3-2017

An Examination of Statistical Disclosure Issues Related to Publication of Aggregate Statistics in the Presence of a Known Subset of the Dataset Using Baseball Hall of Fame Ballots

Gregory J. Matthews
Loyola University Chicago, gmatthews1@luc.edu

Pétala Gardênia da Silva Estrela Tuy
Loyola University Chicago

Robert K. Arthur

Recommended Citation

Matthews, Gregory J.; Silva Estrela Tuy, Pétala Gardênia da; and Arthur, Robert K.. An Examination of Statistical Disclosure Issues Related to Publication of Aggregate Statistics in the Presence of a Known Subset of the Dataset Using Baseball Hall of Fame Ballots. Journal of Quantitative Analysis in Sports, 13, 1: 1-10, 2017. Retrieved from Loyola eCommons, Mathematics and Statistics: Faculty Publications and Other Works, http://dx.doi.org/10.1515/jqas-2016-0085

Gregory J. Matthews*, Pétala Gardênia da Silva Estrela Tuy and Robert K. Arthur

An examination of statistical disclosure issues related to publication of aggregate statistics in the presence of a known subset of the dataset using Baseball Hall of Fame ballots

DOI 10.1515/jqas-2016-0085

Abstract: Each year the members of the Baseball Writers Association of America (BBWAA) vote for eligible former players to be inducted into the Baseball Hall of Fame. The BBWAA tabulates and releases vote totals, but individual ballots remain private. However, many voters forgo their ballot privacy to publish their ballots through various media channels. These publicly available ballots can be aggregated to create a subset of the true ballots. Using these released ballots and the totals released by the BBWAA, this research assesses what can be learned about the group of voters who chose to not disclose their ballot. Attributes of the known and unknown ballot groups are studied by looking at differences in voting preference for individual players as well as voting differences between classes of voters that are defined using latent class analysis (LCA).

Keywords: baseball; latent class analysis; multiple imputation; statistical disclosure.

1 Introduction

Data of a wide range of varieties are important for the advancement of science, governmental and economic decision making, and the improvement of many areas of society. These potential positive outcomes fuel the desire for agencies to collect as much data as possible and, in many cases (e.g. scientific or government data collection) the desire to disseminate data as widely as possible for others to use. While the wide dissemination of data that is beneficial for research purposes is a noble goal, considerations must be made to prevent the disclosure of sensitive information contained in the data. Maintaining individuals' (or organizations') privacy is important for both legal and ethical reasons (among a host of others). As a society, it is largely agreed that privacy is an important part of a citizens life. In fact, privacy is viewed as so important that it is included as a basic human right in the United Nations Declaration of Human Rights (General Assembly of the United Nations, 1948). In the United States, legislative guarantees have been put in place to protect particular types of data including personal medical and educational data (HIPAA, 1996; FERPA, 1974).

So, what manners of data release are appropriate for public consumption? On the one extreme, no data would be released resulting in no possibility of a disclosure of a private piece of information. The other extreme would release all data with no regard for the disclosure of private information. These two often competing extremes must both be considered when disseminating data to the public. Ideally, data are released to the public in ways that are both useful for research and protective of privacy (often by anonymizing the data).

However, even when data are anonymized prior to sharing with other entities, disclosure of private information can still occur. Several high profile examples of these types of disclosures are described in Sweeney (2002), Narayanan and Shmatikov (2008) and Homer et al. (2008). In each of the aforementioned articles, it was demonstrated how to identify individuals in the data, even though the data in each case was believed to be anonymized to protect against the learning of private pieced of information. These types of disclosures, which are not due to unauthorized access to the data, are referred to as statistical disclosures.

A common definition of this type of disclosure (i.e. a statistical disclosure) comes from Dalenius (1977): "If the release of the statistics S makes it possible to determine the value [of confidential statistical data] more accurately than is possible without access to S, a disclosure has taken place." Ideally, the pieces of information that can be learned about an individual would be the same with or without the release of a statistics such as S. However, while this property is highly desirable in terms of individual privacy,

***Corresponding author: Gregory J. Matthews,** Department of Mathematics and Statistics, Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL, USA, e-mail: gmatthews1@luc.edu
Pétala Gardênia da Silva Estrela Tuy: Department of Mathematics and Statistics, Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL, USA
Robert K. Arthur: Five Thirty Eight, 147 Columbus Ave., 4th Floor, New York, NY 10023, USA

Dwork (2006) demonstrates that this is not achievable. The insurmountable hurdle here is auxiliary information. Regardless of what statistics S are released, it is always possible to possess, at least theoretically, auxiliary information that allows a data snooper (i.e. a data user looking to find information that should be private) to learn a private piece of information as a result of the release of the statistics S. As a result of this, much work has focused on developing methods for statistical disclosure control (SDC). Thorough reviews of this topic can be found in O'Keefe and Rubin (2015); Matthews and Harel (2011); Skinner (2009); Hundepool et al. (2006); Willenborg and de Waal (2001).

To give a simple example of statistical disclosure that is relevant to this setting, consider the following example. Imagine a university course with three students, and after the first exam the professor releases the average exam score. If one of the three students makes their exam score public, the two other students, with a small amount of math, now know everyone's exam score exactly. While that is a nearly trivial (but real!) example, the BBWAA Hall of Fame ballots presents a more interesting case for potential data snooping. In this article, we do not focus on SDC methods, rather we play the role of a data snooper in an attempt to learn private pieces of information by combining the release of summary statistics and auxiliary information. The target of our snooping experiment is the private ballots submitted by the members of the Baseball Writers Association of America (BBWAA) when voting for induction into the Baseball Hall of Fame Baseball Writers Association of America (2016). Here we attempt to learn as much as possible about the voters who chose not to make their ballots public.

Submitted ballots are kept private by the BBWAA, but they do release the total votes received by each nominee. It is difficult to learn much about individual voters based on just these aggregate totals, but with the totals and a collection of auxiliary information, it is possible to infer information about the voting habits of individual voters. The auxiliary information possessed in this case consists of a true subset of the full data. Since many of the voters are also writers, a subset of the voters make their complete ballots public, while many others choose not to.

Sources on the internet, such as the website authored by Thibodaux (2016), have aggregated these public ballots to produce a known subset of the full data set. The goal of this experiment is to use the vote totals released by the BBWAA and the known subset of ballots to learn about the voting habits of the voters who did not release their ballots. One type of private information learned here is related to the probability of a particular writer voting for a nominee rather than recreating the exact, unknown ballot of a writer. Further, we can attempt to identify voting patterns that are different between the voters who have public ballots in Thibodaux and the voters whose ballots remain private. This was attempted through the use of multiple imputation restricted based on marginal counts and latent class analysis.

The remainder of the manuscript discusses the process of voting for induction into the Baseball Hall of Fame and the data that were collected for this project; results from both univariate and multivariate analyses of the data; discussion of our findings; and finally a description of possible future work.

1.1 Baseball Hall of Fame voting details

The Baseball Hall of Fame (BBHOF) was created to honor the all-time greatest baseball players. The process by which a player gets elected to the BBHOF currently consists of eligible members of the Baseball Writers Association of America (BBWAA) voting for candidates who meet the requirements for inclusion on the ballot which is created by the Screening Committee. There are approximately 625 eligible members of the BBWAA. Once a ballot is generated, BBWAA electors may vote for up to 10 players on the ballot and write-in ballots are not allowed. Any player who receives votes on 75% of the ballots cast gains election into the BBHOF. Results from the voting process are released in aggregated form, but an individual's ballot remains private and is not released by the BBHOF. The BBHOF does, however, release the names of the voters who cast a ballot. For full details of the election process, see Baseball Writers Association of America (2016).

2 Data and methods

Data were collected from the BBWAA website containing players and their vote totals for 2014, 2015 and 2016. According to BBWAA website 571, 549, and 440 voters cast ballots in 2014, 2015, and 2016, respectively. While the BBHOF does not release individual voter's ballots, many voters, who are also writers, make their ballots public in some form. While these data are publicly available, it is not necessarily easy to collect and aggregate. However, these data are publicly available and have been aggregated by Thibodaux (2016), whose website includes a spreadsheet containing a collection of publicly available ballots for the years 2014 through 2016. 159 out of 571 (27.85%) were publicly released for voting in 2014, 203 out of 549 (36.98%) ballots were known from 2015 voting, and for voting in 2016, 307 out of 440 (69.80%) ballots were known. Vote totals and known vote totals are summarized in Table 1.

(Note: The 2014 and 2015 data were collected on April 30, 2015 and the 2016 data were collected on February 3, 2016. More known ballots may have been collected since then, however, the general concept of this manuscript is unaffected by the particular date that the data were collected.)

3 Univariate disclosure results

Each voter can be labeled as belonging to the known or unknown group based on whether they have voluntarily released their ballot to the public or not. Using the player vote totals released by the BBWAA and the publicly released ballots, it is straight forward to calculate the probability that a voter whose ballot was not released voted for a particular player.

In this setting, we are interested in looking for significant differences in the proportion of votes that a player received between the known and unknown groups of voters. To summarize the differences between the two groups of voters for a specific player, the odds ratio between the two groups is calculated. Specifically, the odds ratio for the j-th player is

$$\theta_j = \frac{\frac{p_{k,j}}{1-p_{k,j}}}{\frac{p_{u,j}}{1-p_{u,j}}}$$

Table 1: Total ballots cast, total ballots known, and percentage of ballots known for the years 2014, 2015 and 2016.

Year	Total ballots	Known ballots	% Known
2014	571	159	27.85%
2015	549	203	36.98%
2016	440	307	69.80%

where $p_{k,j}$ and $p_{u,j}$ are the proportion of voters in the known and unknown groups, respectively, who voted for player j. A Fisher's exact test can then be performed to test the hypothesis that H_0: $\theta_j = 1$ versus H_1: $\theta_j \neq 1$ for $j = 1, 2, \cdots, J$. If the null hypothesis is rejected, that means that there is a statistically significant different in the proportion of votes received by player j between the known and unknown groups. The family-wise error rate was chosen to be $\alpha = 0.05$ and a Holm correction Holm (1979) was used to account for testing multiple hypotheses. Tables 2 and 3 summarize the results of the comparisons between the odds of voting for each player in the known and unknown subsets of the data.

Of the 30 tests performed on the 2014 data, after adjusting for multiple hypothesis testing, the null hypothesis was rejected in five tests: Bagwell, Piazza, Raines, Schilling, and Thomas. In all of these cases, the group of known voters was more likely to vote for the player than a voter in the group that did not release their ballots. In 2015, which had a greater percentage of known ballots, the null hypothesis was rejected in 4 out of the 27 tests when accounting for multiple hypothesis testing. Players included in this group are Bagwell, Piazza, Schilling, and P. Martinez. Again in all of these cases, the group of voters in the known ballot group were more likely to vote for a player than the group of voters with unknown ballots. In 2016, only 3 out of 25 odds ratios were found to be significant: Bagwell, Schilling, and Mussina. These results are graphically summarized in Figures 1–3.

4 Multivariate results

In the univariate analysis, differences in voting patterns between the known and unknown groups were considered individually for each player. However, we also seek to find underlying classes of voters. This requires study of

Table 2: Significant odds ratios and unadjusted p-values for each player for the years 2014, 2015 and 2016.

Players	2014		2015		2016	
	OR	p-values	OR	p-values	OR	p-values
Bagwell	1.96	<0.001[a]	1.83	<0.001[a]	2.39	<0.001[a]
Piazza[b]	2.32	<0.001[a]	2.02	<0.001[a]	1.58	0.11
Raines	2.24	<0.001[a]	1.48	0.030	1.88	0.01
Schilling	2.02	<0.001[a]	1.95	<0.001[a]	2.58	<0.001[a]
Mussina	1.34	0.23	1.85	0.0028	2.09	<0.001[a]
Thomas[b]	4.27	<0.001[a]				
P. Martinez[b]			10.20	<0.001[a]		

[a]Significant after using the Holm correction.
[b]Elected to Baseball Hall of Fame.

Table 3: Non-significant odds ratios and unadjusted *p*-values for each player for the years 2014, 2015 and 2016.

Players	2014		2015		2016	
	OR	*p*-values	OR	*p*-values	OR	*p*-values
Sheffield			0.95	0.89	1.66	0.2
Smith	0.6	0.02	0.70	0.083	0.61	0.03
Morris	0.63	0.02				
Mattingly	0.35	0.03	0.64	0.22		
Kent	1.05	0.94	0.85	0.61	1.67	0.12
Walker	0.89	0.84	0.86	0.68	0.89	0.79
Maddux[a]	5.97	0.09				
Garciaparra			0.32	0.02	0.42	0.4
Glavine[a]	1.64	0.26				
Alou	0	0.28				
Sosa	0.61	0.29	0.73	0.48	0.9	0.96
Martinez	1.25	0.34				
McGriff	0.72	0.36	0.54		0.03	0.49
Gonzalez	0	0.37				
Boone			0	0.53		
Gordon			0	0.53		
Delgado			0.52	0.25		
Percival			0.57	1		
Nomo	0.52	0.88				
Snow	0	0.93				
Gagne	0	0.93				
Rogers	0	1				
Benitez	0	1				
Jones	0	1				
Palmeiro	1.01	1				
Erstad			0	1		
Anderson					∞	1
Eckstein					0.44	1
Edmonds					1.16	1
Griffey[a]					∞	0.03
Hoffman					0.88	0.65
Kendall					0	0.17
Sweeney					0.22	0.45
Wagner					1.11	0.89
Bonds	1.45	0.07	1.46	0.04	1.19	0.47
Clemens	1.29	0.22	1.34	0.12	1.15	0.58
Biggio[a]	1.21	0.44	1.69	0.04		
Trammell	1.16	0.59	0.92	0.76	1.46	0.09
E. Martinez			1.14	0.55	1.54	0.05
Johnson[a]			3.92	0.06		
Smoltz[a]			1.89	0.01		
McGwire	0.79	0.54	1.06	0.88	1.14	0.79

[a]Elected to Baseball Hall of Fame.

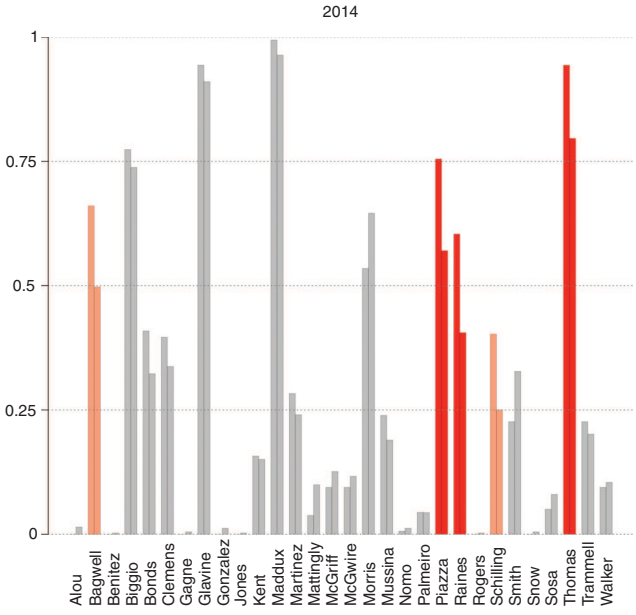

Figure 1: A side by side bar plot comparing the proportion of votes for each player in the known (left bar) and unknown (right bar) groups for the year 2014. *p*-Values from the Fisher exact test were adjusted using the Holm correction to account for multiple testing and significant results are presented with dark and light red representing significance at the $\alpha = 0.01$ and $\alpha = 0.05$ level, respectively.

the two groups from a multivariate perspective. Our goal is to be able to make statements comparing the relative proportions of latent class membership between the known and unknown voter groups.

To accomplish this, our approach is two-fold: First, the unknown ballots were imputed $M = 10$ times, subject to the restrictions of the known vote totals and 10 vote limit per ballot. Using the set of imputed ballots, latent class analysis (LCA) was performed to look for structure in the patterns in the voting, and then results of the LCA were combined across the imputations.

Ballots belonging to voters who kept their ballot private were imputed using the chained equation approach, or fully conditional specification (FCS), (Van Buuren et al., 2006) and implemented with the R package MICE (Van Buuren and Oudshoorn, 2007). Specifically, we view each ballot as an observation and each player on the ballot as a variable. Each row consists of a vector of 0's and 1's with a 1 meaning a voter voted for a player and 0 otherwise. We then build a logistic regression model for each player to predict the probability that a voter voted for a specific player conditional on which other players they voted for. The unknown ballots are imputed using these models, which ensures that combinations of players who are more or less likely to be voted for together are preserved in the unknown ballots. Additionally, this imputation was subject to restrictions on both the row (i.e. voters) and column (i.e. players) totals. The rows are subject to the constraint that each ballot must have at most 10 votes, and the column totals are constrained by the known vote total for each player that are released by the BBWAA. These constraints were imposed using what is essentially a version of an acceptance/rejection algorithm. Basically, an unknown

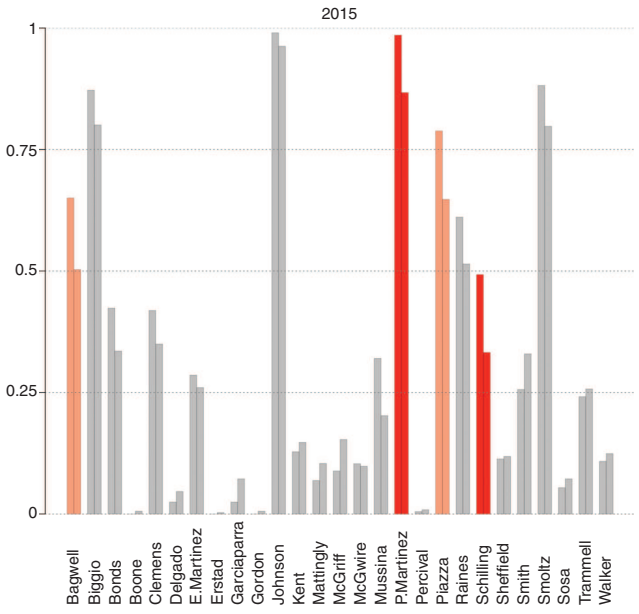

Figure 2: A side by side bar plot comparing the proportion of votes for each player in the known (left bar) and unknown (right bar) groups for the year 2015. *p*-Values from the Fisher exact test were adjusted using the Holm correction to account for multiple testing and significant results are presented with dark and light red representing significance at the $\alpha = 0.01$ and $\alpha = 0.05$ level, respectively.

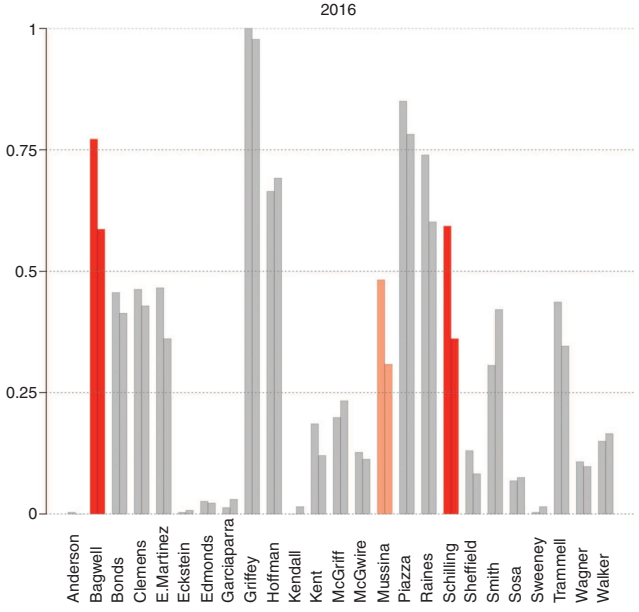

Figure 3: A side by side bar plot comparing the proportion of votes for each player in the known (left bar) and unknown (right bar) groups for the year 2016. *p*-Values from the Fisher exact test were adjusted using the Holm correction to account for multiple testing and significant results are presented with dark and light red representing significance at the $\alpha = 0.01$ and $\alpha = 0.05$ level, respectively.

ballot is imputed with a candidate ballot, and it is rejected if it violates and of the constraints otherwise it is accepted. This procedure is repeated until all of the unknown ballots were imputed and all constraints are satisfied.

For each of the ten imputations, LCA was performed on each imputed data set to classify voters into distinct classes based on the patterns of voters. Here, we performed LCA with a covariate, namely an indicator variable for whether the ballot was known or unknown. LCA was implemented here with the poLCA (Linzer and Lewis, 2011) package in R.

Finally, results of the LCA for each imputed data set were combined across imputations using appropriate combining rules (Little and Rubin, 1987) to reach a final overall estimate of the ratio of latent class odds comparing the public and private ballot groups. An odds ratio near 1 indicates that there is no difference in the prevalences of each latent class between the group of known and unknown ballots, whereas an odds ratio significantly different than 1 indicates that there are different prevalence rates of the two classes between the groups of known and unknown ballots.

4.1 Imputation

Let y_{ij} be the i-th observation (i.e. voter) and the j-th variable (i.e. player on the ballot) where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$ where each variable is partially observed (if a voter released their ballot). $y_{ij} = 1$ if the i-th voter voted for the j-th player and is 0 otherwise. We then define $y_{.j}$ to be the column vector containing 0's, 1's and missing values for the j-th player, and the observed and missing parts of $\mathbf{y}_{.j}$ are $\mathbf{y}_{.j}^{obs}$ and $\mathbf{y}_{.j}^{mis}$, respectively. Thus $\mathbf{Y}^{obs} = (\mathbf{y}_{.1}^{obs}, \mathbf{y}_{.2}^{obs}, \ldots, \mathbf{y}_{.J}^{obs})$, $\mathbf{Y}^{mis} = (\mathbf{y}_{.1}^{mis}, \mathbf{y}_{.2}^{mis}, \ldots, \mathbf{y}_{.J}^{mis})$ and $\mathbf{Y} = (\mathbf{Y}^{obs}, \mathbf{Y}^{mis})$. $\mathbf{Y}^{obs}$ is a $n_{obs} \times J$ matrix and $\mathbf{Y}^{mis}$ is a $n_{mis} \times J$ matrix and $n_{obs} + n_{mis} = N$.

FCS (Van Buuren et al., 2006) for multivariate imputation creates a model for the j-th variable conditional on all of the remaining $J - 1$ variables, and missing values of the j-th variable are imputed using this model. This process is then repeated for each of the J variables imputing values for all of the missing data.

In this setting, we make the assumption that the covariance structure of the public ballots and the private ballots are the same, which we believe to be a reasonable assumption here. Further, the imputations are bound by two restrictions: (1) vote totals received by each player and (2) a maximum of ten votes per ballot. Formally,

$$\mathbf{Y}^{mis}\mathbf{1}_J \leq 10\mathbf{1}_{n_{mis}}$$

and

$$\mathbf{1}'_{n_{obs}}\mathbf{Y}^{obs} + \mathbf{1}'_{n_{mis}}\mathbf{Y}^{mis} = V$$

where $\mathbf{1}_J$ and $\mathbf{1}_{n_{mis}}$ are each a column vector of length J and n_{mis}, respectively, consisting of all 1's, $V = (v_1, v_2, \ldots, v_J)$, and $V_j = \mathbf{y}'^{obs}_{.j}\mathbf{1}_{n_{obs}} + \mathbf{y}'^{mis}_{.j}\mathbf{1}_{n_{mis}}$ is the total votes received by player j for $j = 1, 2, \ldots, J$. We are interested in drawing imputations from

$$P(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{Y}^{mis}\mathbf{1}_J \leq 10\mathbf{1}_J, \mathbf{1}'_{n_{mis}}\mathbf{Y}^{mis} = V - \mathbf{1}'_{n_{obs}}\mathbf{Y}^{obs})$$

These restrictions were incorporated by first generating 100,000 synthetic ballots as candidates to potentially become imputed values for $\mathbf{Y}^{mis}$. Synthetic ballots that did not conform to the restriction $\mathbf{Y}^{mis}\mathbf{1}_J \leq 10\mathbf{1}_{n_{mis}}$ were removed from the potential ballots. From the remaining synthetic ballots that did conform to this restriction, a ballot was randomly sampled and used to impute a value of $\mathbf{Y}^{mis}$. This alone, however, does not satisfy the second restriction $\mathbf{1}'_{n_{mis}}\mathbf{Y}^{mis} \leq V$ and an iterative algorithm was incorporated to satisfy this condition. This worked by iterating through each player j and sampling from the potential candidate ballots that satisfy the first restriction (i.e. vote totals per ballot) until the second condition (i.e. Player j's vote total is equal to v_j). This process was repeated by iterating through the players $j = 1, 2, \ldots, J$ until $\mathbf{1}'_{n_{mis}}\mathbf{Y}^{mis} = V$. The resulting imputation contains imputed ballots with 10 or fewer votes, conforms to the player vote total restrictions, and maintains the covariance structure of the observed public ballots.

This imputation algorithm described previously is implemented here using R (R Development Core Team, 2007) with the function "mice" in the package MICE (Van Buuren and Oudshoorn, 2007). By default the "mice" function uses logistic regression when imputing binary data, and this is the setting that was chosen to impute the unobserved voting dat in this study.

4.2 Latent class analysis

Latent class analysis (LCA) (Collins and Lanza, 2010) is used to identify subgroups, types or classes of individuals. This type of modeling is used to identify patterns of responses based on observed characteristics and relates these patterns to a set of latent classes. The latent variable, or construct, is not observed directly but rather indirectly measured through two or more observed variables.

As before, let $j = 1, 2, \ldots, J$ be the number of observed variables (i.e number of players on the ballot) and the number of response categories for the j-th variable is $R_j = 2$ for all $j = 1, 2, \ldots, J$, and r_j, a specific level of the j-th variable, can take on values of either 0 or 1. Thus, the contingency table containing all possible ballots will have $W = 2^J$ cells with each cell of the contingency table associated with a unique response pattern, $\mathbf{z}_w$, and its frequency. For each voter, their vector of responses (i.e ballot) of the J variables, that is $(r_1, \ldots, r_J)$, is equal to some $\mathbf{z}_w$ in the contingency table with W cells. If we let $\mathbf{Z}$ be a J dimensional vector-valued random variable associated with all W ballots possibilities then for each pattern of response, there exists an associated $P(\boldsymbol{Z} = \boldsymbol{z}_w)$ such that $\sum_{w=1}^{W} P(\boldsymbol{Z} = \boldsymbol{z}_w) = 1$.

The model has two types of parameters: (1) conditional probabilities, that are the probabilities of the response $r_j = 0$ or $r_j = 1$ given the k-th latent class; and (2) prevalences, or unconditional probabilities, that are the probabilities of belonging to the k-th class of the latent variable L. These parameters are estimated using the method of maximum likelihood estimation and, since no closed form solution exists in this case, the likelihood is maximized using numerical optimization techniques. Considering an LCA model with J observed dichotomous variables and one categorical latent variable L with C classes, the marginal probability that $\boldsymbol{Z} = \boldsymbol{z_w}$ is

$$P(\mathbf{Z} = \mathbf{z}_w) = \sum_{c=1}^{C} P(\mathbf{Z} = \mathbf{z}_w|L = c)P(L = c).$$

and the estimated probability of belonging to each class is

$$P(L = c|\mathbf{Z} = \mathbf{z}_w) = \frac{P(\mathbf{Z} = \mathbf{z}_w|L = c)P(L = c)}{P(\mathbf{Z} = \mathbf{z}_w)}.$$

4.2.1 LCA with covariates

Including covariates in LCA is possible through a logistic regression model in which the dependent variable is latent (Agresti and Hoboken, 2008), and we seek to predict the probability of belonging to a particular latent class given some set of covariates. In our setting, we are trying to model the probability that a voter belongs to a particular latent class conditional on the configuration of their ballot. If we let $I(z_{wj} = r_j)$ be the indicator function equal to 1 when $z_{wj} = r_j$ and 0 otherwise, where z_{wj} is the j-th element of the w-th possible ballot configuration and X is the random variable representing whether a ballot is known or unknown, then the latent class model can be expressed as

$$P(\mathbf{Z} = \mathbf{z}_w \mid X = x) = \sum_{c=1}^{C} \gamma_c(x) \prod_{j=1}^{J} \prod_{r_j=0}^{1} \rho_{j,r_j|c}^{I(z_{wj}=r_j)},$$

where $\gamma_c(x) = P(L = c \mid X = x)$ is the probability of belonging to a latent class given the covariates, $\rho_{j,r_j|c}$ is the probability that the j-th value of the vector Z is equal

to r_j given that $L = c$, and $P(\mathbf{Z} = \mathbf{z}_w \mid L = c, X = x) = \prod_{j=1}^{J} \prod_{r_j=0}^{1} \rho_{j,r_j|c}^{I(z_{wj}=r_j)}$. With a single covariate X, $\gamma_c(x)$ can be expressed as

$$\gamma_c(x) = P(L = c \mid X = x) = \frac{\exp^{\beta_{0c}+\beta_{1c}x}}{1 + \sum_{c=1}^{C-1} \exp^{\beta_{0c}+\beta_{1c}x}}$$

with the reference category corresponding to the latent class C.

In our setting, x_i is equal to 0 or 1 if the i-th voter's ballot is unknown or known, respectively. Logistic regression for LCA produces an estimate of the effect of the covariate, x, on each latent class compared to the other latent classes. In LCA with covariates, regression coefficients, β, are estimated rather than prevalences.

4.2.2 Combining rules

Ultimately, based on model fitting criteria, the model with $C = 2$ latent classes was chosen as the best fitting model. In this case, for each imputation there is only one β_0 and β_1 estimate. Using Rubin's combining rules (Little and Rubin, 1987) we can combine the estimates of across imputations as follows:

$$\bar{Q}_M = \sum_{m=1}^{M} \frac{\beta_1^{(m)}}{M}$$

$$B_M = \sum_{m=1}^{M} \frac{(\beta_1^{(m)} - \bar{Q}_M)^2}{M-1}$$

$$\bar{U}_M = \sum_{m=1}^{M} \frac{var(\beta_1^{(m)})}{M}.$$

$\bar{Q}_M$ is the overall estimate of β and $T_M = (1 + \frac{1}{M})B_M + \bar{U}_M$ is the used to estimate the variance of this estimate, $\bar{Q}_M$. In this setting, we argue that each imputed data set is actually a completed population because we are not interested in some larger theoretical population; the collection of all voters who voted in a particular year is our target population. Therefore, when each imputed data set is treated as a population we must set $\bar{U}_M = 0$ giving us $T_M = (1 + \frac{1}{M})B_M$.

4.3 Results

After performing multiple imputation LCA identifies (based on AIC) two latent classes in each of the three years. Figures 4–6 summarize our results by showing the log odds ratio comparing the odds of voting for a player in latent class 1 versus latent class 2. These were computed

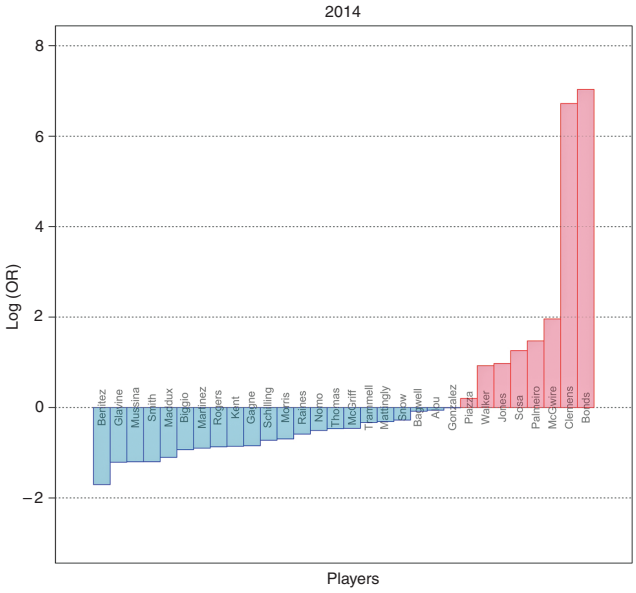

Figure 4: Log odds ratio comparing the likelihood of a voter from class 1 or class 2 voting for a particular player in 2014.

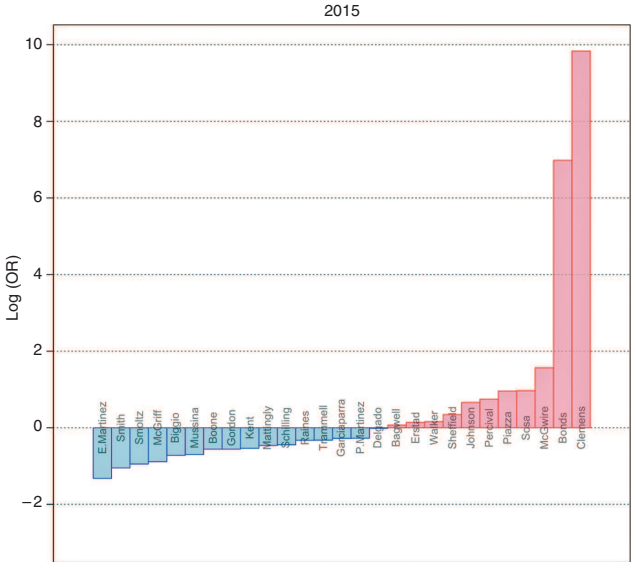

Figure 5: Log odds ratio comparing the likelihood of a voter from class 1 or class 2 voting for a particular player in 2015.

by first looking at the probabilities of voting for a specific player given that a voter belonged to latent class 1 or 2. As an example, consider Barry Bonds in 2016. The estimated probability that a voter in latent class 1 voted for Barry Bonds was 0.9955 whereas the corresponding probability for a voter who belonged to latent class 2 was 0.0172. Therefore, we can calculate the odds ratio between these two groups as follows:

$$\frac{\frac{0.9955}{1-0.9955}}{\frac{0.01722}{1-0.01722}} = 12625.6.$$

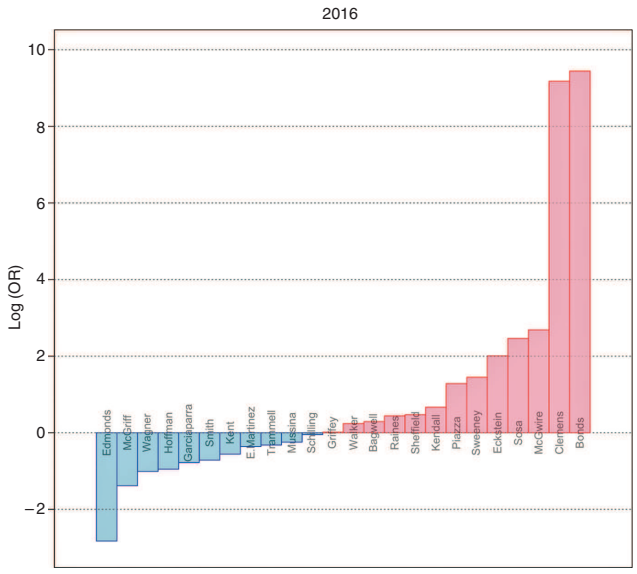

Figure 6: Log odds ratio comparing the likelihood of a voter from class 1 or class 2 voting for a particular player in 2016.

For the sake of displaying all of these odds ratios on a reasonable scale, the log of each of these odds ratios is considered. For Barry Bonds this is $log(12625.6) = 9.4435$ and can be seen in the rightmost bar in Figure 6.

The height of each bar therefore represents the log of the odds ratios comparing the likelihood of voting for a player given that a voter is in latent class 1 versus latent class 2. Red bars indicate that a voter in latent class 1 was more likely to vote for a player than a voter in latent class 2, whereas blue bars represent to opposite (i.e. a voter in latent class 2 was more likely to vote for this player than a voter from latent class 1.).

In all three years, the three largest differences in favor of latent class 1 are Bonds, Clemens, and McGwire. These three players have all been strongly linked to performance enhancing drugs in the Mitchell Report (Mitchell, 2007). Several other players linked to performance enhancing drugs show up in latent class 1 including Sosa, Piazza, Sheffield, and Palmeiro. Latent class 1, therefore, can reasonable be considered to be a latent class related to players associated with performance enhancing drugs (PED's).

Latent class 2 simply seems to be the opposite in that it is characterized by players not associated with PED's or other performance enhancing drugs. The players who had the highest odds ratios in favor of class 2 in 2014 were Benitez, Glavine, Mussina, Smith, Maddux, and Biggio, in 2015 were E. Martinez, Smith, Smoltz, McGriff, Biggio, and Mussina, and in 2016 were Edmonds, McGriff, Wagner, Hoffman, Garciaparra and Smith.

Tables 4–6 show the probability that a voter belongs to latent class 1 or latent class 2 given that they are in the known or unknown voting group for the years 2014, 2015, and 2016, respectively. So, for example, in 2014, the group of voters with known ballots (i.e. they released their ballots publicly) had a 40.63% chance of belonging to latent class 1 (i.e. the pro-PED's group). This is as opposed to a 34.61% chance that a voter with an unknown ballot belonged to latent class 1. Further the odds of falling into latent class 1 vs latent class 2 are 1.293 times larger in the known group compared to the unknown group with a 95% confidence interval of (1.073, 1.557).

In 2015, the odds ratio was slightly larger at 1.387 with a 95% confidence interval (1.234, 1.560) and in 2016 the odds ratio dropped belong both 2014 and 2015 to 1.168 with a 95% confidence interval of (1.077, 1.278). These odds ratios and intervals are summarized in Table 7. Notably, in all three years the 95% confidence interval does not contain 1 indicating that the probability of a voter falling into latent class 1 is significantly between the known and unknown group of voters in all three years.

Table 4: Probability of belonging to class 1 or class 2 given that a voter is in the known or unknown group for 2014.

	Known	Unknown
Class 1 (pro-PED's)	0.4063	0.3461
Class 2 (anti-PED's)	0.5937	0.6539

Class 1 corresponds to the group that is more likely to vote for players related to PED's and class 2 is the group less likely to vote for players related to PED's.

Table 5: Probability of belonging to class 1 or class 2 given that a voter is in the known or unknown group for 2015.

	Known	Unknown
Class 1 (pro-PED's)	0.417	0.3401
Class 2 (anti-PED's)	0.583	0.6599

Class 1 corresponds to the group that is more likely to vote for players related to PED's and class 2 is the group less likely to vote for players related to PED's.

Table 6: Probability of belonging to class 1 or class 2 given that a voter is in the known or unknown group for 2016.

	Known	Unknown
Class 1 (pro-PED's)	0.4469	0.4089
Class 2 (anti-PED's)	0.5531	0.5911

Class 1 corresponds to the group that is more likely to vote for players related to PED's and class 2 is the group less likely to vote for players related to PED's.

Table 7: Odds ratios for belonging to latent class 1 (pro-PED's) vs. latent class 2 (anti-PED's) comparing the group or known and unknown ballots.

Year	Odds ratio (confidence interval)
2014	1.293 (1.073, 1.557)
2015	1.387 (1.234, 1.560)
2016	1.168 (1.077, 1.278)

An odds ratio of 1 here indicates that an individual voter in the known group is more likely to be in latent class 1 than latent class 2.

5 Discussion

BBWAA members cast private ballots each year to decide which players get inducted into the Baseball Hall of Fame. The BBWAA aggregates these ballots and releases vote totals for all players, but does not release individual ballots. However, some of the voters choose to release their ballots to the public. In this manuscript, the aggregate voting results released by the BBWAA along with a subset of ballots voluntarily released by certain voters were used to attempt to learn as much as possible about the voting behaviors of the writers who wished to keep their ballots private. We first examined the difference in the odds of voting for individual players between the two groups of voters (i.e known ballots vs unknown ballots). We then went on to use multiple imputation to create plausible ballots for the unknown group of voters based on the correlation structure of voters in the known group and the restriction that each ballot can contain a maximum of 10 players. With ballots imputed, latent class analysis was then used to look for class structure in the voting habits of the BBWAA. Results of latent class analysis were combined across imputations, and differences based on class membership between individuals whose ballots were known were compared to those individuals whose ballots were unknown in an effort to learn about the group of individuals who did not publicly release they ballots.

5.1 From a privacy standpoint

While in this example, nothing specific can be learned about an individual voter with certainty, we do discover some major differences between the groups pertaining to individual players and further we find that the group of voters who did not release their ballots are less likely than the group of voters with known ballots to vote for players actually or perceived to be connected to PED's. While specific voters who do not release their ballot maintain a high degree of privacy, their privacy is certainly weakened by learning the probability that they voted one way or the other for a specific player in a smaller group of people, and with the addition of latent class analysis we can learn about the likelihood of an individual in a specific group (i.e. known or unknown ballots) belonging to one of two well defined classes (i.e. pre-PED's vs anti-PED's). Further, this weakened privacy is a direct result of the individuals who participated in the voting deciding to relinquish their right to a private ballot and publishing their ballots through various media channels. This raises the question about what an individual participant in a database who does not value their own privacy owes to another participant in the same database who does value they privacy, since an individuals' wish to maintain the privacy of their own data is dependent in some way on the other participants maintaining the privacy of their own data.

5.2 From a baseball standpoint

Players identified in latent class 1 were overwhelmingly associated with performance enhancing drugs, both in the Mitchell Report and in other investigative journalism. While no player can be shown authoritatively not to have used PEDs, players in latent class 2 seemed to be among those with the cleanest reputations (e.g. Greg Maddux, Tom Glavine). For this reason, the primary difference between classes seems to come from the reputation of the player as a PED user.

The difference in support for these two classes depending on whether a voter released their ballot or not is consistent with older voters taking more conservative attitudes toward PED users, which is also noted in Pollis (2015). While some voters consider all players independent of any allegations of PED use, others strongly consider a player's resume in the context of their reputation for steroid use. Voters who have publicly denounced steroid users or announced that they would not support them tend to have covered baseball for longer periods of time. Similarly, voters who do not release their ballots tend to have be longer-serving members of the BBWAA. As one of the most divisive issues of the last decade, it makes sense that PED use would emerge as the major explanatory variable in the LCA, and that older writers would separate out from younger ones in this analysis.

References

Agresti, A. and J. W. Hoboken. 2008. "An Introduction to Categorical Data Analysis." *Journal of Applied Statistics* 35:283–291.

Baseball Writers Association of America. 2016. "BBWAA Election Rules." http://baseballhall.org/hall-of-famers/bbwaa-rules-for-election. Accessed on January, 2016.

Collins, L. M. and S. T. Lanza. 2010. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. New York: Wiley.

Dalenius, T. 1977. "Towards a Methodology for Statistical Disclosure Control." *Statistik Tidskrift* 15:429–444.

Dwork, C. 2006. "Differential Privacy." Pp. 1–12 in *ICALP*. Berlin: Springer.

FERPA. 1974. "Family Educational Rights and Privacy Act of 1974." 20 *U.S.C.* § *1232*.

General Assembly of the United Nations. 1948. "Universal Declaration of Human Rights." *United Nations Resolution 217 A (III)*.

HIPAA. 1996. "Health Insurance Portability and Accountability Act." *Pub.L. 104-191, 110 Stat. 1936*.

Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6: 65–70.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. 2008. "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-density SNP Genotyping Microarrays." *PLoS Genetics* 4:e1000167.

Hundepool, A., J. Domingo-ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. S. Nordholt, G. Seri, and P. Paul De Wolf. 2006. "A CENtre of EXcellence for Statistical Disclosure Control Handbook on Statistical Disclosure Control Version 1.01".

Linzer, D. and J. Lewis. 2011. "poLCA: An R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software* 42:1–29. http://www.jstatsoft.org/v42/i10/.

Little, R. J. A. and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Matthews, G. J. and O. Harel. 2011. "Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy." *Statistics Surveys* 5:1–29.

Mitchell, G. 2007. "Report to the Commissioner of Baseball of an Independent Investigation into the Illegal Use of Steroids and Other Performance Enhancing Substances by Players in Major League Baseball." http://files.mlb.com/mitchrpt.pdf. Accessed on January, 2016.

Narayanan, A. and V. Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets." Pp. 111–125 in *Proc. of the 29th IEEE Symposium on Security and Privacy*. IEEE Computer Society. http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf. Accessed on January, 2016.

O'Keefe, C. M. and D. B. Rubin. 2015. "Individual Privacy Versus Public Good: Protecting Confidentiality in Health Research." *Statistics in Medicine* 34:3081–3103. http://dx.doi.org/10.1002/sim.6543.

Pollis, L. 2015. "Ninety Percent Mental: Are Secret Ballots Ruining Cooperstown?" *Baseball Prospectus*. http://www.baseballprospectus.com/article.php?articleid=25306. Accessed on May, 2016.

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. greg. ISBN 3-900051-07-0.

Skinner, C. 2009. "Statistical Disclosure Control for Survey Data." Pp. 381–396 in *Handbook of Statistics Vol. 29A: Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C. R. Rao. Amsterdam: Elsevier.

Sweeney, L. 2002. "k-anonymity: A Model for Protecting Privacy." *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 10:557–570.

Thibodaux, R. 2016. "Baseball Hall of Fame Vote Tracker." http://www.bbhoftracker.com/. Accessed on April 30, 2015 and February 3, 2016.

Van Buuren, S., J. Brand, C. Groothuis-Oudshoorn, and D. Rubin. 2006. "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation* 76:1049–1064.

Van Buuren, S. and C. Oudshoorn. 2007. *mice: Multivariate Imputation by Chained Equations. R package version 1.16.* http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm. Accessed on May, 2015.

Willenborg, L. and T. de Waal. 2001. *Elements of Statistical Disclosure Control*. Berlin: Springer-Verlag.