An analysis of voting gerrymandering in matrix aspect

Xiangyun Ella Xu

September,2008

Abstract

In an election, people vote for the one they want. So we think that the one with more ballots wins. However, different choice of weights or voting method do make difference; different reasonable approach won't necessarily give the same election outcome. Bill Clinton, for instance, failed to receive a majority vote during the 1992 Presidential election. The election taught us that ballots cast need not be ballots counted; that voting machines may provide a rapid count, but not necessarily a full and accurate one.

1 Introduction

Since election day, the public has been treated to an array of proposed ways to correct our elections. Some are interesting, and each might correct some form of abuse. But, quite frankly, most of these suggestions have absolutely nothing to do with the real difficulties of the Presidential election. The real problem concerns our archaic election procedures. As it will be shown, by almost any objective measure, our standard tools of democracy are seriously flawed. They need not deliver as promised. The problem is so severe that even in settings not hindered by the complications of an Electoral College, even those election which are free from controversy, election outcomes need not mean what we commonly think they do. The bothersome reality is that our election outcomes can fail to reflect the wishes of the voters.

The 2000 U.S Presidential election taught us several important lessons. We learned the value of emphasizing patience and accuracy over speed in making statistical projections; all major television networks were acutely embarrassed on election night by forecasting a Florida victory for Vice President Al Gore, then withdrawing that premature prediction, next predicting that Governor George W. Bush would win, and finally retracting even that forecast.

More than a month later, the outcome remained in doubt. We learned the value of voting when an extra vote per Florida precinct, or maybe even a couple more forceful punches of the ballots per precinct, might have determined already on election night, one way or the other, the next U.S President. This electoral difficulty is not restricted to the historical Presidential election of November 2000; it is a recurring problem that stands ready to plague all of our elections on an uncomfortably regular basis. Like in 1998, with only about 37% of the vote, Jesse Ventura beat Hubert Humphrey and Norm Coleman to become the governor of the state of Minnesota.

The point which will be made is that our basic voting procedures can generate problems so worrisome that it is reasonable to worry about the legitimacy of most election outcomes. This is not a conjecture; mathematical support will be provided.

2 The rule

In this paper, I use the following rule to run an election. Each entry of the matrix represents a person's will. There are two candidates, 0 and 1. The closer each entry towards 0, the stronger the person's will to vote for candidate 0. In the same way, the closer each entry towards 1, the stronger the person's will to vote for candidate 1. Here is a matrix I made up for instance:

$$\begin{bmatrix} 0.3 & 0.1 & 0.7 \\ 0.7 & 0.6 & 0.8 \\ 0.9 & 0.1 & 0.3 \end{bmatrix}$$

Now here is the rule for deciding who wins. Firstly, round each entry then we get such a matrix:

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Add each row or column and divide by 3 and then round it. So each row or column give a result, add them, so we get the number of votes of this matrix. If we divide this number of votes by 3 and round it, we can see the final result which gives the winner, i.e, either 0 or 1. In this case, if we do the adding in rows, we get

$$round(1/3) = 0 \quad round(3/3) = 1 \quad round(1/3) = 0$$

So the result is $round((0 + 1 + 0)/3) = 0$. So candidate 0 wins. However, if we do the

adding by column, we can see that

$$round(2/3) = 1 \quad round(1/3) = 0 \quad round(2/3) = 1$$

The candidate 1 wins. In this case, different ways of counting give totally different winner.

3 Larger sample

Now we investigate larger matrix with more runs. Let me have an 55 by 55 matrix with entries extracting from a uniform distribution between 0 and 1. With the rules mentioned above , we get a value for either way of adding, let me call them 'sum of rows' and 'sum of columns' respectively. So then I can get the larger one of them and call it 'sum of maximum'. We repeat this process for the 55×55 matrix for 50,000 times and see how the votes distribute. With the following Matlab code:

```
N=55
for i=1: 50000
A=round(rand(N));
F=sum(A,1);
R=sum(round(F/N));
result(i)=R;
end
for i=1: 50000
A=round(rand(N));
G=sum(A,2);
C=sum(round(G/N));
result1(i)=C;
end
for i=1: 50000
A=round(rand(N));
F=sum(A,1);
G=sum(A,2);
R=sum(round(F/N));
C=sum(round(G/N));
M=max(R,C);
result2(i)=M;
end
i=0:1:N;
subplot(3,1,1);
hist(result,i) subplot(3,1,2);
hist(result1,i) subplot(3,1,3);
hist(result2,i)
```

we can get the following histograms Fig 1

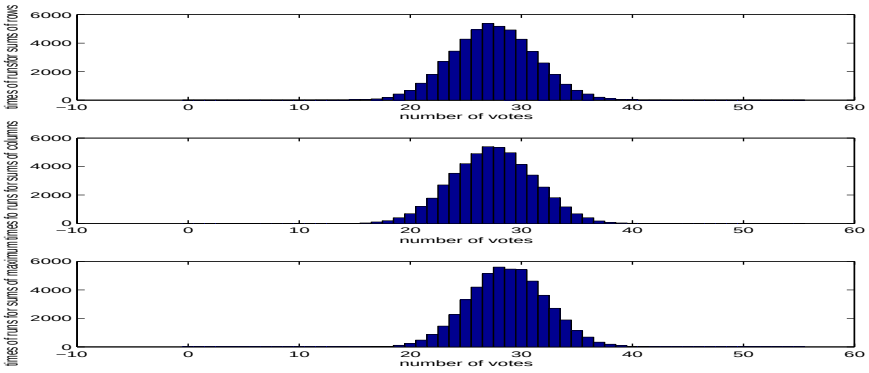

Fig 1

From the histograms we can see the distribution of the votes is normal. The plots of sums of rows and sums of columns are quite similar, with both mean at $x = 27.5$. However, the mean of the sums of maximum occurs at $x = 29$. This indicates that the curve of distribution of sum of maximum shifts right to the sum of rows/columns. But how much is the shift? How can we calculate the shift? What's more, in this case, any single vote greater or equal to $k = 0.5$ stands for candidate 1, vice versa. But what if we set the different boundaries k, say, 0.3 or 0.8, how will the difference between the third plot and first/second plot change? In the next section we are going to answer these questions.

4 Relationship between the 'shift' and k

Firstly, let us have a look at the histograms of the three different sums when we have different k. When we have $k = 0.6$, i.e, any entry of a matrix greater or equal to 0.6 stands for candidate 1. The following are the matlab code:

```
for  k=0.6  case: N=55;
k=0.6;
for  i=1: 50000
A=rand(N);
A(find(A>k))=1;
A(find(A<k))=0;
F=sum(A,1);
R=sum(round(F/N));
result(i)=R;
G=sum(A,2);
C=sum(round(G/N));
result1(i)=C;
F=sum(A,1);
```

G=sum(A,2);

R=sum($round$(F/N));

C=sum($round$(G/N));

M=max(R,C);

$result2$(i)=M;

end

i=0:1:N;

$subplot$(3,1,1);

$hist$($result$,i) $subplot$(3,1,2);

$hist$($result1$,i) $subplot$(3,1,3);

$hist$($result2$,i)

Then we can get the following histograms Fig 2:

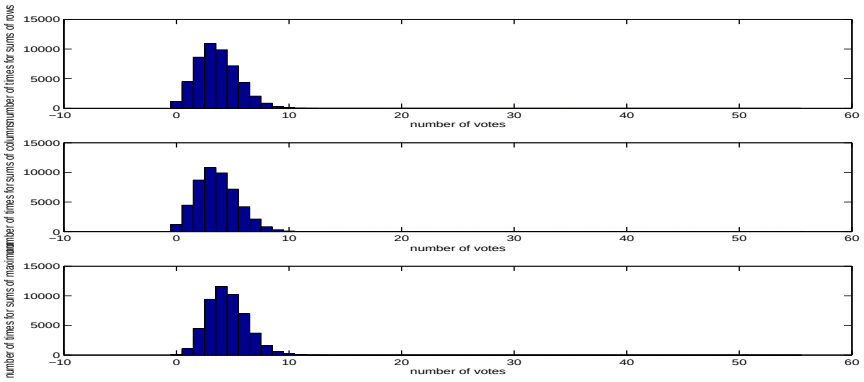

Fig 2

From fig 2 we can see that the curves all shift to the left comparing to fig1, with smaller means at 3,3 and 4 respectively. Also, the maximum curve shifts right to the row/clumn curve. Similarly, let us set k=0.4 and we can get the following histograms Fig 3:

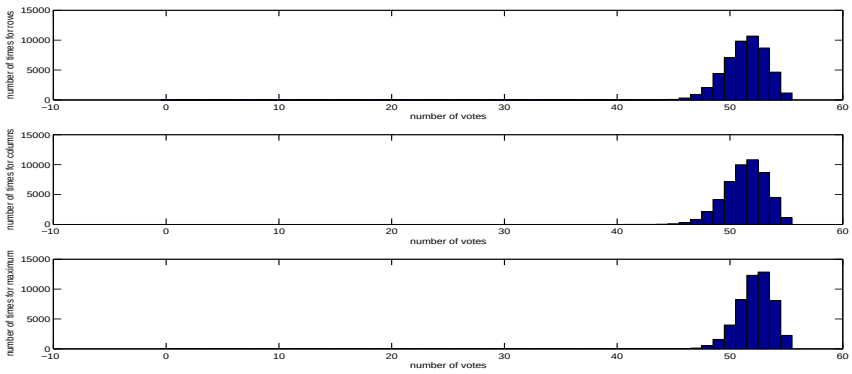

Fig 3

From fig 3 we can see that the curves all shift to the right comparing to fig 1, with smaller means at 52, 52 and 53 respectively. Also, the maximum curve shifts right to the row/clumn curve. Next, we will investigate the relationship between k and the bias between the maximum curve and row/column curve. Let us define the bias to be the average mean squared distance from the distribution of sums of rows/columns and sums of maximum. So I write up the matlab code like this:

```
N=55;
len=0.01;
w=1/len;
P=ones(w,N);
Q=ones(w,N);
T=ones(w,N);
for k=0.01: 0.01: 1
for i=1: 10000
A=rand(N);
A(find(A>k))=1;
A(find(A<k))=0;
F=sum(A,1);
R=sum(round(F/N));
result(i)=R;
G=sum(A,2);
C=sum(round(G/N));
result1(i)=C;
F=sum(A,1);
G=sum(A,2);
R=sum(round(F/N));
C=sum(round(G/N));
M=max(R,C);
result2(i)=M;
end
i=1:1:N;
[yout,xout]=hist(result,i);
[yout1,xout1]=hist(result1,i);
[yout2,xout2]=hist(result2,i);
m=fix(k*100);
P(m,:)=yout;
Q(m,:)=yout1;
T(m,:)=yout2;
end
PP=((P-T).*(P-T));QQ=((Q-T).*(Q-T));
k=0.01:0.01:1
plot(k,mean(PP,2)',-r')
hold on
plot(k,mean(QQ,2)')
```

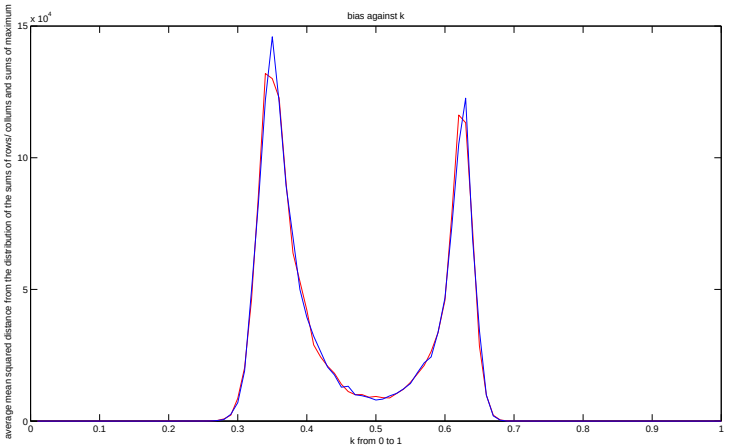

Fig 4

Then we get the following histograms Fig 4.

In Fig 4, the red curve indicates the average mean squared distance from the sums of rows against k, while the blue curve indicates the average mean squared distance from the sums of columns against k. There are two peaks at $k = 0.35$ and $k = 0.65$ respectively in this figure. However the bias is almost zero at other values of k. So we want to see the what happens at these peaks.

5 Investigation on the 'peaks'

Now let us investigate the peaks at different size of matrix. Here are some figures at different N:

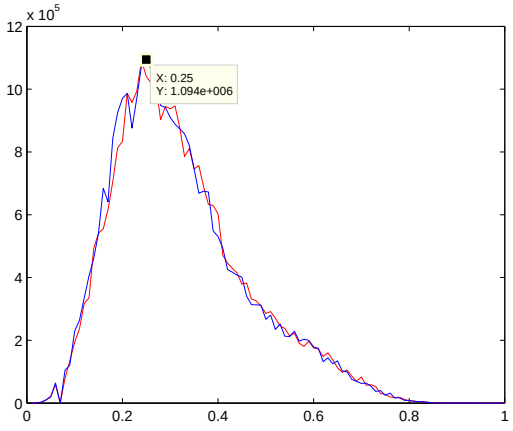

Fig 5: $N = 3$

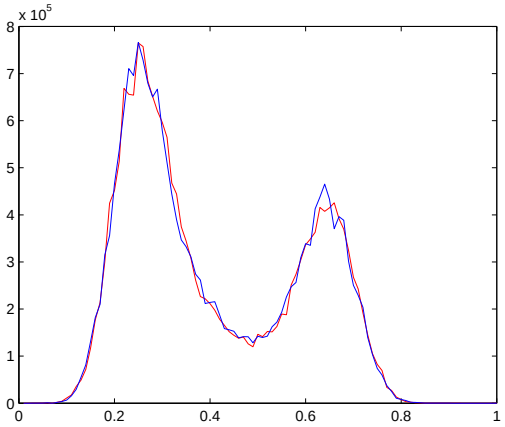

Fig 6: $N = 7$

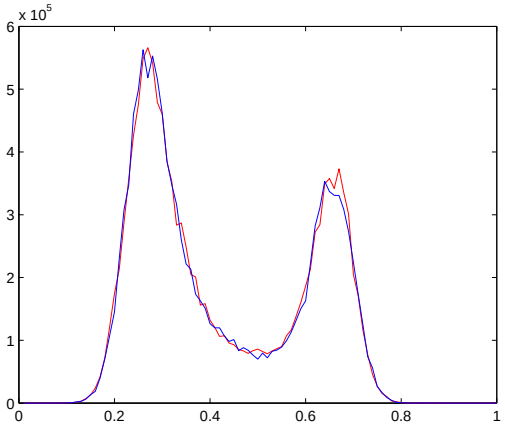

Fig 7: $N = 11$

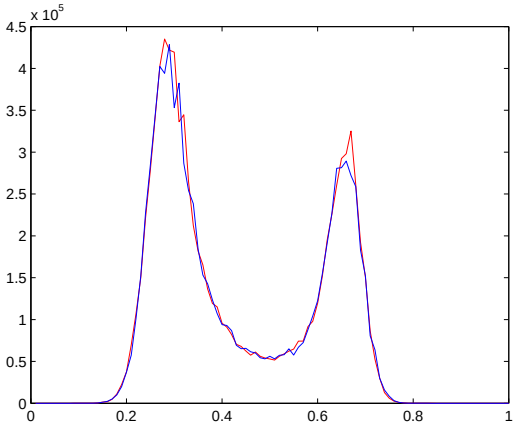

Fig 8: $N = 15$

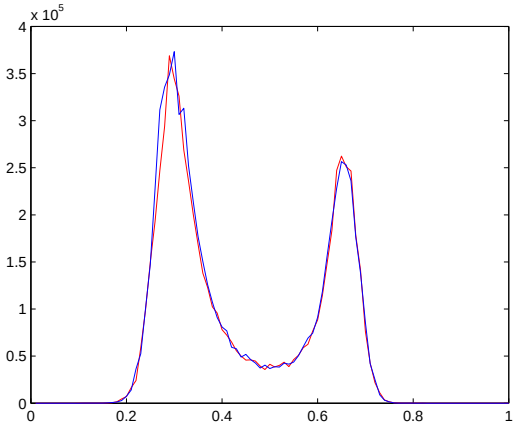

Fig 9: $N = 19$

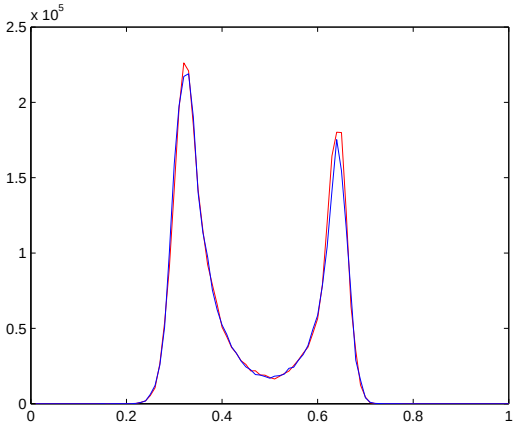

Fig 10: $N = 33$

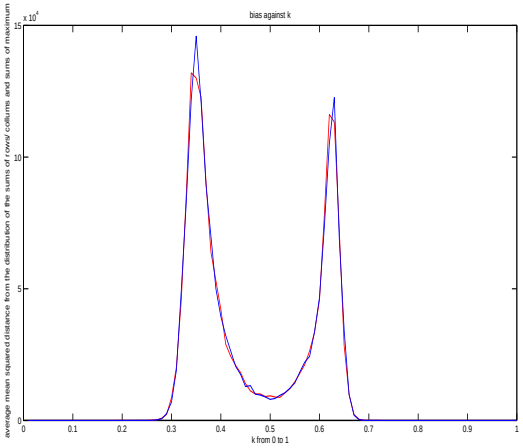

Fig 11: $N = 55$

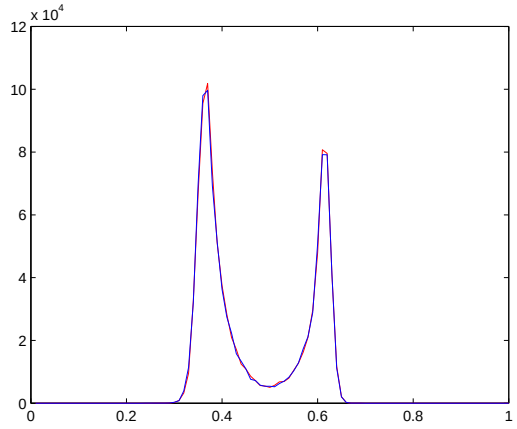

Fig 12: $N = 77$

We can see that at different N, peaks occur at the same value of k. But as N gets bigger, the value at the left peak gradually decreases, while the value at the right peak increases, i.e. the difference between the height of the two peaks get smaller. To investigate this change, I plot more graphs at different N and write down the value at the two peaks and calculate the difference. The following is the table of the statistics:

N	y at left peak	y at right peak	difference
3	1094000	144200	949800
7	766200	465400	300800
11	565900	373200	240000
15	435200	325200	110000
19	373300	256600	116700
23	314300	241400	72900
27	277000	206900	70100
33	226200	180200	46000
39	206500	158200	48000
55	145900	122600	23300
62	141100	104900	36200
69	130600	99640	30960
77	101800	90730	21070

According to the above table, I plot a diagram with difference against N and get the following figure:

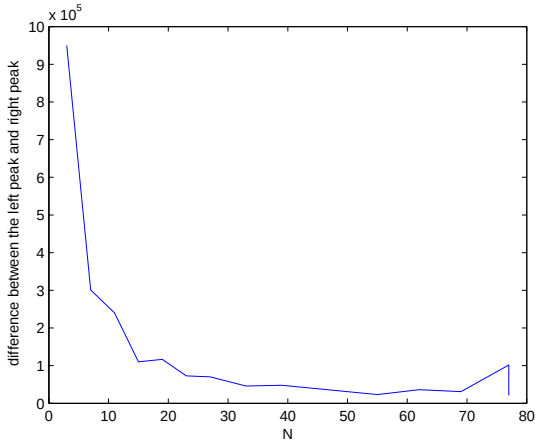

Fig 13: difference against N

Conclusion

After a series of experiment and investigation, we find that, with a small matrix, if we set k at 0.25 to 0.35, the bias is quite large. This means that candidate '1' has a larger probability to win even with a smaller sum of votes. As the sample gets larger, we are more sure to have candidate '1' win if we set k at around 0.3, 'left point', say, or k at 0.65, 'right point', say. However when N not big enough, to have k at the 'left point' is a better choice, because the left peak is still higher than the right one. Moreover, fig13 shows that with a larger sample, the difference between the two peaks are getting smaller, which means that we can choose either the left or right point.

Acknowledgements

I would like to thank my tutor Mason Porter and Nick Jones for his invaluable help and support (and additional resources). I would like to thank my fellow James Wall for providing guidance.

Reference