# A Simple Explanation for Bias at the 50-50 Threshold in RDD Studies Based on Close Elections[1]

James M. Snyder, Jr.
Department of Government
Harvard University
NBER

Olle Folke
School of International and Public Affairs
Columbia University
IFN

Shigeo Hirano
Department of Political Science
Columbia University

May, 2011

[1] We thank Andrew Gelman, Gary King, David Lee, Adam Glynn, Arthur Spirling, Justin Grimmer, and Jas Sekhon for their helpful comments on an earlier draft of this paper.

# 1. Introduction

Recent papers have criticized RDD studies that use the outcome of "close" elections as a quasi-random treatment variable, because the electoral outcomes sometimes exhibit substantial "sorting" near the 50-50 threshold that distinguishes winners from losers. That is, an observable attribute of one of the candidates – such as incumbency status, or whether the candidate has the same party affiliation as the officials who are presumed to control key features of the electoral process – appears to help predict which candidate wins even in very close elections. Jason Snyder (2005) shows that in U.S. House elections between 1926 and 1992, incumbents win noticeably more than 50% of the very close races. Caughey and Sekhon (2010) investigate this further, and show among other things that winners in close U.S. House races raise and spend more campaign money. Grimmer et al. (2011) show that U.S. House candidates from the party in control of state offices, such as the governorship, secretary of state, or a majority in the state house or state senate, hold a systematic advantage in close elections.

One assumption underlying these critiques is that if the outcomes in close races are "random," then we should not expect to see any sorting near the 50-50 threshold. That is, for any observable $X_i$, near the threshold we should expect $Pr(Win_i|X_i) = .5$.

In this note we show that this assumption is false. In particular, if an electoral constituency is biased toward one party – say, the Democrats – in terms of voter party identification or ideological affinities, then even in "close" races we *expect* to see the Democratic candidate winning more than 50% of the time. This follows simply from the shape of the normal (or any strictly unimodal) density function.

Of course, *at* the threshold we expect the Democratic candidate to win exactly 50% of the time. But this is not true *near* the threshold. Given data limitations, researchers are typically forced to use windows around the 50-50 threshold of 1% or 2% or even larger. We show that under very plausible conditions, there will be a surprisingly large amount of sorting even inside a small window such as 1% or even 0.5%.

This does not mean that the critiques in the papers above are incorrect. However, it does mean that we can provide a simple theoretical explanation for some of the sorting that these papers identify. Incumbents, as well as candidates whose party won the governorship or state legislature in the previous election, tend to be from the favored party in a district or state, rather than the disadvantaged party. Thus, we actually expect these candidates to win more than 50% of the time, even in close races. Observing this is not necessarily a sign of electoral fraud (as suggested by Snyder, 2005). Also, we might not need complicated models of campaign resource allocation (as in Grimmer et al., 2011) to account for the patterns.

## 2. Model and Results

Consider a simple two-party model of one electoral district in which the outcome in any given election is determined by a long-term "normal vote" and a short-term "shock." Let $\mu_D$ denote the normal vote and let $\epsilon$ denote the shock, where $\mu_D$ is a real number and $\epsilon$ is a random variable. The vote-percentage for the Democratic candidate is $v = \mu_D + \epsilon$. Suppose $\mu_D > 50$, so the district tends to favor Democratic candidates. Even though Democrats are favored, if $\epsilon$ is negative and large enough in magnitude then the Democrat might lose. Also, if $\epsilon$ is near $50 - \mu_D$ then the outcome will be near 50-50, i.e., the race will be "close." Note that $\epsilon$ incorporates all factors other than the normal vote that affect election outcomes, such as partisan tides, candidates' relative qualities, incumbency advantages, national and local economic shocks, cross-cutting or wedge issues, and campaign strategies and tactics.

Suppose $\epsilon$ has a normal distribution, i.e., $\epsilon \sim N(0, \sigma_D^2)$. Then $v \sim N(\mu_D, \sigma_D^2)$.[1] Suppose we are researchers with access to a large number of election outcomes for the district, and we attempt an RDD with a window of $[50-\delta, 50+\delta]$, where $\delta$ is "small," say 1, 2 or 3. Figure 1 presents an example, with $\mu_D = 60$ and $\delta = 3$. As the figure makes clear, we do not expect to see the Democratic candidate winning 50% of the time in this window. The shaded area to the right of the line at the 50% threshold shows where the Democratic candidate wins, and the shaded area to the left of the threshold shows where the Republican candidate wins. Since the area on the right is clearly larger than that on the left, we expect to see the Democrat win *more* than 50% of the time.

[1] Of course, this cannot literally be true since the normal distribution has unbounded support and $v$ must be between 0 and 100. But this does not matter for our analysis, particularly since the focus is on close elections. The results below do not require that $\epsilon$ has a normal distribution, only that it is strictly single-peaked. If $\epsilon$ has a uniform distribution then the results do not hold.

How much more? Consider a linear approximation to the density function of $v$ around 50%. The density of $v$ is $f(v) = (2\pi\sigma_D^2)^{-1/2}exp(-(v-\mu_D)^2/2\sigma_D^2)$. The slope of this density at $v = 50$ is $f'(50) = (\mu_D - 50)f(50)/\sigma_D^2$. The probability the Democratic candidate wins given that $v \in [50-\delta, 50+\delta]$ is then approximately

$$\begin{aligned} P_D &= \frac{\delta f(50) + \delta^2 f'(50)/2}{[\delta f(50) + \delta^2 f'(50)/2] + [\delta f(50) - \delta^2 f'(50)/2]} \\ &= \frac{\delta f(50)[1 + \delta(\mu_D - 50)/\sigma_D^2]}{4\delta f(50)} \\ &= \frac{1}{2} + \frac{\delta(\mu_D - 50)}{4\sigma_D^2} \end{aligned}$$

For U.S. House elections during the period 1980-2008, the average within-district standard deviation of the Democratic percentage of the two-party vote is about 5.8 – to be conservative, set $\sigma_D = 6$. Suppose, for example, the threshold is 2% ($\delta = 2$). Then in a district that is 60% Democratic ($\mu_D = 60$), the probability the Democratic candidate wins is approximately 0.5 + 20/144 = 0.64. So, in this case we should expect to see Democrats winning about 64% of the races in a 2% window around the 50-50 threshold, not 50% of the races. Even in a small window of 49% to 51%, the Democrats are expected to win about 57% of the races.

Figures 2a and 2b show exact calculations based on the normal density function, rather than linear approximations. Figure 2a presents the calculations for $\sigma_D = 9$ and Figure 2a presents the calculations for $\sigma_D = 6$. (The first value is slightly larger than the within-state standard deviation of the Democratic percentage of the two-party vote over the period 1980-2008). Evidently, many of the numbers are much larger than 50%, especially when $\delta \geq 2$.

## 3. Implications

Does this matter? It probably depends on the application. Consider typical questions for which the RDD approach seems suited. Does party affiliation affect roll call voting independently of constituency preferences? Do Republican governors, or Republican-controlled state legislatures, promote more "pro-economic-growth" policies than Democratic governors or legislatures? Vastly simplified, the underlying model assumed in attacking these questions is typically of the form:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \epsilon_i \tag{1}$$

where $D_i = 1$ if the Democrat candidate wins constituency $i$ and $D_i = 0$ if the Republican candidate wins, and $X_i$ is another relevant variable such as the average or median preference of voters in constituency $i$. The outcome variable $Y_i$ might be a roll call voting score, a measure of tax policy, or economic growth. The parameter of interest is $\beta_1$.

The identification strategy underlying the RDD is that $D_i$ is approximately independent of $X_i$ (and everything else) once we limit attention to close elections. If this assumption is correct, then OLS estimates of $\beta_1$ will be consistent.

Unfortunately, given the sorting described above $D_i$ and $X_i$ will often be strongly correlated even in close elections. For example, if $X_i$ is the median preference of voters in constituency $i$ on a liberal-conservative ideology scale, then it is probably correlated (positively) with the percentage of voters in the constituency who are Democrats. And, as shown above, the probability that $D_i = 1$ is positively correlated with the percentage of voters in the constituency who are Democrats, and therefore with $X_i$. Thus, an OLS regression of $Y_i$ on $D_i$ will yield an estimate of $\beta_1$ that is biased upward, even if this regression is conducted only on a sample of close races.[2]

How can we address this problem? One obvious idea is to control for $X_i$ in the regression analysis. Of course, in many cases $X_i$ is unobservable – indeed, the difficulty or impossibility of measuring $X_i$ is often one motivation for using an RDD in the first place. If measuring $X_i$ is impossible, another idea is to control flexibly for $\mu_{Di}$. In many cases, however, even this is difficult or impossible – e.g., the available measures of the normal vote for U.S. congressional districts are often poor, and measures for smaller constituencies such as state legislative districts are generally even poorer. If measuring $\mu_{Di}$ is also impossible, a third idea is to control flexibly for $v_i$, which may also substantially reduce the bias. This leads to specifications similar to those suggested in Imbens and Lemieux (2008).

[2] Put differently: even in a close election, learning that the Democratic candidate won the election provides information about the underlying partisan composition of the constituency, and, therefore, probably about other characteristics of the constituency.

## 4. An Application

Grimmer et al. (2011) show that U.S. House candidates from the party that controls the governorship in a state win significantly more than 50% of the "close" races in which they are involved. As we show below, this is also true for candidates in statewide races – i.e., candidates for the U.S. senate and governor, and "down ballot" statewide executive offices such as attorney general, secretary of state, treasurer, and auditor. In this section we show that candidates for these offices also win significantly more than 50% of the "close" races if they are from the party that voters in the state appear to favor. This is exactly what is predicted in section 2 above. More significantly, we show that when we compare the relative impact of "state voter partisanship" and "control of the governors' office" on the probability that a candidate wins in a close race, the first variable dominates.

Consider the following specification:

$$D_i = \theta_0 + \theta_1 G_i + \theta_2 P_i + \nu_i \tag{2}$$

where the dependent variable $D_i = 1$ if the Democratic candidate wins race $i$ and $D_i = 0$ if the Republican candidate wins; $G_i = 1$ if the Democrats control the governorship in the state where race $i$ is held at the time of the election, and $G_i = 0$ if the Republicans control the governorship; and $P_i$ is a measure of the partisanship of voters in the state where race $i$ is held at (or at least near) the time of the election. We estimate equation (2) limiting the sample to the set of "close" races, where the winner's vote percentage is below $(50+\delta)\%$, for $\delta \in \{1, 2, 3, 4, 5\}$.

While we restrict attention to close races when estimating (2), we use *lopsided* races to measure $P_i$. More specifically, suppose race $i$ is held in state $j$ in year $t$. Consider all statewide races in state $j$ in years $t-6$ to $t-1$ in which the winner won by *more* than 60%. Let $P_i = 1$ if the Democrats won 3 or more of these contests and the Republicans won no more than 1 of them, or the Democrats won 2 of these contests and the Republicans won none. Symmetrically, let $P_i = 0$ if the Republicans won 3 or more of these contests and the Democrats won no more than 1 of them, or the Republicans won 2 of these contests and the Democrats won none. Drop all other cases – i.e., treat these as "ambiguous" cases in which the voters do not lean clearly one way or the other. Call state-years with $P_i = 1$ "Democratic -leaning" cases, and call state-years with $P_i = 0$ "Republican-leaning" cases.

We study the period 1880-2009, and the following offices: U.S. senator, governor, lieutenant governor, attorney general, secretary of state, treasurer, auditor/comptroller/controller, superintendent of education, commissioner of agriculture, public utility commissioner, corporation commissioner, and lands commissioner. We also include at-large elections for U.S. representative.

Figure 3 presents a scatterplot of the fraction of races won by Democratic candidates in "close" races as a function of $\delta$. The points labelled with $D$'s are for Democratic-leaning cases, and those labelled with $R$'s are for Republican-leaning cases. As predicted by the simple model in section 2, the Democrats win more than 50% of races in the Democratic-leaning cases, and this percentage grows strongly as the window size grows. Similarly, the Republicans win more than 50% of races in the Republican-leaning cases, and this percentage grows strongly as the window size grows. For $\delta \geq 2\%$ the differences between the Democratic-leaning and Republican-leaning cases are quite large and statistically significant.

Table 1 shows estimates of equation (2). For making comparisons, it also includes estimates of models with $G_i$ or $P_i$ alone as regressors.[3] The top panel uses all available data and shows estimates when $G_i$ is the only independent variable. This panel shows that the findings reported in Grimmer et al. (2011) for the U.S. House also hold for statewide offices, at least for wide enough windows around the 50% threshold. For $\delta \geq 2$ the Democrats win more often when they control the governorship at the time of the election than when they do not, and the effect is large and highly statistically significant for $\delta \geq 3$. The second panel again shows estimates when $G_i$ is the only independent variable, but restricts the sample to cases where $P_i$ is not missing. This panel is important for making comparisons with the third and fourth panels, which use the same sample. Again, for $\delta \geq 2$ the Democrats win more often when they control the governorship at the time of the election than when they do not, and for $\delta \geq 3$ the effect is large and highly statistically significant. Note that the estimated coefficients are even larger than in the top panel.

[3] In the interest of space we do not show the estimates of the constant term, $\theta_0$.

The third panel again shows estimates when $P_i$ is the only independent variable. Essentially, it is a tabular version of Figure 3. It shows that for all values of $\delta \geq 1$ the Democrats win more often in Democratic-leaning state-years than in Republican-leaning state-years, and that for $\delta \geq 2$ the difference is large and statistically significant. Note also that for all $\delta$ the estimated effect is always larger than the corresponding estimated effect for $G_i$ in the second panel, and for $\delta = 2$, 3, and 4 it is roughly twice as large.

The bottom panel shows estimates of equation (2), with $G_i$ and $P_i$ both included as independent variables. Note that the estimates for $G_i$ are cut approximately in half relative to those in the second panel. In fact, the estimated value of $\theta_1$ is not statistically significant at the .05 level except for the relatively large window of $\delta = 5\%$. By contrast, the estimates for $P_i$ are only slightly smaller than those in the third panel, and the estimated values of $\theta_2$ are statistically significant for all $\delta \geq 2$. Moreover, for $\delta = 2$, 3, and 4 the estimated value of $\theta_2$ is more than three times as large as the estimated value of $\theta_1$, and for $\delta = 5$ it is nearly twice as large.

We experimented with other definitions of $P_i$ based only on control of down-ballot offices. The simplest of these measures uses just one or two offices – e.g., control of the attorney general's office or of the offices of state treasurer or state auditor/controller/comptroller (henceforth, simply auditor). Interestingly, we find qualitatively similar results – often even stronger results – using these alternative definitions. For example, we defined $P_i = 1$ if Democrats controlled the state treasurer's office and the office of auditor (for states that elect both), or if Democrats controlled the state treasurer's office (for states that do not elect the auditor), or if Democrats controlled the office of auditor (for states that do not elect the treasurer); $P_i = 0$ if Republicans controlled the state treasurer's office and the office of state auditor (for states that elect both), or if Republicans controlled the state treasurer's office (for states that do not elect the auditor), or if Republicans controlled the office of auditor (for states that do not elect the treasurer); and we omitted cases where both offices were elected and the Democrats controlled one and the Republicans controlled the other. Importantly, this office has nothing to do with elections, so it seems impossible that control of this office would allow a party to skew electoral outcomes in its favor. Using this definition of $P_i$, we find that the estimates of $\theta_2$ in equation (2) are statistically significant and always much larger than the estimates of $\theta_1$.[4]

## 5. Conclusions

Although our paper focuses on regression discontinuity designs that rely on elections, it raises a general question about RDDs. Given that *all* empirical studies have finite sample sizes, do we *always* need a model of the process that generates the forcing variable in order to assess what constitutes an "adequately small" window around the threshold?

[4] Thus, it would appear that controlling the office of treasurer or auditor has a much larger impact on winning close races than controlling the governors office. It seems unreasonable to attribute this to the treasurer or auditor's officer per se, so we prefer the interpretation that control of these offices is a good proxy for state partisanship.

# References

Caughey, Devin M. and Jasjeet S. Sekhon. 2010. "Regression-Discontinuity Designs and Popular Elections: Implications of Pro-Incumbent Bias is Close U.S. House Races." Unpublished manuscript.

Grimmer, Justin, Eitan Hersh, Brian Feinstein, and Daniel Carpenter. 2011. "Are Close Elections Random?" Unpublished manuscript.

Imbens, Guido and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142(2): 615-635.

Snyder, Jason. 2005. "Detecting Manipulation in U.S. House Elections." Unpublished manuscript.

**Table 1: State Partisanship vs. Governor Control**

Dep. Var. = Democratic Victory Dummy

| | Margin = 1% | Margin = 2% | Margin = 3% | Margin = 4% | Margin = 5% |
|---|---|---|---|---|---|
| Democratic Governor | -0.007 | 0.033 | 0.066 | 0.090 | 0.128 |
| | (0.030) | (0.021) | (0.018) | (0.016) | (0.014) |
| R-squared | 0.000 | 0.001 | 0.004 | 0.008 | 0.017 |
| N | 1150 | 2244 | 3193 | 4053 | 4909 |
| Democratic Governor | -0.011 | 0.045 | 0.082 | 0.110 | 0.170 |
| | (0.048) | (0.034) | (0.028) | (0.025) | (0.022) |
| R-squared | 0.000 | 0.002 | 0.007 | 0.012 | 0.029 |
| N | 433 | 861 | 1232 | 1577 | 1968 |
| Democ Leaning State | 0.053 | 0.100 | 0.146 | 0.200 | 0.236 |
| | (0.048) | (0.034) | (0.028) | (0.025) | (0.022) |
| R-squared | 0.003 | 0.010 | 0.021 | 0.039 | 0.055 |
| N | 433 | 861 | 1232 | 1577 | 1968 |
| Democratic Governor | -0.030 | 0.016 | 0.039 | 0.050 | 0.101 |
| | (0.051) | (0.036) | (0.030) | (0.026) | (0.023) |
| Democ Leaning State | 0.062 | 0.095 | 0.134 | 0.183 | 0.201 |
| | (0.051) | (0.036) | (0.030) | (0.026) | (0.023) |
| R-squared | 0.004 | 0.010 | 0.022 | 0.042 | 0.064 |
| N | 433 | 861 | 1232 | 1577 | 1968 |

The sample in the top panel includes all available cases within the stated winner vote margin. The remaining panels include all cases where *Democ Leaning State* is not missing. Standard errors in parenthesis.

Figure 1

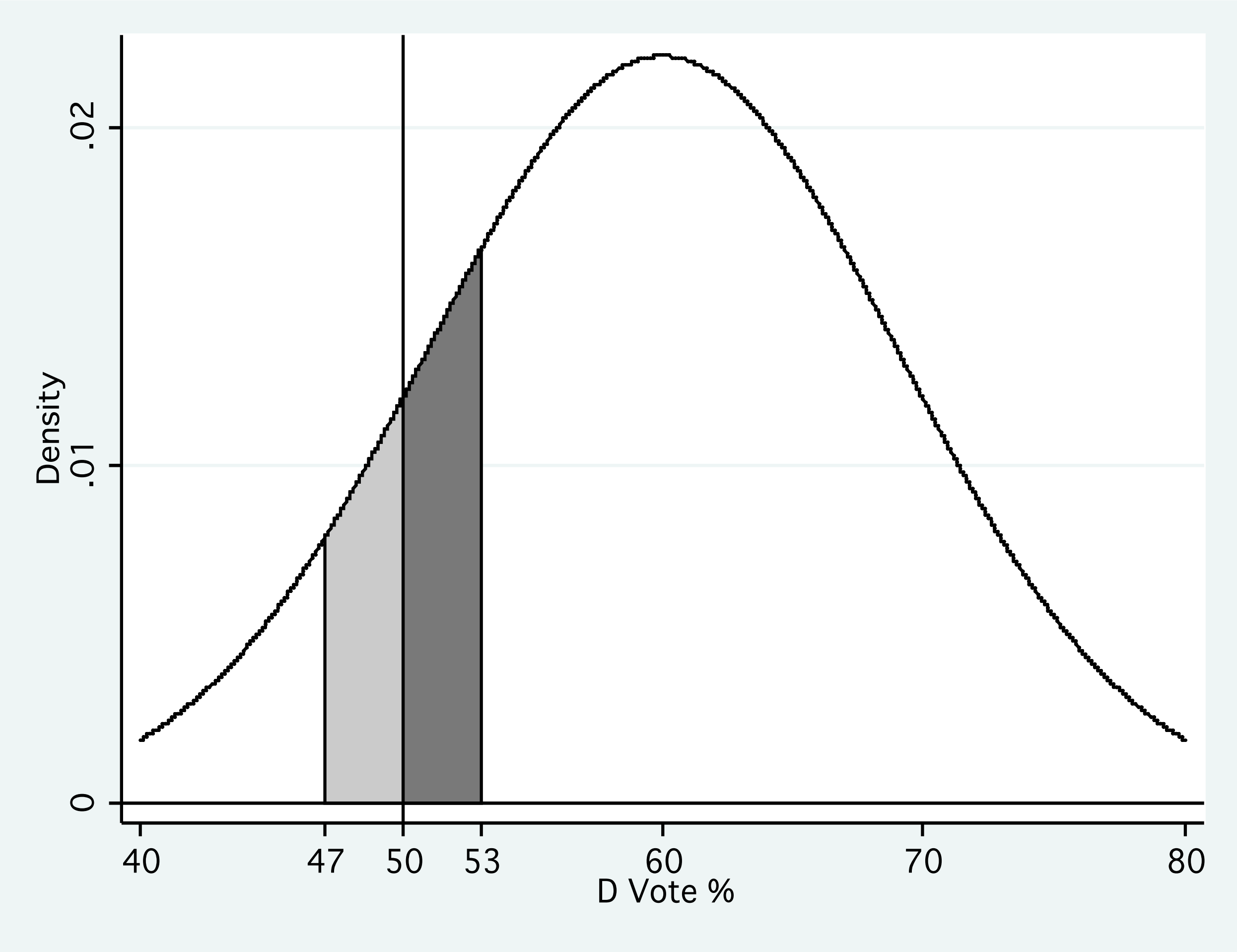


Figure 2a

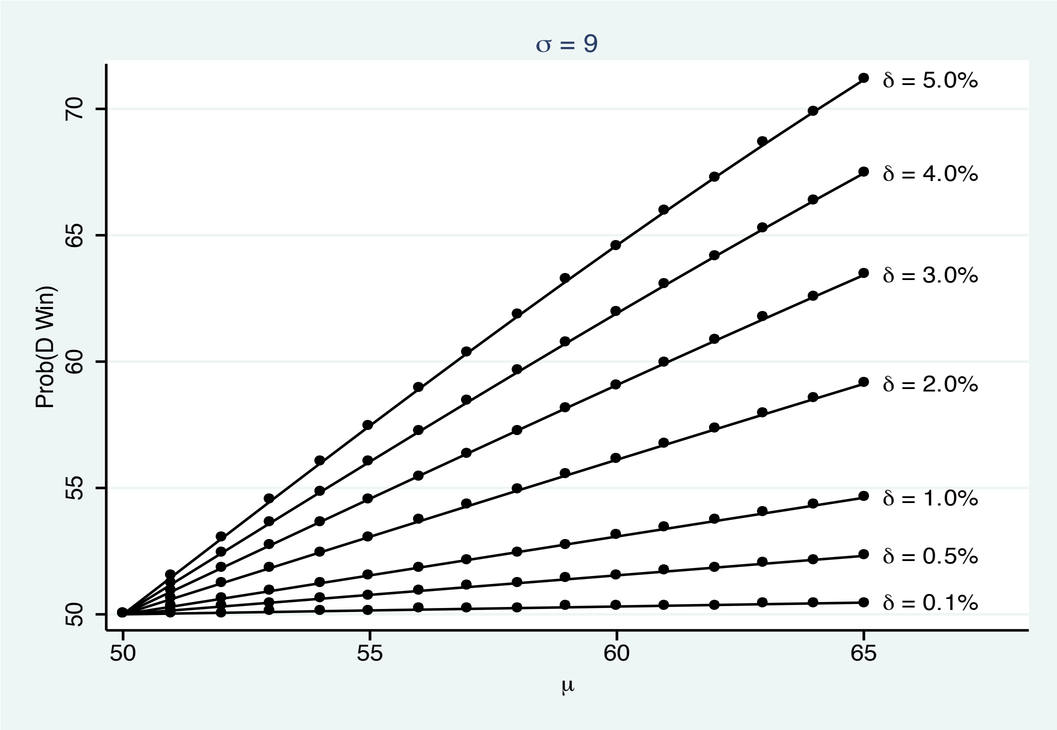


Figure 2b

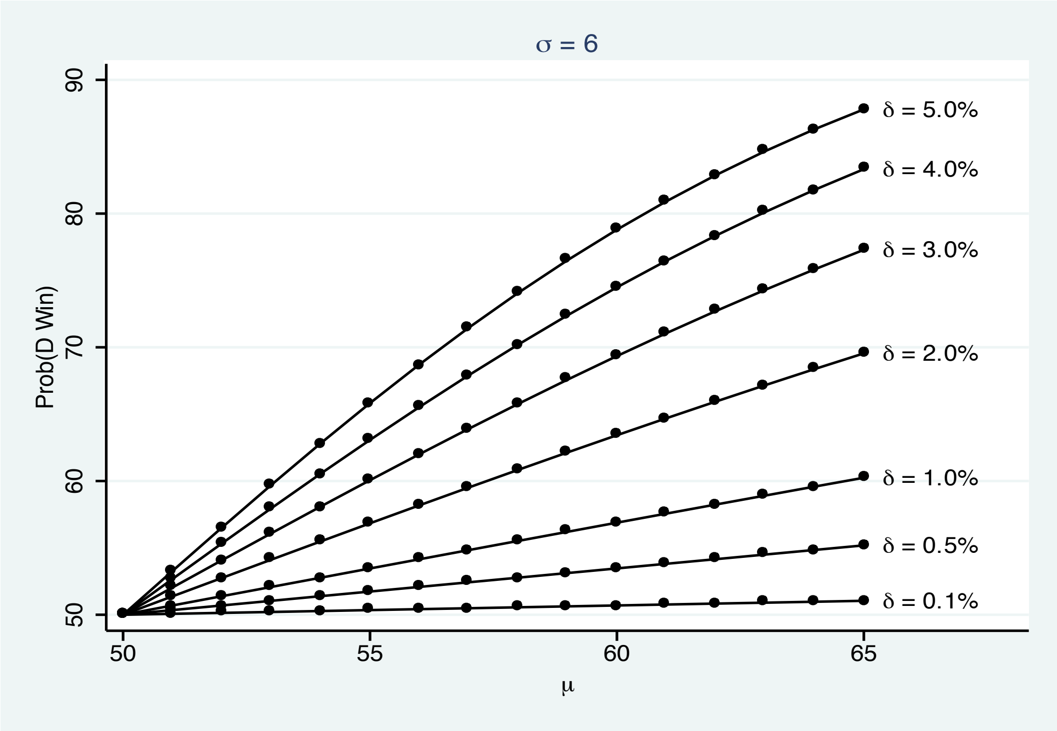
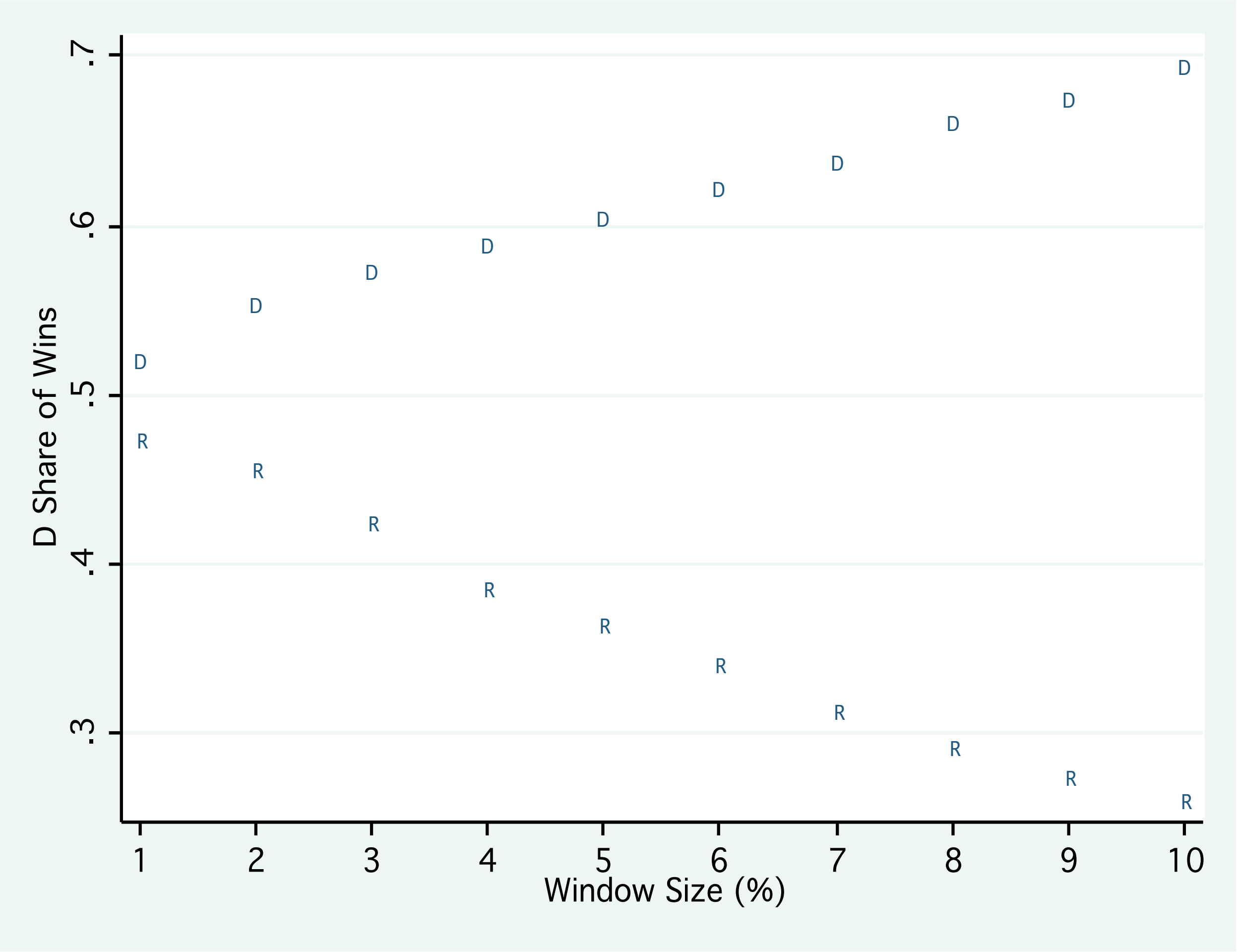


Figure 3


D Share of Wins
D
R
Window Size (%)