Spring 2017

Telephone Polls and PPS Sampling: A Potential Boon to the Polling Industry

Jade McKay Burt
Utah State University

Follow this and additional works at: http://digitalcommons.usu.edu/etd
Part of the Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation
Burt, Jade McKay, "Telephone Polls and PPS Sampling: A Potential Boon to the Polling Industry" (2017). All Graduate Theses and Dissertations. 6835.
http://digitalcommons.usu.edu/etd/6835

This Thesis is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact dylan.burns@usu.edu.
Telephone Polls and PPS Sampling: A Potential Boon to the Polling Industry

by

Jade McKay Burt

Thesis submitted in partial fulfillment of the requirements for the degree of

DEPARTMENTAL HONORS in

Statistics in the Department of Mathematics & Statistics

Approved:

Thesis/Project Advisor
Dr. Damon Cann

Departmental Honors Advisor
Dr. David Brown

Thesis Reader
Dr. Richard Cutler

Director of Honors Program
Dr. Kristine Miller

UTAH STATE UNIVERSITY
Logan, UT

Spring 2017
Abstract

In the wake of the 2016 election, the polling industry has no shortage of critics. While these are difficult times for the industry as a whole, there are exciting innovations happening that will serve to benefit and revitalize the industry for years. One of these exciting innovations is Probability Proportional to Size (PPS) sampling. I will elaborate on what PPS sampling is and provide a mathematical foundation for its use in polling. I also discuss what some of the myriad of issues plaguing the polling industry are and then show how PPS sampling can be used to remedy many of these ills. Finally, I look at a real-world application of PPS sampling. The Mia Love internal polling team, Y2 Analytics, granted me access to their PPS data. I use it to show that we can accurately model the electorate using PPS samples and that polls conducted by this method are at least as accurate as other polls using simple random samples.
Acknowledgements

First of all, I’d like to thank Dr. Damon Cann. None of this would have been possible if not for him. He’s been an incredible mentor and also had the connection that allowed me to do an internship with Y2 Analytics from which I got the idea for this thesis. I’d like to thank Y2 Analytics and particularly Scott Riding and Quin Monson for helping me to give this project more definition and for allowing me to access their data including past polls and voter files freely. In addition, I’d like to thank the Department of Mathematics and Statistics at Utah State for allowing me the flexibility to take this class as both a 5000-level elective and have Dr. Cann as my mentor for the project.
# Table of Contents

Final Written Project .................................................................................................................. 1

Introduction ............................................................................................................................... 1

Problems with Contemporary Polling ....................................................................................... 1

An Introduction to PPS Sampling ............................................................................................. 4

The Mathematical Explanation for PPS Sampling ...................................................................... 5

The Effectiveness of PPS Polling in Real Life Applications ....................................................... 7

Predicting Turnout using a PPS Model ....................................................................................... 8

The Accuracy of PPS Polling Results ......................................................................................... 11

How PPS Sampling Remedies Contemporary Problems with Polling ..................................... 15

Conclusion .................................................................................................................................. 17

Reflective Writing Section ......................................................................................................... 18

Bibliography .............................................................................................................................. 22
List of Figures

Figure 1. Histogram of Predicted Turnout Probability for all Voters on the Voter File October 2014 ........ 9
Figure 2. Model Prediction vs. Performance ................................................................. 10
Figure 3. Mia Love Vote Percentage Prediction and Outcome ........................................... 12
Figure 4. Doug Owens Vote Percentage Prediction and Outcome ...................................... 13
Figure 5. Mia Love Margin of Victory Predictions and Outcome ....................................... 14

List of Tables

Table 1. Comparison of Predicted Turnout vs. Actual Turnout by Decile ............................. 10
Introduction

The polling industry has taken a lot of flak for some supposed high-profile polling disasters in the last few years. People blame them for missing the Brexit vote, the 2016 election, the 2014 midterms, the Israeli Parliamentary Elections, the British Parliamentary Elections, and more. Many high-level politicians have taken to saying that polling is lies or is rigged. So, with all of this vitriol directed toward the industry, why does polling continue? What purpose does it fill? Polling helps legislators understand where their constituents are on a variety of issues so that they can effectively represent them and be their mouthpiece. Polling is also used to help candidates identify persuadable voters, effectively allocate resources, and target messages. Another reason for polling is to understand the demographic composition of a nation or an electorate or for research on an entirely non-political issue. Companies engage in polling to better understand their market and to test potential advertising messages.

Because polling is so important to business, government, and many other organizations such as non-profits, I think it is important to combat this negativity and the approach I have chosen is to recommend a potential method of sampling that addresses many of polling’s problems and may help it to rehabilitate its reputation. My purpose in writing this paper is to discuss a sampling method known as Probability Proportional to Size (PPS) sampling that has the potential to serve as a remedy for many of the problems facing the polling industry. I plan to show its successes in combatting many of these problems using polling data acquired by the company Y2 Analytics and firmly believe that PPS sampling for polls is a viable option for getting lower-cost, but reliably accurate polls.

Problems with Contemporary Polling

Despite the abundance of applications, the polling industry of today faces ever-increasing challenges. One well-documented challenge is declining response rates – particularly when surveys are conducted over the phone (Sheehan, 2001) (Leeuw, 2001). Declining response rates cause increased
costs because polling companies must dial more people to get the same number of respondents as before. Adding to the challenge is the existing regulatory scheme regarding the legality of dialing cell phones. The way the Telephone Consumer Protection Act has been interpreted is also detrimental to the polling industry. In 2015, the Federal Communications Commission issued a Declaratory ruling saying the following: “It shall be unlawful... to make any call (other than a call made for emergency purposes or made with the prior express consent of the called party) using any automatic telephone dialing system or an artificial prerecorded voice... to any telephone number assigned to a... cellular telephone service” (Federal Communications Commission, 2015). Although at first glance this may seem innocuous, the problem arises from the fact that the terms “prior express consent” and “automatic telephone dialing system” are not defined. Furthermore, the penalty for violation of this law is between $500 and $1500 for each violation. Each contact attempt irrespective of whether it was completed counts as a violation, so the potential fines for companies are enormous. It is not uncommon to make 1000 attempts in order to secure a sample of 100 cell phone interviews, so fines could be as high as $1.5 million for non-compliance in a single survey effort. The law’s current interpretation is that it is not enough to simply dial cell phones – it’s necessary to make sure that cell phone numbers aren’t loaded into software that has autodial capability. An iPhone has autodial capabilities and would therefore be covered by this interpretation of the law. Since there are now more people abandoning their landlines for cell phones, this strict interpretation of the law is making life difficult for the survey research industry and requiring greater efficiency in making calls because dialing cell phones, since it now takes even longer than before, has become more expensive. That leads to the next thing harming the polling industry. In addition to declining response rates and suffocating regulation, polling companies face increased costs of conducting high-quality polling. Despite research showing that online polls can be representative and accurate (Barber, 2014) (Ansolabehere, 2014), people and news organizations are still skeptical of internet polls. While it is true that there are many internet polls that are unrepresentative and engage in
convenience sampling, such as those that allow individuals to vote repeatedly or opt-in to the poll, there are many reputable polls conducted online. In fact, online polls offer some advantages that live-interviewer telephone polling doesn’t offer. Studies have shown that respondents are more willing to express views society would frown upon if they don’t have to admit it to a live human being (Morning Consult, 2015). Despite the fact that online polling can be reputable, because of the abundance of shoddy online polls, The New York Times refuses to publish online polls to this day (The New York Times, 2008). Sure, polls can be done cheaply by telephone, but those who do them on the cheap get what they pay for. Generally the product is a totally unrepresentative Interactive Voice Response (IVR) poll also known as a robopoll. So, through a combination of good reasons and unfounded tradition, there is a great need for a cheaper, more efficient, yet accurate method of sampling that saves enough in costs to allow pollsters to conduct accurate and representative telephone surveys with live voice interviewers. While some of these things are beyond their control, there are some solutions that have helped them stay abreast of the changing seas of the public opinion industry. Finally, it is difficult for traditional polls to correctly identify likely voters. The traditional method pollsters have used to attempt to solve this problem is simply to ask respondents how likely they are to vote in a particular election. The question goes something like this: “What are the chances of you voting in the election for president in November?” This is called a respondent’s self-reported likelihood of voting. Generally, allowed responses include a scale such as: “Almost certain to vote, will probably vote, the chances are 50-50, and the respondent doesn’t think they will vote. However, each pollster does the wording of this question and the set of answer responses a little differently. The pollster will set an arbitrary threshold on that scale (generally those 50-50 or below) and if a respondent answers below it, then the interview is terminated and they are not considered a likely voter. Studies have examined how effective self-reported likelihood of voting as an indicator of likelihood to vote is using voter-validation data available through the American National Election Studes (ANES). They showed that the major source of error was
misreporting by respondents who vote regularly but not always. Those individuals are prone to exaggerate their likelihood to vote because of the social desirability of being a voter (Petrocik, 1991) (Silver, 1986). In a study of post-election vote validation involving self-reported vote likelihood in polling data from the 2008 US Presidential Election, the 2009 New Jersey General Election, and the Wisconsin Recall Elections in 2011, the findings were that “a meaningful fraction of those who say they will vote do not (i.e., they “flake-out”), and that a large fraction of those who say they will not vote, in fact, do (i.e., they “flake-in”)” (Rogers, 2013). They found that for the 2008 general election, the flake-out rate was 13.3%, for the 2009 New Jersey General Election, the flake-out rate was an astounding 54.2%, and for the 2011 Wisconsin Recall, the flake-out rate was 17.3%. The rate of voters flaking in is even more astounding. For the 2008 General Election, the flake-in rate was 54.2%. For the 2009 New Jersey General Election, the flake-in rate was 29.3%. For the 2011 Wisconsin Recall Election, the flake-in rate was 39.5% (Rogers, 2013). Obviously, there are some serious methodological problems with simply asking people whether they are going to vote and being able to predict what the electorate will look like is of tantamount importance to producing an accurate poll.

**An Introduction to PPS Sampling**

The method I will discuss in this paper that has the potential to remedy many of the problems mentioned before is Probability Proportional to Size (PPS) sampling. PPS sampling has been studied in relation to internet polling (Barber, 2014), but lacks research in relation to telephone interviews.

The PPS samples I will discuss in this paper were acquired in the manner subsequently described. The first step is to produce a probability of voting for each individual voter in the voter file employing variables such as vote history, age, and gender, found therein by using a logistic regression model. As I will discuss later, this is actually a far more accurate predictor of who will actually turn out than self-reported likelihood to vote which is used by most pollsters today (Rogers, 2013). The second step is to draw a sample from the list of voters with probabilities proportional to the likelihood of voting.
This is known as the first stage of the sampling process. Then, from the voters that were sampled in the first stage, we divide them up into quintiles and draw a simple random sample from a list of all the voters in each quintile. This is the second stage of the PPS sampling process and those individuals selected in the simple random sample are considered the final sample and are the ones that will receive telephone calls to be a part of the survey. When the call house is collecting survey interviews, they are instructed to ensure that each quintile is equally represented in the final completed interviews.

This process makes intuitive sense. The first step basically takes the entire electorate found on the voter file and gives us a sample that more closely mirrors the electorate that will turn out on Election Day. It functions like a likely voter screen. From there, the second-stage sample reduces the number of individuals in the first stage of the sample and gives us a sample of a manageable size so that they can actually be dialed in a reasonable amount of time. After the survey is conducted, any additional weighting done is to match the demographics of the sample that was drawn in stage one because that sample is reflective of what we predict the population being studied (the likely electorate) looks like.

**The Mathematical Explanation for PPS Sampling**

Chapter six of Sharon L. Lohr’s book *Sampling: Design and Analysis* provides an excellent explanation of the mathematical basis for PPS sampling. She writes, “When first presented with the idea of unequal-probability sampling, some people think of it as unnatural or contrived. On the contrary, for many populations with clustering, unequal-probability sampling... produces a sample that mirrors the population better than an equal probability sample.” It provides a fantastic example of sampling with unequal probabilities and how it can yield an equal probability of every unit being sampled while preserving resources. PPS and cluster samples are particularly useful when there is a shortage of resources; for example, when a survey needs to be conducted in person. They give an example of an interviewer seeking to gather a sample of as many as 37,652 patients at 294 nursing homes. However, the patients were not spaced out equally between the nursing homes and there aren’t enough
interviewers to send at least one to each nursing home. If they took a simple random sample of nursing homes and then another simple random sample of patients within each selected home, it would artificially inflate the likelihood of sampling the individuals at nursing homes with low numbers of patients. To ensure that each patient has an equal chance of being sampled, you have to sample nursing homes in proportion with the size of the population of patients or beds at the home. So, the first stage of probability proportional to size sampling involves randomly drawing a sample of nursing homes with probabilities of being selected for the initial sample proportional to the number of beds. Then, you would take a simple random sample of a pre-specified number of occupants from a list of all the beds within the nursing home. If this technique is followed and one assumes that the number of residents equals the number of beds and assumes that a home has the same number of beds when visited as are listed in the sampling frame which are both reasonable assumptions, then the sampling design results in every resident having the same probability of being included in the sample. This technique can be generalized to situations that don’t involve hospitals as long as the same general procedures are followed. Major advantages of this technique are that the cost is known before selecting the sample and the same number of interviews are taken at each home which equalizes the work load among interviewers. This helps combat one of the major problems facing polling discussed earlier: skyrocketing costs. With PPS samples, it is possible to know how much acquiring a representative sample will cost with much more accuracy than other sampling methods allow.

In William G. Cochran’s book “Sampling Techniques”, he goes into more detail about the math behind PPS Sampling. He explains the following: Let $M_i$ be the total number of elements in the $i$th unit. If all of the $M_i$ are known, it allows us to use PPS sampling which involves selecting the units with probabilities proportional to their sizes $M_i$. The technique was first developed by Hansen and Hurwitz in 1943. However, the earliest methods of performing this type of sampling were only functional for small numbers of units. Let $N$ represent the number of units we are sampling from. Lahiri in 1951 provided an
alternate method that streamlines the process of acquiring this sample. He says let $M_{\text{max}}$ be the largest of the $M_i$. Draw a random number between 1 and $N$; suppose this is $i$. Now draw another random number $m$ between 1 and $M_{\text{max}}$. If $m$ is less than or equal to $M_i$, the $i$th unit is selected. If not, try another pair of random numbers.

That all works fine for sampling one unit, but since we will be performing survey research, we will need to sample large numbers, $n$, of units. We will also need to sample without replacement which complicates PPS sampling. Doing sampling without replacement means that keeping selection probabilities proportional to the chosen sizes is more difficult. Thankfully, although the $n$ for survey research is large, the number of units $N$ is very large. In a state, $N$ is incredibly large and often over a million. Pulling a sample of size $n=5000$ or so does little to shift selection probabilities in proportion the sizes in question (the likelihood of turning out to vote).

The method of PPS sampling most closely aligned with what we did at Y2 Analytics was suggested by Madow (1949). One of its biggest advantages is that the sample is easy to draw for any $n$. This is because the method deals with the issue of shifting probabilities as we draw each successive element in the sample by retaining $\pi_i = nz_i$ true for all $n$. To understand that equation, let $\pi_i$ be the probability that the $i$th unit is in the sample having been selected on any draw. Recall that $n$ is the size of our sample, and let $z_i$ be defined as the probability that the $i$th unit is drawn first. Basically, in lay terms, we’re saying that the chance of being included in the sample is the same on the first draw as on any other draw.

**The Effectiveness of PPS Polling in Real Life Applications**

I will show PPS sampling in action in telephone surveys using data provided by Y2 Analytics, a survey research firm based out of Salt Lake City. They are the company with which I did my internship in the summer of 2016 and performed Representative Mia Love’s internal polling for her 2014 and 2016 election campaigns. They have graciously granted me access to their internal polls for the purposes of
writing this paper. I have selected the tracking polls conducted in the week before the 2014 election. I would have liked to have selected the 2016 election and Y2 Analytics granted me access to their data for that election, but the state of Utah delayed in producing a voter file that was updated after the 2016 election and I didn’t have access to vital information contained in the voter file to do this analysis for the 2016 election. The first thing I want to look at is the predicted turnout probabilities. I will analyze how accurately they can predict someone’s likelihood of turning out to vote based on the factors that go into a predicted turnout likelihood which determines how likely that individual is to be selected for a PPS sample. We will look at the list of individuals registered to vote on the voter file and the probability that our model assigned them of turning out to vote. We will sort the voter file into deciles based on how likely people were to turn out. Then, I will match those individuals against the October 2015 voter file which contains accurate information about who actually voted in the 2014 election. I will calculate what percentage of the sampled individuals in each turnout decile actually turned out to vote and compare that to the average percentage of predicted turnout in each decile. We should see a lower voter turnout percentage with each successive decile and we should also observe that the predicted voter turnout likelihood will be close to the actual percentage of people that turned out to vote.

Predicting Turnout Using a PPS Model

For starters, the model used by Y2 Analytics to predict turnout in Utah’s 4th district in 2014 is based on a logistic regression where the dependent variable is turnout with two outcomes (yes and no). The independent variables they used as predictors in this model were found in the voter file and are age, gender, vote history, and how many years the individual has been registered to vote in Utah. Their model became a little more sophisticated for the 2016 election, but those were the only predictors they used to create the 2014 turnout model. By vote history, I mean that they looked at whether individuals had turned out to vote in previous elections and based the model largely on that information found in the voter file. The number of years an individual has been registered to vote in Utah is an indicator of
stability which is correlated with likelihood of turning out to vote. The chart below is a histogram showing the distribution of turnout probabilities across all voters in the congressional district based on the model designed by Y2 Analytics for the 2014 election.

Figure 1. Histogram of Predicted Turnout Probability for all Voters on the Voter File October 2014

Now that I have explained how the turnout probabilities were calculated, I will show how accurate Y2 Analytics was at predicting turnout validating its use of these predictions in its PPS sample. I divided the electorate up into deciles based on turnout probability so that the lowest decile contains the lowest 10% of predicted turnout probabilities. I averaged the turnout probabilities in those deciles and found the percentage of people that actually voted in each decile according to the voter file. Those results are described in the table and graphic below.
Table 1. Comparison of Predicted Turnout vs. Actual Turnout by Decile

<table>
<thead>
<tr>
<th>Decile</th>
<th>Average Predicted Turnout Probability</th>
<th>Actual Turnout Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05417687</td>
<td>0.06833572</td>
</tr>
<tr>
<td>2</td>
<td>0.10121362</td>
<td>0.13138138</td>
</tr>
<tr>
<td>3</td>
<td>0.15067247</td>
<td>0.18406728</td>
</tr>
<tr>
<td>4</td>
<td>0.23412324</td>
<td>0.28005078</td>
</tr>
<tr>
<td>5</td>
<td>0.33131091</td>
<td>0.32719496</td>
</tr>
<tr>
<td>6</td>
<td>0.45591116</td>
<td>0.45116385</td>
</tr>
<tr>
<td>7</td>
<td>0.61946386</td>
<td>0.59121137</td>
</tr>
<tr>
<td>8</td>
<td>0.76605145</td>
<td>0.73275525</td>
</tr>
<tr>
<td>9</td>
<td>0.89564345</td>
<td>0.84424306</td>
</tr>
<tr>
<td>10</td>
<td>0.97634050</td>
<td>0.93154256</td>
</tr>
</tbody>
</table>

Figure 2. Model Prediction vs. Performance

In the graph above, the color teal represents predicted turnout by Y2 Analytics and orange represents the actual voter turnout. As you can see in the lower deciles, predicted turnout is not high.
enough and in the higher deciles, it is too high. When asked what he thought caused this discrepancy between his company’s predictions and actual turnout within deciles, Scott Riding said, “We overpredicted [turnout in] Utah County which is why [there are] errors on the top end of the model.” He noted that there was no geography variable involved in the model and that affected its performance (Riding, 2017). Such a geography variable was added to the 2016 model. One reason why turnout in Utah county was so different from the rest of the district is because in 2014, Mia Love was the only person running for any major office. There was no Senate race going on in the state that year, nor was there a governor’s or presidential race to help drive turnout. Despite the fact that polls narrowed considerably toward the end of the 2014 race, since many voters in Utah County considered Mia Love likely to win her election and there was no one higher up on the ballot than her to drive turnout, they saw no reason to turn out to vote. Since many of the district’s high turnout voters are found among Utah County republicans, that pushed down turnout for the top deciles. So, what explanation was there for the model underpredicting how likely people in the lower deciles were to vote? Previous research has found similar underprediction of the lowest deciles’ rates of voting for general elections (Barber, 2014). Perhaps it is a reflection that campaigns are successfully getting low-turnout voters to show up at the polls on election day because it is on these voters that campaigns concentrate their get out the vote resources. Despite small discrepancies, the table shows that we were never more than about points away from predicting the actual turnout in the decile and usually did much better. That indicates that we can actually effectively model the probability of a voter showing up to vote and validates our use of this as a measure of size for our PPS sample model.

The Accuracy of PPS Polling Results

Now that we’ve established that we can accurately predict how likely people are to vote, let’s discuss survey accuracy. I’ve collected all the publicly released surveys from October 2014 and
compared them to both the actual election outcome and Rep. Mia Love’s internal polling conducted using PPS sampling from the same time period. The charts below show how the polls compared.

**Figure 3. Mia Love Vote Percentage Prediction and Outcome**

- Y2 Analytics (October 25-30)
- Utah Policy/Dan Jones (October 7-9)
- Utah Policy/Dan Jones (October 20-29)
- FM3 (D) (October 6-7)
- BYU CSED (October 15-22)
- Actual Result
Figure 4. Doug Owens Vote Percentage Prediction and Outcome

- Y2 Analytics (October 25-30)
- Utah Policy/Dan Jones (October 7-9)
- Utah Policy/Dan Jones (October 20-29)
- FM3 (D) (October 6-7)
- BYU CSED (October 15-22)
- Actual Result

Percentage of the Vote
Figure 5. Mia Love Margin of Victory Predictions and Outcome

The graph in red reflects the percentage of the vote that each poll predicted Mia Love would get. Y2 Analytics nailed her vote percentage on the head at 51%. One other poll released at the same time had her polling as low as 42%. The others fell somewhere in between. While the poll conducted using PPS sampling didn’t exactly match the percentage of the vote received by Doug Owens, it was only two percentage points off which is well within the margin of error. The chart in blue illustrates how the PPS poll did against the others at predicting Doug Owens’s percentage of the vote. One poll, the BYU CSED poll, did get Doug Owens’s percentage exactly right, but none of the rest did any better than the poll conducted by Y2 using PPS sampling and some did worse. However, at the end of the day, what
most readers of polls look at is not the percentage of the vote each candidate received – it is the margin of victory signaled by the poll. The purple chart shows the margin of victory each poll predicted. Again, one poll, this one conducted by Utah Policy/Dan Jones, got the margin of victory exactly right, but it was not the same poll that got Doug Owens’s percentage of the vote correct. That poll, the BYU Center for the Study of Elections and Democracy poll, was actually the farthest off and the only poll to show Doug Owens ahead. I concur that the most important metric to look at is not the exact percentage of the vote predicted by a poll for either candidate, but the predicted margin of victory for one candidate over another. Most of the polling conducted in October of 2014 was not too far off, but this comparison shows that, at a minimum, PPS sampling performs no worse than traditional polling and often performs better – as illustrated by the fact that it was one of the closest polls to the actual margin of victory for Mia Love in all measured categories.

One important thing to note about these polling comparisons is the sample sizes because the sample size affects the margin of error on a poll. All the polls had very comparable sample sizes except for the BYU CSED poll. Its sample size was 236 while all the other polls had sample sizes between 400 and 425. This may have contributed to its large miss of the margin of victory for Mia Love, but even considering the size of the sample it was based on, it was still the worst performing poll. The fact that all the polls that clustered around the true election outcome had nearly identical sample sizes reinforces the idea that PPS sampling performed as well as any other Simple Random Sample Poll (within the margin of error of the actual outcome) and better than the BYU CSED poll which despite its small sample size was still way off.

**How PPS Sampling Remedies Contemporary Problems with Polling**

In addition to accuracy, PPS samples allow us to experiment with different turnout scenarios. Many people argued whether correctly or incorrectly that the 2016 campaign was activating voters who had lain dormant for years. PPS sampling allows us to shift our turnout model to see what would happen
if those people actually did turn out at higher rates than their vote history would predict. We can then see how the electorate and corresponding outcome would shift if more (or less) low turnout voters were to show up than expected. This is a feature of PPS sampling that allows us to produce best case and worst case scenarios if the electorate shifts dramatically. Yet another benefit of PPS samples in polling is that they, unlike most public polls which are generally designed for media headlines, are designed to handle the complex designs and questionnaires needed by scholars to model attitudes and behavior (Barber, 2014).

The issue of low response rates is also alleviated by PPS samples. In a study, two samples of the population were drawn for both a general and primary election. One of the samples was a simple random sample (SRS) and the other was a PPS sample. For the general election, the survey response rate for those sampled by the SRS method was 6.6% while the PPS sampling response rate was 8.5%. This means that the PPS sample increased people’s likelihood of answering the survey by roughly 33%. Although those numbers are for the general election, PPS samples are particularly well-suited to low-turnout elections and we would expect it to perform even better for the sample drawn to predict a primary election. That was the case. The response rate for the primary survey was 6.1% for those sampled using SRS and the response rate was 12.5% for those who were sampled using PPS sampling. (Barber, 2014). That means that people sampled using the PPS sample were more than twice as likely to take the survey than those sampled via a simple random sample. This is indicative of large available gains in telephone efficiency. Paying for data collection via telephones is not cheap so doubling efficiency of your sample would halve costs alleviating in a big way one of the major pressures on contemporary survey research. You can conduct quality telephone survey research as long as you use PPS sampling to be smart about what sample gets dialed.

So, while I have shown that PPS sampling for telephone surveys is, if not better than traditional sampling methods, at least equivalent in terms of quality, I also believe that it is the answer to many of
the problems facing contemporary polling that I discussed at the beginning of this paper. The biggest issue it solves is predicting who will be in the electorate. As discussed, enormous shares of survey respondents either flake-in or flake-out and self-reported vote likelihood is not a good indicator of whether someone will actually cast a ballot. However, using a simple logistic regression model, we can predict with remarkable accuracy how likely people are to vote without even asking them. The fact that we don’t have to ask is also a plus because anyone who has worked in the survey industry will tell you that you should cut any procedural question you don’t have to ask because time is money. Call centers get paid more for longer surveys and cutting questions saves on costs or frees up questionnaire space for another question aimed at answering clients’ or researchers’ questions. That leads in to the next thing afflicting the polling industry that PPS sampling helps to solve: rising costs. As I said earlier, you get what you pay for. However, PPS sampling is a way of getting a sample relatively cheaply that accurately models the electorate and still leaves enough budget for live interviewers rather than robopollers.

Conclusion

While I believe I have made a convincing case for the usefulness of PPS sampling, I admit that there are limits to generalizing based on the work I’ve done here. With regard to the weaknesses of what I’ve looked at here, I recognize that I have not conducted an experiment. To do so would have required having the same company conduct both a PPS poll and a poll based on a simple random sample. That way, I could have compared apples to apples when looking at the accuracy of each poll. Doing so also would have allowed me to look at survey response rates in surveys conducted by Y2 Analytics and compare the response rates of the PPS survey against the SRS survey. Future steps for research would include setting up an actual experiment rather than just observing the results of polling conducted by different polling firms. It would also be useful to compare a broad sample of PPS surveys against simple random sample surveys of the same races rather than just conclude that PPS performs as well based only on a single congressional race survey conducted in Utah’s 4th district. Widening our
sample of races would increase the validity of our conclusion that PPS samples are as accurate as what
the survey industry has traditionally been relying on.

I see PPS sampling as a step in the right direction for the future of polling. Not only is it cost
effective, but it’s accurate and allows for intriguing research possibilities that were previously
unavailable when sampling using simple random samples. While many internet polls are already being
conducted using PPS samples and have been shown to be largely accurate, it’s time that the world of
telephone phone interviews catches up.
Reflective Writing

Creating a Suitable Undergraduate Capstone Experience

The experience I had writing this capstone thesis was a really suitable capstone to my undergraduate education. My entire undergraduate education has straddled the companion fields of statistics and political science. It seemed fitting to me that my Honors capstone project would do the same. By writing about PPS sampling and its application to public opinion polling, I wrote my thesis about a subject critical to political science, but from a very quantitative perspective and learned about and a technique that has only recently emerged to combat declining survey response rates and gain greater precision in modeling the electorate. I was first introduced to PPS sampling while working in a summer internship at Y2 Analytics in Salt Lake City. The concept intrigued me as it seemed like a technique that all pollsters should be applying because it was such a good idea. Therefore, it is the capstone, because it is work in an emerging statistical discipline that I learned about during a summer internship, that is highly applicable to the field of public opinion research which overlaps both the areas I have studied during my undergraduate career: statistics and political science.

Substantially Adding to my Overall Education and Future Goals

The most obvious place that writing this thesis helped me progress to future goals and improved the overall quality of my education wasn’t evident to me until I went to visit grad schools in the process of selecting which one I would like to attend to get my Ph.D. One question I was frequently asked by faculty was “What is the topic of your senior thesis?” I was glad that I actually had a response for them. Neither of my two majors is requiring me to write a thesis and if it weren’t for the honors capstone, I wouldn’t have had to write any thesis at all. I did often clarify to these professors that it was an honors
thesis rather than a senior thesis. Having written something like this gave me lots of credibility with the political science programs I was applying to. It showed that I was used to writing papers that required research to these top-notch graduate schools.

Create a Positive, Meaningful Mentor Relationship

I’m lucky to have the best honors capstone mentor in the world. Dr. Damon Cann is someone that I feel I can approach to discuss anything and someone whose advice I can rely on to help make good decisions about my future educational plans. He’s been nothing but supportive during the past two years of my undergraduate career. I began doing research with him because he was someone that I felt I could approach. I knew him from several years of being his neighbor. We began doing simple work such as data collection, but over time, he provided me with tasks that helped to push me and grow my abilities. He took the time to explain those tasks in weekly meetings we had together. When I began the process of applying for grad school, he not only wrote me a letter of recommendation, but helped me identify a writing sample topic and then coached me through the process of preparing the first draft of a paper. He’s been a fantastic mentor and someone that I know I can rely on to be in my corner when push comes to shove.

Deepen Research Experience within my Major

For starters, my research experience was deepened specifically because I wouldn’t have otherwise had to write a thesis of any type. However, to address how it was deepened specifically within my major(s), I feel that in both the stats major and the political science major, they did a good job of giving me a very broad education. With regard to statistics, I feel that I’ve had a taste of basically everything leaving this department. I’ve seen some machine learning. I’ve gained some proficiency with R. I’ve done things I considered complicated like proofs and analysis. I’ve done the basic things like hypothesis testing covered in introductory statistics and have learned about distributions. I’ve done
categorical data analysis and have even had a basic introduction to Bayesian estimation. The point is that I’ve constantly been pleasantly surprised at both the breadth and depth of my statistics education at Utah State University. However, this capstone deepened that education because it allowed me to explore sampling which is something I really hadn’t understood particularly well beforehand and look at some applications of it. I feel like I dove pretty deep into the math behind much of the statistical sampling used for Probability Proportional to Size Sampling. This was courtesy of Richard Cutler.

**Require Critical Thinking about Topics in that Major**

This may be brief, but I feel that I have already covered this to some degree. Being able to write a paper that includes a section on how PPS sampling works under the hood required critical thinking on my part. I also had to struggle to get R to work properly to produce my desired graphics. However, as always seems to happen, we learn more from our struggles than we do from the times that we are coasting. I feel that critical thinking about statistical software and sampling theory really helped me to gain understanding about those two subjects within the Statistics major.

**Broaden the Student’s Experience across Disciplines**

This is a case where my project is clearly interdisciplinary. Up until this point, despite having two majors they had largely remained separate from each other. Once, for STAT 5600, a class taught by Richard Cutler, I did a very interdisciplinary project that I designed. In short, I classified people into one of 7 partisan strength categories based on information found in the CCES data set. I used tree-based methods and R for that project. However, the point is that my experience mixing these two disciplines is very limited. This is a project that very clearly allowed me to play in the sandboxes of the two disciplines at the same time.

**Engage the Student in his Local or Global Community**
The best way this paper engaged me with the local community (if Salt Lake counts as local) is through my internship. I did a lot of polling using PPS sampling for local communities such as South Salt Lake, Provo, and Orem. I believe that allowing these cities to become aware of the wishes of their citizens (and allowing citizens to be heard) through conducting a public opinion survey counts as engagement in the local community. I hope that other national pollsters will someday adopt the PPS sampling method used by Y2 Analytics during my time there. That would be how there’s a global connection. Although I didn’t specifically work on any National Republican Senatorial Committee (NRSC) projects, they saw our method and began to show lots of interest while I was there for my internship. At a minimum, this has given me a much better idea of how a survey should be set up and how a sample should be drawn. When I engage in survey research in grad school and in my future career that may affect people much more broadly, I’ll be sure to consider using a PPS sample especially if funds are limited or the election scenario I’m seeking to model has the potential to be low on turnout.
Bibliography


Morning Consult (2015). Why does Donald Trump perform better in online versus live telephone polling?


Professional Author Bio

Jade Burt is graduating from Utah State University in the spring of 2017 with a double major in Statistics (in which he did departmental honors) and political science. He will also be receiving a minor in mathematics. He was awarded the outstanding undergraduate student of the year award for both the statistics department and the political science department.

While attending Utah State, he was heavily involved in scholarly work and eventually became a Merrill Scholar and a USU Research Foundation Student Scholar. He was the chairman of the USU Research Foundation Student Scholar Committee for about a year. He worked as a research assistant with Dr. Damon Cann for two years of his undergraduate time. That work eventually led to an internship with the polling company Y2 Analytics from which the idea for this thesis originated.

His future plans include continuing to work for Y2 Analytics until August of 2017. At that time, he will leave to attend the University of Michigan in Ann Arbor in pursuit of a Ph.D. in Political Science. He will focus on American Politics. His career aspirations include working as a professor at a university.