PLSC 502: "Statistical Methods for Political Research"

Exercise Five
October 6, 2016

Introduction and Data

The subject of this exercise is sampling and sampling distributions. Immediately before the 2008 presidential election (specifically, on September 18, 2008), there were 92,854 registered voters in Centre County, PA. Today we're going to use those voters as data – that's right, every one of them, including many of your professors, possibly your landlord, and some of our former (but hopefully no current) graduate students. Registered voters in Centre County constitute a "population of interest" for the purposes of this exercise. `PLSC502-2016-ExerciseFive.csv` contains all of them, albeit with many of the interesting tidbits (names, addresses, etc.) stripped out to protect the guilty. The variables we *do* have include:

- `ID` – an arbitrary (and indecipherable) voter identification number,
- `DateOfBirth` – the voter's date of birth,[1]
- `RegDate` – the date on which the voter first registered to vote,
- `ZipCode` – the five-digit ZIP code in which the voter lives,
- `Precinct` – the numeric identifier ($\in [1, 89]$) for the precinct in which the voter lives,
- `Active` – whether (=1) or not (=0) the voter is listed as "active" on the rolls,
- `LastVoteDate` – the date on which the voter last voted,
- `PartyID` – what the name suggests, coded 1 = Democrat, 2 = Republican, and 3 = other, and
- `Female` – a naturally-coded gender indicator.

[1] R has various ways of handling data that are encoded as dates. Note that `read.csv` typically reads dates as either factor or character variables; a straightforward way to convert those into date-formatted objects is:

```
> Data$DOB <- with(Data, strptime(DateOfBirth, "%d%b%Y"))
```

See `?strptime`, `?format.Date` and/or the `chron` package for more details.

Exercise

1. Begin by creating a variable (`Primary`) that indicates whether (=1) or not (=0) a voter voted in the 2008 Pennsylvania primary (which was held on April 22, 2008; this was the last election held prior to the date on which the data were collected).

2. Draw a single simple random sample of 150 voters from the Centre County data. Discuss briefly how your sampled data compare to the population as a whole, particularly with respect to the `DateOfBirth`, `Active`, `PartyID`, `Female`, and `Primary` variables.

3. Draw a cluster sample of $N_c = 10$ clusters, using precincts as your PSU. Again, briefly discuss the similarities and differences between the clustered sample and the population for the five variables mentioned above.

4. Draw a stratified random sample of 1% of the population. Stratify on `PartyID`, and sample such that half of your sample comprises Republicans and the other half Democrats (that is, oversample from the two major parties, and undersample (to zero) from "Other" parties). Again, discuss briefly how your sample differs from the population on `DateOfBirth`, `Active`, `Female`, and `Primary`.

5. Finally, illustrate the empirical sampling distribution of the sample mean and variance of the proportion of Republican voters where the (simple random) sample size (N) is 20, and again where the sample size is .800.

As always, this exercise is worth 50 possible points; it is due by 5:00 p.m. ET on Thursday, October 13, 2016.