

Analysis of Voting Irregularities in Palm Beach County in the 2000 Presidential Election

M.C. Spruill

October 11, 2004

At one time soon after the U.S. presidential election of 2000, upon consulting the web site whose URL is

<http://madison.hss.cmu.edu/#bush>

one would have found several engaging analyses of the voting patterns in the presidential election of 2000 in the state of Florida. Alas, the ephemeral nature of the web has removed the opportunity of consulting this site, but perhaps some of the sites below still exist. Most of the analyses (see, for example, <http://www.econ.jhu.edu/people/ccarroll/carroll.html>) concluded that in Palm Beach County, Buchanan received far more votes than would be expected based upon an analysis of voting patterns in the other counties in Florida. Other analyses (see, for example, <http://www.princeton.edu/shimer/election.html>) suggested that the voting patterns are not so irregular as portrayed. Some analyses employed raw vote totals, some log votes, and others proportions (vote shares). The ones employing raw votes tended to show Palm Beach County as an outlier (and hence irregularities occurred) while those that employ vote share tended to show that it is not an outlier. There is clearly some question about the proper normalization of the votes received by the candidates for the purposes of analyses and it is clear that this normalization must involve the sizes of the counties.

We point out here what none of the other analysts observed and that is that the correct normalization is \sqrt{n} ; the quantities entering into a regression should be $vote_i/\sqrt{n}$, where n is the number of votes cast in the county for all candidates and $vote_i$ is the number of votes cast in that county for candidate i . Thus correct analyses involve either \sqrt{n} times vote share, number of votes divided by \sqrt{n} , or weight the regression in some way. Neither the analysis of raw vote totals by conventional regression techniques, nor the analysis of proportions by the same are correct; both are flawed. A correct regression analysis is based upon the variables $vote_i/\sqrt{n}$. Furthermore, it is shown that based upon these variables, there is unequivocal evidence that there were irregularities, of undetermined nature or origin, in Palm Beach County.

Denote the vector of numbers of votes for the k candidates in some particular county by

$$\mathbf{Y}' = (Y_1, \dots, Y_k).$$

Then this is (approximately) a multinomial random variable $Mn_k(n, \mathbf{p})$ where \mathbf{p} is a probability vector representing the proportion of voters who will vote for candidate i in the i^{th} coordinate and n is the total number of persons in the county who vote. By standard techniques of probability one has that for large n ,

$$\sqrt{n}(\mathbf{Y}/n - \mathbf{p})$$

is distributed approximately as a k -dimensional normal random vector with mean $\mathbf{0}$ and covariance Σ , where Σ has u, v^{th} entry $p_u q_u$ if $u = v$ and $-p_u p_v$ if $u \neq v$. It follows that for any two coordinates j (Buchanan) and i (Bush) the conditional distribution of $\sqrt{n}(Y_{j,n}/n - p_j)$ given $\sqrt{n}(Y_{i,n}/n - p_i) = u$ is normal with mean $-up_j/q_i$ and variance $(p_j/q_i)[q_i q_j - p_i p_j]$. The usual linear regression model will apply, therefore, to the individual data points consisting of normalized votes for Bush and Buchanan for each county, if one assumes that they are homogeneous with respect to the proportions \mathbf{p} . This is a key feature; if one presumes the counties to be homogeneous with respect to the \mathbf{p} 's then the variance is constant over all counties. Furthermore, denoting by $V_r = Y_r/\sqrt{n_r}$ the normalized vote for candidate Buchanan in county r , we have

$$V_r = \sqrt{n_r} p_{Buchanan}/q_{Bush} - (p_{Buchanan}/q_{Bush})U_r + e_r,$$

where U_r is the normalized vote for Bush in county r and the e_r are iid

$$N(0, p_{Buchanan}[q_{Buchanan}q_{Bush} - p_{Buchanan}p_{Bush}]/q_{Bush}).$$

Plainly, if we ignore the votes for the other candidates, the ordinary linear regression model for the variables V_r on the variable $\sqrt{n_r} - U_r$ is appropriate here. Also, it is clear that none of the variables $Y_{Buchanan}$, $Y_{Buchanan}/n$, $\log Y_{Buchanan}$, or $\log Y_{Buchanan}/n$ share this property of being appropriate for an ordinary regression analysis; in each case the variances depend upon n , the size of the county, so that only a weighted analysis would be appropriate.

We undertook a linear regression analysis of the data linked to the URL ([http:// madison. hss. cmu. edu/ #bush](http://madison.hss.cmu.edu/#bush)) utilizing the variables $\sqrt{n} - U$ and V . Omitting Palm Beach County one obtains the least squares equation

$$v = 0.004652(\sqrt{n} - u)$$

The mle of the number of votes for Buchanan in Palm Beach county is 1299.95 votes. Furthermore, a 95% confidence interval is (1299.4, 1300.5) and we see that the actual number of votes for Buchanan of 3407 is far outside this interval. One concludes that there were irregularities; if all of the counties had the same proportions of voters for each candidate then the results we observed would be extremely improbable. Other links at ([http:// madison. hss. cmu. edu/ #bush](http://madison.hss.cmu.edu/#bush)) explain in greater detail caveats pertinent to these ruminations. The maximum likelihood estimate of the overvote to Buchanan there is $3407 - 1300 = 2107$. Other links at ([http:// madison. hss. cmu. edu/ #bush](http://madison.hss.cmu.edu/#bush)) analyze the source of the overvote.

The voting patterns in Florida show irregularities in Palm Beach County. An estimate of the number of votes, based upon the correct normalization of the voting figures assuming an approximate multinomial model and homogeneity across counties, Buchanan received in excess of what he should have is 2,107 votes.