

# The Statistics of Baseball and Politics

Can Baseball be used to predict to outcome of elections?  
An introduction to Nate Silver, the 2<sup>nd</sup> Chicago whiz-kid

(after Howard Katch, of course).

Mark Berron, CQE, CSQE, SSBB  
Perot System Government Services  
Mark.Berron@psgs.com



# Background

- First off, the title is misleading. There is really minimal correlation between baseball World Series winners and who wins the Presidential elections. It went well for a few series after World War II, but it has broken down. Now that you're here....
- The elections of 2004 proved beyond a doubt that exit polls are misleading, but why? They had been used for years with some amount of success. Were they lucky?
- The **methods** use in predicting baseball outcomes can be used in predicting the Presidential races, and local and statewide ones as well. We will concentrate on the Presidential race.
- Baseball is full of statistical data and is basically a statistician's dream. The methods used in coaxing out reliable variables that can predict a teams success or the success of a game have for years been rather unreliable. And unlike football, where the success of a team is highly reliant on one uncontrollable variable, the quarterback, baseball has many controllable factors that can be reduced down to create some amount of predictability.
- Politics is like baseball in a sense. It is competitive, there are a number of factors that influence the outcome, such as incumbent (home team) advantage (not in play this year), economy, world affairs, likeability of the candidates, and many others. And there are a few controlled factors that can help predict with some degree of reliability the outcome of the election.

# Who is Nate Silver ???

- Nate is a native of Chicago and is all of 30 years old.




- He has a passion for baseball and invented a technique that enables him to more or less instantly gauge a players potential.
- It is called PECOTA, or Player Empirical Comparison and Optimization Test Algorithm.
- It essentially takes players of this generation, compares them to like players of older generations, finds the ones with close correlations, and sees how those older players did, and uses that to base his predictions of this generation.

# The baseball techniques

- Assumption : You know something about baseball statistics.
- PECOTA is not the only thing Nate Silver uses in determining how a player and his team will do. It is just the one most heavily weighted. How much is his secret.
- QERA, or Quick ERA. This is a ratio that Nate uses that takes out certain, “noise” factors that influence a pitcher’s ERA, but are not his fault. Such as errors committed behind him, the ballpark conditions, and other things.
- $QERA = (2.69 + K\% * (-3.4) + BB\% * 3.88 + GB\% * (-0.66))^2$ .
- K is % strikeouts over 9 innings, BB is walks, GB % is Ground balls to Fly balls. Thus for say, Andy Pettitte, for example, has a 19.6% K rate, a 7.9% BB rate, and a 62.7% GB rate, giving him a QERA of 3.68. Note that QERA is exponential, which is appropriate since run scoring is not linear.

# Some baseball team stats...

East	W	L	PCT	GB	E#	L10	STRK	HOME	ROAD	X W-L	LAST GAME	NEXT
# Tampa Bay	97	65	.599	-	-	6-4	W1	57-24	40-41	91-71	10/13 @ BOS, W 9-1	10/14 8:07P
& Boston	95	67	.586	2.0	E	6-4	W1	56-25	39-42	95-67	10/13 v TB, L 1-9	10/14
New York	89	73	.549	8.0	E	8-2	L1	48-33	41-40	87-75	9/28 @ BOS, L 3-4	-
Toronto	86	76	.531	11.0	E	5-5	W1	47-34	39-42	93-69	9/28 @ BAL, W 10-1	-
Baltimore	68	93	.422	28.5	E	1-9	L1	37-43	31-50	73-88	9/28 v TOR, L 1-10	-
Central	W	L	PCT	GB	E#	L10	STRK	HOME	ROAD	X W-L	LAST GAME	NEXT
# Chicago	89	74	.546	-	-	4-6	W3	54-28	35-46	89-74	10/6 v TB, L 2-6	-
Minnesota	88	75	.540	1.0	E	5-5	L1	53-28	35-47	89-74	9/30 @ CWS, L 0-1	-
Cleveland	81	81	.500	7.5	E	6-4	L1	45-36	36-45	85-77	9/28 @ CWS, L 1-5	-
Kansas City	75	87	.463	13.5	E	7-3	L1	38-43	37-44	72-90	9/28 @ MIN, L 0-6	-
Detroit	74	88	.457	14.5	E	3-7	L2	40-41	34-47	78-84	9/29 @ CWS, L 2-8	-
West	W	L	PCT	GB	E#	L10	STRK	HOME	ROAD	X W-L	LAST GAME	NEXT
# Los Angeles	100	62	.617	-	-	7-3	W1	50-31	50-31	88-74	10/6 @ BOS, L 2-3	-
Texas	79	83	.488	21.0	E	4-6	L1	40-41	39-42	76-86	9/28 @ LAA, L 0-7	-
Oakland	75	86	.466	24.5	E	4-6	L5	43-38	32-48	76-85	9/28 @ SEA, L 3-4	-
Seattle	61	101	.377	39.0	E	4-6	W3	35-46	26-55	67-95	9/28 v OAK, W 4-3	-
National League 												
East	W	L	PCT	GB	E#	L10	STRK	HOME	ROAD	X W-L	LAST GAME	NEXT
# Philadelphia	88	76	.538	7.0	E	7-3	W3	48-33	41-40	88-76	10/13 @ LAD, W 7-5	10/14

What statistics stick out??

# Does home field advantage translate into the postseason??

- **Overall Results**

Without looking at the individual games within each series the results for home field advantage for each overall series are:

- Divisional Series: Home - 55.5%, Road - 44.5%
- Championship Series: Home - 52.5%, Road - 47.5%
- World Series: Home - 55.9%, Road - 44.1%

• Thus it tends to prove that the home team has the advantage in post season play.

# But how did they get home field advantage??

## All Star Game Winners

1995	National, 3-2	Texas
1996	National, 6-0	Philadelphia
1997	American, 3-1	Cleveland
1998	American, 13-8	Colorado
1999	American, 4-1	Boston
2000	American, 6-3	Atlanta
2001	American, 4-1	Seattle
2002	7-7 tie, 11 innings	Milwaukee
2003	American, 7-6	Chicago
2004	American, 9-4	Houston
2005	American, 7-5	Detroit
2006	American, 3-2	Pittsburgh
2007	American, 5-4	San Francisco

## World Series Winners

	National	American	Home?
<b>1996</b>	Atlanta Braves	🍀 New York Yankees	N
<b>1997</b>	🍀 Florida Marlins	Cleveland Indians	N
<b>1998</b>	San Diego Padres	🍀 New York Yankees	Y
<b>1999</b>	Atlanta Braves	🍀 New York Yankees	Y
<b>2000</b>	New York Mets	🍀 New York Yankees	Y
<b>2001</b>	🍀 Arizona Diamondbacks	New York Yankees	Y
<b>2002</b>	San Francisco Giants	🍀 Anaheim Angels	Y
<b>2003</b>	🍀 Florida Marlins	New York Yankees	N
<b>2004</b>	St. Louis Cardinals	🍀 Boston Red Sox	Y
<b>2005</b>	<a href="#">Houston Astros</a>	🍀 Chicago White Sox	Y
<b>2006</b>	🍀 St. Louis Cardinals	Detroit Tigers	N
<b>2007</b>	Colorado Rockies	🍀 Boston Red Sox	Y

🍀 = winner


Clearly, without statistics, home field advantage gained from the All-Star game gives the American League the advantage. They have won 8 out of the last 12 series, while having HFA 10 times.

# So, how good are Nate's methods?

Prediction for 2006 season for the AL East on March 26<sup>th</sup>, 2006

AL East	W	L	RS	RA	BatDelta	PitDelta
Yankees	94	68	910	777	+117	+2
Red Sox	93	69	913	783	+82	+28
Blue Jays	79	83	788	809	-36	+23
Orioles	77	85	777	817	-11	-21
Devil Rays	69	93	744	869	-24	-81

## Final standings 2006

American League 

East	W	L	PCT	GB	E#	L10	STRK	HOME	ROAD	X W-L	LAST GAME	NEXT GAME
New York	97	65	.599	-	-	5-5	L2	50-31	47-34	95-67	10/7 @ DET, L 3-8	-
Toronto	87	75	.537	10.0	E	7-3	W2	50-31	37-44	86-76	10/1 @ NYY, W 7-5	-
Boston	86	76	.531	11.0	E	5-5	W1	48-33	38-43	81-81	10/1 v BAL, W 9-0	-
Baltimore	70	92	.432	27.0	E	4-6	L1	40-41	30-51	69-93	10/1 @ BOS, L 0-9	-
Tampa Bay	61	101	.377	36.0	E	3-7	L4	41-40	20-61	65-97	10/1 @ CLE, L 3-6	-
Central	W	L	PCT	GB	E#	L10	STRK	HOME	ROAD	X W-L	LAST GAME	NEXT GAME
Minnesota	96	66	.593	-	-	6-4	W1	54-27	42-39	93-69	10/6 @ OAK, L 3-8	-

The year after the Chicago White Sox won the World Series in 2005, he predicted the next year they would finish with a record of 72 and 90.

In 2006, they finished with **72 and 90.**

# Election exit polls

In 2004, these were very bad, to the point where CNN quit using them. All tend to favor Kerry to some degree.

	Exit Poll	Final Result
Alabama	Gore by 1.2	Bush by 14.9
Arizona	Gore by 3.6	Bush by 6.3
Colorado	Gore by 3.1	Bush by 8.4
Georgia	Bush by 4.7	Bush by 11.7
North Carolina	Gore by 3.0	Bush by 12.8

The final tally was Bush 51.9% and Kerry 48.1%, with Bush ahead by about 3 million votes. Clearly, something is up. There are two trains of thought on the polls:

- 1) The polls are correct and the actual vote count was “tampered” electronically.
- 2) The vote count is correct and the polls are off.

# The vote count is wrong...

X	Mean	Standard Error	p-value	Odds (1:X)	Method Used			
					Simon_Baiman	Rounding	CI	2-tail
0.508	0.481	0.005686	0.0000010259	974772	X			
0.505	0.481	0.005686	0.0000121746	82138		X		
0.514	0.481	0.005686	0.0000000033	307416049		X		
0.505	0.481	0.005102	0.0000012768	783198		X	X	
0.514	0.481	0.007653	0.0000080938	123551		X	X	
0.505	0.481	0.007653	0.0008563397	1168		X	X	
0.514	0.481	0.005102	0.0000000000	20049235521		X	X	
0.508	0.481	0.005686	0.0000020518	487386	X			X
0.505	0.481	0.007653	0.0017126793	584		X	X	X
0.514	0.481	0.005102	0.0000000001	10024617760		X	X	X
0.505	0.481	0.005102	0.0000025536	391599		X	X	X
0.514	0.481	0.007653	0.0000161877	61775		X	X	X

This chart indicates the ranges of what the polls arrived at versus the mean, which is the actual vote count. 0.481 means Kerry's 48.1% of the vote. The p-values indicate that the polls results are different and therefore "error-prone" compared to the actual result. They are all below .05, the usual "alpha" risk.

# The polls are wrong...

- It is fairly well established by both Democrat and Republican pollsters that since polling began in 1982, there has been a slight Democrat leaning in the polls.

National Exit Poll Estimates, 1992  
from Mitofsky and Edelman (1995, p 91)

Release	Clinton	Bush	Perot	Clinton-Bush	Estimate-Actual
8:10 pm	46.2%	35.9%	18.0%	10.3%	4.4%
9:10pm	45.6%	35.5%	18.9%	10.1%	4.2%
10:10pm	44.8%	36.1%	19.1%	8.7%	2.8%
11:10pm	44.3%	36.5%	19.2%	7.8%	1.9%
12:10pm (sic)	44.2%	36.7%	19.1%	7.5%	1.6%
End of Night Vote Count	43.4%	37.5%	19.1%	5.9%	--

This pretty much stayed the same in 1996 and on. But because the result was not close, Reagan in the 80's and Clinton in the 90's, issues weren't raised with the polling.

# Who is right ??

- In the beginning of the year, a person appeared on a political blog named “Poblano”. He started to publish predictions about what the upcoming results from the May North Carolina Democratic primary would be.
- The predictions were based on demographics and the results from similar demographic districts that had voted in South Carolina the week before.
- All the polls were predicting a dead heat between Obama and Clinton, but “Poblano” predicted a 17% victory for Obama.
- Obama won by 14% and “Poblano” was a red, hot, chili pepper.
- “Poblano” of course, was Nate Silver, who revealed himself as a dedicated baseball statistician who feels he can just as readily predict political elections.

# Who is right ??

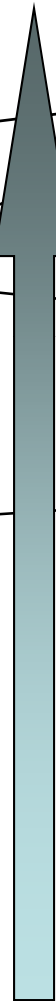
- Since that time, he has made numerous other predictions of the primary results with only 1 major miss, in which he predicted a slight Obama victory, but Clinton won by 6.
- Thus it is beginning to appear that indeed the polls have been wrong.
- The pollsters swear by their method but it has to be called into question. Many question the use of a small error of only 1.1%, when there are many variables associated with polling. For instance, it was found that on average 50 people were polled per precinct. Though technically you are in a “normal” situation, the error rate has been estimated at closer to  $\pm 5\%$ . If done on an overall voting population, that would mean Kerry’s CI was actually 50.1(the sample mean)  $\pm 5\%$ , or between 45 and 55%.

# What is Nate Silver doing different ??

- It turns out that he is using a combination of the current polls being published nationally, putting more weight on the more accurate ones (based on history), and combining in the established demographics of the districts.
- These established voting districts that are compared historically and with similar voting districts, account for 72% of his predictions.
- Let's look at some of the analysis...

# The statistical backbone

Agency	Weight	Date	Size	DEM	GOP	Margin
FL Rasmussen	1.71	10/12	1000	51	46	Obama +5
FL Strategic Vision	1.37	10/7	1200	52	44	Obama +8
FL PPP	0.98	9/28	941	49	46	Obama +3
FL Quinnipiac	0.94	9/28	836	51	43	Obama +8
FL Mason-Dixon	0.89	10/5	625	48	48	Obama +2
FL SurveyUSA	0.86	9/28	599	47	48	McCain +1
FL CNN	0.78	9/29	770	51	43	Obama +8
FL Rasmussen	0.75	10/8	700	50	47	Obama +3
FL Rasmussen	0.74	10/5	1000	52	45	Obama +7
FL Suffolk	0.66	9/29	600	46	42	Obama +4
FL InsiderAdvantage	0.66	9/30	532	49	46	Obama +3
FL SurveyUSA	0.65	9/17	707	45	51	McCain +6
FL Quinnipiac	0.61	9/24	1161	49	43	Obama +6
FL Miami Herald	0.59	9/16	800	45	47	McCain +2
FL ARG	0.57	9/24	800	47	46	Obama +1
FL Research 2000	0.57	9/17	600	45	46	McCain +1
FL Strategic Vision	0.44	9/22	1200	45	48	McCain +3
FL Mason-Dixon	0.37	9/17	625	47	45	Obama +2
FL National Journal	0.34	9/13	402	44	44	Tie
FL PPP	0.30	9/7	968	45	50	McCain +5
FL CNN	0.29	9/15	907	48	44	Obama +4
FL Rasmussen	0.27	9/24	700	47	48	McCain +1
FL Rasmussen	0.25	9/28	500	47	47	Tie
FL InsiderAdvantage	0.24	9/10	511	42	50	McCain +8
FL ARG	0.24	9/16	600	46	46	Tie
FL Quinnipiac	0.21	9/7	1032	43	50	McCain +7
FL SurveyUSA	0.18	8/2	679	44	50	McCain +6
FL Zogby Interactive	0.17	9/11	995	41.8	52.1	McCain +10.3
FL Mason-Dixon	0.15	8/26	625	45	44	Obama +1
FL Rasmussen	0.15	9/21	500	46	51	McCain +5
FL Rasmussen	0.11	9/14	500	44	49	McCain +5
FL Strategic Vision	0.11	8/23	1200	42	49	McCain +7
FL WR Logistics	0.10	7/8	629	47.2	44.5	Obama +2.7
FL Quinnipiac	0.10	8/21	1069	43	47	McCain +4
FL Rasmussen	0.09	9/7	500	48	48	Tie
FL ARG	0.07	8/19	600	46	47	McCain +1
FL InsiderAdvantage	0.07	8/11	419	44.2	47.8	McCain +3.6
FL PPP	0.07	8/1	807	44	47	McCain +3
FL Rasmussen	0.07	8/18	700	46	48	McCain +2
FL Polling Average				48.1	46.0	Obama +2.1
FL Trend-Adjusted				50.1	43.2	Obama +6.9
FL 538 Regression	0.71			49.5	43.3	Obama +6.2
FL Snapshot				50.1	43.2	Obama +6.8
FL Projection		MOE ±4.9		51.9	46.6	Obama +5.2



Trends – Look at same agency (CNN) at different times

Techniques – Why is the smaller sized Rasmussen poll higher ranked?

Techniques – Why is the older Rasmussen poll higher ranked?

What overall trend can be seen?

# So for every state a projection..

Agency	Weight	Date	Size	DEM	GOP	Margin
FL Rasmussen	1.71	10/12	1000	51	46	Obama +5
FL Strategic Vision	1.37	10/7	1200	52	44	Obama +8
FL PPP	0.98	9/28	941	49	46	Obama +3
FL Quinnipiac	0.94	9/28	836	51	43	Obama +8
FL Mason-Dixon	0.89	10/5	625	48	46	Obama +2
FL SurveyUSA	0.86	9/28	599	47	48	McCain +1
FL CNN	0.78	9/29	770	51	43	Obama +8
FL Rasmussen	0.75	10/8	700	50	47	Obama +3
FL Rasmussen	0.74	10/5	1000	52	45	Obama +7
FL Suffolk	0.66	9/29	600	46	42	Obama +4
FL InsiderAdvantage	0.66	9/30	532	49	46	Obama +3
FL SurveyUSA	0.65	9/17	707	45	51	McCain +6
FL Quinnipiac	0.61	9/24	1161	49	43	Obama +6
FL Miami Herald	0.59	9/16	800	45	47	McCain +2
FL ARG	0.57	9/24	600	47	46	Obama +1
FL Research 2000	0.57	9/17	600	45	46	McCain +1
FL Strategic Vision	0.41	9/22	1200	45	48	McCain +3
FL Mason-Dixon	0.37	9/17	625	47	45	Obama +2
FL National Journal	0.34	9/13	402	44	44	Tie
FL PPP	0.30	9/7	968	45	50	McCain +5
FL CNN	0.29	9/15	907	48	44	Obama +4
FL Rasmussen	0.27	9/24	700	47	48	McCain +1
FL Rasmussen	0.25	9/28	500	47	47	Tie
FL InsiderAdvantage	0.24	9/10	511	42	50	McCain +8
FL ARG	0.24	9/16	600	46	46	Tie
FL Quinnipiac	0.21	9/7	1032	43	50	McCain +7
FL SurveyUSA	0.18	8/2	679	44	50	McCain +6
FL Zogby Interactive	0.17	9/11	995	41.8	52.1	McCain +10.3
FL Mason-Dixon	0.15	8/26	625	45	44	Obama +1
FL Rasmussen	0.15	9/21	500	46	51	McCain +5
FL Rasmussen	0.11	9/14	500	44	49	McCain +5
FL Strategic Vision	0.11	8/23	1200	42	49	McCain +7
FL WR Logistics	0.10	7/8	629	47.2	44.5	Obama +2.7
FL Quinnipiac	0.10	8/21	1069	43	47	McCain +4
FL Rasmussen	0.09	9/7	500	48	48	Tie
FL ARG	0.07	8/19	600	46	47	McCain +1
FL InsiderAdvantage	0.07	8/11	419	44.2	47.8	McCain +3.6
FL PPP	0.07	8/1	807	44	47	McCain +3
FL Rasmussen	0.07	8/18	700	46	48	McCain +2
FL Polling Average				48.1	46.0	Obama +2.1
FL Trend-Adjusted				50.1	43.2	Obama +6.9
FL 538 Regression	0.71			49.5	43.3	Obama +6.2
FL Snapshot				50.1	43.2	Obama +6.8
FL Projection				51.9	46.6	Obama +5.2
					MOE ±4.9	

Demographics (weight 71%) That never changes. Don't have any data on how it was obtained. But it is historical and doesn't vary much.

FL Polling Average		48.1	46.0	Obama +2.1
FL Trend-Adjusted		50.1	43.2	Obama +6.9
FL 538 Regression	0.71	49.5	43.3	Obama +6.2
FL Snapshot		50.1	43.2	Obama +6.8
FL Projection		51.9	46.6	Obama +5.2
			MOE ±4.9	

What does the MOE indicate?

# All states together...

UPDATED 10/7/2008 12:15

**New England**

State	EV	Obama	McCain
MA	12	98%	2%
CT	7	97%	3%
ME	4	94%	6%
NH	4	82%	18%
RI	4	98%	2%
VT	3	100%	0%

**Acela**

State	EV	Obama	McCain
NY	31	99%	1%
NJ	15	92%	8%
MD	10	99%	1%
DC	3	100%	0%
DE	3	99%	1%

**South Coast**

State	EV	Obama	McCain
FL	27	73%	27%
GA	15	16%	84%
NC	15	60%	40%
VA	13	84%	16%
SC	8	7%	93%

**Gulf Coast**

State	EV	Obama	McCain
TX	34	11%	89%
AL	9	2%	98%
LA	9	11%	89%
MS	6	9%	91%

**Highlands**

State	EV	Obama	McCain
MO	11	56%	44%
TN	11	5%	95%
KY	8	7%	93%
OK	7	1%	99%
AR	6	13%	87%
WV	5	33%	67%

This is a synopsis of all the statewide polls as of October 7<sup>th</sup>.

The percentages under each candidate give the statistical probability of that candidate winning the state's electoral votes.

The red means favors McCain, the blue, Obama.

Colors in-between are leaning one way or another, but can't really be called.

State	EV	Obama	McCain
PA	21	89%	11%
OH	20	71%	29%
MI	17	91%	9%
IN	11	54%	46%

**North Central**

State	EV	Obama	McCain
IL	21	99%	1%
WI	10	90%	10%
MN	10	91%	9%
IA	7	96%	4%

**Prairie**

State	EV	Obama	McCain
KS	6	6%	94%
NE	5	4%	96%
ND	3	21%	79%
SD	3	12%	88%

**Southwest**

State	EV	Obama	McCain
AZ	10	9%	91%
CO	9	85%	15%
NM	5	89%	11%
NV	5	72%	28%

**Big Sky**

State	EV	Obama	McCain
UT	5	0%	100%
ID	4	2%	98%
MT	3	24%	76%
WY	3	2%	98%
AK	3	5%	95%

**Pacific**

State	EV	Obama	McCain
CA	55	98%	2%
WA	11	96%	4%
OR	7	98%	2%
HI	4	100%	0%

So who will win??

# Baseball

*by Clay Davenport*

---

One million trials

A Monte Carlo simulation uses random numbers to simulate the playoff series. Each game of the series is given a separate outcome probability, based on the home team, the team's performance during the regular season, and the expected starting pitcher matchup. Since a different sequence of random numbers is used every day, the results may change slightly from one day to the next, even if there was no game and no change in the pitching matchups.

DS = Division series  
CS = Championship Series  
WS = World Series

```
----- October 14 -----
      Win DS   Win CS   Win WS   Yesterday
LAA   0.0000   0.0000   0.0000
TB    100.0000  87.5770  47.2104  Won 13-4, lead series 3-1
CWS   0.0000   0.0000   0.0000
BOS   100.0000  12.4230   7.9769   Lost 13-4, trail series 3-1
CHC   0.0000   0.0000   0.0000
PHI   100.0000  85.0143  39.8423   Lead series 3-1
LAD   100.0000  14.9857   4.9704   Trail series 3-1
MIL   0.0000   0.0000   0.0000
```

Source: Baseball Prospectus

# Politics

