A NonParametric Analysis of UnderVotes in the Palm Beach Presidential Vote: Implications of a Recount

Bruce E. Hansen
Stockwell Professor of Economics
Department of Economics
University of Wisconsin
http://www.ssc.wisc.edu/~bhansen
bhansen@ssc.wisc.edu

November 17, 2000
Updated: November 19, 2000

1 Summary

A precinct-level nonparametric analysis of the November 2000 Presidential vote in Palm Beach County, Florida, shows that undervoting – not marking the ballot sufficiently for the counting machines to read – cost Gore approximately 3104 more votes than Bush.

Our analysis also predicts that a hand recount in Palm Beach will net approximately 294 votes for Gore, but the actual gain can range from 184 to 394. This is an estimate which only takes into account the so-called "hanging chads" and not the so-called "dimples" in the undervotes.

The estimates are based on a Binomial model which allows voting probabilities to vary by precinct. This is a nonparametric problem because there are 494 precincts in the sample, implying 494 distinct probabilities. This methodology may be of general interest because it gives precise yet accurate confidence intervals using only precinct-level vote data.

In brief,

- Due to undervoting, we estimate Gore lost 5982 votes with a 95% confidence interval of [5903, 6074]. We estimate Bush lost 2885 votes with a confidence interval of [2801, 2969].
- The estimated difference in lost votes (to Gore relative to Bush) is 3104.
- If there is a hand recount, we estimate that Gore will pick up 568 votes, Bush 274 with a net increase of 294 for Gore.
- Our confidence interval for the net increase is [184, 394]. This is the plausible range of the likely net increase to Gore's vote count over Bush's due to a recount in Palm Beach county.

Our estimates are conservative because they are based on the yield rate from the four-precinct hand recount, which only counted ballots as valid if they had been punched through. The Wednesday court ruling called for election officials to consider "dimpled" ballots as valid votes. This will increase the yield rate, although the extent is unclear.

The data is available on my webpage.

2 Methodology

2.1 Confidence Intervals for Lost Votes

Let i index a voting precinct. Let T_i be the number of counted votes in the precinct, and let V_i be the number of votes for a particular candidate. Let P_i be the probability that a vote in the precinct is cast for that candidate. The estimate of P_i is

$$\hat{P}_i = V_i / T_i.$$

The distribution of V_i is $B(T_i, P_i)$, the Binomial distribution with parameters T_i and P_i. Thus $E\left(\hat{P}_i\right) = P_i$ and $Var\left(\hat{P}_i\right) = P_i\left(1 - P_i\right)/T_i$.

Let U_i be the number of undervotes in the precinct, which we treat as unknown out-of-sample votes. Let V_i^0 be the uncounted votes for the candidate (out of the U_i undervotes). We assume that that the probability that undervotes have the same distribution as counted votes. Thus the distribution of V_i^0 is $B(U_i, P_i)$. Note that $EV_i^0 = P_i U_i$ and $Var\left(V_i^0\right) = P_i\left(1 - P_i\right)U_i$.

Aggregating over n precincts, the total number of undercounted votes is

$$V^0 = \sum_{i=1}^n V_i^0.$$

The estimate of this is

$$\hat{V}^0 = \sum_{i=1}^n \hat{P}_i U_i.$$

The forecast error is the difference

$$\hat{V}^0 - V^0 = \sum_{i=1}^n \left(\hat{P}_i - P_i\right) U_i - \sum_{i=1}^n \left(V_i^0 - P_i U_i\right).$$

The two sums on the RHS are independent (conditional on U_i), so

$$\begin{aligned} Var\left(\hat{V}^0 - V^0\right) &= Var\left(\sum_{i=1}^n \left(\hat{P}_i - P_i\right) U_i\right) + Var\left(\sum_{i=1}^n \left(V_i^0 - P_i U_i\right)\right) \\ &= \sum_{i=1}^n Var\left(\left(\hat{P}_i - P_i\right) U_i\right) + \sum_{i=1}^n Var\left(V_i^0 - P_i U_i\right) \\ &= \sum_{i=1}^n U_i^2 Var\left(\hat{P}_i\right) + \sum_{i=1}^n Var\left(V_i^0\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} U_i^2 P_i \left(1-P_i\right)/T_i + \sum_{i=1}^{n} P_i \left(1-P_i\right) U_i \\
&= \sum_{i=1}^{n} P_i \left(1-P_i\right) U_i \left(U_i/T_i+1\right).
\end{aligned}
$$

Hence the forecast standard error for $\hat{V}^0$ is

$$
s(\hat{V}^0) = \sqrt{\sum_{i=1}^{n} \hat{P}_i \left(1-\hat{P}_i\right) U_i \left(U_i/T_i+1\right)}.
$$

2.2 Confidence Interval for Yield Rate

The yield rate is the percentage of undervotes that when examined in a hand recount are found to be hand-readable. Let r denote the true yield rate. Let $\hat{r} = R/S$ denote an estimate of the yield rate, where S is a sample of undervotes and R is the number recovered. We see that $E(\hat{r}) = r$ and $Var(\hat{r}) = r(1-r)/S$.

Palm Beach had four precints recounted by hand (precincts 6B, 162E, 193, and 193E). In the machine count the four precincts had 496 undervotes. In the hand recount, there was an increase of 47 votes (33 for Gore and 14 for Bush). This is yield of 47 out of 496, or an estimate of 9.5% for the yield rate, or $\hat{r} = .095$. We estimate its standard deviation as 0.0132 for a 95% confidence interval of [0.069, 0.121].

2.3 Confidence Interval for Yields

As above, let V^0 be the number of uncounted votes for a candidate throughout Palm Beach County. After a recount, some number C will be counted. Assuming the process is random, we see that $EC = rV^0$ and $Var(C) = r(1-r)V^0$.

The estimate of C is $\hat{C} = \hat{r}\hat{V}^0$. The error is

$$
\begin{aligned}
\hat{C} - C &= \hat{r}\hat{V}^0 - C \\
&= r\hat{V}^0 + (\hat{r}-r)\hat{V}^0 - C \\
&= r(\hat{V}^0 - V^0) + (\hat{r}-r)\hat{V}^0 + (rV^0 - C)
\end{aligned}
$$

These three components represent the randomness due to the estimate $\hat{V}^0$ of V^0, the estimate $\hat{r}$ of r, and the randomness in the actual recounting procedure. They are independent and hence the variance of $\hat{C}$ is the sum of the variances of the three components:

$$
Var(\hat{C}) = r^2 Var\left(\hat{V}^0 - V^0\right) + \left(V^0\right)^2 Var(\hat{r}) + Var(C).
$$

Hence the forecast standard error for $\hat{C}$ is

$$
s(\hat{C}) = \sqrt{\hat{r}^2 s(\hat{V}^0)^2 + \hat{r}(1-\hat{r})\hat{V}^0 \left(\hat{V}^0/S+1\right)}.
$$

2.4 Differences Between Candidates

Let V_1^0 and V_2^0 denote the undervotes for candidates 1 and 2, respectively, and so on. The estimated difference in votes is $\hat{V}_1^0 - \hat{V}_2^0$ which has a standard error of $2s(\hat{V}_1^0)$. From a recount, C_1 of V_1^0 will be counted, and C_2 of V_2^0 will be counted. Let $D = C_1 - C_2$ denote the difference. Note that $ED = r\,(V_1^0 - V_2^0)$ and $Var(D) = r(1-r)\,(V_1^0 + V_2^0)$.

Our estimate of D is $\hat{D} = \hat{C}_1 - \hat{C}_2$. The error is

$$\begin{aligned}\hat{D} - D &= \hat{r}\left(\hat{V}_1^0 - \hat{V}_2^0\right) - D \\ &= r(\left(\hat{V}_1^0 - \hat{V}_2^0\right) - \left(V_1^0 - V_2^0\right)) + (\hat{r} - r)\left(\hat{V}_1^0 - \hat{V}_2^0\right) + \left(r\left(V_1^0 - V_2^0\right) - D\right)\end{aligned}$$

which has variance

$$Var(\hat{D}) = r^2 Var\left(\hat{V}_1^0 - \hat{V}_2^0\right) + \left(\hat{V}_1^0 - \hat{V}_2^0\right)^2 Var(\hat{r}) + Var(D)$$

and standard error

$$s(\hat{D}) = \sqrt{4\hat{r}^2 s(\hat{V}^0)^2 + \hat{r}(1-\hat{r})\left(\left(\hat{V}_1^0 - \hat{V}_2^0\right)^2/S + \hat{V}_1^0 + \hat{V}_2^0\right)}.$$

3 Data Description

The county of Palm Beach is divided into 531 voting precincts. In addition, there are 106 precincts for collection of absentee ballots. Our analysis is confined to the voting precincts.

The county of Palm Beach has posted[1] the results of the Presidential race for each precinct. These are the initial count, not the recount.

Some listed precincts reported zero votes. Others had just a few. I deleted all precincts which reported less than 20 votes. This left 494 observations.

All our variables are percentages; that is, are ratios multiplied by 100.

4 Analysis of Yield Rate

Our estimate and confidence interval for the yield rate is as follows. Palm Beach had four precints recounted by hand (precincts 6B, 162E, 193, and 193E). In the machine count the four precincts had 496 undervotes. In the hand recount, there was an increase of 47 votes (33 for Gore and 14 for Bush). This is yield of 47 out of 496, or an estimate of 9.5% for the yield rate. If we treat this as a random sample, we can construct the log-likelihood of a Binomial distribution, and use the log-likelihood statistic with an asymptotic critical value of 3.86 to obtain a 95% confidence interval for the true yield rate. In percentage, this interval is [7.12 , 12.26].

[1] I obtained this data from the webpage of Peter and Jonathan Orszag, www.sbgo.com/election.htm.

4.1 Analysis of Votes by Precinct

Using the formulas presented in the methodology section, we find the following.

From undervoting,

- Gore lost 5982 votes with a standard error of 44 and 95% confidence interval of [5903, 6074].
- Bush lost 2885 votes with a standard error of 43 and 95% confidence interval of [2801, 2969].
- The difference is 3104, with confidence interval [2928, 3280].

From a hand recound, we estimate:

- Gore will gain 568 votes with a standard error of 82, for a confidence interval of [404, 732].
- Bush will gain 274 votes with a standard error of 41, for a confidence interval of [192, 356].
- The net change (in favor of Gore) will be 294, with a standard error of 50, for a confidence interval of [184, 394].