STATISTICAL METHODS II
ASSIGNMENT 10
DUE: 29 MARCH 2011

I was not entirely sure which dataset to have all of you analyze: terrorism or votes. looking at the structure of the terrorism dataset, I thought the votes would be more appropriate. Also, we have already been fitting votes in class, so you will have a better feel for vote counts and what patternsare interesting.

⋆ ⋆ ⋆

In 2009, a presidential election was held in Afghanistan. Over thirty people ran for that position, although there really were just two candidates who had a possibility of winning: Dr. Abdullah Abdullah and Hamed Karzai. In the 34 provinces of Afghanistan, the people turned out to vote, although the vote was depressed in some areas due to fear of terrorist attacks.

When the vote ended, the independent electoral commission began tallying the counts. Karzai, according to official counts, won the election. Because of certain electoral rules which are unimportant here, Dr. Abdullah Abdullah accused Karzai of falsifying the vote.[1]

Our job is to determine if we can detect electoral hijinx. To do this, we need to determine if there is a statistically significant relationship between the proportion of the vote for Karzai and the proportion of the vote declared invalid. This is what we have been doing with Sri lanka throughout much of this semester.

PS: Getting no statistically significant results supports Karzai's assertion that the election was fair; getting statistically significant results supports Abdullah's contention of an unfair election.

[1]It turns out that the 'independent' electoral commission is not so independent.

⋆⋆⋆

The dataset we will use for this assignment is available from

`http://www.electoralforensics.org/datasets/afg2009pres.csv`

The format of this dataset is "long," as opposed to the "wide" datasets to which we are accustomed. In this dataset, there are just three variables: `PROVINCE`, `CANDIDATE`, and `VOTES`. The first variable is the `PROVINCE` from which the `CANDIDATE` received the specified number of `VOTES`. Get acquainted with the second variable. In this dataset, the unit of analysis (record, row) is the canidate-province; that is, each row represents the votes for the specified candidate in the specified province. This is a hallmark of long data.[2]

To obtain the variables of interest, you will have to select `VOTES` corresponding to specific `CANDIDATES`. As this is new to you (actually, we did it once, a long time back), I am providing the first part of the script for this assignment (linked on the assignments page). Look through the section in the script where the values are obtained from this dataset; it is important for the future (hint).

Your grade for this assignment will depend on how well you answer the initial question, how well you test the assumptions, and which model you select — and why. (Did you try the Poisson? Just the Binomial? What about the Gaussian? Any quasi models?) The writeup should be a narrative of what you did and why you did it. Still, do not include any `R` code in the write-up; provide names for your tests, not functions. Make sure you supply the appropriate graphs — presentation quality for those that illustrate your ultimate answer — rough for those that you use for testing. As always, attach your `R` script to the end of the homework packet as an appendix.

⋆⋆⋆

As always, get started early and ask if you are unsure. *Ciao*!

[2]We are accustomed to a row representing a province or an electoral division, as is true in `sri2010pres`.