View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Columbia University Academic Commons

ISERP

Working Paper 06-10

Forecasting House Seats from General Congressional Polls

JOSEPH BAFUMI
Dartmouth College

ROBERT S. ERIKSON
Department of Political Science
Columbia University

CHRISTOPHER WLEZIEN
Temple University

Columbia University in the City of New York
November 2006

INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH AND POLICY

PIONEERING SOCIAL SCIENCE RESEARCH AND SHAPING PUBLIC POLICY

FORECASTING HOUSE SEATS FROM GENERIC CONGRESSIONAL POLLS

Joseph Bafumi
Dartmouth College (jbafumi@gmail.com)

Robert S. Erikson
Columbia University (rse14@columbia.edu)

Christopher Wlezien
Temple University (Wlezien@temple.edu)

ISERP Working Paper 06-10

November 2006

~1,800 Words
3 Tables

According to the frequent polling on the generic ballot for Congress, the Democrats hold a large advantage leading up to the vote on November 7. But does this Democratic edge mean that the Democrats will win a majority of House seats? Doubts are often expressed about the accuracy of the generic ballot polls. And even if the polls are correct in indicating a majority of *votes* going to Democratic candidates, further doubts are expressed about whether the Democrats' vote margin will be sufficient to win the most *seats*.[1]

This paper is intended to provide some guidance for translating the results of generic congressional polls into the election outcome.[2] Via computer simulation based on statistical analysis of historical data, we show how generic vote polls can be used to forecast the election outcome. We convert the results of generic vote polls into a projection of the actual national vote for Congress and ultimately into the partisan division of seats in the House of Representatives. Our model allows both a point forecast—our expectation of the seat division between Republicans and Democrats—and an estimate of the probability of partisan control. *Based on current generic ballot polls, we forecast an expected Democratic gain of 32 seats with Democratic control (a gain of 15 seats or more) a near certainty.* The details follow.

The easy part is forecasting the vote from the generic polls. To properly interpret the generic polls, we estimate a regression equation predicting the vote in the 15 most recent midterm elections, 1946-2002, from the average generic poll result during the last 30 days of each campaign. (Details are shown in the appendix.) Based on this analysis, we can confidently offer the following rule of thumb for predicting the national vote based on polls over the last 30 days before the election:

1. convert the percentage point lead, e.g., Democrats 51% Republicans 41%, in the generic poll to a percent Democratic of the two party vote, e.g., 51-41 converts to 55% Democratic or 5% more Democratic than 50-50;

2. if the poll is based on registered voters rather than "likely" voters, subtract 1.5 percentage points—thus a 56%-44% Democratic lead in a registered voter poll converts to a narrower 54.5%-45.5% lead in terms of likely voters;[3]

3. cut this lead in half; and

4. add a percentage point to the Democrats as a reward for being the *non*-presidential party.

---

[1] See for example Jonathan Kastellec, Andrew Gelman, and Jamie Chandler, "Seeking 50% of Seats, Needing More than 50% of Votes: Predicting the Seats-Votes Curve in the 2006 Elections" http://www.stat.columbia.edu/~gelman/research/unpublished/house2006.pdf

[2] Respondents typically are asked which party they plan to vote for (or who they want to win) "if the election were being held today", though there is variation in question wording. For instance, some organizations use the wording "Looking ahead to the Congressional elections in November." Other organizations use "Thinking about the next election for US Congress."

[3] The registered voter correction represents the average adjustment necessary from generic polls over past midterms. Democratic registrants in the past have turned out with slightly greater frequency than today. There is no certainty that this correction will hold exactly in 2006.

From the regression analysis, our 95% confidence interval for the forecast using this formula is +/-3.7 percentage points.[4]

Now consider the polls over the final thirty days of the 2006 campaign. As of early October 24, PollingReport.com listed the results of 6 generic ballot polls conducted during the final 30 days of the campaign, by CNN (2), USA Today/Gallup, ABC/Washington Post, Fox/Opinion Dynamics, and Newsweek.[5] The average Democratic two-party share in these polls is 57.7%. Applying our formula as described above, the Democrats should win 55% of the two-party vote with a confidence interval from 51.3 to 58.7 percent, implying that the Democrats almost certainly would win a majority of the votes cast.

But would this mean that the Democrats would also win the most *seats?* If the Democrats were to win 55% of the vote, this would represent a 6.4 percentage point swing from 2004, when they received 48.6%. If Democrats were to win *exactly* 6.4% more of the 2006 vote in every district than they won in 2004, they would win 228 seats. However, an average swing of 6.4% percentage points will be spread unevenly—sometimes more than 6.4% and sometimes less. We must take the degree of uniformity into account. Moreover, as we already discussed, the prediction that the average vote swing will be 6.4% is itself subject to error. We take these considerations into account by a set of simulations described below. The simulations suggest that a *predicted* national vote surge of 6.4 percentage points would yield the Democrats 235 seats, for a 32-seat gain. This is 7 seats more than with uniform swing.

The simulations are constructed as follows. For each possible generic ballot integer value from 50% Democratic to 60% Democratic, we compute 1,000 simulations of the 435 seat outcomes. Each simulation includes:

(a) a random draw from the density of the possible vote outcomes from our generic poll regression equation, based on the predict from the generic poll plus forecast error; and

(b) a set of 435 random draws of district level predictions conditional on the 2006 national shock (from [a]) plus district-level characteristics and shocks based on a regression model from the 2004 election.[6]

---

[4] The exact equation is:

Dem Vote Share = 24.38 + 0.51 * Dem Poll Share – 1.07 * Presidential Party

Adjusted R-squared = 0.75; Root MSE= 1.90,

where Presidential Party takes the value "1" under a Democratic President and "-1" under a Republican. For more detail and analysis, see Joseph Bafumi, Robert S. Erikson and Christopher Wlezien, "Ideological Balancing, Generic Polls and Midterm Congressional Elections," at http://www.temple.edu/ipa/workingPapers/.

[5] All but the ABC/Washington Post poll were listed as among likely voters. The results for the ABC/Washington Post poll were among registered voters. The likely voter results were obtained from the ABC/Washington Post news release posted on realclearpolitics.com.

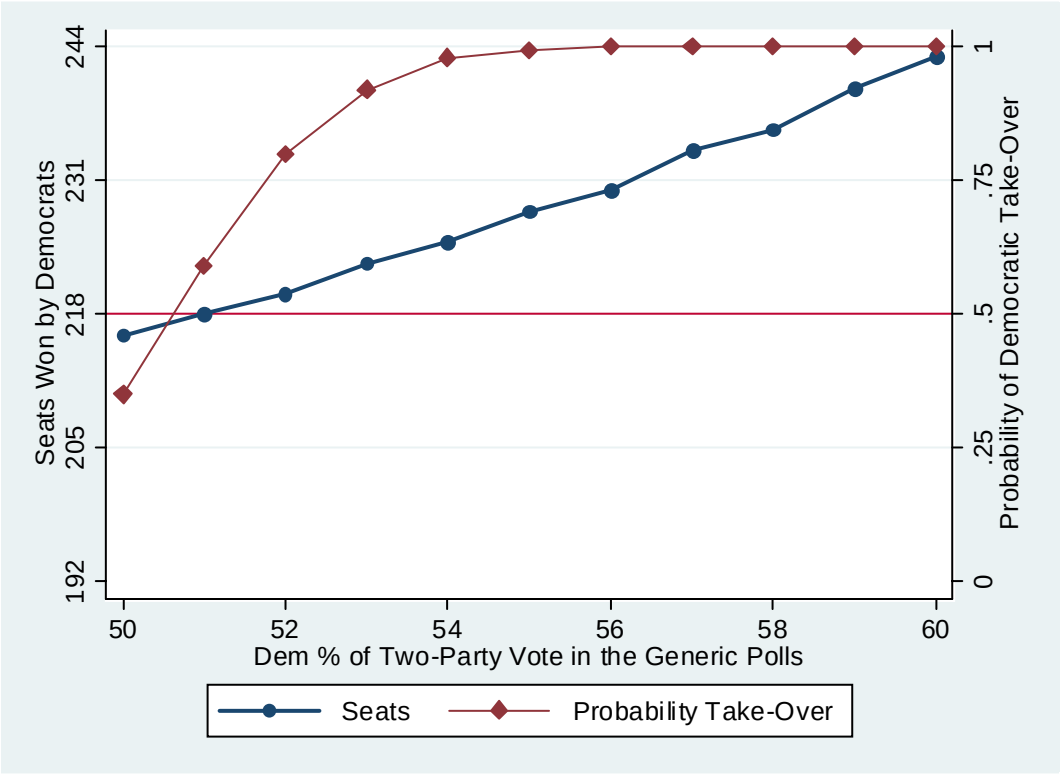
[6] Actually, the number of random draws is 374. The remaining 61 districts are assigned automatically to the current party holding the seat due to unopposed candidacies in 2004.

Details are presented in the appendix.

For each generic ballot integer value from 50% Democratic to 60% Democratic, we have conducted 1000 simulations of the seat division as per the methodology discussed in the previous paragraph. For each of the possible generic poll outcomes (50% Democratic, 51% Democratic, etc), we provide both a seat forecast (as an expectation) plus a probability estimate regarding whether the Democrats will win a majority (218 seats or more) of the seats.

Figure 1 shows the translation of the generic vote into a likely seat outcome with an accompanying probability of a Democratic majority of seats. As can be seen, the threshold on the generic vote where the Democrats are favored to win the seat majority is 51 percentage points (in likely voter surveys) or between 52 and 53 percent (in registered voter surveys). In terms of point spreads, the cut point where the odds of partisan control are even is a spread of about 2 points in *likely* voter surveys (or about 5 points in registered voter surveys).

FIGURE 1: SUMMARY OF SIMULATED ELECTION OUTCOMES PREDICTED FROM VARYING GENERIC BALLOT POLL RESULTS



Thus we see that the Democrats need more than the plurality of the generic vote in order to expect to control the House. In the range where the Democrats have between 50 and 52 percent of the "two-party" generic vote, majority control is up for grabs. But if the Democrats have 53% or more of the vote by the generic ballot question, as they do as of this writing, they are the heavy favorites (probability > .90) to win the House of Representatives.[7]

Our forecast based on current generic polls is a Democratic gain in the range of 32 Democratic seats, an amount that exceeds most current forecasts. Do we exaggerate Democratic prospects? A useful reality check is to consult available district level polls and compare them with our predictions for the same districts from our simulations assuming a 58%-42% split in the generic ballot polls. Averaged across 32 Republican-held districts with October polling, the average district level poll margin is 51.7% Democratic, 48.7% Republican. For these same 32 districts, our average prediction is 50.3% Democratic, 49.7% Republican. The two sets of numbers match nicely. If anything, our simulations might underestimate Democratic strength in the sampled districts.[8]

Of course if the poll numbers to the generic ballot question shift as the election nears, the forecast should be revised according to the weight of new polling information. Figure 1 provides the guidance. If current trends in the congressional generic ballot polling persist (which they have in past election campaigns), the Democrats are near certain to win control of the House. But this assumes a continued Democratic lead of 8 or more points among likely voters in the generic ballot. If the lead dips below this level, the Republicans can rekindle their hope of holding the House.

[7] Readers conditioned to the idea that their districting advantage would allow the Republicans to govern with a minority of votes cast might be surprised that the threshold in terms of the national vote at which control is likely to revert to the Democrats is only slightly greater than 50%--about 51% of the generic vote and about 50.5 of the national vote. (This threshold is at least one percentage point higher, however, in terms of the *mean* district vote.) The explanation is the partisan asymmetry in 2006 retirements. Among retirees who had faced major-party competition in 2004, 19 were Republicans and only 6 were Democrats. Strategic Republican retirements in anticipation of a Democratic wave would cause an electoral ripple even if the larger wave does not arrive. Our calculations are that if there is no vote swing whatsoever from 2004 to 2006, the Democrats would pick up 5 or more seats just from the greater number of Republican than Democratic retirements.

[8] Like with the generic polls, we should discount the district leads in the polls by about half. Pollsters are most likely to sample preferences in the districts thought to be competitive. Thus, sampled districts might have larger than average Democratic gains.

APPENDIX

Simulating Votes and Seats

For each integer value of the vote from 50% Democratic to 60% Democratic in generic ballot polls, we obtain 1000 computer simulations of the seat distribution. Each batch of 1000 simulations is based on a value of the national vote, N_{gj}.

$$N_{gj} = P_g + e_j,$$

where P_g = the projected national Democratic percent of the vote given the prediction from the generic ballot result *g*, from the 15-election regression equation displayed below as Equation 1

and e_j = a random draw (j = 1 to 1000) from the distribution of error, signified by the root mean squared error (RMSE) in Equation 1.

EQUATION 1

Nat'l % Dem Vote = 24.38 + 0.51 * *Generic Ballot %* – 1.07 * *Presparty* + e_j,

Adjusted *R*-squared = 0.75; Root MSE= 1.90, *N*=15 midterm elections, 1946-2002

where *Presparty* = 1 if a Democratic President and -1 if a Republican President

For each simulated value of the national vote, we need to simulate the outcome in 435 congressional districts. The district vote (D_{gjk}) is:

$$D_{gjk} = N_{gj} + L_k + u_k$$

$$= P_g + L_k + e_j + u_k,$$

where L_k = the district (local) component of the expected district vote and u_k = the simulation of the district *k* error.[9] The formula for L_k differs for open seats and incumbent races. For open seats (no incumbent running), the district vote simulation is from Equation 2. For seats where the incumbent seeks reelection, the district vote simulation is from Equation 3.

[9] Actually, all simulated errors e_i and u_k are drawn fresh with each iteration.

EQUATION 2: Open seats

$D_{gjkm} = -42.61 + 0.89 * \%Kerry_k + P_g + e_j + u_{k_open}$,

where $\%Kerry_k$ = the Kerry percent of the two-party vote in district k in 2004

EQUATION 3: Incumbent races

$D_{gjkm} = -45.91 + 0.94 * \%Dem(2004)_k + 6.58 * Frosh_k + P_g + e_j + u_{k_incumbent}$,

where $\%Dem(2004)_k$ = the Democratic vote for the House in district k in 2004;
$Frosh_k$ = 1 if a Freshman Democrat in 2006 and -1 if a 2006 Freshman Republican, otherwise 0.

The numerical parameters of equations 2 and 3 are derived from regression equations predicting the 2004 district vote for 2004 open seats and 2004 incumbent races.[10] The constant terms of Equations 2 and 3 (-42.61 and -45.91) offset P_g, the projected national vote from the generic polls. The constants represent the 2004 regression intercepts minus the 2004 national vote.[11]

Importantly, the root mean squared errors from the two 2004 regression equations provide the standard deviations for u_{k_open} and $u_{k_incumbent}$. For open seats, this value is 7.56. For incumbent races, it is 5.20. These values represent the uncertainty around the district vote after taking into account the projected district vote from the 2004 presidential vote (open seats) or 2004 congressional vote (incumbent seats).

The final step is to tally the 1000 simulated outcomes for each of eleven values of the generic vote, 50% Democratic to 60% Democratic. For generic ballot result g, we obtain 1000 estimates of the national vote N_{jg}, each with 435 estimates of the district vote D_{gjk}. For each value of the generic vote j, we record the mean seat outcome, the distribution of 1000 seat

[10] For incumbent races, the incumbent's past vote is an obvious benchmark. For open seats, we substitute the district presidential vote as a measure of underlying district partisanship.

[11] For the adjustment, we assume that the mean district vote swing (for districts contested in both 2004 and 2006) equals the net 2004-2006 national vote swing in terms of votes counted. The 2004 open seat equation, based on the 2004 presidential vote, provides a baseline equation. With no net swing 2004-2006, this equation would apply unmodified for 2006. For specific hypothetical values of the vote (P_g), the constant is the original intercept minus 48.6 (the national 2004 congressional vote). Note that the equation adds the hypothetical 2006 national vote (P_g). For incumbent races, the procedure for determining the constant term is the same except that some algebra must be applied so that the net swing of the mean district vote is consistent with the national vote. For incumbent races, the constant equals the intercept for the 2004 incumbent race equation minus 48.6 (2004 national vote) minus 1.72. The 1.72-point adjustment is necessary so that the projected mean district swing across all districts equals the difference between the projected national vote (P_g) and 48.6. Given our methodology, all seats that were uncontested in 2004 must be assigned to the 2004 winner. Bernie Sanders's former seat as an Independent is assigned to the Democratic candidate.

outcomes, and the frequency with which the seat outcome for the Democrats equals or surpasses 218, the number for a numerical majority. This provides the basis for the graph of Figure 1.

We show some examples of the distributions in the following graphs. The six histograms below show the seat expectation, the distribution around it and the probability that the Democrats will win enough seats to win control of the house for generic polls with Democratic votes equal to 50%, 52%, 54%, 56%, 58% and 60%.

FIGURE 2: GENERIC POLLS = 50% DEM 50% REP

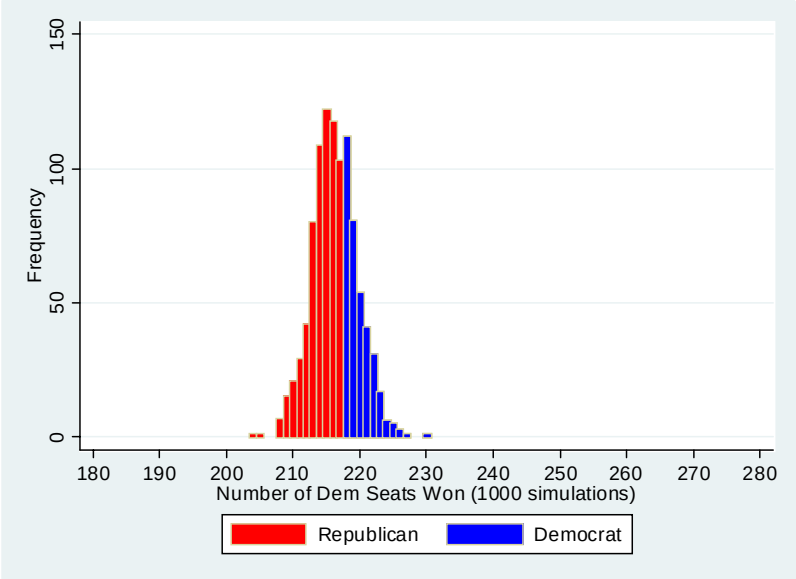


FIGURE 3: GENERIC POLLS = 52% DEM 48% REP

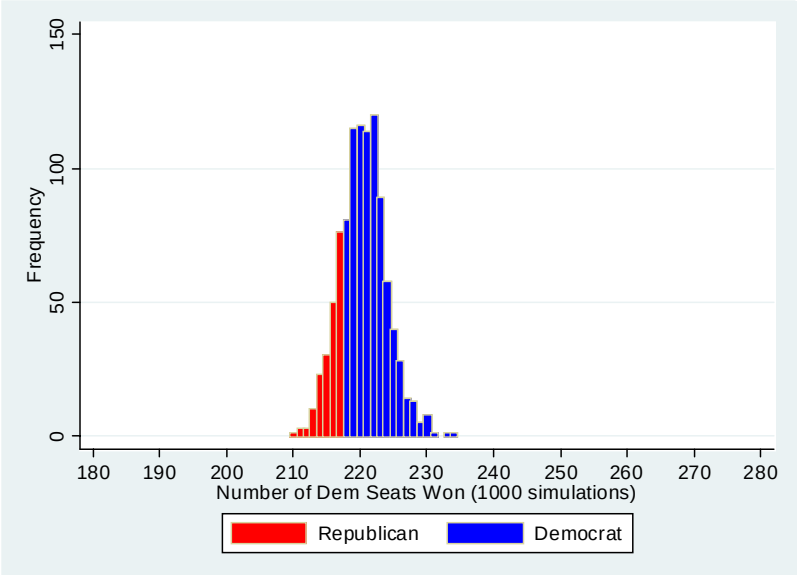


FIGURE 4: GENERIC POLLS = 54% DEM 46% REP

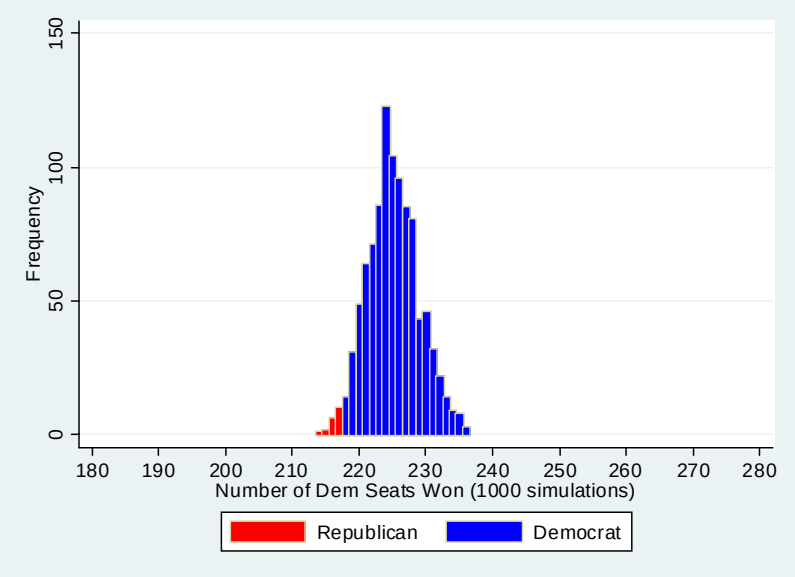


FIGURE 5: GENERIC POLLS = 56% DEM 44% REP

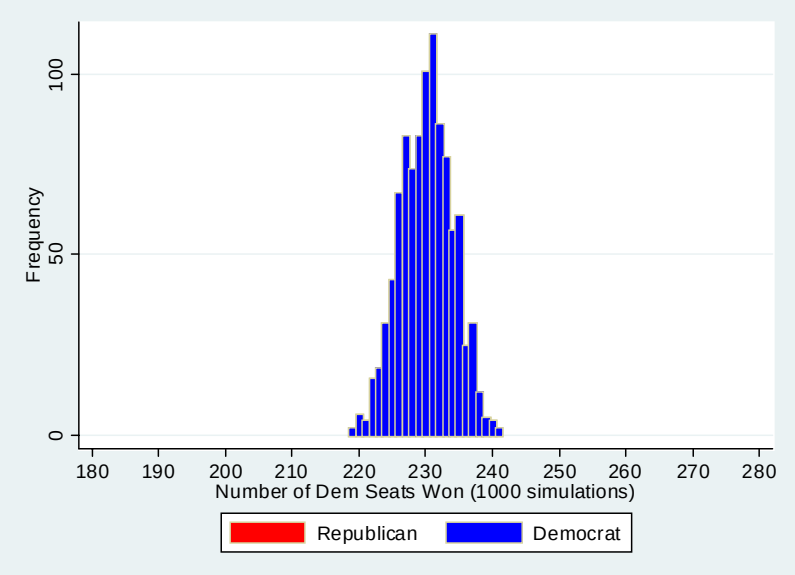


FIGURE 6: GENERIC POLLS = 58% DEM 42% REP

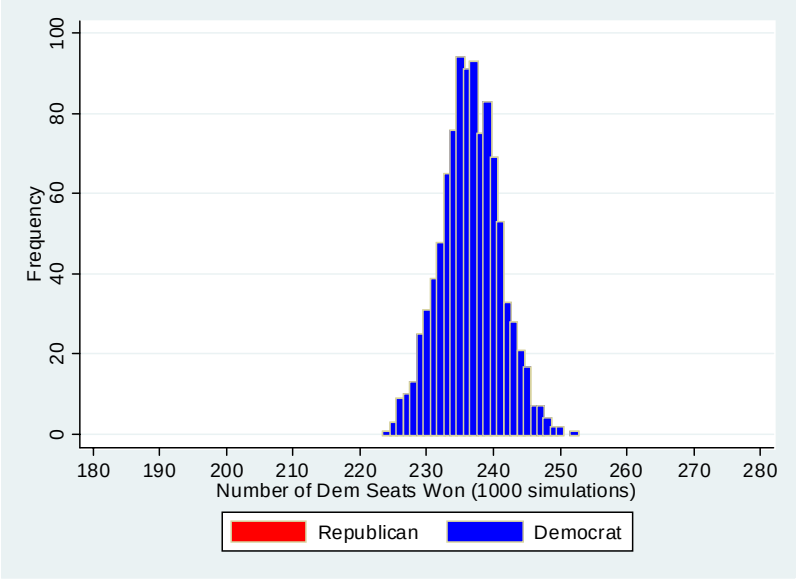
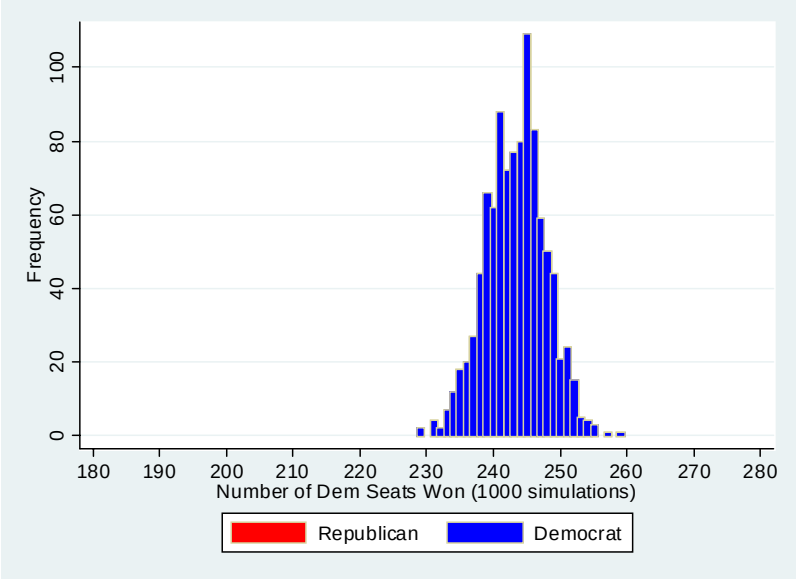


FIGURE 7: GENERIC POLLS = 60% DEM 40% REP



ISERP
INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH AND POLICY

Institute for Social and Economic Research and Policy
Columbia University in the City of New York
420 West 118th Street
8th Floor, Mail Code 3355
New York, NY 10027
Tel: 212-854-3081
Fax: 212-854-8925
Email: iserp@columbia.edu
www.iserp.columbia.edu

Editorial Board

Karen Barkey, Sociology
Peter Bearman, Sociology/ISERP
Alan Brinkley, History
Alessandra Casella, Economics
Ester Fuchs, Political Science/SIPA
John Huber, Political Science
Ira Katznelson, Political Science/History
Herbert Klein, History
Mary Clare Lennon, Sociomedical Sciences
Mahmood Mamdani, Anthropology/SIPA
Marianthi Markatou, Biostatistics
William McAllister, ISERP
Kathryn Neckerman, ISERP
Richard Nelson, Business/SIPA
Elliott Sclar, Urban Planning/SIPA
Seymour Spilerman, Sociology
Charles Tilly, Sociology
Harrison White, Sociology

Administration

Peter Bearman, Director
Kathryn Neckerman, Associate Director
Leah Lubin, Assistant Director

NOVEMBER 2006