Butterfly Ballot Blunder: An Investigation of the 2000 Presidential Election

Dani Justo and Zoe Khan

2025-03-05

Introduction

The 2000 U.S. presidential race was a highly contested election between Al Gore and George W. Bush, with the latter eventually winning by a hair due to Florida voters. However, something strange was afoot: people noticed that the "butterfly" ballot design for Palm Beach county in Florida was associated with a very high vote percentage for another candidate, Pat Buchanan. It was theorized that the ballot design led people to inadvertently vote for Buchanan instead of Gore. Our goal in this report is to analyze the relationship between votes for Bush and Buchanan in order to predict how many votes were potentially miscast.

Data Description

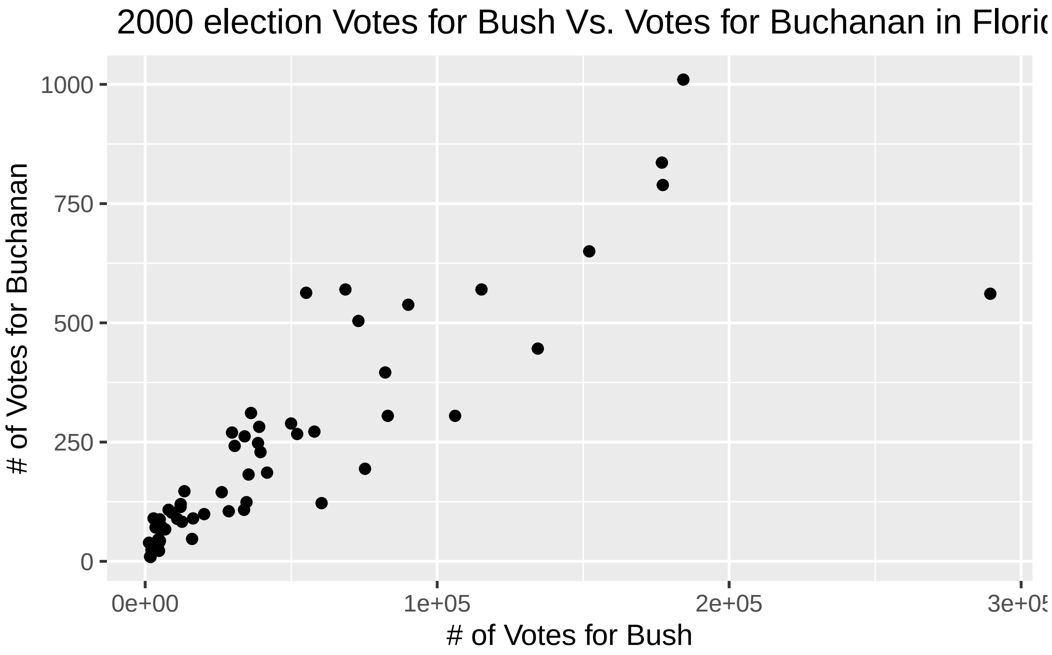
Ask about what summary statistics should be included!!!

For our data analysis, we used the The "Dramatic U.S. Presidential Election of 2000" data set from the `Sleuth2` package. This data set includes the number of votes for Buchanan and Bush in all 67 counties in Florida during the 2000 U.S. presidential election. We want to see if the confusing "butterfly" layout of the ballot used in palm beach county caused the discrepancy in votes between Bush and Buchanan, so we've removed palm beach county from our data set. This allows us to observe relationship between Bush and Buchanan's votes under "normal" circumstances, namely, if the election really hadn't been affected by the ballot.

I'll talk about these summary statistics here

mean_bush	mean_buch	median_bush	median_buch
41696.82	210.7576	18300	111

The scatter plot displayed below has a correlation coefficient of 0.867, suggesting a strong positive association between the number of votes for Bush and the number of votes for Buchanan. The R^2 of the linear regression model is 0.862 indicates that the number of votes for Bush account for 86.2% of the variability in the number of votes for Buchanan.



Modeling Process

When we explored an initial linear model, we found violations of inference and decided to use a transformed model to rectify this issue. After some experimentation, we found that a double-log transformation satisfied conditions and would allow us to continue with our inquiry. Let $Bush_i$ denote the votes for Bush, and $Buchanan_i$ denote the votes for Buchanan. Our final transformed linear model is

$$E[log(Buchanan_i)|log(Bush_i)] = \beta_0 + \beta_1 log(Bush_i)$$

The null hypothesis is that there is no relationship between votes for Bush and Buchanan. The alternative hypothesis is there there is a relationship:

$$H_0 = \beta_1 = 0$$

$$H_A = \beta_1 \neq 0$$

The estimates and standard errors of the model parameters are below:

	Estimate	Std. Error	t value	P-value
(Intercept)	-2.34	0.35	-6.61	0
log(Bush2000)	0.73	0.04	20.32	0

We get a very small p-value ($= 0.05$), so we reject the null hypothesis and conclude that there is a relationship between votes for Bush and Buchanan. In our prediction interval, we are 95% confident that an individual county with 152,846 votes for Bush would have between 250 and 1,394 votes for Buchanan. However, the true number of votes for Buchanan was 3,407, which is well out of range of the prediction interval. Taking the difference between the true estimate and the interval values, we might predict that between 1,763 and 3,157 votes were miscast.