raise public and government awareness of the size of the problem.
2. Identify the complex resource needs of anti-slavery activities.

3. Assess the success or failure of specific anti-slavery activities.

High-level precision is particularly important for evaluating the effectiveness of anti-slavery activities that are targeted to a specific area before the activities are judged worthy of more general deployment. Such measures might include enhanced enforcement of existing laws; passage of specific new laws; vigorous information campaigns designed to educate people about how to identify, and then report, activities that could involve modern slavery. It is important to realize, however, that just because a measure “sounds good,” that does not prove its effectiveness. In fact, such measures might even prove to be counterproductive.

While there have been attempts to do this in the past, concerns have been raised that such studies focused only on specific populations, such as minors or foreign nationals, and on specific industries, such as commercial sexual exploitation of minors or agricultural labor. Some may argue that these industries are easier to identify or target, but that there are substantial issues related to networks such as domestic servitude and traveling sales crews that are more difficult to identify.

However, there are potential solutions to this. The following five methods are among the most promising opportunities to develop a national human trafficking prevalence estimate for the United States in the short to medium-term future. Some of these efforts have been executed successfully in other developed countries, and other pilot or smaller studies indicate that national-level replication could be possible in the near future.

1) Multiple Systems Estimation (MSE) has been conducted successfully in the United Kingdom in 2014 and in the Netherlands in 2016. This method primarily applies a form of capture-recapture from administrative data from multiple concurrent lists of identifiable victims to estimate a victim population size. Basically, this method involves analyzing three to four concurrent lists of identifiable victims for apparent overlap—incidences where a specific victim was named on one or more lists in the same time period. This helps statisticians estimate how likely it is that the known list of victims collected by these groups is identifying the same victims and who they may be missing.

Several U.S. stakeholders are interested in applying MSE to administrative data in the United States, but are currently hindered by multiple jurisdictional, confidentiality, and systems-based challenges. However, this method is generally considered to be the most reliable and ideally suited for developed countries, where generally lower vulnerability conditions may result in overall lower prevalence rates, while stronger law enforcement and rule of law makes modern slavery more difficult to detect.

These conditions render nationally representative household surveys, such as those commissioned by the Walk Free Foundation in partnership with Gallup World Poll, less useful in the United States context. While developing countries may have survey respondents who are more likely to be personally affected by modern slavery, this is very difficult to detect with the standard sample sizes of 1,000 to 2,000 people in developed countries. Given these limitations and the fact that many developed countries maintain reliable and identifiable victim data, MSE becomes an attractive option in that context.

2) Respondent Driven Sampling (RDS) was originally introduced by Douglas Heckathorn (2002) to provide sufficient statistical parameters for chain-referral methods to determine statistically valid indicators and to counteract issues of initial sample selection, volunteerism, referral methods, homophily or non-random referral, and potential overrepresentation by those with larger personal network biases.

Originally, RDS—a process where initial survey respondents are identified and then asked to identify other respondents who share similar characteristics, in an attempt to locate difficult-to-find populations—was challenged by a lack of statistical parameters that would allow statisticians to infer the total population from this nonrandom sample. However, Heckathorn’s contributions allowed researchers to start using this effective method of identifying respondents, as well as to statistically infer information about the total population.

Sheldon Zhang has championed this work in the anti-slavery field among migrant workers in San Diego and in North Carolina, and improved upon the referral methods to further reduce the biases that have traditionally reduced the statistical validity of such approaches. While promising, this type of chain-referral approach also requires incentives for participation and may be difficult and costly to replicate on a national scale.

3) Taking online-based surveys of U.S. stakeholders is an important and relatively low-cost possible method of obtaining more information on human trafficking prevalence data and related issues, but would be difficult to use for statistical inference. There have been several promising efforts, such as the pilot online survey of United Against Slavery. The organization plans to begin integrating questions into its survey design related to human trafficking data and prevalence.

It is also worthwhile to consider methods for influencing the general population’s understanding of,
interest in, and willingness to help eliminate contemporary slavery.

This could be done almost for free, using some form of social media, and at a relatively low cost with local advertising. Evaluating the impact of this type of information campaign would be relatively inexpensive, since traditional survey methods work with the general public.

4) National aggregation of statewide prevalence estimates is another possible method to determine human trafficking prevalence in the United States. A recent study from the University of Texas in Austin used available administrative data and targeted qualitative interviews in Houston, Texas, to estimate that 313,000 people are enslaved in Texas. This is a promising step for a large and populous state, but for this method to gain traction, other states will have to follow suit with their own statewide estimates.

It was noted earlier that prevalence estimates tend to be imprecise. As the following numbers show, “imprecise” is an understatement. Scaling that number of 313,000 modem slaves up to the world’s population, there would have to be about 86 million modern-day slaves on Earth. That number compares with the International Labor Organization’s global estimate of about 20 million and the Walk Free Foundation’s of about 46 million, of whom only 60,000 are in the United States.

Moreover, it is reasonable to assume that Texas has a below–world-average slavery rate. A recent study by Martin and Smith (2015) estimated that there were about 148,000 modern slaves in Texas, less than a half of the 313,000 estimate. These estimates were derived using different definitions of slavery, different data sources, and different analytic methods, and resulted in an extraordinarily wide range.

5) Gallup public opinion surveys on human trafficking awareness and willingness to report to hotlines and/or law enforcement is a potential method that has already yielded some compelling preliminary results. In early 2016, the Gallup World Poll included a set of items on the Gallup Daily Poll for the United States related to human trafficking awareness and the general public’s willingness to report human trafficking. This information is useful to the field due to the valuable, publicly accessible, and anonymous case and call information that is published online by the National Human Trafficking Hotline, as well as the importance of having a victim-centered and non-law-enforcement hotline known to survivors.

While the survey in 2016 indicated that many respondents would report a crime of human trafficking to law enforcement directly, an encouraging number of respondents were both aware of the National Human Trafficking Hotline and indicated that they would report crimes there. This method can be further developed alongside a more robust vulnerability model to try to determine the statistical validity of using the publicly available information on calls to the National Human Trafficking Hotline as an approximation of the incidence of human trafficking in the United States.

These efforts should also reflect the best practices in applied research and replication standards, including transparent and reproducible methods, as well as the development of theoretically grounded models. The development of a theoretically grounded model, in particular, is necessary to advance our understanding of vulnerabilities to human trafficking in the United States and around the world. To date, there has been moderate success to this effect by employing the United Nations–derived human security framework.

It is important to identify a large set of potential survey prevalence predictors for which local data are typically available—a challenge for slavery experts. A related, but different, version of this approach had been implemented by the Walk Free Foundation when it developed its global vulnerability model and prevalence estimates for more than 160 countries using the global slavery index, and also by Martin and Smith (2015), when they used regression analysis to estimate slavery statistics for the 50 U.S. states.

In general, showing the effectiveness of any intervention is complicated. It requires:

1. Conducting pre-intervention surveys in two or more reasonably similar areas.
2. Implementing the intervention in some of the area(s), but not in the “comparison” area(s).
3. Conducting post-intervention surveys in the same areas.
4. Comparing (average) pre-post change between intervention and comparison areas.

Surveys involved in such demonstration projects have to be well–designed. As noted earlier, RDSs, if they are carefully implemented, can produce unbiased estimates even for hard–to–reach hidden populations. In designing such studies, it is especially important to obtain final sample sizes that guarantee adequate power for detecting any meaningful change in slavery prevalence.

If it were possible to obtain the requisite number and type of victim lists from law enforcement agencies and/or from other organizations throughout the nation, MSE methods could also be used. The advantage of MSE over RDS is that MSE is simpler and costs less. The disadvantage is that MSE may not be able to provide all relevant details about the nature of local
enslavement. In contrast, protecting respondent privacy by adequate methods of data anonymization presents an added challenge for RDSs.

Traditional survey methods can be employed to assess the general population’s understanding of, interest in, and willingness to help eliminate contemporary slavery. However, given the rapid changes in public awareness of human trafficking and diversity of thought on vulnerability modeling to human trafficking, this approach will require much more academic and practitioner collaboration and discussion, as well as possibly an updated survey.

The U.S. federal government is making significant progress in evaluating the feasibility of MSE for the U.S. context and human trafficking data. Many stakeholder groups and nonprofits, such as the McCain Institute and United Against Slavery (UAS), continue to convene workshops and efforts to address these issues.

Unfortunately, statistically valid direct methods, such as MSE and RDS, are unlikely to be practical for estimating prevalence at the national level. However, we think that a promising estimation strategy could be constructed by regressing statistically valid prevalence estimates that were obtained for a representative sample of local areas on a set of potential predictors of slavery prevalence that are available at appropriate local levels, possibly including markers inspired by the human security theory.

This approach involves major challenges for statisticians, as well as for experts on modern slavery. Given that obtaining survey-based unbiased prevalence data is likely to be the most time-consuming and expensive component of the proposed approach, it is important to design a minimally adequate set of local areas for survey-based prevalence estimation. This is the statistician’s challenge.
On October 24th, 1929, the stock market crashed. Black Thursday slid into Black Tuesday when, on the 29th, the market crashed again. Panic spread.

President Herbert Hoover was not entirely convinced that things were so bad. His pet theory was that this was a blip in the economy and that soon—very soon—things would settle down and simply go back to normal.

To prove it, his administration began collecting information on unemployment. Hoover was certain this new data would silence his adversaries. At the end of only a single week, seeing a promising twitch in the statistics, he grandiosely declared his theory correct.15

Unfortunately for Hoover, things did not get better. The market crashes were omens for the Great Depression. Hoover’s inability to square his theory with the ground realities ultimately meant he lost the next election to Franklin D. Roosevelt in 1932.

Now, every month, the Bureau of Labor Statistics (BLS), a federal statistical agency, publishes the stiffly named “The Employment Situation.” It contains lots of statistics including the number of jobs created, the labor force participation rate, and the unemployment rate. It’s used as a description of the state of the economy. Markets move on the release of this information; political fortunes hinge on it.

The third statistic in the list above, the unemployment rate, is considered a key economic indicator. As firms lay off workers only if times are already bad, the unemployment rate is regarded as a lagging indicator. It reflects changes in the economy that are already occurring.
Provided you know what people do for a living, calculating the unemployment rate does not appear, at first glance, to be onerous. Basically, we have to know how many people there are and count how many of them are unemployed. Once we know those two numbers, we just divide them. (This is called a ratio statistic because we need to determine both parts of the fraction.)

Let’s start with who counts as unemployed. For example, say Ezra just moved to Ohio and is currently looking for a nursing job. Is he unemployed? Clearly, yes. How about Claire who just quit her job at a law firm to start a new one next month? She is technically not working today, but she doesn’t fit our mental picture of an unemployed person. Finally, what about Milo who is working part-time at a university but hopes to land a more permanent position? How should we count people like Milo who feel they are underemployed?

Suddenly we find ourselves bogged down by lots of details. Perhaps things will be simpler if we switch to counting the total number of people. We can start by restricting our tally to people who are adults, or close to adulthood. In modern times, we aren’t expecting most ten-year-old Ottos and Ethels to be toiling away in dank coal mines or clanking factories. People in school full-time should probably be left out as well since they are otherwise occupied. Possibly retired people too. As we pare down this list, we are slowly deciding who in the population should be part of the “labor force.”

WHO COUNTS?

It’s a steamy day in 1790. You’ve been galloping around town on your horse. You arrive at a house and knock on the door. It opens and you cautiously peer inside. Quill at the ready, you introduce yourself and begin tallying and cataloging the residents. This is your job as an enumerator for the U.S. census: counting who lives where and (presumably) with great care.

A census is required every ten years by the U.S. Constitution and so, is called the decennial census (“dec”=10). The goal is to literally count every person who lives in the country (not just citizens). This count is used primarily to determine how many representatives each state can send to Congress, a process called apportionment. (Enumerators were used extensively until 1960, the first year when all census forms were mailed to homes. That said, enumerators still picked up the completed forms from each household that year.)

The first census in 1790, however, had two extra objectives. First, the population count could be redeployed to “fairly” dis-
tribute expenses from the Revolutionary War against the British. It also contained an embryonic count of a specific type of labor force: Males who were free, white, and at least sixteen years old. That is, males who were eligible to serve in a fledgling military. In 1790, there were around 813,000 such males out of a population of just over 3.9 million free and enslaved people. (Age was deemed irrelevant for women, slaves, or non-white free people even though many worked and obviously slaves worked without pay.)

Many government officials felt this number was unexpectedly low because it didn’t square with previous population estimates. They even suspected that families pretended to contain fewer people when the enumerator came knocking to avoid being taxed. In the end, the count was accurate; rather, it was the colonial-era population estimates which were too high.

By 1800, many people were pushing for additional questions to be added to the census. They hoped the extra information would help the government pass better laws. For example, the fourth census in 1820 included the first questions about employment, classifying workers as in agriculture, commerce, or manufacturing.

As the years progressed, the instructions to enumerators became increasingly elaborate. For example, in 1870 enumerators were issued a stern warning: “The inquiry, ‘Profession, occupation, or trade,’ is one of the most important questions of this schedule. Make a study of it.” The government wanted them to be specific: “Instead of saying ‘packers,’ indicate whether you mean ‘pork packers’ or ‘crockery packers,’ or ‘mule packers.’” They even went so far as to insist doing some detective work if the resident was unhelpful.

By the time the 1890 census rolled around, enumerators were charged with distinguishing veterinary surgeons from the usual kind; chemists from metallurgists; actors from showmen. Even hucksters and peddlers would now be categorized based on their wares. And as the Industrial Revolution swept through, new jobs such as railroad officials or telegraph office messenger boys were added to the list of occupations.

All of these shifting and burgeoning categories were attempts to measure the absurd concept known as the “gainful worker.” The technical definition of a gainful worker was someone who had a paid profession they usually did. It didn’t matter if they were doing something else at the moment or hadn’t done that “usual” job in many years. This definition lumbered along until the Great Depression when, amidst staggering unemployment, reporting something you usually did was a demonstrably silly exercise.

And so the 1940 census switched gears. From gainful workers, people began talking about the “labor force.” This change in ter-
minology was accompanied by a clearer query: Were you working or looking for work last week? If you were giving piano lessons this week, even if you were trained to be an engineer, then you would be recorded as a music teacher. If you were looking for a new job this week even though you were a cashier last week, you would be listed as seeking work. In both cases, you would be regarded as part of the labor force. This was an easier question to answer.³

Eventually, a new statistic materialized: the labor force participation rate. This is the ratio of people in the labor force divided by everyone. (And then multiply by 100 to convert the fraction into a percentage.)

"Everyone" is defined as those who are sixteen or older, not in the military, and not institutionalized (e.g., nursing homes, prisons). This is the group of people who could be eligible to work in a (civilian) job. Basically, the "labor force" is everyone who has some kind of (again, civilian) job or has been looking for one in the past four weeks.

The labor force participation rate indicates what percent of the U.S. population is working or available to work. It is computed monthly from Current Population Survey (CPS) data. The graphs in Figure 3.1 show how this rate has changed since 1948.⁴ (The graphs in this figure are examples of time series data because time (day, month, year, etc.) is on the horizontal axis of the graph.)

Figure 3.1A displays two versions of the labor force participation rate. The black, spiky line is the original statistic. With this version of the measure, it can be difficult to interpret changes from one month to the next. A shift could have occurred because there was an underlying change in the economy. Or, it could be because it’s June and high school students are getting summer jobs at the pool or it’s November and shops are hiring temporary workers to accommodate the Christmas rush.

A statistical technique called seasonal adjustment removes the second type of fluctuation so we can focus on fundamental shifts to the labor force. It results in a smoother line like the blue one in Figure 3.1A, which is the seasonally adjusted labor force participation rate.

In Figure 3.1B, the blue line again represents the seasonally adjusted labor force participation rate. (It’s the exact same blue line as in the first plot, but the two plots have different vertical axis scales making it look like the lines are not the same.)

This rate has hovered between 60% and 65% since the late 1940s. There was a dip after the housing bubble burst in 2007, but it is slight if we take a longer-term perspective. (That said, even small reductions in the labor force participation rate can represent a large number of individuals.)
Figure 3.1 Plot A compares the choppier not seasonally adjusted monthly labor force participation rate (black) with the smoother seasonally adjusted one (blue). Plot B shows the same seasonally adjusted monthly rate (again in blue) along with the rate for men (gray) and women (black) separately. The blue lines in both plots are the same; however, the vertical axis scales don’t match, which is why they appear different in the graphs. The vertical line at 2007 marks the burst of the housing bubble. (Source: Current Population Survey.)
The roughly 40% of people who are not in the labor force include housewives and househusbands, students, and retirees who aren’t in nursing homes. It also includes people who would like a job but have given up looking for one (more on this later).

World War II brought many women into the workforce and plenty of them decided to stay once the war was over, permanently altering the workplace. We can see this trend from the other two lines in Figure 1.1B. Males are shown in gray and women in black. The rate of women in the labor force has steadily increased from around 30% to nearly double that in only 70 years. In fact, rates for men and women appear to be converging and will possibly match each other eventually.

When we group people who have a job and people who are looking for a job together, we have deferred the problem of what counts as a job. We turn to that next by looking at unemployment, a more hotly debated statistic.

**BUREAUCRATS ABOUND**

After experimenting with various questions about employment, queries about unemployment were added to the decennial census in 1880. The first results on this topic were published for the 1890 census, and occupational data was analyzed from every conceivable angle filling nearly a hundred pages with tables. One of those angles was race.

The 1890 census tables meticulously split whites into three categories: white with American parents, white with foreign parents, and white but foreign born. Everyone else was under the “colored” category with those of “Negro descent” counted as a subset. This example shows how choosing what data to collect and how to display it reflects the issues and prejudices of the day. The 1800s brought with them a lot of tension regarding immigration. In the middle of the century, many Irish escaped the potato famine by moving to the U.S., giving rise to the anti-Irish sentiments, among other ethnic groups. Federal immigration policy was beginning to be codified, starting with the 1875 Page Act which was expanded in 1882 with the Chinese Exclusion Act.

In subsequent censuses, the questions regarding unemployment changed substantially (including those about race). The minimum age drifted from 10 to 16 years. Unmarried daughters who helped around the home were reclassified as doing “housework—without pay” as opposed to having no occupation at all. All of these changes, large and small, made it difficult to compare unemployment rates from one census to the next.
Nowadays, the U.S. uses a consistent definition of unemployment and the labor force. Furthermore, it gathers this type of data frequently using a set data collection procedure. Getting to this stage shadowed the rising concern about the vast fortunes made in the Gilded Age while laborers worked under dire conditions in factories and mines. This apprehension induced many people from workers to government officials to act.

We start with Ethelbert Stewart who began his professional life as a lowly worker at the Decatur Coffin Company in Illinois. After being disgusted with how workers were treated in his factory, he began writing about it, hoping for change. Eventually, he became the head of the Bureau of Labor Statistics (BLS) in 1921 where he made his mark by helping to define and estimate unemployment.15

In the early 1920s, President Warren G. Harding, through then Secretary of Commerce (and later president) Herbert Hoover, organized a conference on unemployment. Everyone came away realizing that no one had a clue what the actual unemployment rate was. And so, the American Statistical Association set up the Committee on Governmental Labor Statistics to assist with wording the questions about joblessness for the 1930 census.6

By 1937, Roosevelt had unseated Hoover as president. Frances Perkins, a cabinet member, chose Isador Lubin to replace the retiring Stewart as the head of BLS because he would, “remember that statistics were not numbers but people coping or failing to cope with the buffetings of life.”15 He helped make the shift from “gainful worker” to “labor force” in the 1937 Census of Unemployment. To encourage people to participate in this census, Roosevelt spoke to Americans through the radio in one of his fireside chats.6

Finally, as part of Roosevelt’s Works Progress Administration (WPA), a systematic procedure was set up to collect unemployment data in 1940 and was called the Current Population Survey (CPS). Those at the WPA argued that taking a census—that is, asking everyone about their employment status—was too time-consuming and too expensive to do frequently.17 Therefore, only a subset of Americans (a sample) would be surveyed in the CPS instead of everyone in the country (a census).

The federal government still uses this survey today, although it is now administered by the Census Bureau on behalf of BLS. In the CPS, 60,000 households are in the sample at any given time. Households are selected to represent the entirety of the U.S., both rural and urban. For reliability reasons, each household is surveyed multiple times across a few months.

The people surveyed within each household are asked about their employment status, among other things. If they aren’t cur-
rently working, they are asked what types of things (if any) they have done to find a job within the previous four weeks.8

In addition to collecting data from households, BLS contacts 142,000 business and governmental agencies (who employ people too!) every month. As these organizations may have offices in multiple locations, the survey ends up including around 689,000 worksites around the country. This survey is called the Current Employment Statistics (CES) program.

The CES gathers information on wages and salaries, hours worked, and so forth, which supplements the unemployment and labor force information provided by CPS.3 All of this information is rolled into “The Employment Situation.”

Pheww. Now that we’ve reached modern times, let’s look at how these surveys help to produce the unemployment statistics.

POSSIBILITIES FOR U

Every month, people wait eagerly for the economic news contained in “The Employment Situation,” or, more colloquially, the monthly jobs report. While the report is lengthy, two numbers immediately ricochet around the media, the stock markets, and in politicians’ speeches: (a) the number of jobs added and (b) the unemployment rate. These numbers often turn into a mini-referendum on an administration’s policies.

To calculate these statistics, CPS sorts people who are 16 years or older who are neither in the military nor institutionalized (e.g., prison, nursing home) into six categories:

1. people working full-time,
2. people happy working part-time,
3. people working part-time but who would rather work full-time,
4. people not working, but have been looking for a job in the past four weeks,
5. people not working and would like to, but are not actively looking for a job (this includes people who quit looking for a job because they became discouraged), and
6. everyone else who is not looking for a job.

The unemployment rate is the number of people considered unemployed divided by the number of people in the labor force. There is not just one unemployment measure but six, charmingly named U1 to U6. Who counts as unemployed and who counts as the labor force depends on the categories above.8
U1 measures long-term unemployment. U2 tracks people who either just lost their job or finished a temporary one. That is, U2 zooms in on the newly-unemployed.

The official unemployment rate is U3. The labor force consists of the first four categories in the list, the fourth group representing the unemployed. (The labor force participation rate is calculated from this definition of the labor force.)

Moving from U3 to U6, we steadily absorb more people into the labor force and classify more people as unemployed. This is visible when comparing U3 and U6 in Figure 3.2. The boxes correspond to the six divisions of the population. The thick, black outline identifies the groups that comprise the labor force. Finally, the shaded boxes are the groups of people considered unemployed.

The most expansive definition of unemployment is U6, which includes everyone who would like a job but isn’t currently looking for one—at least not in the last four weeks—to the labor force. Moreover, who counts as unemployed is larger as well: those who want a job regardless of whether they are looking for one along with those who are working part-time but want a full-time job.

While U3 can be calculated back to 1948, U6 appeared only in 1994. (Various U-monikered statistics existed before 1994, but their definitions differed from those in use today.)

We can compare seasonally adjusted U3 and U6 statistics in Figure 3.3. (This is another example of a time series graph.) Remember that seasonal adjustment means we’ve smoothed out the bumps so that we can better see general trends over time.

The first feature which jumps out is that during its entire existence, U6 has always been higher than U3. (Actually, all six unemployment measures generally move in tandem.) It’s also not surprising both measures soared during each recession. What is interesting in this graph is that the gap between U3 and U6 widened substantially during the last recession. And so, the next section will focus on people counted as unemployed in U6, but not in U3.

**WHAT’S IN A NAME?**

Not all jobs are created equal. Some jobs require a lot of training, others do not. Some jobs have regular hours whereas others are assigned at the last minute. One effect of counting jobs is they are treated equally. Employed? Great. Unemployed? Not great. However, that obscures the variety of jobs which exist in an economy. Tomato pickers. Lawyers. Waiters. Teachers. Photographers.
Figure 3.2  The full rectangle represents everyone who is 16 or older, not in the military, and not institutionalized. These people are then divided into six categories which are the boxes within the rectangle. The U3 and U6 statistics drawn here are the most commonly reported unemployment rates, computed as the ratio of unemployed people to the size of the labor force. Here, the thick, black outline represents the labor force and the shaded box(es) represent the unemployed. (Source: Bureau of Labor Statistics.)
What's in a Name?

Figure 3.3  Monthly, seasonally adjusted U.S. unemployment rates using the U3 (blue) and U6 (black) definitions. The shaded areas represent the major recessions in the U.S.: after the 1973 and 1979 oil crises and the Great Recession. (Sources: Current Population Survey, National Bureau of Economic Research.)

Jobs have different remuneration schemes as well. From regular salaries to hourly wages. From zero-hour contracts to unpaid internships. Treating them the same way when computing labor statistics unfortunately papers over these differences. Whether or not people are making a living wage is irrelevant.

A primary division between types of jobs is between full-time and part-time work. For the latter, a key question is why they are only working part-time. In CPS, these people are split into various categories, three of which we will focus on here. First, some people decide to work fewer hours “by choice.” That could mean they are taking care of children or sick relatives, are in school, or simply want to work part-time. The second and third groups are comprised of those with part-time jobs, but would rather be working full-time. Some part-timers couldn’t find a full-time one whereas others have reduced hours because of “unfavorable business conditions.”

In Figure 3.4, we can compare these three groups from 1995 to the present. We are focusing only on people who do not work on a farm (i.e., non-agricultural) because such work is heavily seasonal, subject to other rules, and is therefore often its own category.
Figure 3.4 Three types of part-time, non-agricultural workers are graphed here: those who were part-time by choice (gray); those who work part-time because of forced reduced hours and other kinds of business conditions (black); and those who were unable to find full-time work (blue). The percentage in each category is graphed with the burst of the housing bubble in 2007 marked. This monthly data has been seasonally adjusted. (Source: Current Population Survey.)

Nearly 80% of part-time workers say they work fewer hours by choice. However, the most interesting aspect of this graph is what happened after the housing bubble burst in 2007. There was a sharp increase in the percentage of part-timers who had their hours cut (black line). It is possible (though we are speculating here) that these individuals were previously working full-time and suddenly found themselves working part-time instead. As the economy improved, this percentage slowly decreased.

This is one benefit of having multiple unemployment measures. U3 would count everyone in Figure 3.4 as employed, even though some of them may not be happy with their circumstances. U6, however, distinguishes between these types of part-time workers.

With the rise of the gig economy, the nature of employment is shifting. People who work at these types of jobs, such as a driver for Uber or someone on a temporary contract, are technically called “contingent workers.” We may need new types of U statistics in the future to learn more about these types of workers.
Industriousness is a celebrated quality in a capitalist society, as this line from Sandburg’s famous poem reflects. Accordingly, there are many ardent debates about why people don’t have a job. For starters, the Industrial Revolution began replacing many jobs done by humans with machines. On the optimistic side, economist John Maynard Keynes wrote in 1930 that “technological unemployment” was temporary. Eventually, he reasoned, it would be replaced by people working fewer hours and spending their free time on artistic pursuits. The chase for larger piles of money would end!

The other end of this discussion veered into murky areas of who ought to be working and why they were not doing so. A WPA employee observed that people often disparagingly claimed that, “Anyone who really wants a job can find one.” Others enquired why a person experienced what was uncharitably called a “period of idleness.”

The reasons for people being out of work are often more complicated than being lazy. Some possible explanations from the 1930 census include:

<table>
<thead>
<tr>
<th>Reason</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>voluntary lay-off</td>
<td>plant closed</td>
</tr>
<tr>
<td>personal disability</td>
<td>part-time</td>
</tr>
<tr>
<td>family reasons</td>
<td>substitute workers</td>
</tr>
<tr>
<td>weather conditions</td>
<td>machines introduced</td>
</tr>
<tr>
<td>breakdown</td>
<td>reduction of force</td>
</tr>
<tr>
<td>off-season</td>
<td>cheaper labor substituted</td>
</tr>
<tr>
<td>lack of orders</td>
<td>worker too old</td>
</tr>
<tr>
<td>job completed</td>
<td>laid off</td>
</tr>
<tr>
<td>shortage of materials</td>
<td>dissatisfaction</td>
</tr>
</tbody>
</table>

The length of this list is a good sign as it incorporates a wide range of plausible reasons for being unemployed. There is, however, a critical problem: The jobless person is self-diagnosing his reason
for being unemployed. Therefore, many of these choices are purely speculative, leaving ample room for inaccurate results. And so, by 1945, those probing questions were dropped from the surveys.\textsuperscript{6}

Now, people are only asked if they would like a job. If so, they are asked if they are “available to work” and whether they have given up looking for a job. The advantage these questions have is that they are clear. The downside (and this is a big downside) is that they provide little information about an important group.

The U3 statistic omits two groups of people from the labor force: those who aren’t looking for a job (e.g., retired, studying, disabled) and those who want a job but aren’t actively looking. We focus on the latter bunch in Figure 3.5.

Figure 1.5A shows that in 1995, around 10% of the people U3 drops from the labor force are those who want a job but aren’t actually searching for one. While that percentage has generally been falling, we can see an uptick during the Great Recession after the burst of the 2007 housing bubble.

We can look at a subset of this group more closely in Figure 1.5B. Here, those who gave up on their job search are graphed as a percentage of those who want a job. Before 2007, this statistic fluctuated around 7%, but after that point it shot up to 22%. Thanks to the subsequent recession, some of these people probably lost their job and then, disheartened, quit looking for a new one.

From a measurement perspective, these are people who stopped being counted in the U3 unemployment statistic once they gave up looking for a job. Handily, these individuals are then picked up in U6.

**AROUND THE WORLD**

The International Labour Organization (ILO) was launched as part of the Treaty of Versailles after World War I in 1919. It was eventually absorbed into the United Nations after World War II. The ILO has many missions. For example, it works to ensure people have living wages and to improve working conditions. It also publishes manuals to help countries produce their own official statistics.

The ILO proposes a statistic similar to U3, counting as unemployed only people at least 15 years old who are not working but looking for a job. Many countries have adopted a similar measure. The United Kingdom and Japan, for example, both use the 15 years and older threshold for their unemployment rates; the U.S., as we’ve seen, starts its count at 16 years. (The minimum age is 16 in the U.S. because of child labor laws and schooling requirements.)
Figure 3.5  Plot A shows the percentage of people who want a job among those not counted as part of the U3 labor force. Plot B shows the percentage of people who have given up looking for a job among those from the group in Plot A. The data is not seasonally adjusted and was collected monthly. The 2007 burst of the housing bubble is marked and was the precursor to the Great Recession. (Source: Current Population Survey.)
Unemployment Rate for People 15 and Over

<table>
<thead>
<tr>
<th>Year</th>
<th>China</th>
<th>India</th>
<th>Japan</th>
<th>United Kingdom</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>2%</td>
<td>2%</td>
<td>4%</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>2005</td>
<td>4%</td>
<td>4%</td>
<td>6%</td>
<td>8%</td>
<td>10%</td>
</tr>
<tr>
<td>2010</td>
<td>6%</td>
<td>6%</td>
<td>8%</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>2015</td>
<td>8%</td>
<td>8%</td>
<td>10%</td>
<td>12%</td>
<td>15%</td>
</tr>
</tbody>
</table>

These differences in definitions can make it hard to compare unemployment rates across countries. Consequently, the ILO attempts to create comparable statistics by adjusting annual unemployment rates reported from different countries using additional demographic information with statistical models.

We can see its estimates in Figure 3.6 for five countries: the U.S. (our comparison point), the United Kingdom, Japan, India, and China. The U.S. unemployment rate is the most rocky. In particular, the unemployment rate roughly doubles after the 2007 crisis in the U.S. and the United Kingdom compared to the Asian countries in our graph.¹⁴

These are very different countries, with very different governmental structures and social programs. Being unemployed in the U.S. differs from not having a job in the United Kingdom or in India. Labor conditions are different. So are labor protections. Moreover, for those who are counted as employed, we can’t tell whether those jobs pay a living wage, are temporary, or require specialized skills. We need to make sure we don’t read too much into cross-country comparisons.
SUMMARY

Widespread unemployment during the Great Depression spurred the federal government to track the unemployment rate on a regular basis. As a result, every month since 1948, the official unemployment statistic—or U3—is reported as part of the closely monitored “The Employment Situation.”

U3, one of six unemployment measures, is the ratio of the unemployed to the labor force and is considered a lagging economic indicator. Only people who do not have a job and have been looking for one in the past four weeks are counted as unemployed. The labor force participation rate is the ratio of people in the labor force to the total number of potential workers.

The various statistics in “The Employment Situation” are published by the Bureau of Labor Statistics using the Current Population Survey and the Current Employment Statistics program. These include the labor force participation rate and the number of jobs created, in addition to the unemployment rate.
IS THE GINI INDEX OF INEQUALITY OVERLY SENSITIVE TO CHANGES IN THE MIDDLE OF THE INCOME DISTRIBUTION?
Is the Gini Index of Inequality Overly Sensitive to Changes in the Middle of the Income Distribution?

Joseph L. Gastwirth

Department of Statistics, George Washington University, Washington, DC

ABSTRACT

The Gini index is the most commonly used measure of income inequality. Like any single summary measure of a set of data, it cannot capture all aspects that are of interest to researchers. One of its widely reported flaws is that it is supposed to be overly sensitive to changes in the middle of the distribution. By studying the effect of small transfers between households or an additional increment in income going to one member of the population on the value of the index, this claim is re-examined. It turns out that the difference in the rank order of donor and recipient is usually the most important factor determining the change in the Gini index due to the transfer, which implies that transfers from an upper income household to a low income household receive more weight that transfers involving the middle. Transfers between two middle-income households do affect a higher fraction of the population than other transfers but those transfers do not receive an excessive weight relative to other transfers because the difference in the ranks of donor and recipient is smaller than the corresponding difference in other transfers. Thus, progressive transfers between two households in the middle of the distribution changes the Gini index less than a transfer of the same amount from an upper income household to a lower income household. Similarly, the effect on the Gini index when a household in either tail of the distribution receives an additional increment is larger than when a middle-income household receives it. Contrary to much of the literature, these results indicate that the Gini index is not overly sensitive to changes in the middle of the distribution. Indeed, it is more sensitive to changes in the lower and upper parts of the distribution than in the middle.

1. Introduction

In his seminal article developing the relationship between measures of income inequality and an underlying social welfare function, Atkinson (1970, p. 255) noted that the Gini index was one of three measures that were sensitive to transfers at all income levels. After analyzing the effect of an “infinitesimal” transfer of income from a household to one with lower income, he concluded (on p. 256) that the Gini index gives more weight to transfers in the center of the distribution than at the tails, that is, the Gini index attaches more weight to transfers affecting the middle class. Alison (1978) provided a thorough review of the properties of several commonly used measures of inequality. Examining the effect of a transfer of an amount from the household with the jth largest income to the ith largest, he concludes that for a typically shaped income distribution the Gini index tends to be most sensitive to transfers around the middle of the distribution and the least sensitive to transfers among the very rich and very poor. Jasso (1979) pointed out that the formula for the effect of a transfer in Alison (1978) did not consider the possibility that a transfer would change the order, so the formula is only valid when the rank-order of the households is unaffected by the transfer (1979). Neither Jasso (1979) nor Alison (1979) comment on the effect of this on the conclusion that the Gini index is most sensitive to transfers in the middle of the distribution. Since the 1970s many authors (Ahn 1997; Jones and Weinberg 2000; Madden 2000 p. 76; Borghi 2005; DeMaio 2007; Callan and Keane (2009); Cobham and Sumner 2013; OECD 2013; Pressman 2013; Schmid and Stein 2013; Chang 2014; Bird and Zolt 2015; Thewissen et al. 2015) have noted the sensitivity of the Gini index to changes in the middle of the distribution. Similarly, Jenkins (2009) and Pak et al. (2016) refer to the Gini index being sensitive to income differences around the mode. Others state that it is relatively insensitive to changes in the top and bottom part of the distribution and more sensitive to changes in the middle (Roberts and Willits 2015) or more sensitive to changes in the middle than in the lower or higher tails of the distribution (Williams and Doessel 2006). Referring to the Gini index, Green et al. (1994, p. 59) wrote “An increase or decrease in the middle of the distribution will have a greater impact on the index than a similar change at either end, since there are more earners in the middle ranks,” Krozer (2015), Cobham et al. (2015), and the Wikipedia (n.d.) entry on income inequality metrics described it as being overly sensitive to changes in the middle of the distribution. The guide by the Australian Bureau of Statistics (2015) states “The Gini coefficient is sometimes criticized as being too sensitive to
relative changes around the middle of the income distribution. This sensitivity arises because the derivation of the Gini coefficient reflects the ranking of the population, and ranking is most likely to change at the densest part of the income distribution, which is likely to be around the middle of the distribution.” In their analysis of the growth in inequality in France, Fremeaux and Piketty (2014) observed, “the increase in income inequality during the 2000s is sharper for indices sensitive to the middle of the distribution like the Gini coefficient.”

Recently, Aaron (2015) and Gale et al. (2015) reached different conclusions about the effect of increasing taxes on the rich on income inequality. Gale et al. (2015) calculated that an increase in the top tax bracket from 39.6% to 50% would only lower the Gini index from about 0.560 to 0.556. Aaron (2015) argued that the Gini index is relatively insensitive to changes at the top and bottom of the distribution and that the ratio of the incomes of taxpayers at the 99th percentile to those at the 10th percentile was preferable to assess the impact of a tax increase on the very rich. He also noted that both sides agreed that the Gini index is more sensitive to changes in the middle of the distribution.

This article re-examines this claim and shows that it is incorrect. Consider a typical income distribution with a density that first increases, reaches its mode, and then decreases. The effect of a change in the distribution due to a transfer or addition to the income of one household will depend on whether it preserves or changes the order or ranks of the households, the difference in the ranks of the donor and recipient of a transfer, and who receives the additional income. In the case of a mean preserving transfer, one from the highest income recipient to the lowest has a larger impact on the value of the Gini index than a transfer of the same amount from any other household to one with less income. Transfers in the middle of the distribution, especially around the mode, do change the relative ranking of a higher proportion of the population. This does not necessarily imply that such a transfer has the largest impact on the numerical value of the Gini index because the change in its numerator also depends on the difference in the ranks of the donor and recipient. When one household receives an additional amount of income and the incomes of all others are unchanged; the average income (the denominator of the Gini index) is slightly increased. The number of households the recipient passes over to reach its higher rank does affect the numerator of the Gini index but in a manner that gives slightly less weight to changes in the middle.

The transfers of primary interest in economics obey the Pigou–Dalton criteria (Thon and Wallace 2004), which states that transfers from a poorer to a richer household increase inequality, while transfers from a richer household to a poorer one that does not reverse their relative ranking decreases inequality. Although the article will emphasize these transfers, in a few illustrative examples this principle will not hold.

Section 2 reviews several useful representations of the Gini index. One of them will be used in Section 3 to examine the effect on the Gini index of a small order-preserving transfer, e.g., an infinitesimal transfer from a high income household to one with less income, discussed by Atkinson (1970) or a small additional increment given to one household (Hoffman 2001). In the first situation, the effect of a transfer depends on the difference between the rank of the donor and the rank of the recipient; the largest decrease in the Gini index occurring when the highest ranked household transfers the small amount to the lowest ranked. The effect of one household receiving a small increment in income, which does not change the rank-order, depends on the rank of the recipient and again the Gini index decreases (increases) most when the poorest (richest) household receives the small increment. Section 4 examines the effect of a transfer of income that does not preserve the order, for example, the recipient of the transfer now has more income than several households whose incomes previously were greater than the recipient was and the rank of the donor might decline. In this situation, the rank order of a larger fraction of the population around a middle-income recipient or donor will change when that household is involved in the transfer; however, the change in the Gini index also depends on the difference in the ranks of the donor and recipient. The relative weight of these two components depends on the magnitude of this difference in ranks. When both the recipient and donor are in the middle of the distribution, the component of the change in the Gini index due to the number of households affected is relatively more important than the difference in the ranks, which should be small. When the donor is in the upper income region, the difference in after transfer ranks will be more important. The trade-off in the relative weights of the two components of change in the value of the Gini index indicates that the Gini index reflects the effect of non-order preserving transfers in various parts of the distribution and is not overly sensitive to those involving the middle. Section 5 focuses on how the Gini index changes when there is a small increase in the income of one household. It turns out that the Gini index decreases (increases) the most when the lowest (highest) income household receives the increment, while the magnitude of the change is smaller when the recipient is in the middle of the distribution. The final section discusses the implications of the results.

The Gini index has also been criticized (Hesse 2016) because different income distributions can have Gini indices with the same value. Newly proposed indices based on the ratio the share of total income of a fixed percent, for example, 10 or 20, at the upper end of the distribution to the share of income of a fixed percent, for example, 40 or 20 have been used by Palma (2011) and Dorling (2014) in studies of inequality. The Appendix shows that this family of indices can also have the same numerical values for different underlying distributions.

2. Formulas for the Gini Index

The Gini index of an income distribution, $F(x)$, is the ratio of the area between the line of equality and the Lorenz curve; however, it will be more convenient to use the fact that it is the ratio of the mean difference ($\Delta$) to twice the mean ($\mu$). The mean difference, $\Delta$, of the distribution is the expected absolute difference of two independent observations from $F(x)$. The empiric estimate ($d$) of $\Delta$ is the average of the absolute value of the differences in all pairs of incomes. Formally, given a sample $x_1, \ldots, x_n$ of observations from $F(x)$ it is defined (Kendall and Stuart 1977, p. 46) as

$$d = \left(\frac{1}{n(n-1)}\right) \sum_{i,j=1, i \neq j}^{n} |x_i - x_j|.$$

Sometimes $n(n-1)$ is replaced by $n^2$, but formula (1), which includes comparisons of an observation with itself, is commonly
used to estimate $\Delta$ (Sudheesh and Dewan 2013). The Gini index, $G$, of the distribution underlying the data is $\Delta/2\mu$. Denoting the sample mean by $\bar{x}$, the estimate of $G$ is $g = d/(2\bar{x})$.

Thon (1982) reviewed several expressions for $G$ and $g$. To assess the impact of transfers and changes, a convenient form first orders the observations by their size, that is, $x_1 < x_2 < \cdots < x_n$ so

$$g = \left[ \frac{\sum_{i=1}^{n} (2i - n - 1)x_i}{\bar{x}n(n-1)} \right].$$  \hspace{1cm} (2)

The numerator of (2) is one-half the $d$ (David 1968; David and Nagaraja 2003) and gives weight $(2i - n - 1)$ to the $i$th observation. For the Gini index the largest observation, $x_n$, receives weight $(n - 1)/\bar{x}n(n - 1)$ and the smallest receives weight $-1/\bar{x}n(n - 1)$. Indeed, starting from the smallest, the weight given to each successive order statistic increases by $2/\bar{x}n(n - 1)$. If $n = 2m - 1$, so the median is the $m$th largest observation, it receives weight zero. When $n = 2m$, the median is the average of $x_m$ and $x_{m+1}$ and the numerator of (2) gives weight $-1/\bar{x}n(n - 1)$ to $x_m$ and $+1/\bar{x}n(n - 1)$ to $x_{m+1}$. Thus, the numerator of $g$ gives more weight to both extremes and less weight to the observations in the middle observations. The mean in the denominator weights each observation equally, so the relative weight given to each of the ordered observations increases with their distance from the median. Yitzhaki (2003) and Yitzhaki and Schectman (2013) discussed the properties and a wide variety of statistical applications of the mean difference and Gini index.

3. The Effect of a Small Order Preserving Transfer or Increment on the Gini Index

3.1. Transfers to an Individual

Following Atkinson’s (1970) examination of the effect of an “infinitesimal” transfer from a higher income recipient to a lower one, consider the effect of a small transfer of $\epsilon$ from $x_i$ to $x_j$, where $x_i < x_j$, which is not large enough to change the order of the observations. The mean, $\bar{x}$, is unchanged and only the terms in the numerator of (2) that will change involve $x_i$ and $x_j$. Thus, the change in the numerator is

$$\left(2j - n - 1\right)[(x_j - \epsilon) - x_j] - \left(2i - n - 1\right)[(x_i + \epsilon) - x_i]$$

$$= -2(j - i)\epsilon$$ \hspace{1cm} (3)

and $g$ changes by $-2(j - i)\epsilon/\bar{x}n(n - 1)$. In particular, the decrease in $g$ due to a transfer of $\epsilon$ from any observation to the one immediately below, that is, $j - i = 1$ is $-2\epsilon/\bar{x}n(n - 1)$; so the effect of such a transfer is the same throughout the distribution. From (3), it is clear that the largest decrease occurs when the highest income recipient transfers $\epsilon$ to the lowest and the magnitude of the change depends only on $\epsilon$ and twice the difference in the ranks of the donor and recipient. An order-preserving transfer from the highest income holder to the median receives one-half the weight as a transfer to the poorest. These considerations demonstrate that the Gini index is not overly sensitive to small order preserving transfers in the middle of the distribution as transfers to or from the middle do not have as large an effect on $g$ as transfers from the upper end to the lower end of the distribution.

Comment: Schmid (1991) studied the sensitivity of a variety of indices of inequality to small transfers that preserve the rank order. In large samples his result is equivalent to (3), where $j/n$ and $i/n$ approach the fractions $\beta$ and $\alpha$, respectively. Interestingly, the sensitivity of the coefficient of variation for the Theil-Atkinson family depends on the population quantiles, $F^1(\beta)$ and $F^1(\alpha)$ rather than on $\beta$ and $\alpha$.

3.2. Increase to a Single Unit

Next, consider the situation where one unit receives an order preserving increase of size $\epsilon$, for example, the $j$th so $x_j$ becomes $x_j + \epsilon$ and the overall mean becomes $n\bar{x} + \epsilon/n$, where $\bar{x}$ is the mean of the original observations. From formula (2), it follows that the Gini index ($g_{1i}$) of the new data is

$$g_{1i} = \frac{\sum_{i=1}^{n}(2i - n - 1)x_i + \epsilon(2j - n - 1)}{(n - 1)(n\bar{x} + \epsilon)}.$$ \hspace{1cm} (4)

The part of formula (4) that depends primarily on $j$ is the second term in the numerator, which increases linearly in $j$. When $j = 1$, that is, the smallest additional income goes to the poorest household, the numerator decreases by $(n - 1)\epsilon$. When $n = 2j - 1$, the $j$th ranked member of the population is the median, and as a result the contribution to the numerator is zero, and when $j = n$, that is, the richest household receives the additional income, the numerator increases by $(n - 1)\epsilon$. Since the increase in the denominator of (4) is the same regardless of which household receives the increase, the largest decline in the Gini index occurs when $j = 1$, that is, the poorest household receives the additional income because the numerator decreases the most and the denominator increased. When the median household receives the additional income, the Gini decreases slightly because the mean (the denominator) has slightly increased. Furthermore, considering formula (4) as a function of epsilon, routine calculation shows that for $j > (n + 1)/2$ the increase in the numerator has a greater effect than the increase in denominator, so the largest increase in the Gini index occurs when the additional increment goes to the top-ranked household.

3.3. The Subset of the Population Whose Receipt of an Increment Decreases the Gini Index

It is interesting to determine the percentage of households who can receive an order preserving increment ($\epsilon$) that leads to a decrease in the Gini index of the population. The first step is to calculate the difference between $g_{1i}$ and the original $g$. From (2), it follows that $\sum_{i=1}^{n}(2i - n - 1)x_i = g\cdot n\bar{x}(n - 1)$. Substitution in (4), yields $g_{1i} = g\cdot n\bar{x}(n - 1) + \epsilon(2j - n - 1)$, which implies that

$$g_{1i} - g = \frac{\epsilon(2j - n - 1 - \epsilon(n - 1))g}{(n - 1)n\bar{x} + (n - 1)\epsilon}$$

$$= \frac{\epsilon}{(2j - n - 1)n\bar{x} + (n - 1)\epsilon} - 1 + \frac{n - 1}{(2j - n - 1)n\bar{x} + (n - 1)\epsilon}. \hspace{1cm} (5)$$
Thus, when the \( j \)th ranked household receives the additional small increment, \( g_j < g \) or inequality decreases when
\[
2j - n - 1 < g(n - 1) - 1 < (1 + g)n + (1 - g))/2
\]
and \( g_j > g \) or inequality increases if
\[
2j - n - 1 > g(n - 1) - 1 < (1 + g)n + (1 - g))/2.
\]

When \( n \) is large, Equation (5a) implies that inequality will still be reduced as long as \( j/n \) is less than \((1 + g)/2\), that is, the recipient (\( j \)th ranked) is at or below the \((50 + 50)g\)th percentile. When the original Gini index is \(0.50\) (0.30), this implies that as long as the recipient is in the lower \(75\%\) (65\%) of the distribution, inequality as measured by the index will decrease. On the other hand, if the small increment goes to a household in the upper \(25\%\) (35\%), inequality will increase by an amount that increases with the rank of the recipient. For any value, \( g \), of the original Gini index, as expected the largest increase in inequality occurs when the household with the largest income receives the small increment and the greatest decrease occurs when the household with the lowest income receives it. Regardless of the original value of the Gini index, it will decrease as long as a household in the lower half of the distribution receives it; however, the magnitude of the decrease is greatest when a household at the lower end receive the increment.

### Comment
The result that in large samples the Gini index will decrease as long as the original percentile of the recipient is below the \((50 + 50)g\)th appears in Hoffman (2001, their sec. 2), who considered the effect of an “infinitesimal increment” that does not change the income order. Later, Hoffman presented a similar formula (p. 245, (16)) for the effect of a small increase given to the \( j \)th ranked household in a sample of \( n \) from a continuous distribution. Letting \( j/n \) approach \( a \) as \( n \) increases in the second formula in (5), that is, the recipient is at the \(100a\)th percentile of the distribution, the change in the Gini index is
\[
e \frac{2a - (1 + G)}{n}.
\]

Formula (6) is equivalent to Hoffman’s formula in large samples (there is a minor typo in his formula as the denominator should be \( n \) times the mean). In place of \( j/n + 1 \) in (5) or its limit \( a \) in (6) Hoffman considered \( F(x_j) \) and concludes that the relationship between the change in the Gini index and \( x_j \) depends on the form of the distribution while (6) does not depend on the underlying distribution. Because \( x_j \) is the \( j \)th-order statistic in the sample of \( n \), the sampling distribution of \( F(x_j) \) is the same as the \( j \)th-order statistic in a sample of \( n \) from the uniform distribution (David and Nagaraja 2003, p. 14), so its expected value is \( j/(n + 1) \), which converges to \( a \) as \( n \) increases. As the empiric percentiles of a random sample from a uniform distribution are consistent estimators of the population percentiles, in large samples the formulas are equivalent. However, the change in the Gini index depends on the original percentile of the recipient, rather than the form of the underlying distribution.

Although the analysis leading to formulas (4) and (5) does not require the order preserving transfer or increment to be small, in practice, the size \( n \) of the population will be reasonably large so the order-preserving requirement restricts the possible magnitude of the transfer or increment. Thus, the results in this section are applicable to the “infinitesimal” transfers considered by Atkinson (1970) and the “small ones” discussed by Allison (1978). In the order-preserving context, neither transfers or an addition to one member of the population in the middle receive excessive weight compared to other parts of the distribution, so the Gini index is not especially sensitive to these types of change in the middle part of the distribution.

### 4. The Effect of a Transfer that Changes the Ranks While Preserving the Mean
Consider the case where the amount, \( a \), transferred by the \( j \)th highest income recipient to the \( i \)th, where \( i < j \), is sufficient to change the ordering of the \( n \) incomes. This means that either \( x_i + a \) is larger than some of the observations, \( x_r \), where \( r > i \) or \( x_j - a \) is now smaller than some of the observations \( x_r \) that previously were below \( x_j \), or both occur.

#### 4.1. The Rank of the Recipient Increases While the Rank of the Donor is Unchanged
In this case, the income of the recipient now equals \( x_i + a \), which becomes the \( k \)th largest. Thus, the \( k - i \), observations, \( x_r \), \( r = i + 1, \ldots, k \) that were larger than \( x_i \) were less than \( x_i + a \). After the transfer, the ranks of each of these \( k - i \) observations decrease by one, so each of their contributions to the numerator of \( g \) decreases by \( 2x_r \). The contribution of \( x_i + a \), which is now the \( k \)th largest observation is \( (2k - n - 1)(x_i + a) \), however, this replaces its previous value \( (2i - n - 1)x_i \), so the transfer increases the contribution of \( x_i \) by \( (2k - n - 1)(x_i + a) - (2i - n - 1)x_i = 2(k - i)(x_i + a) - 2(k - n - 1) \). Thus, the contribution of the first \( k \) observations to the numerator has changed by

\[
- \sum_{r=i+1}^{k} 2x_r + 2(k - i)x_i + a(2k - n - 1)
\]

\[
= - \sum_{r=i+1}^{k} 2(x_r - x_i) + a(2k - n - 1).
\]

If the rank \( j \) of the donor is unchanged, its contribution to the numerator of \( g \) is reduced by \( (2j - n - 1)a \), so the net change in the numerator is

\[
- \sum_{r=i+1}^{k} 2(x_r - x_i) - 2a(j - k).
\]

As there are \( k - i \) terms in the summand, adding and subtracting \( 2a(k - i) \) implies that the change in the numerator of the Gini index is

\[
2 \sum_{r=i+1}^{k} [a - (x_r - x_i)] - 2a(j - i).
\]

It follows that the after transfer Gini index is related to the original one \( g \) by

\[
g_t = g - \frac{2}{n(n - 1)} \sum_{r=i+1}^{k} (x_r - (x_i + a)) + j - i a.
\]

Formulas (8a) and (8b) show that the effect of a transfer of size \( a \) from the \( j \)th ranked household that increases the \( i \)th
ranked household to rank $k$, where $i < k < j$ but does not change the rank of the donor, depends on:

1. The size ($a$) of the transfer.
2. The difference ($j - i$) between the original ranks of the donor and recipient and
3. The number ($k - i$) of households in the interval $[x_i, x_j]$.

The incomes, $x_r$, $r = i+1, \ldots, k1$ of these $k - i$ households satisfy $x_i < x_r < x_j + a$, that is, $x_r - x_i < a$ implying that each term in the summand in (8a) is positive but less than $a$, so the summand in (8a) is less than $a(k - i)$. Thus, expression (8a) is negative whenever the Pigou–Dalton condition holds (Thon and Wallace 2004), that is, after the transfer the rank of the donor remains greater than the rank of the recipient. Consequently, $(j - i) > (k - i)$ and (8b) implies that the Gini index decreases.

Consider the implications of these results when a high-income donor transfers money to a low-income recipient. Because the density function of most income distributions is relatively small in the tails of the distribution, the number ($k - i$) of terms in the summand will be less than the number of terms involved when the recipient is in the central part of the distribution. This means that the difference $(j - i)$ in the ranks of the donor and recipient is much larger than the difference $(k - i)$ in the before and after transfer ranks of the recipient. Thus, the positive summand is much smaller than the absolute value of the second term in (8a), so the term in (8b) reflecting the effect of the transfer will be negative and the Gini index will decrease.

The density function underlying most income distributions has a mode near or below median, so a transfer from a high-income donor to a recipient (4th household) in the middle changes the ranks of a larger fraction of the population than when the recipient is in the lower part of the distribution. This implies that the number of terms in the summand in (8a) and (8b) is larger when the recipient is in the middle of the distribution than when it is in the lower portion of the distribution. Thus, the contribution of the summand in (8a), which is positive, should be larger when the recipient is near the middle of the distribution than when it is in the lower end. This first term, however, is offset by the negative term $-2a(j - i)$, which reflects the difference between the original rank of the donor and recipient. This term is smaller when the recipient is in the middle; therefore, the decrease in the numerator of the Gini index as well as the Gini index should be less when the recipient is in the central portion than when the recipient is in the lower tail of the distribution. Thus, the Gini index is more sensitive to transfers from a wealthier household to one in the lower portion of the distribution than it is when the recipient is in the middle of the distribution.

When the transfer is between two households in the central portion of the distribution the summand in Equations (8a) and (8b) will increase because the number of households $(k - i)$ who had somewhat higher income than the recipient did but now have less will be larger. However, the term $2a(j - i)$ will be smaller than it is when the donor is in the upper income range as the rank $j$ of the donor now is in the middle. Thus, the net effect of a transfer between two middle-income households decreases the Gini index by an amount that is less than a transfer from a high-income donor to a middle-income recipient. The Gini index will decrease even more; however, when the recipient is in the lower part of the distribution as the density function is lower than it is in the middle so the lower income recipient will pass over fewer households than when the recipient is in the middle part. This implies that the first term in (8a) and (8b) will be smaller in this case than when the recipient is in the middle. Thus, transfers from an upper or middle income household to a middle-income one does not affect the Gini index more than a transfer of the same size from the same donor to a low income household. In fact, the index is more sensitive to transfers to the lower end than it is to transfers involving the middle.

Comment: For small transfers obeying the Pigou–Dalton condition reasonable approximations to (8b) provide insight into the trade-off between the value of the density function of the income distribution when the $i$th ranked household is the recipient, that is, $f(x_i)$, and the difference between the original ranks of donor and recipient. When $a$ is small, the probability an observation falls in $[x_i, x_i + a]$ is approximately $f(x_i) a$. Assuming the density function is uniform in this small interval implies that the expected number, $k - i$, of incomes in the interval is $n f(x_i) a$. This assumption also implies that the average value of the $k - i$ terms, $(x_i - (x_i + a))$ is approximately $-a/2$. Thus, the second term in (8b) is approximately

$$-rac{2}{n(n-1)\bar{x}} \left(-nf(x_i)\frac{a^2}{2} + (j-i)a\right)$$

$$= \frac{2}{(n-1)\bar{x}} \left(f(x_i)\frac{a^2}{2} - \left(\frac{j-i}{n}\right)a\right).$$

The term $(\frac{j-i}{n})$ in (8c) reflects the difference between the original percentiles of donor and recipient. The values of the density function $f(x)$ of the income distribution are quite small, e.g., in 2014, the income interval [20,000, 24,999] contained 5.51% of the U.S. households. Assuming the density is nearly uniform in this interval implies that the density is approximately 0.000011, or slightly over $10^{-5}$. For small transfers, $a$, the absolute value of the term $(\frac{j-i}{n})a$ will exceed the term involving the density; especially when the transfer is from a high income donor to a low-income recipient. Furthermore,

$$(\frac{j-i}{n}) = \int_{x_i}^{x_j} f(t)dt = f(t^*)[x_j - x_i] \int_{x_i}^{x_i} f(t)dt$$

where $t^*$ is in $[x_i, x_j]$. (8d)

The second equality follows from the mean value theorem for integrals. The Pigou–Dalton criteria implies that $x_j - x_i \geq 2a$, which implies that $(\frac{j-i}{n})a \geq 2f(t^*)a^2$. For small progressive transfers between middle-income households, the approximate decrease (8c) in the Gini index is greater or equal to

$$\frac{1}{(n-1)\bar{x}} \left(f(x_i)\frac{a^2}{2} - 4f(t^*)a^2\right).$$

In the region around the mode, typical income density functions do not decline very rapidly as $x$ moves away from the mode. For example, in the 2014 U.S. income data, incomes in the interval [15,000, 19,999] just below the modal interval contained 5.44% of the households, while the interval [25,000, 29,999] contained 5.11%. Thus, $4f(t^*)$ should exceed $f(x_i)$. Indeed, the first intervals of length 5000, that contains about one-half or one-fourth the proportion of households as the modal interval are [75,000, 79,999] and [120,000, 124,999], respectively. For a donor in one of these intervals and a recipient in the modal region, the value of $x_j - x_i$ is much larger than $2a$, so (8e) does
not approximate the decrease \((8c)\) in the Gini index resulting from such a transfer.

### 4.2. Transfers that Increase the Rank of the Recipient and Decrease the Rank of the Donor

Next, assume that the size \((a)\) of the transfer from the \(j\)th ranked household in the original distribution is sufficiently large that its rank after the transfer will decrease to \(t\), where \(t < j\); again the recipient’s rank increases from \(i\) to \(k\), where \(k < t\). This means that the observations \(x_r\), where \(r = t, t + 1, \ldots, j - 1\), are less than \(x_j\) but larger than \(x_i\). The contributions of each of these observations to the numerator will increase by \(2x_r\), while contribution of \(x_j\) will change by \((2t - n - 1)(x_j - a) - (2j - n - 1)x_j\). This causes the numerator of the Gini index of the new data to differ from the numerator of the Gini index of the original data by

\[
2 \sum_{r=t}^{j-1} (x_r - x_j) - a(2t - n - 1) = -2 \sum_{r=t}^{j-1} (x_j - x_r) - a(2t - n - 1). \tag{9}
\]

Thus, when the transfer of the amount \(a\) from the \(j\)th ranked to the \(i\)th ranked household increases the rank of the recipient to \(k\), while the donor’s rank is reduced to \(t\), it follows from \((7)\) and \((9)\) that the Gini index, \(g_1\), of the new data is

\[
g_1 = \frac{\bar{x}n(n-1) - \sum_{k=i+1}^{k} 2(x_i - x_k) - 2 \sum_{j=1}^{j-1} (x_j - x_i) - 2a(t - k)}{\bar{x}n(n-1)}. \tag{10}
\]

Equivalently,

\[
g_1 = g - \frac{2}{n(n-1)\bar{x}} \left( \sum_{k=i+1}^{k} [x_i - (x_k + a)] + \sum_{r=t+1}^{j-1} [(x_j - a) - x_r] + (j - i)a \right). \tag{10a}
\]

When the after transfer rank of the donor is larger than that of the recipient, the first two terms in the parenthesis of \((10a)\) are negative and would increase the Gini index; however, their effect is offset by the positive third term, which decreases the Gini index. Recalling that the terms \(x_i - x_r, r = i+1, \ldots, k\) and \(x_j - x_r, r = t, \ldots, j - 1\) are less than \(a\), it follows that

\[
\sum_{r=i+1}^{k} 2(x_i - x_r) < 2a(k - i) \quad \text{and} \quad 2 \sum_{r=t}^{j-1} (x_j - x_r) < 2a(j - t). \tag{11}
\]

Hence, the magnitude of the contribution of the first two of the three terms in the parentheses in \((10a)\) is less than \(2a\) times the number of households not involved in the transfer whose ranks changed by one. This number, \(k - i + j - t = (j - i) - (t - k)\) is less than \((j - i)\) since \(t > k\). Thus, the Gini index decreases for transfers from a richer to poorer household that satisfy the Pigou–Dalton criteria. Only when the after transfer ranks are close, that is, \(t - k\) is small, will the two terms in \((10)\) or \((10a)\) reflecting the effect of the number of households whose ranks changed by one as a result of the transfer have a major impact on the change in the Gini index.

For the small transfers considered by economists (Atkinson 1970), when the donor is in the upper portion of the distribution while the recipient is in the lower part both \(k - i\) and \(j - t\) are much less than \((j - i)\), so the magnitude of the effect of the households whose rank changed is relatively small. In this situation, the decrease in the Gini index due to the difference in the ranks of the donor and recipient will have a greater impact and offset the increase in the Gini index due to the number whose ranks changed by one because of the transfer.

When the donor (\(j\)th ranked) is in the upper part of the distribution and the recipient (\(i\)th ranked) in the modal region of the distribution, for example, near the mode \(m\), the number, \(k - i\) of households passed over by the recipient is larger and the difference, \(j - i\), in the original ranks is smaller than the case of a low-income recipient. Thus, the first of the three terms in \((10a)\), which increase the Gini index, is larger, while the third term, which decreases the index, is smaller. Thus, the net decrease in the index when a donor in the upper region makes a small transfer \((a)\) to a recipient in the modal region is less than when the recipient is at the lower part. When both the recipient and donor are in the middle of the distribution, the higher density of the income distribution in the central region implies that the number of terms in the first and second terms in the parenthesis in \((10a)\) is larger than in the situation where a high-income donor makes a transfer to a middle or low-income one. The term \((j - i)\) reflecting the difference in the original ranks of the donor and recipient is smaller than in the previous two types of transfer. Thus, the net decrease in the Gini index arising from a progressive transfer between two households in the middle will be smaller than that resulting from a transfer from a high-income donor to a low or middle-income household.

**Comment:** The approximations used in the previous comment imply that for small transfers of size \(a\), the change in the Gini index in formula \((10a)\) is approximately

\[-\frac{1}{(n-1)\bar{x}} \left( f(x_i)a^2 + f(x_j)a^2 - 2f(t^*)[x_j - x_i]a \right). \tag{12}\]

Again, \(t^*\) is in \([x_i, x_j]\) and \(x_j - x_i \geq 2a\). When the difference, \(x_j - x_i\) between the original incomes of donor and recipient is large, the third term in \((12)\) that decreases the Gini index, is larger in magnitude than the first two terms that increase the index. When the difference, \(x_j - x_i\) in the original incomes of donor and recipient is relatively small as in the case of a transfer between two middle-income households, the approximate decrease \((12)\) in the Gini index is greater or equal to

\[-\frac{1}{(n-1)\bar{x}} \left( f(x_i)a^2 + f(x_j)a^2 - 4f(t^*)a^2 \right). \tag{12a}\]
This lower bound \((12a)\) to the decrease is larger than \((8e)\) because the ranks of more households change when the transfer lowers the donor’s rank and increases the recipient’s.

Some readers might be interested in a simple example where the rank-orders of the population change because of the transfer. Based on the mean incomes of the five quantiles reported by the Current Population Survey, consider the following nine lifetime incomes: 11,651, 30,509, 48,322, 50,322, 52,322, 54,322, 83,519, 137,600, and 185,206. Their mean is 72,641.44, Gini index is 0.41678, and mean difference is 60,549.28. Table 1 presents a few examples of the change resulting from a transfer of $10,000, from a higher income household to one with a lower income. This large amount will enable us to illustrate the effect of the transfer increasing the recipients rank or reducing the donor’s rank.

The transfers between the sixth and fifth as well as the fifth to fourth violate the Pigou–Dalton condition as the donor’s after rank and increment is more than the recipient’s.

Examine Table 1, the largest change in the Gini index occurs when the richest (rank 9) donates to the poorest (rank 1). As the rank of the recipient of a donation from the richest increases, the magnitude of the change in the Gini index and mean difference decrease as expected. When the sixth ranked household donates to the poorest, the Gini index changes more than when the recipient is the second ranked; however, when the sixth ranked donates to the fifth ranked, the Gini index and mean difference are greater than their values on the original set of nine. The absolute values of the change in the Gini index and mean difference, however, are smaller than when the sixth ranked household transfers money to the lowest ranked.

Transfers between the fourth and fifth ranked households also create a slight increase in the Gini index, due to the lowered rank of the donor; however, the absolute value of this increase is smaller than the absolute value of the decrease resulting from transfers from the ninth ranked household to the poorest or the median one. Again, transfers involving the two households in the central part of the distribution do not create the largest changes in the Gini index.

In sum, the change in the Gini index due to a nonorder preserving but mean preserving transfer depends on the difference between the before transfer ranks of the donor and recipient and the number of other households whose rank changed. From \((8b)\) and \((10a)\), the effect of the transfer, \((j - i)a\) is to decrease the Gini index, while the terms in the summands reflecting the change in the rank order and increment given to the recipient increase the Gini index. When an upper income household makes a transfer to one in the lower or central portions of the distribution, the term involving the difference in ranks is the dominant one, while when the transfer is between two households in the middle; the number of other individuals whose rank changed in the process has a larger role. Thus, the decrease in the Gini index is larger when a high-income household makes a transfer to a low or middle-income recipient than when the transfer is between two households in the modal region. Thus, the Gini index is not overly sensitive to transfers in the middle. As Atkinson (1970) noted, the Gini index is sensitive to changes at all levels and a reason for this is that the change has two components, one of which is more sensitive to transfers when the donor and recipient are from different parts of the distribution and the other when they are both in the modal region.

5. The Effect on the Gini Index When One Household Receives a Small Increment

Finally, consider the effect of the \(i\)th ranked household receiving an additional amount, so the overall mean becomes \((nx + a) / n\). The denominator of \(g_i\) replaces the original mean, \(x\) by the new mean and is the same regardless of which household receives the additional amount, \(a\). Thus, both the sensitivity of the numerator and Gini index to which member of the population receives the increment are of interest.

5.1. Increments that do not Change the Rank of the Recipient

If the rank of the recipient is unchanged, which is likely when \(a\) is small, formula (2) implies that the numerator of the Gini index will change by \((2i - n - 1)a\). If \(i = 1\), the numerator changes by \(-a(n - 1)\) and if \(i = n\), it increases by \(a(n - 1)\). When \(n = 2m - 1\) and the median receives the increment the numerator does not change. Indeed, the absolute value of \((2i - n - 1)\) declines from \(n - 1\) to 0 as \(j\) ranges from 1 to \(m\) and then increases to \(n - 1\) as \(i\) ranges from \(m\) to \(n\). Clearly, the numerator of the Gini index changes less when the recipient is in the middle part of the distribution than when a low or upper income household receives the increment.

The after increase Gini index \((g_i)\) is given by

\[
g_i = g - \frac{a}{nx + a} \left( g + \frac{n + 1 - 2i}{n - 1} \right). \tag{13}\]

The new index is less than the original when the term in the parenthesis is positive, that is, when

\[(n - 1)g + n + 1 - 2i < 0 \text{ or } 2i < (n - 1)g + (n + 1).\]

For a large population of size \(n\) as \(n \to \infty\) and \(i/n \to p\), that is, the recipient is the 100\(p\)th percentile, it follows that the increment will reduce inequality, as measured by the Gini index whenever \(p\) is less than \((1 + g)/2\). Thus, the Gini index will decrease if a small increment that does not change the rank of the recipient if that recipient is below the (50 + 50\(g\))th percentile.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
Rank of & Rank of & Gini index & Change in the Gini & Mean difference \\
 donor & recipient & & index & \\
\hline
9 (highest) & 1 (lowest) & 0.3862 & -0.0306 & 56105 \\
9 & 2 & 0.3900 & -0.0268 & 56660 \\
9 & 5 & 0.4045 & -0.0122 & 58772 \\
9 & 8 & 0.4129 & -0.0038 & 59994 \\
6 & 1 & 0.4045 & -0.0122 & 58772 \\
6 & 2 & 0.4084 & -0.0084 & 59327 \\
6 & 5 & 0.4237 & 0.0069 & 61549 \\
5 & 4 & 0.4344 & 0.0076 & 61660 \\
5 & 5 & 0.4267 & 0.0099 & 61994 \\
\hline
\end{tabular}
\caption{The effect on the Gini index and mean difference of a transfer of $10,000 from the household \(i\)th ranked to the \(i\)th ranked household.}
\end{table}
5.2. Increments that Increase the Rank of the Recipient

If the increment $a$, is sufficient to affect the order, the rank of the recipient increases by the number ($b$) of observations between $x_i$ and $x_i + a$. Then, the rank of the recipient increases to $i + b = k$. The reasoning in the previous section shows that the numerator of the Gini index of the new data differs from that of the original data by

$$-\sum_{r=i+1}^{k} 2(x_r - x_i) + a(2k-n-1).$$  \hspace{1cm} (14)

The $x_r$, $r = i + 1, \ldots, k$ fall in the interval, $[x_i, x_i + a]$ so the first term in (14) is negative but $\geq -2a(k-i).$ The second term is negative if the final rank, $k$, of the recipient is less than $(n+1)/2$ and positive otherwise. The first term is influenced by the number of observations between $x_i$ and $x_i + a$, which is largest when the rank, $i$, of the recipient is in the middle region of the distribution because the density is highest there. The second term in (14) tells us that when the final rank, $k$, of the recipient, is less than $(n+1)/2$, the numerator of the Gini index will decrease as will the index as the mean in the denominator has increased.

From (14), it follows that the Gini index, $g_1$ after the increment is

$$g_1 = g - \frac{1}{n-1} \cdot \frac{\sum_{r=i+1}^{k} (x_r - x_i) - a(2k-n-1)}{(n-1)(nx+a)}.$$

Calculations similar to those in Section 4, yield

$$g_1 = g - \frac{n}{n-1} \cdot \frac{1}{nx+a} \cdot \left( \frac{n-1}{n} \cdot (x_r - x_i) - a \cdot \left( 2 \frac{k}{n} - \frac{n+1}{n} \right) \right).$$  \hspace{1cm} (15)

Because the first term in (14) is negative, the numerator of the Gini index will decrease for some values of $k > (n+1)/2$, provided they satisfy

$$\sum_{r=i+1}^{k} 2(x_r - x_i) > a(2k-n-1).$$  \hspace{1cm} (17)

For the remaining values of $k$, all of which are in the upper half of the distribution, the Gini index will increase. When the household with the lowest income is the recipient both terms in (14) are negative provided the amount ($a$) is not so large that their after transfer rank ($k$) is in the upper half of the distribution and no longer satisfies (17). Thus, for small increments analogous to the transfers considered by both Atkinson (1970) and Allison (1978), the Gini index will decrease the most when the household with the lowest income is the recipient.

For a large population, as $n$ increases the term for the change in the Gini index in (16) becomes

$$-\frac{a}{nx+a} \left( 1 + g + \frac{2}{na} \sum_{r=i+1}^{k} (x_r - x_i) - 2 \frac{k}{n} \right).$$  \hspace{1cm} (18)

When $a$ is small, $2 \frac{k}{r=i+1} (x_r - x_i)$ approximately $= -a(k-i)$ as the $x_r$ are approximately uniformly distributed in the small interval $[x_i, x_i + a]$ so the change in the Gini index is approximately

$$-\frac{a}{nx+a} \left( 1 + g - \left( \frac{i}{n} + \frac{k}{n} \right) \right).$$  \hspace{1cm} (19)

Formula (19) implies that the decrease in the Gini index resulting from a small increment is largest when $\frac{n}{n} + \frac{k}{n}$ is small. From (8d), it follows that $k/n \sim i/n + f(t^*)a$, where $t^*$ is in $[x_i, x_i + a]$. Thus, $\frac{i}{n} + \frac{k}{n} \sim 2\frac{i}{n} + f(t^*)a$. Because the density function in the modal region is higher than the density in the very low-income part of the distribution, both $i/n$ and $f(t^*)$ are smaller when a low-income household receives the increment than when a household in the modal region receives it.

The above considerations also imply that the Gini index will increase when

$$\frac{i}{n} + \frac{k}{n} \sim 2\frac{i}{n} + f(t^*)a > (1 + g).$$

Thus, the index will decrease when the recipient is below the $50(1 + g - f(t^*)a)$th percentile. This result is very similar to the situation when the increment does not change the rank of the recipient, that is, the Gini index decreases when the recipient can be in the $(50 + 50)$th percentile. However, it shows that the fraction of households who can receive the increment and the Gini index decrease is slightly smaller when the increment increases the rank of the recipient than when its rank is unchanged.

Formula (19) also shows that when the highest income household receives the small increment the increase in the Gini index is largest, as minus the term in the large parenthesis in (19) becomes $1 - g$, which is $> 0$. Furthermore, this increase is larger in absolute value than the decrease in the Gini resulting when the median income household receives the increment. This follows from comparing the changes in the numerator (14) of the Gini as the effect of the second term is $a(n-1)$ when the highest income household receives the increment, while it near zero when the median household receives it. While the first term in (14) is negative, it is $\geq -2a(k-i)$ and $n-1$ is much larger than $2(k-i)$ because only a small fraction of households have their ranks changed as a result of a small increment going to a household in the middle region.

The conclusion that additional small increments given to households in the lower (upper) part of the distribution increase (decrease) the Gini index is quite intuitive. The result that the effect on the Gini index of an additional increment given to a household in the middle sufficient to change the ordering is less than the effect of the same increment given to a household in some parts of the distribution, such as the extreme lower and upper, is less obvious. These changes in the Gini index arising from one household receiving additional income are not consistent with the Gini index being more sensitive to changes in the middle of the distribution than elsewhere.

For readers interested in a numerical example, consider the example of nine incomes and assume that one of them receives an additional $24,000$. This value ensures that the rank of the recipient would increase in some cases. For each of the nine possible recipients, Table 2 reports the new values of the Gini index.
and the resulting change and mean difference. The number of households who had more income originally but the recipient now exceeds is the number passed. Recall that the Gini index, mean difference, and mean of the original nine are: 0.41677, 60,549, and $72,641.60, respectively. The ratios of the new Gini index to the original one are 0.8842, 1.0192, and 1.0495 if the first (lowest), median, or ninth (highest) income receives the addition. In percentage terms, the Gini index changes the most when the poorest member of the population is the recipient and the change when the median member receives it is less than the change when the richest member receives it. Thus, the Gini index is more sensitive to additional income going to a household in the two extreme regions of the distribution than it is to a similar change in the middle of the distribution.

### 6. Summary and Discussion

Although the Gini index yields one number summarizing the entire income distribution or Lorenz curve and cannot capture all the changes in the income distribution that economists or policy makers are interested in, the scenarios studied here indicate that the criticism that it gives undue weight to changes in the middle of the distribution is inaccurate. The recent book by Foster et al. (2013, p. 21) did note that although the Gini coefficient is considered to be most sensitive to changes involving incomes at the middle this “is not entirely accurate.” It continues, however, with “The effect of a given-sized transfer on the Gini coefficient depends on the number of people between giver and receiver, not on their respective income levels. Because, empirically, there tend to be more observations bunched together in the middle of the distribution, the effect of a transfer near the middle tends to be larger.”

In Section 3, where a transfer or increment preserves the order, for example, when it is small, the opposite is true, that is, transfers from the rich to the poor had a greater effect on the Gini index. When transfers or an increment did not preserve the order, discussed in Sections 4 and 5, in most situations the difference in the ranks of the donor and recipient, before or after the transfer or addition is the main contributor to a change in the Gini index. Transfers or an additional increment involving a middle-income household do change the ranks of a higher fraction of the population because of the density function being higher in that region, however, the ranks of these households decrease by just one. Only when the transfer is between two households in the middle of the distribution did the number of individuals whose rank changes because of the transfer have an important role. This effect on the change in the Gini index does not offset the relatively small difference between the after transfer ranks of the donor and recipient when both are in the central region. Thus, a small transfer or additional increment affecting the middle of the income distribution does not have undue weight on the resulting change in the Gini index. Indeed, in the scenarios examined here a transfer or addition to a middle-income household had a smaller impact on the Gini index than a transfer or addition to a low-income household. In view of the larger magnitudes of the weights in the numerator of the Gini index in formula (2) assigns to the extreme ordered incomes than to the middle ones, in the context of small transfers or additions, these conclusions are not that surprising.

When the total income of the population and consequently the mean increase, because twice the mean is the denominator of the Gini index it does not fully reflect a shift in favour of the upper end (Gastwirth 2014). Indices such as the ratio of the share of income received by the top 20% to the lower 20%, used by Dorling (2014), the ratio of the share of the top 10% to the lower 40% introduced by Palma (2011) or the median based Gini index (Gastwirth 2014) increase more than the Gini index in response to such a change. Like the Gini index, however, these indices can have the same numerical value for data from two distributions even though the Lorenz curves intersect. A method for constructing two different distributions with same value of a measure that is the ratio of the top 100b% to the bottom 100b%, where b < 1 − u, is described in Appendix A.

From a statistical viewpoint, it is unreasonable to expect one summary measure to capture the features of an entire distribution. Thus, the relationship between the choice of measure and its underlying social welfare function, stressed by Atkinson (1970), Newbery (1970), Sheshinski (1972), and Sen (1974) remains very important. Jenkins (2009) noted that the ability to calculate several indices, which focus on changes in different income ranges, is very useful for the analysis of income and earnings distributions. By using different weights than those in the numerator of formula (2), following Mehran (1976) one can create a summary measuring placing increased weight on the part of the income distribution most relevant to the purpose of a study. Like the numerator of the Gini index, these measures are linear combinations of order statistics and there is a large literature deriving their large sample distributions (David and Nagaraja 2003; Greselin et al. 2009; Giorgi and Gigliarano 2016). Similarly, the Lorenz curve and related functions or transformations of it (Sordo et al. 2014; Arnold 2015; Gastwirth 2016) can be used to emphasize the region of the income distribution of primary concern.

Other measures of inequality, for example, the generalized entropy family (Cowell 2011) or Atkinson’s family may be superior to the Gini index for some analytic purposes because the formulas for decomposing them across subpopulations are analogous to the decompositions in classical analysis of variance. Lambert and Decoster (2005), however, showed that the Gini index is useful for some types of inequality decomposition analyses. The Gini index is a well-studied index with a long history (Giorgi 1990) and is associated with the area between the line of equality and the Lorenz curve (Gastwirth 1972); providing a graphical summary of the distribution that economists and policymakers have found useful. Furthermore, it may be difficult

<table>
<thead>
<tr>
<th>Recipient</th>
<th>Number passed</th>
<th>Gini index</th>
<th>Change in the Gini index</th>
<th>Mean difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>First (lowest)</td>
<td>1</td>
<td>0.3685</td>
<td>-0.0483</td>
<td>55502</td>
</tr>
<tr>
<td>Second</td>
<td>4</td>
<td>0.3802</td>
<td>-0.0366</td>
<td>57258</td>
</tr>
<tr>
<td>Third</td>
<td>3</td>
<td>0.4064</td>
<td>-0.0303</td>
<td>60216</td>
</tr>
<tr>
<td>Fourth</td>
<td>2</td>
<td>0.4086</td>
<td>-0.0081</td>
<td>61549</td>
</tr>
<tr>
<td>Fifth</td>
<td>1</td>
<td>0.4101</td>
<td>-0.0066</td>
<td>61712</td>
</tr>
<tr>
<td>Sixth</td>
<td>0</td>
<td>0.4109</td>
<td>-0.0059</td>
<td>61882</td>
</tr>
<tr>
<td>Seventh</td>
<td>0</td>
<td>0.4197</td>
<td>0.0029</td>
<td>63216</td>
</tr>
<tr>
<td>Eighth</td>
<td>0</td>
<td>0.4286</td>
<td>0.0118</td>
<td>64549</td>
</tr>
<tr>
<td>Ninth (highest)</td>
<td>0</td>
<td>0.4374</td>
<td>0.0207</td>
<td>65883</td>
</tr>
</tbody>
</table>
for a government agency producing income statistics to choose between members of a family, which place greater emphasis on different parts of the distribution. While the U.S. Census Bureau reports three different versions of Atkinson’s index and the Theil in addition to the Gini (DeNavas-Walt and Proctor 2014), the Gini index and the income shares of the quintiles are the primary ones used in their discussion of changes in inequality.

Yitzhaki (1982) gave a necessary condition for first- and second-order stochastic dominance of two distributions in terms of the mean and mean difference. This connection between the Gini index and second-order stochastic dominance has an important role in many areas of economics and public policy, for example, in studies of investment decisions under uncertainty (Lutkebohmert 2009; Levy 2016), economic growth and recessions (Bishop et al. 1991) and measuring possible discrimination (Le Breton et al. 2012; Hoy and Huang 2017; Salas et al. 2017).

When Government agencies publish the Gini index along with the mean, median, and quintiles, if they could include the deciles and 95th and 99th percentiles of the distribution and the shares (or average income) of each decile and the top 5 and 1 percent, researchers would be able to accurately estimate most measures of inequality. Providing this information in grouped form would also preserve the privacy and confidentiality of the incomes of the survey respondents.
CHAPTER 3

MEASURING ASPECTS OF SEXUALITY AND GENDER: A SEXUAL HUMAN RIGHTS CHALLENGE FOR SCIENCE AND OFFICIAL STATISTICS

This article is excerpted from *Chance*

by Matt Jans, Bianca D. M. Wilson and Jody L. Herman

© 2018 Taylor & Francis Group. All rights reserved.
Sexual orientation and gender identity (SOGI) are two demographic and personal dimensions that have received relatively little attention in large-scale surveys and official statistics, despite recent advances in social, cultural, and legal equality for lesbian, gay, bisexual, and transgender (LGBT) people.

About 10 million adults in the US (4.1%) identify as LGBT, according to polling conducted by Gallup, up almost 2 million from 2012. Notably, LGBT rights and recognition have seen major advances over the past two decades. For example, according to the Pew Research Center on Religion and Public Life, nationwide support of same-sex marriage has continued to increase. It is currently at 55%, passing the percentage opposing in 2012, and continuing to climb since the 2015 legalization of same-sex marriage. Across religious and demographic groups, 63% of Americans say that homosexuality should be accepted, with acceptance rates of 66% among Catholics, 76% among white mainline Protestants, and 80% among religiously unaffiliated people.

Further, polling on bathroom use rights shows a slight majority support for transgender people using public restrooms in accordance with their gender identity, as opposed to their sex assigned at birth.

Less-quantitative evidence of social and cultural progress of the LGBT community is reflected in the number and types of LGBT characters in mainstream popular entertainment. From the shows “Ellen” and “Will and Grace” in the 1990s, which included representations of gays and lesbians, to the expansion of representation to transgender people and LGBT people of color through “Orange is the New Black,” “Transparent,” “Shameless,” “Black-ish,” “Modern Family,” and “Better Things,” it’s clear that positive and realistic portrayals of LGBT people are now more common than in the past.

In addition to the increasing presence of ethnically and socioeconomically diverse LGBT people in fictional shows, multiple documentary and news segments have been devoted to the issues of transgender people specifically, such as Katie Couric’s NatGeo special “Gender Revolution,” HBOs “Vice Special on Trans Youth,” and NBC’s “Special on Transgender
Youth.” All of this indicates a meaningful socio-cultural shift.

With this developing recognition and acceptance, it is surprising that more large-scale and official surveys do not measure these important personal and demographic characteristics.

A “Special Population” or “Part of the General Population”? Specification and Measurement Challenges

Perhaps one reason for scant SOGI measurement is the perspective that LGBT people are part of a “special population” to be targeted with sampling methods that are sometimes at odds with the goals of general population surveys (e.g., non-random convenience samples). Because the rate of LGBT identification in the general population is low, it is usually impractical to use general population survey methods for one-time surveys that focus primarily on LGBT respondents. However, surveys with large samples, or that recur regularly and can be combined over time, are often excellent vehicles for sampling the LGBT population.

A second challenge to the broad inclusion of LGBT people in “standardized surveys” is the seeming complexity of deciding how to measure SOGI. LGBT people have always been sampled in general population surveys, but they have been a “hidden population” because the questions needed for them to identify themselves accurately often have not been used. Alfred Kinsey’s mid 20th-century sexuality research may be partly to blame. Kinsey’s method used open-ended sexual history interviews with complex scoring systems to obtain very nuanced measures of sexual identity, attraction, and behavior that are not only infeasible in general population surveys, but have large reliability and reproducibility concerns. However, measuring SOGI need not be so complex.

At the other extreme, and a much better representation of SOGI measurement best practices for surveys, the Gallup statistic cited earlier uses an all-inclusive definition, asking one question that uses common identity labels that have sociopolitical and cultural meaning to most people: “Do you, personally, identify as lesbian, gay, bisexual, or transgender?” Following Gallup, a survey could include SOGI measurement by adding only one question. However, gains in sensitivity are balanced by losses in specificity and the inability to separate the sometimes overlapping subgroups of lesbian, gay, bisexual, and transgender individuals for analysis.

There is general agreement among sexuality researchers that sexual orientation has three primary dimensions: sexual attraction (i.e., whether a person is physically attracted to people of the same sex, other sex, or both), sexual behavior (whether a person has had sexual relationships with people of the same sex, other sex, or both), and sexual identity (the term a person uses to define their sexual orientation, such as gay, lesbian, bisexual, asexual, straight, or something else).

Each measure has its own strengths and limitations. In the mid-1980s, public health researchers began to realize that some men who had sex with men (MSM) did not identify themselves as “gay,” meaning that survey questions using terms such as “gay” or “homosexual” would not accurately capture their orientation. Thus, asking the sex of sexual partners was the right way to identify this sexual minority group.

Of course, sexual behavior measures do not capture the full spectrum of sexual orientation, particularly among people who have not been sexually active.
during the timeline defined by the survey. In contrast, sexual attraction measures offer the most inclusive definition, and tend to identify the largest number of people who vary from the heteronormative framework, including unexpressed attraction. However, looking across general population surveys, particularly those used for public health surveillance, measures of sexual identity and sexual behavior tend to be the most common, probably due to the direct public health uses of these variables for risk and exposure analyses.

The situation for gender identity measurement is somewhat different. Most surveys already have a measure of sex (e.g., “Are you male or female?”), sometimes referred to as “gender” for numerous reasons. Nuances of labeling and polite terminology aside, such a binary perspective on gender is overly simplistic and fails to capture transgender identification and other gender nonconformity (as well as physical conditions sometimes called “intersex” or “differences of sexual development” [DSD]).

People whose gender identity matches their sex assigned at birth (i.e., for whom the traditional sex question above would be easy to answer) are sometimes called cisgender, while the term transgender (refers to people whose internal perception of their gender does not match the sex they were assigned at birth). Gender nonconforming is sometimes also used by people who feel they do not fit traditional sex-based gender expression roles in some way. Many more terms describe specific gender identities and expressions, and new terms develop all the time, particularly among younger people. Nonetheless, survey items about gender identity that include transgender identification have been used in health research, particularly HIV research, for decades. Yet, questions about gender identity have not been part of large-scale surveys until relatively recently and are only now, slowly, becoming a common part of the dialogue for survey planning.

### How Many Surveys Measure SOGI

The Federal Interagency Working Group on Improving SOGI in Federal Surveys of the U.S. Office of Management and Budget (which sets statistical policy for U.S. federal statistics) recently conducted a review and summary of the federal surveys that measure SO and GI. Across all federal surveys and studies, 11 surveys (and one study) included some measure of SOGI, most within the Department of Health and Human Services. While each of the 11 surveys included a measure of self-identified sexual orientation, three also collected sexual attraction and four collected sexual behavior.

Gender identity, on the other hand, was measured much less often, with only six of these 12 surveys/studies (50%) measuring gender identity (in perspective, 100% of the surveys asked about self-identified SO, 25% also asked about sexual attraction, and 33% asked about sexual behavior).

Some of the most-interesting developments in the U.S. federal statistical system are in the Behavioral Risk Factor Surveillance System (BRFSS), which has included an official optional SOGI module since 2014, but also allows individual states to include their own SOGI measures. The first state to ask about transgender status and sexual orientation in BRFSS was Vermont in 2000, although this question’s wording was quite inaccurate and insensitive, focusing on cross-dressing and using the term “transvestite,” which do not necessarily indicate a transgender identity. The question has since been changed to a more-inclusive and -accurate terminology. The BRFSS as a whole is a pioneer in the formalization of SOGI measurement in population surveys. With its range of questions and national scope, it offers an interesting opportunity to look at the effect of question wording on regional variation in identification rates.

While federal surveys are some of the largest and most-used general population surveys, they are not the only ones to ask SOGI questions. The General Social Survey (GSS) also has a long history in SO measurement and has asked about sexual behavior (sex of sex partners) since 1988 and a sexual identity question since 2008. The California Health Interview Survey (CHIS) has asked about sexual identity and sexual behavior since 2001, and GI since 2015. Notably, many surveys only ask these questions of adults, although CHIS now asks gender expression of teens. (For other surveys and studies that have measured SOGI, see the Further Reading recommendations.)

### General Advice and Best Practices for Measuring SOGI and Sampling the LGBT Population

The OMB reports discussed above focus on current U.S. federal statistical practice and future directions for including SOGI measures in officials surveys. For readers seeking best practices in SOGI measurement, we recommend reviewing the SMART and GenIUSS reports published by the Williams Institute. Drawing on those and our related experience in testing and implementing SOGI questions, we provide some
general best practices, followed by questions here and answers about SOGI measurement.

1) Clearly define the LBGT population to measure within the target population of the survey and goals of the study: Most “general population” surveys begin with a definition of the target population to sample. By definition, these surveys will include LGBT and non-LGBT respondents. If large numbers of LGBT respondents are required for analyses, over-sampling plans or cross-year combining (if the survey will be repeated) should be considered. Further, for SOGI measurement within general population surveys, researchers and survey administrators should take care to make sure the specific SOGI measures they use are pertinent to the goals of the survey and statistical reporting.

For example, if the point of the survey is studying sexually transmitted diseases, a sexual behavior measure will be more useful than a sexual attraction or sexual identity measure. If the purpose of the study is to document rates of harassment by police among transgender sex workers, gender expression may be more important to measure than gender identification. An investigator may sometimes wish to assess all of these possible dimensions, but every researcher has to consider which dimensions of SO and GI are needed most for the aims of their study.

2) Make sure questions are population- and age-appropriate: Some peoples’ sexual and gender identities are experienced early in life, while others unfold later. Similarly, a person’s sexual experiences tend to shape their identity over time. A youth may be able to indicate with certainty that they are attracted only to members of the same sex, but may not have had the opportunity to express that attraction. Particularly with younger people, it can be better to ask about sexual identity and attraction than sexual behavior.

3) Pre-test measures, even if they are taken from existing surveys or follow best practices: While there are now comprehensive best practices for SOGI measurement, every survey should pre test selected measures and interviewing procedures with people similar to those who will be in its sample. There is still much to learn about the role of survey mode and question placement in measurement error and item non-response on SOGI questions.

Particularly in the context of general population surveys, the terminology used to measure SOGI must be understood equally by respondents who identify as a sexual/gender minority and those who do not. Thus, questions must be written in relatively plain language, so respondents can answer whether or not they understand SOGI-specific terminology. For example, the Williams Institute and the California Health Interview Survey experimentally tested gender identity questions in an adult population, and found that a two-question measure using simple language performed better than a one-question measure using complex terms. Without that pre-testing, we might have been tempted to select the one-question measure simply because it was a single question. However, our pilot test clearly showed that the two-question measure was the right choice.

4) Consider the data collection context: This includes the mode of data collection, physical context of data collection, and issues such as the age of respondents sampled for the survey. For example, will questions be asked of each member of the household in a fairly open setting where all respondents (and perhaps even those who are not selected to respond) can hear the questions and answers? Will questionnaires be administered in a classroom or school library, where classmates could observe students’ responses (even if the questions are self-administered)?

Such methodological details could seriously affect accuracy of response and missing data rates, and should be addressed conscientiously in the survey design. In general, it is recommended to administer SOGI questions in a self-administered mode, or other private mode, when possible to make the respondent more comfortable when answering, and reduce the risk of disclosure to others present.

Frequently Asked Questions (FAQs) about SOGI Measurement

Question #1: What specific questions should I ask to measure sexual orientation accurately?

According to the Williams Institute’s SMART report, the recommended way to ask about self-identified sexual orientation is, “Do you consider yourself to be heterosexual or straight, gay or lesbian, or bisexual?”

Some experts advise against using the technical terms “heterosexual” and “homosexual” because it requires a relatively high level of literacy to understand them. Pre-testing can help address whether this is a concern for the specific population being studied, although including them sometimes facilitates translation into other languages that do not have a term for “straight.”
To measure sexual behavior, the recommended approach is to ask, "In the past [DEFINE TIME PERIOD], who have you had sex with? Men only? Women only? Both men and women? I have not had sex in [DEFINE TIME PERIOD]." While a questionnaire design purist would place a screener question asking about sexual activity in the defined time periods first, then skipping out people who have not had sexual partners in that period (see method used in CHIS), this question is otherwise simple and straightforward. Because it does not ask for number of sexual partners, this question could as easily be answered for the past year, past decade, or lifetime. Modifications would obviously have to be made for interviewer-administered modes.

Finally, the recommended method of measuring sexual attraction is to ask, "People are different in their sexual attraction to other people. Which best describes your feelings? Are you only attracted to females, mostly attracted to females, equally attracted to females and males, mostly attracted to males, only attracted to males, or not sure?" For both attraction and behavior, if possible, the response options should be ordered based on the respondent’s sex so other sex options are asked first (e.g., "attracted to women" would appear first for men).

Question #2: What questions should I ask to measure gender identity accurately?
Following the GenIUS report about gender identity measurement, and pilot testing conducted on the California Health Interview Survey, a two-step (i.e., two-item) measure appears to be the simplest way to ask about gender identity while having a minimal impact on the overall duration of the survey (compared to a one-step version that provides a definition of transgender and asks respondents whether they consider themselves to be transgender).

Gender Identity Measure Step 1:
What sex were you assigned at birth, as shown on your original birth certificate?
- Male
- Female

Gender Identity Measure Step 2:
How do you describe yourself? (check one)
- Male
- Female
- Transgender
- Do not identify as female, male, or transgender

In interviewer-administered surveys, or when there is space in a self-administered questionnaire, respondents can be allowed to provide their own gender identity terms, allowing for a fully accepting and affirming respondent experience while producing no additional burden for most respondents. Capturing such data on recurring surveys also facilitates tracking the evolution of gender identity and expression terms over time. This can be done for sexual orientation questions as well, and can help give a voice to hidden sexual and gender minorities.

Question #3: Where should I place the questions in the survey, and what mode should I use?
If possible, SOGI questions should be placed in self-administered sections of surveys (e.g., paper and pencil, or computer-based instruments where an interviewer is not present). There is some evidence that asking questions this way results in obtaining more-accurate answers. However, using item non-response as a measure of sensitivity, there does not seem to be a problem with asking SOGI questions in telephone surveys, where the interviewer is not physically present with the respondent.

Specific recommendations can vary depending on the specific measure used. For GI, despite the need for empirical testing, it is generally recommended to place the two-step measure in the same location as the traditional sex question (essentially replacing that single question with the two-step measure). For sexual orientation, the advice is slightly more complex. For general surveys (i.e., where the main topic of the survey isn’t sexuality), sexual identity measures can be placed in demographics sections.

However, if the interview or questionnaire includes a section or questions about sexual behavior or sexual history, then questions can be placed there. One caution is not to place SO questions in a section that is so sensitive or personal that it might experience higher rates of item nonresponse (e.g., sexual violence), leading respondents to skip the relatively innocuous SO questions.

As with most questionnaire design decisions, context and order of the entire questionnaire must be taken into account. When sexual identity is considered a demographic characteristic, it fits comfortably with questions on age, race, and education. However, questions about the gender or sex of one’s sex partners, or the people to whom one is attracted, probably would seem out of place with typical demographic questions.
**Question #4:** There seem to be a lot of different terms for different sexual and gender identities. I don’t feel like I know all those terms. I don’t want to offend LGBT respondents by excluding terms or using words that are out-of-date, but also don’t want confuse non-LGBT respondents by including a lot of terms that don’t apply to them. Help!

Remember that in general population surveys, the goal is to write questions that work well for both LGBT and non-LGBT respondents. The goal is to make sure that every respondent has a place to, as comfortably as possible, describe their situation with respect to the question asked. Further, there is much heterogeneity in terminology among the LGBT community, with age, race/ethnicity, and location influencing the terms people use to describe themselves. This means that in most general population surveys, providing lists of LGBT-specific terms isn’t possible. However, when respondents volunteer other terms, they should be recorded.

Particularly if one of the objectives of the study is to capture contemporary terminology, then “other, specify” options can be included. The challenge with these is that they then must be “back coded” into the standardized response categories for reporting. This is

### Visual explanation of terminology for gender identities cross-classifying sex at birth and current gender identity.

![Graphic used with permission from Sari Reisner. See GenIUSS Group report (p. 48) in Further Readings.](image-url)

<table>
<thead>
<tr>
<th>Assigned Sex at Birth</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Infant designated a male sex on original birth certificate</td>
<td>Infant designated a female sex on original birth certificate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Current Gender Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Non-Transgender Male</td>
</tr>
<tr>
<td>Cross-Sex Identified Transgender Male</td>
</tr>
</tbody>
</table>

| Female                    |
| Cross-Sex Identified Transgender Female | Male birth sex, female gender identity |
| Non-Transgender Female    | Female birth sex, female gender identity |

| Male-to-Female (MTF)      |
| Male-to-Female (MTF)      | Male birth sex, MTF gender identity |
| Potential Measurement Error | Female birth sex, MTF gender identity |

| Female-to-Male (FTM)     |
| Potential Measurement Error | Male birth sex, FTM gender identity |
| Female-to-Male (FTM)     | Female birth sex, FTM gender identity |

| Other Gender (Specify)  |
| Other Transgender Identity | Male birth sex, other gender identity |
| Other Transgender Identity | Female birth sex, other gender identity |
particularly important with low-incidence characteristics like SO and GI.

**Question #5:** My survey doesn’t have any questions about sexuality in it at all, and it seems like such a sensitive topic to add. Won’t respondents find these questions offensive, because they deal with such personal and sensitive topics? Will they abandon my survey?

There is no evidence that SOGI questions obtain particularly high levels of item nonresponse or induce break-off (unit nonresponse). In fact, income regularly produces the most item nonresponse across surveys, modes, and wordings, and most surveys still include some measure of income in their demographics sections. Drug and alcohol use also have higher levels of item nonresponse than SOGI measures, and these questions are asked in many surveys.

While there is some evidence that sexual orientation item nonresponse can be higher in racial/ethnic minorities and respondents with limited English proficiency, this may be due to question wording and translation as much as sensitivity of the topic. It is important to remember that sensitivity in the survey interview context, with the right protections for privacy, can be more like sensitivity in a doctor’s office than sensitivity in the workplace.

**Question #6:** If plain language is to be used, and questions are supposed to be simple, how can I be sure that all respondents understand terms like “gay” and “transgender” equally? Shouldn’t I use technical and “proper” terms like “homosexual” and “heterosexual” instead of “gay” and “straight”?

Respondents’ understanding of technical terminology is a common problem in survey measurement. While some survey best practices recommend that all respondents receive any definition given, others recommend that interviewers be trained to provide definitions as needed. The most error-prone situations arise when respondents must ask for a definition, because those who do not realize they need the definition to answer accurately will not request it. Thus, the safest approach is to phrase questions as simply as possible to avoid the need for definitions.

Pre-testing, such as cognitive interviewing, can help establish whether the terms chosen are understood equally by a cross-section of the population surveyed. Based on our experience in testing gender identity questions on CHIS, there appears to be no need to provide a detailed definition of transgender; in fact, providing such definitions seems to make a question harder to answer than asking two simple questions (sex assigned at birth and current gender identity).

With respect to using popular language (e.g., gay/straight) versus technical terms (e.g., heterosexual/homosexual), there is good reason to consider using common terminology rather than technical terminology.

**In Conclusion**

It has never been easier to measure SOGI on large-scale surveys. Successes in multiple federal surveys, state-based large-scale surveys like CHIS, and polling conducted by Gallup are great examples of what is possible. There is no evidence of risk to overall data quality by asking SOGI measures and the quality of the resulting data is substantially enhanced with respect to studying sexual and gender human rights.

Based on social, economic, and health research showing disparities among sexual/gender minorities, it is imperative that more surveys add SOGI measures to ensure that these marginalized populations are a regular part of the data-based decision-making world in which we live.
BONUS JOURNAL CONTENT
From *The American Statistician*

**Statistical Interference in the 21st Century: A World Beyond p < 0.05**

**Getting to a Post "p < 0.05" Era**

*Editorial:* Moving to a World Beyond "p < 0.05", 10.1080/00031305.2019.1583913

*Article:* What Have We (Not) Learnt from Millions of Scientific Papers with P Values?, 10.1080/00031305.2018.1447512

*Article:* Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics, 10.1080/00031305.2018.1558111

*Article:* Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary, 10.1080/00031305.2018.1497540

*Article:* Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p-Values, 10.1080/00031305.2018.1518271

*Article:* Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness, and Relevance, 10.1080/00031305.2018.1543138

*Article:* The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making, 10.1080/00031305.2018.1529623

*Article:* Expert Knowledge Elicitation: Subjective but Scientific, 10.1080/00031305.2018.1518265

*Article:* Before p < 0.05 to Beyond p < 0.05: Using History to Contextualize p-Values and Significance Testing, 10.1080/00031305.2018.1537891

*Article:* The Limited Role of Formal Statistical Inference in Scientific Inference, 10.1080/00031305.2018.1464947

*Article:* Large-Scale Replication Projects in Contemporary Psychological Research, 10.1080/00031305.2018.1505655

129
Bonus Content

Interpreting and Using $p$

**Article:** Valid $P$-Values Behave Exactly as They Should: Some Misleading Criticisms of $P$-Values and Their Resolution With $S$-Values, 10.1080/00031305.2018.1529625

**Article:** The $p$-Value Requires Context, Not a Threshold, 10.1080/00031305.2018.1529624

**Article:** Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty, 10.1080/00031305.2018.1537889

**Article:** Putting the $P$-Value in its Place, 10.1080/00031305.2018.1470033

**Article:** Evidence From Marginally Significant $t$ Statistics, 10.1080/00031305.2018.1518788

**Article:** The $p$-value Function and Statistical Inference, 10.1080/00031305.2018.1556735

**Article:** $p$-Values, Bayes Factors, and Sufficiency, 10.1080/00031305.2018.1502684

**Article:** Limitations of $P$-Values and $R$-squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment, 10.1080/00031305.2018.1518269

Supplementing or Replacing $p$

**Article:** An Introduction to Second-Generation $p$-Values, 10.1080/00031305.2018.1537893

**Article:** A Proposed Hybrid Effect Size Plus $p$-Value Criterion: Empirical Evidence Supporting its Use, 10.1080/00031305.2018.1564697

**Article:** Three Recommendations for Improving the Use of $p$-Values, 10.1080/00031305.2018.1543135
**Bonus Content**

**Article:** The False Positive Risk: A Proposal Concerning What to Do About \( p \)-Values, [10.1080/00031305.2018.1529622](https://doi.org/10.1080/00031305.2018.1529622)

**Article:** Moving Towards the Post \( p < 0.05 \) Era via the Analysis of Credibility, [10.1080/00031305.2018.1543136](https://doi.org/10.1080/00031305.2018.1543136)

**Article:** Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels, [10.1080/00031305.2018.1518268](https://doi.org/10.1080/00031305.2018.1518268)


**Article:** Abandon Statistical Significance, [10.1080/00031305.2018.1527253](https://doi.org/10.1080/00031305.2018.1527253)

**Article:** Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science, [10.1080/00031305.2018.1518264](https://doi.org/10.1080/00031305.2018.1518264)

**Adopting More Holistic Approaches**

**Article:** Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don’t Expect Replication, [10.1080/00031305.2018.1543137](https://doi.org/10.1080/00031305.2018.1543137)


**Article:** How Large Are Your \( G \)-Values? Try Gosset’s Guinnessometrics When a Little “\( p \)” Is Not Enough, [10.1080/00031305.2018.1514325](https://doi.org/10.1080/00031305.2018.1514325)

**Article:** Predictive Inference and Scientific Reproducibility, [10.1080/00031305.2018.1518270](https://doi.org/10.1080/00031305.2018.1518270)

**Article:** Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing, [10.1080/00031305.2018.1513377](https://doi.org/10.1080/00031305.2018.1513377)

**Article:** Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II, [10.1080/00031305.2018.1543617](https://doi.org/10.1080/00031305.2018.1543617)
Article: Frequentist, Bayes, or Other?, 10.1080/00031305.2018.1459317


Article: Multiple Perspectives on Inference for Two Simple Statistical Scenarios, 10.1080/00031305.2019.1565553

Reforming Institutions: Changing Publication Policies and Statistical Education

Article: Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p < 0.05$ Universe, 10.1080/00031305.2018.1537888

Article: The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration, 10.1080/00031305.2018.1505658

Article: Coup de Grâce for a Tough Old Bull: “Statistically Significant” Expires, 10.1080/00031305.2018.1543616

Article: The World of Research Has Gone Berserk: Modeling the Consequences of Requiring “Greater Statistical Stringency” for Scientific Publication, 10.1080/00031305.2018.1555101

Article: Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their $p$-Value Ban, 10.1080/00031305.2018.1537892

Article: Content Audit for $p$-value Principles in Introductory Statistics, 10.1080/00031305.2018.1537890

Article: Beyond Calculations: A Course in Statistical Thinking, 10.1080/00031305.2018.1505657
There is a growing recognition of the importance of statistical reasoning across many different aspects of everyday life. This is the case now more than ever, in our data-rich world, where the volume, availability, and types of data have increased significantly. It is crucially important that statistical reasoning is introduced to students early in their education, giving them key skills for nearly any career path they choose. Professionals in every field encounter data throughout their working lives, and the ability to reason statistically will enable them to make better evidence-based decisions. For members of the general public, learning how to reason statistically enables them to better understand risk, make decisions in the face of uncertainty, and become more informed citizens.

This exciting book series features:

- A forum highlighting the important role of statistical and probabilistic reasoning in many areas.
- Concepts presented while assuming minimal background in Mathematics and Statistics.
- A broad audience including professionals across many fields, the general public and courses in high schools and colleges.
- Topics include Statistics in wide-ranging aspects of professional and everyday life, including the media, science, health, society, politics, law, education, sports, finance, climate, and national security.
- Short and inexpensive books of 100-150 pages that can be written and read in a reasonable amount of time.

All proposals or requests for information should be sent to: asacrc@crcpress.com
Visit the ASA Portal today:
bit.ly/ASA_Portal