American Journal of Applied Mathematics and Statistics, 2017, Vol. 5, No. 3, 101-105
Available online at http://pubs.sciepub.com/ajams/5/3/3
©Science and Education Publishing
DOI:10.12691/ajams-5-3-3

Bias Correction by Sub-population Weighting for the 2016 United States Presidential Election

Bumjun Park[*]

International Department, Hankuk Academy of Foreign Studies, Yongin, South Korea
*Corresponding author: bumjunpark99@gmail.com

Abstract The 2016 Presidential Election was an international surprise, as President Donald Trump came back from a seemingly large deficit in the pre-election opinion polls. As most, if not all, of the major polls missed the election results, the public started to doubt the credibility of pre-election polls. This article proposes that there was a methodological error in the polls. The polls used the census data of American population to weigh their data. However, population may not have a correlation with turnout, meaning that a certain population group may not vote much; not contributing to the electorate. For this reason, the polls based on population might systematically over or underestimate a particular candidate. Thereby, the proposition is that the polling agencies should consider the electorate, not the population for modifying the polling results. The proposition is substantiated with a series of statistical simulations supporting the claim that a poll conducted based on the electorate resembles the actual result more accurately. Conclusively, it argues that, as the polls play a pivotal role in affecting the media and the electorate, the improvement of polls is necessary for well-informed forecasts to be available.

Keywords: *election polls, bias correction, sub-population weighting, turnout rate, simulation, prediction*

Cite This Article: Bumjun Park, "Bias Correction by Sub-population Weighting for the 2016 United States Presidential Election." *American Journal of Applied Mathematics and Statistics*, vol. 5, no. 3 (2017): 101-105. doi: 10.12691/ajams-5-3-3.

1. Introduction

Prior to the United States 2016 Presidential Election, most of the polling agencies predicted an easy win for Secretary of State Hillary Clinton, only to find out that the predictions were substantially wrong. Although Clinton did win in popular votes by a margin of 2 percent points, Trump won by a big margin in the electoral votes. Even though most political forecasters disregard pre-election polls due to their non-probability samples, the error in this election seems substantial to the extent that most, if not all, of the projections predicted a landslide Clinton victory.

The erroneous predictions of the elections, along with other worldwide missed predictions such as the prediction of Bremain, led to a global trend of discrediting political opinion polls. But there were always missed predictions, and they did not become a global issue. In fact, political opinions polls were not always accurate; historically, the 1936 Presidential Election and the 1948 Presidential Election both deviated significantly; In 1948, serious enough to the extent that *Chicago Tribune* published their papers with the headline reading, "Dewey Defeats Truman." In summary, this trend of misleading polls is not a new one.

But after these infamous incidents, the polling agencies gradually improved their polling methodology. The 1936 election polls were conducted during the Great Depression to those who owned private phones, which led to polling a more wealthier sample who generally favor the Republican candidate. After Franklin Delano Roosevelt, the Democratic candidate won the 1936 election, the agencies started randomly sampling the entire public more thoroughly. For the 1948 election polls, the agencies had not set sampling quotas for each division such as race, gender, age, leading to a biased poll. Therefore, after the 1948 election, the agencies started setting quotas based on the population in the sampling process. And after quota sampling began, the opinion polls seemed to improve, only to drastically fail in the 2016 Presidential Election.

This election showed the shortcomings of the present quota sampling methods. Under the current methods, the Census data of the population is considered for setting quotas. But the loophole of such method is that the turnout of each cleavage may be ill-considered. Ultimately, what matters for elections is the electorate, not the populace. Therefore studying the trend of turnout and composition of the electorate and setting quotas based on the electorate seems necessary. Especially, the turnout rate based on age shows a convincing trend; the younger voters are turning out significantly less compared to older voters, leading them to have relatively less significant effect on the results of an election. Overall, this study focuses on the statistical improvement that considering the turnout rate of each division may bring about.

2. The Insignificance of Polls

Historically, polls have not been considered viable predictors. Gelman and King said that polls give relatively useless predictions [1]. One reason they referred to was that high non-response biases present in the current polling methods lead to misrepresenting the public opinion. Instead, they proposed that economic measures or other indicators are better for predictions. Allan Lichtman, a professor at American University, who was famous for correctly predicting the presidential election for 30 years, also did not use polls in his prediction model [2]. He rather used indicators such as approval ratings or control of the Congress. Also, another proposition was made that the non-probability sampling for modern Internet polls lead to even more biases [3]. After all, Hillygus says that the proliferation of various inaccurate political polls are undermining the impact of election polls [4].

However, pre-election polls are not important because of the ability to predict results; rather, they are important because of their frequented appearances in the media. Broh suggested that the polls' results and the leaders in the polls were nearly always reported on major media sources [5]. Moreover, he said that the importance lies in the fact that the figures of the polls are subject to the journalists' interpretations. By emphasizing, neglecting, or even distorting the results, the journalists may affect the electorate variably with the same polling results.

Table 1. Opinion Polls from Five Polling Agencies

Poll	Clinton	Trump	Johnson	Stein
Fox News	48	44	3	2
NBC News	47	41	6	3
Rasmussen	45	43	4	2
CBS News	45	41	5	2
ABC	47	43	4	1
Actual	48.06	45.97	3.28	1.06

3. Deviation in Opinion Polls

Table 1 summarizes the polling data of 5 major polling agencies. To estimate the extent of deviation of pre-election polls from the actual results, Frederick Mosteller devised six measures of estimation in his 1949 book, *The Pre-election Polls of 1948: Report to the Committee on Analysis of Pre-election Polls and Forecasts* [6]. Of these 6 measures, Mosteller 3, the average deviation in percentage points between predicted and actual results for each candidate, is used most frequently [4]. To map the measures in order to best understand the accuracy of polls, Shipman and Leve developed a method of creating a contour map [7]. In the contour map, each grid represents 1 percentage point, and the polls' expected popular vote for Clinton is on the x-axis and the expected popular vote for Trump is on the y-axis, with the actual result of the election in the center. Since Mosteller 3 computes the average deviation from the actual result, it can be measured based on the distance from the center point.

Each of the 5 opinion polls' results were plotted accordingly on the plot and based on this contour map, it can be inferred that most of the polls were outside of the center contour, which means none of the 5 polls were within a 2 percent point range. All 5 polls underestimated the support for President Trump by at least 2 percent points, while most of them also underestimated support for Clinton on the national level.

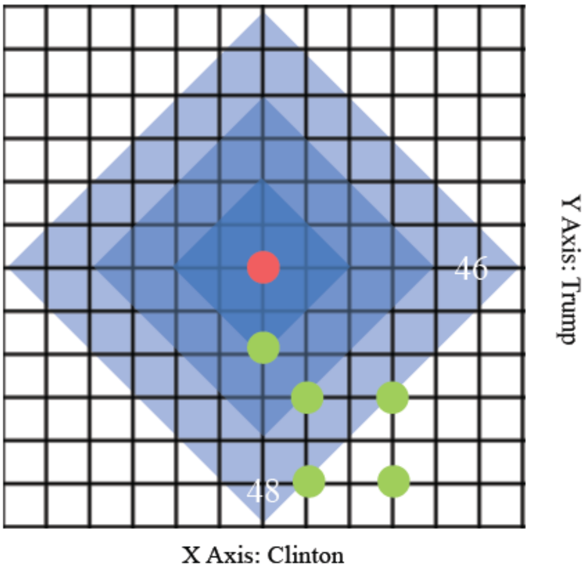

Figure 1. Mosteller Three Contour Map for the Poll Results of Donald Trump and Hillary Clinton

Additionally, to estimate the accuracy of the polls on the state level, a binomial z-score distribution was used. By expressing support for a candidate as a binomial, either supporting one candidate or not, the accuracy of state polls for predicting the support for a particular candidate can be assessed using Equation 1. The variable P for the equation is the actual popular vote percentage for the candidates, p is the expected popular vote percentage based on the opinion polls and n is the total ballot count per state. With regard to this equation, a negative value meaning underestimation and a positive value meaning overestimation. Based on each state's opinion polls provided by RealClearPolitics, and the ballot data provided by United States Election Project, Equation 1 was applied to each state's election results, and the results are given in Table 2. The z-Rep column shows the result of applying Equation 1 for the support for the Republican party while z-Dem applies the equation for the Democratic Party.

$$\frac{P-p}{n\times p\times(1-p)}. \quad (1)$$

Observing the results, it is first notable that state polls generally missed the result generally by a large margin, which again proves the inaccuracy of polls. In particular, for most of the states, the results were biased towards a particular candidate; generally overestimating support for Clinton while underestimating support for Trump. Especially in the 5 states won by Obama in 2012 that Trump won for this election (Florida, Michigan, Ohio, Pennsylvania, and Wisconsin), the underestimation of the support for Trump was severe, all having a z-score around a negative 200. At the same time, they generally overestimated the support for Clinton; Three of the five states were overestimated while the other two were underestimated relatively less.

Table 2. Result of Equation 1 by State

State	z-Rep	z-Dem
Alabama	-176	42
Alaska	-123	50
Arizona	-82	-46
Arkansas	-204	72
California	-151	-390
Colorado	-111	-108
Connecticut	-159	-114
Delaware	-156	-31
Florida	-191	-111
Georgia	-94	-66
Hawaii	-26	-51
Idaho	-207	-11
Illinois	-114	-252
Indiana	-340	3
Iowa	-172	20
Kansas	-172	39
Kentucky	-215	68
Louisiana	-234	-70
Maine	-90	-15
Maryland	-276	81
Massachusetts	-248	-103
Michigan	-246	31
Minnesota	-152	107
Mississippi	-225	139
Missouri	-240	69
Montana	-166	-60
Nebraska	-80	-96
Nevada	32	-40
New Hampshire	-57	-45
New Jersey	-69	-153
New Mexico	0	-5
New York	-194	-424
North Carolina	-196	-31
North Dakota	-260	55
Ohio	-286	-24
Oklahoma	-134	29
Oregon	-144	-186
Pennsylvania	-234	20
Rhode Island	-104	-45
South Carolina	-376	-83
South Dakota	-181	56
Tennessee	-425	36
Texas	-276	-266
Utah	-170	-38
Vermont	-153	-153
Virginia	-40	-76
Washington	-74	-180
Washington D.C.	-111	-1927
West Virginia	-160	29
Wisconsin	-271	3
Wyoming	-135	-30

4. Methodological Error

This study proposes that such error may be attributed to the current polling agencies' methods of setting quotas based on population. Directly citing from NBC News and SurveyMonkey's Methodology explanation, it says "[the polls] have been weighted for age, race, sex, education, region, and voter registration status using the Census Bureau and Bureau of Labor Statistics's Current Population Survey…" The problem of weighting polls based on the populace is that it cannot reflect whether a certain group will vote more or less.

Turnout rate is an important factor that most of the polls might be systematically putting out of consideration. About the importance of turnout rate, Burden said that "Turnout is the most common and the most important component of an individual's participation in the political process" [8]. Additionally, Burden pointed out that the turnout rate showed a clear trend of decrease. The clear decrease can be attributed to various factors. For example, it has been proven that lower education and lower income tends to lead to lower turnout [9]. Another viable factor could be growing political apathy among the populace [10]. Also, less partisan motivation among people and the increase of people indicating themselves as independents could be leading to lesser and lesser people showing up on election day [11].

For this study, the most important cause for decreasing turnout is the changing age distribution of the electorate. According to the United States' Census Bureau's Current Population Reports, the population of ages 65 and older will double by 2060 while the increment of younger population is relatively slight leading to an older electorate. The problem of an aging electorate is that older voters tend to vote more for the Republican party, and generally have a higher turnout rate, favoring the Republican party in the elections. Since 2004, of the states that changed parties in between two presidential elections, the states that changed from Republican to Democrat had higher share of electorate for young voters than the states that changed from Democrat to Republican (p=.004). This means that a higher older voter share tends to benefit the Republican candidate.

And recently, according to the Trend Analysis Plot for the younger share of electorate provided by Figure 2, the share of electorate for younger voters has shown a clear trend of decline. At the same time the population of younger and older voters are both increasing. Thus, if a poll weighs or sets quotas with the population data, it may lead to overestimating the younger voters, who increased in population but decreased in the actual turnout; thereby overestimating the support for the Democratic candidate.

Such trend of the younger voters, that the polls expected to vote, not actually voting might have contributed to overestimating Clinton and underestimating Trump. After all, Clinton may not have had the clear lead that polls suggested in the first place. The polls that did not consider the decreasing voting of younger voters contributed to this seriously misleading prediction. Not only that, but close elections tend to generate higher turnout rate [10]. Therefore, the biased prediction of a landslide victory for Clinton that the majority of the press reported of might have undermined the turnout rate even more. Ultimately, all factors combined, led to an unexpected and seemingly unlikely victory of Donald Trump.

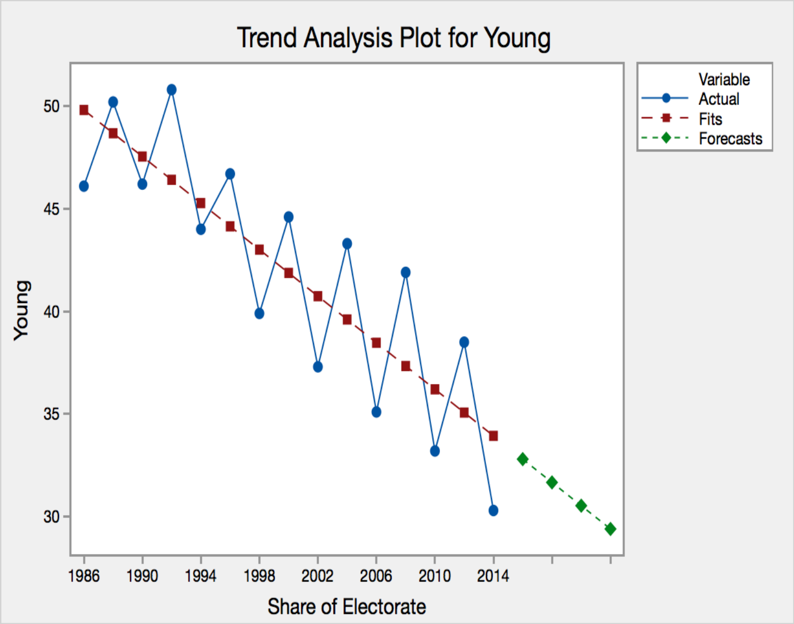

Figure 2. Time Series Analysis of the Younger Share of Electorate

Table 3. Demographics for the Simulation

Age	VAP	Turnout Rate	Electorate	n_p	n_e
18~24	31200000	0.585	18252000	252	207
25~34	44100000	0.664	29282400	356	333
35~44	40600000	0.699	28379400	328	322
45~54	43200000	0.735	31752000	349	361
55~64	40900000	0.766	31329400	330	356
65~74	27600000	0.781	21555600	223	245
75~	20200000	0.766	15473200	163	176

4.1. Method

To prove that considering the expected turnout rate in setting age quotas for polling, a simulation was conducted. The simulation analyses were conducted in R (R Core Team, 2017). The Voting Age Population by age bracket was set based on the U.S. Census Bureau's data. And based on the average turnout rate for every age bracket, the model electorate was formed. To represent the population based polling methods, n_p divides a sample of 2000 people of different age groups based on the proportion of VAP. On the other hand, to represent the electorate based polling method proposed by this study, n_e divides a sample of 2000 people of different age groups based on the proportion of Share of Electorate.

4.2. Results

Considering that most polling agencies consider those under 44 as young, the most evident fact is that the new polling method has significantly fewer younger voters in its sample because younger voters show a lower turnout rate. Based on CNN's exit polls, nationally, 53% of the younger voters favored Clinton, while 39% of them favored Trump. On the other hand, 44% of the older voters favored Clinton, while 52% of them favored Trump. Assuming that Trump and Clinton supporters were distributed among the voting age population based on these percentages, 100 simulation samplings were conducted for each polling method. The smoothed distribution graphs of the polls were plotted on Figure 3.

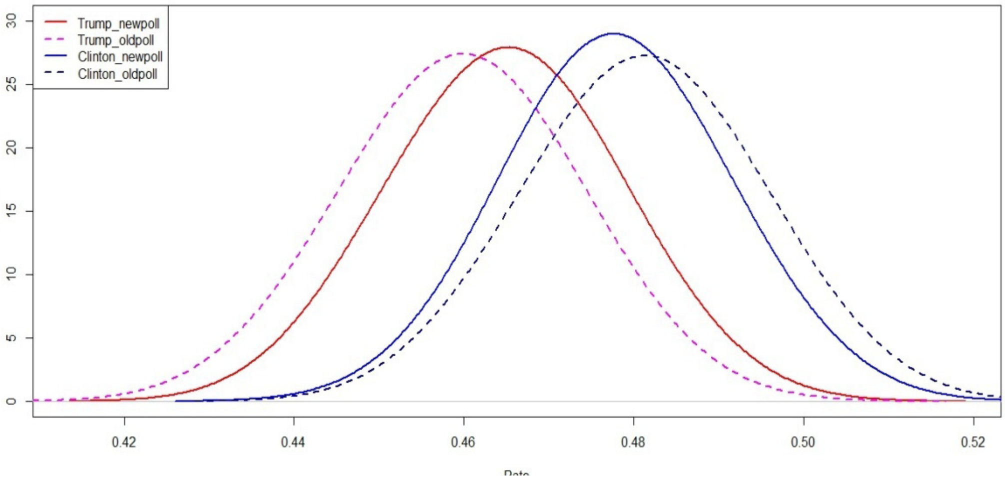

Figure 3. Distribution Plots of the Simulation Results

Table 4. Descriptive Statistics of the Simulation Results

	Trump		Clinton	
	M	SD	M	SD
Electorate Based	0.465	0.010	0.478	0.009
Population Based	0.460	0.010	0.481	0.011

The blue lines represent the support for the Democratic candidate while the red lines represent while the dotted lines represent the support for the Republican candidate. Additionally, while the dotted lines represent the population based method, the solid lines represent this study's method. In Figure 3, it is notable that the solid lines show a slighter gap compared to the dotted lines, resembling the actual result relatively better. The difference between Trump and Clinton using this study's polling method is around 1.3 percentage points while the population based method shows a difference around 2.1 percentage points. In order to visually show the improvement, the results of this study's polls were applied to Figure 1's Mosteller 3 contour map with the point *E* showing the electorate based approach. Compared to the other 5 opinion polls, it becomes clearer that the electorate based method is more accurate.

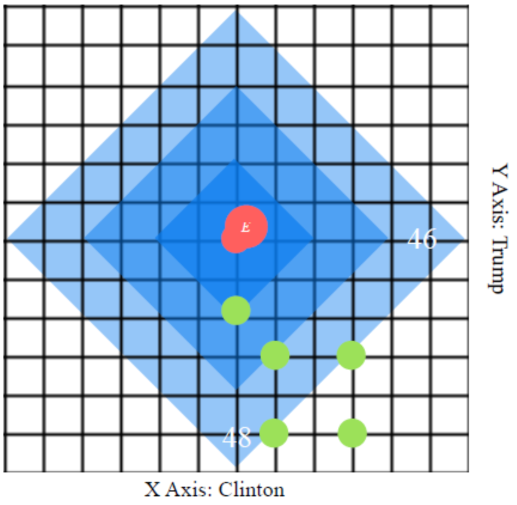

Figure 4. Mosteller Three Contour Map for the Poll Results of Donald Trump and Hillary Clinton Including the New Method

4.3. Conclusion

Ultimately, to prevent such unexpected surprises afterwards, this study proposes that the polling agencies might be able to better their polls by conducting electorate based polls. Not just the absolute number of members of a group, but also that groups tendency to vote more or less, and how that voting tendency affects the composition of the electorate should be considered. Despite the fact that professional forecasters do not use the possibly biased, even meaningless pre-election opinion polls, the election polls affect the public the most. Media reports about the leader in the polls and the journalists' interpretations can be significant enough to nudge the electorate and veer the result of the final ballot.

After all, improving the polling method may increase the opinion polls' utility as an indicator for presidential election predictions. As the position of President of the United States is pivotal to the country, correctly predicting the winner of that presidency should be important as well. And such methodological improvements suggested by this study might contribute to preventing such a pivotal event happening unexpectedly.

References

[1] Gelman, A., and King, G., "Why are American presidential election campaign polls so variable when votes are so predictable?" *British Journal of Political Science, 23* (4), 409-451. Oct. 1993

[2] Lichtman, J. A., "The keys to the white house: An index forecast for 2008," *International Journal of Forecasting*, 24 (2), 301-309. Jun. 2008.

[3] Silver, N. "The Worst Pollster in the World Strikes Again." *FiveThirtyEight*, Mar. 2009. fivethirtyeight.com/features/worst-pollster-in-world-strikes-again/.

[4] Hillygus, D. S., "The evolution of election polling in the United States". *Public Opinion Quarterly, 75* (5), 962-981. Dec. 2011

[5] Broh, A. C., "Horse-race journalism: Reporting the polls in the 1976 presidential election," *Public Opinion Quarterly,* 44 (4), 514-529. Jan. 1980.

[6] Mosteller, F., *The Pre-election Polls of 1948: The Report to the Committee on Analysis of Pre-election Polls and Forecasts*. Social Science Research Council, New York, 1949.

[7] Shipman, J., & Leve, J. H., "A New "Interval" Measure of Election Poll Accuracy," Paper presented at the annual meeting of the American Association For Public Opinion Association. May. 2005.

[8] Burden, B. C., "Voter turnout and the national election studies," *Political Analysis, 8* (4), 389-398. Jul. 2000.

[9] Reiter, H. L., "Why is turnout down?" *Public Opinion Quarterly, 43* (3), 297-311. Jan. 1979.

[10] Shaffer, S. D., "A multivariate explanation of decreasing turnout in presidential elections 1960-1976", *American Journal of Political Science, 25* (1), 68-95. Feb. 1981.

[11] Ladd, E. C., "The Shifting Party Coalitions, 1932-1976," *Emerging Coalitions in American Politics*, 81-102, Jan. 1978.