The 2000 Presidential Election A Statistical Postmortem

Last year's presidential election was like none other in our nation's history. Although comparisons have been made to the controversial 1876 election between Rutherford B. Hayes and Samuel Tilden, that election did not hinge upon the question of recounts and was ultimately decided by an electoral commission appointed by the United States Congress, not (in effect) by the Supreme Court.

This remarkable election also provided many interesting statistical lessons. In this column I will touch upon two aspects of the presidential campaign that received particular attention—the accuracy of the polls in predicting the outcome, and the now infamous "butterfly ballot."

Polls, Polls and More Polls

Never have there been so many polls involved in a presidential election as in 2000. Not surprisingly these polls reported figures that varied significantly over time, as well as from poll to poll at any particular time. Below you can see how the final nationwide polls looked just before the November 7 election. Each poll's sample size N represents the number of "likely voters" interviewed. Each poll's percentages add up to less than 100% because of the presence of undecided voters.

When these polls were reported, as throughout most of the campaign, the race was declared "too close to call." Even though Governor Bush was ahead in twelve of the fourteen polls this seems like a reasonable assessment given that neither candidate had over 50% support and voters often change their minds at the last minute. The phrase "too close to call" is, however, traditionally based not on those considerations but on the theory of statistical variation and sampling error.

Specifically, a simple mathematical model for an opinion poll regards each response as a *Bernoulli* (0–1) *random variable*, with 1 representing a vote for Candidate A and 0 representing a vote for any other candidate. With N respondents in the poll, the total number of individuals favoring Candidate A is then a binomial random variable with parameters N and p, where p is the percentage of *all* likely voters who support Candidate A. From basic statistical theory, the percentage $\hat{p}$ of respondents to the poll who prefer Candidate A is a *binomial proportion* that from poll to poll will vary around p according to a bell-shaped, or *normal*, distribution having standard deviation $\sigma = \sqrt{p(1-p)/N}$.

You may know that when values are drawn from a normal distribution, generally about 95% of them will fall within two standard deviations of the mean. For opinion polls the mean is the true population percentage p,

	Percentage favoring:				
Poll:	Bush	Gore	Nader	Buchanan	N
CBS	44	45	4	1	1091
CNN/USA Today/Gallup	47	45	4	1	2350
ABC	48	45	3	1	1801
Marist College	49	44	2	1	623
NBC/Wall Street Journal	47	44	3	2	1026
Newsweek	45	43	5	0	808
Pew Research Center	45	43	4	0	1301
Fox/Opinion Dynamics	43	43	3	1	1000
ICR	46	44	7	2	1103
Christian Science Monitor	48	42	3	1	1292
Hotline Bullseye	47	40	4	1	1000
Reuters/MSNBC	46	44	5	1	1200
CBS/New York Times	47	42	5	1	1158
Voter.com	46	41	4	0	1000

MARK SCHILLING is Professor of Mathematics at California State University, Northridge.

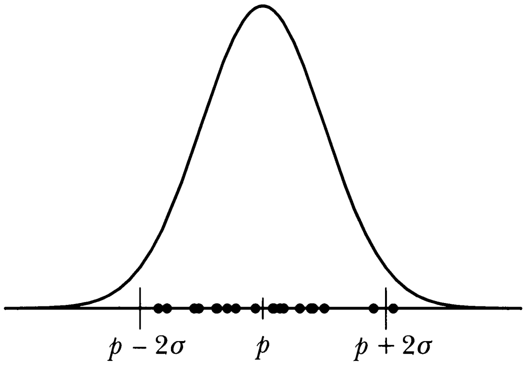


Figure 1

and so approximately 95% of all polls should yield a result within 2σ of p (see Figure 1).

Note that in order to compute 2σ we need to know the value of p—but if we know the value of p there is no need to conduct a poll to estimate it! As p is in fact *unknown* in any polling situation, normally the value $\hat{p}$ obtained from the poll is used in its place, yielding the expression

$$2\hat{\sigma} = 2\sqrt{\hat{p}\left(1-\hat{p}\right)/N}.$$

This quantity is known as the *margin of error* of the poll, and is commonly reported in the media when the poll results are announced. If you try some of the values of $\hat{p}$ and N in the list of polls above, letting $\hat{p}$ be the percentage favoring Bush or Gore, you will find that the margins of error vary between two and four percent, with most being around three percent.

If all of the presidential polls had a margin of error of 3% and were conducted using proper survey methodology, we would therefore expect that in around 95% of these polls the percentage for a given candidate would lie within ±3% of that candidate's true level of support. These polls would therefore fall within a range of about 6% of each other. Looking at the table above we see that is indeed the case (Bush: 43%–49%, Gore: 40%–45%). (It should be noted, however, that not all of the polls *did* use well-established survey methods. Voter.com, for example, was an on-line poll; its results diverged consistently from the other polls throughout the campaign, suggesting that on-line polling is at present a biased and unreliable polling procedure.)

Many individuals—including some prominent media figures—have had a misconception about the margin of error, believing it relates to the *difference* between the two leading candidate's percentages as opposed to a single candidate's level of support. In fact, the support levels for Bush and Gore had a very strong negative correlation—when one of them fell in the polls the other generally went up by about the same amount. As a result, the margin of error for the difference of the correlated proportions $\hat{p}_{\text{Bush}} - \hat{p}_{\text{Gore}}$ is roughly *double* the margin of error indicated above, being therefore around *six* percent for most polls.

This, then, gives us the true technical meaning of the phrase "too close to call"—it means that the difference in the percentages for two candidates is less than the margin of error of that difference. It is easy to check that all but two of the polls shown above are "too close to call"— the results given in the poll could reasonably have occurred regardless of which of the two major candidates was ahead among all likely voters. Ironically, the two polls that could have made a call (*Christian Science Monitor* and *Hotline Bullseye*) got it wrong, as Gore actually finished with a slightly higher popular vote total than Bush. This may have been due to a last minute move by many undecided voters and some Nader supporters to Gore.

The Case of the Butterfly Outlier

The "butterfly" ballot used in Palm Beach County, Florida will go down in history as possibly having helped determine the result of the 2000 Presidential Election. You will recall that this ballot had the names of the various candidates listed on both sides of the page, with punch holes down the middle in an arrangement that may have confused some voters. Bush's name was first on the ballot and the hole the voter needed to punch to vote for Bush was the first hole on the ballot. Gore's name was second on the ballot, but the voter needed to punch the third hole. The second hole represented the Buchanan ticket. The juxtaposition of the punch hole for Buchanan with the Gore/Lieberman names may have led to many accidental votes for Buchanan that were intended for Gore. The key questions are (i) did this actually occur, and if so, (ii) how many such votes were cast in error?

Statistical analysis provides some rather compelling evidence. Consider Figure 2, which plots the total votes cast for Buchanan against the total votes for all candidates for each of the sixty-seven counties in Florida. Palm Beach county appears to be an extreme outlier with its 3407 votes for Buchanan, whereas based on the results from other counties we might have expected only 1000 or so.

Before making a rush to judgment, however, note that only the fifteen or so larger counties (>100,000 votes) show clearly in the plot; any discrepancies among the small to medium counties may be hidden by the considerable clustering that occurs near the origin. To address this problem it is advisable to transform the variables plotted to better spread out the displayed points. Figure 3 shows how the plot looks when common logs are taken of each variable. Palm Beach is still an outlier, but the magnitude of the effect seems very much smaller than before.

Continued on p. 34.

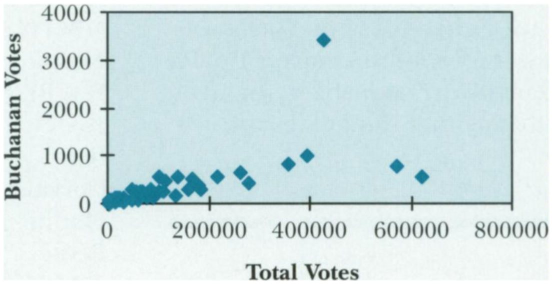


Figure 2

or

$$4 = 3\sin^2 A + 1 - \cos(B + C)\cos(B - C)$$

or

$$3\cos^2 A = \cos A \cos(B - C).$$

Hence either $A = \pi/2$ or $3\cos A = \cos(B - C)$. Since the latter equation is equivalent to $2R\cos A = 2R\cos B \cos C$, $AH = HD$ or H bisects AD.

The above steps can be reversed and thus the converse holds and the claim follows.

Also solved by Xiuoxuan Jin (student), and S. Smith.

S-52. (Quickie) Condition for a Parallelogram

Consider the figure imbedded in the complex plane with A as the origin and $B = z$, $B' = w$. Then $D = iz$ and $D' = -iw$ so that $DD' = -i(z + w)$. Hence $E = z + w$ which proves that $ABEB'$ is a parallelogram.

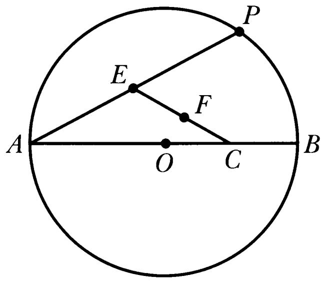
Comment. This is a converse problem to one in the Jan. 2000 Wisconsin Mathematics, Science & Engineering Talent search where one was given $ABEB'$ was a parallelogram and one was to prove DD' was equal and perpendicular to AE. This also follows from the above proof.

S-53†. (Quickie) Variable Segment of Constant Length

As in the preceding problem, we use complex numbers. In the figure below, the complex representation for O, A, B, and P, are given by 0, $-2a$, $2a$, and $2ae^{i\theta}$, respectively. Then $C = a$, $E = a(e^{i\theta} - 1)$ so that $F = ae^{i\theta}/2$. Finally,

$$PF = \frac{3a|e^{i\theta}|}{2} = \frac{3a}{2}.$$

Also P, F, and O are always collinear.



Late solutions. S-45, Mary Megrant. Problem 138, David Hill (undergraduate).

Continued from p. 29.

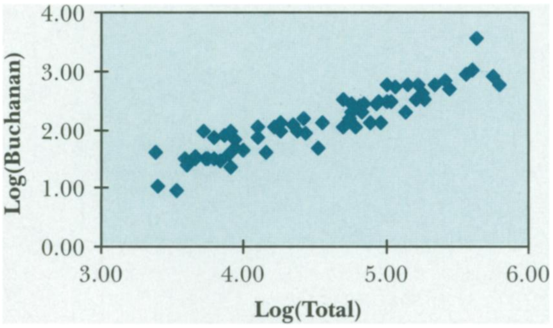


Figure 3

While the plots are quite suggestive, a more thorough analysis of the issue of miscast votes requires the use of *multiple regression analysis*. The basic idea in the present context is to find a linear equation that predicts the percentage of votes that Buchanan would receive in a county (and from that, the number of votes) from other relevant information known as *explanatory variables*. Obvious explanatory variables here are the percentages of votes in each county for the other presidential candidates. Another possibly significant factor is the percentage of votes that Buchanan received in the 1996 Florida Republican primary.

Using data available from Professor Christopher Carroll of John Hopkins (`http://www.econ.jhu.edu/people/ccarroll/carroll.html`), I performed a regression analysis using data from all Florida counties except Palm Beach. For reasons suggested by the two plots above, taking logs of some variables is preferable to using the original variables themselves. Here is the predictive equation I obtained:

$$\begin{aligned}\text{Log(Buchanan\%)} = {} & 9.484 - 10.724\ \text{Bush\%}\\ & - 11.371\ \text{Gore\%}\\ & - 0.117\ \text{Log(Total County Votes)}\\ & + 0.957\ \text{Log(Buchanan'96\%)}.\end{aligned}$$

The signs of the regression coefficients indicate that Buchanan's percentage of the vote was higher (i) where Bush and Gore did poorly, (ii) in smaller (more rural) counties, and (iii) where he fared well in the 1996 primary. All of these indications agree with common sense.

We can now apply this equation to predict Buchanan's performance in Palm Beach County based on his performance in the other sixty-six counties. Plugging in the values of the explanatory variables for Palm Beach, we obtain Log(Buchanan%) = –2.864, which indicates that Buchanan should have received approximately 1.4% of the vote in Palm Beach County, or 590 votes, compared to the nearly 8% (3,407 votes) that he did receive there. The indication is that as many as five out of every six Buchanan votes in Palm Beach County were cast in error. Had these votes gone to Gore instead of Buchanan, Gore would have been ahead in the original popular vote count and we might have a different president in office today.

This should not be considered a definitive analysis. Voting data is a complicated and highly interesting subject for statistical examination. As I write this article, a complete review of every single ballot cast in the 2000 Florida election is underway. There will be many additional statistical investigations to come, and I encourage you to keep up with them for both their political *and* statistical interest. ■