PLSC 502: "Statistical Methods for Political Research"

Exercise Four
October 9, 2023

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 the "population of interest" for the purposes of this exercise. `PLSC502-2023-ExerciseFour.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 the base package's `read.csv` command typically reads dates as either factor or character variables; a straightforward way to convert those into date-formatted objects uses the `lubridate` package:

```
> Data$DOB <- dmy(Data$DateOfBirth)
```

See `?strptime`, `?format.Date` and/or the `chron` or `lubridate` packages 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. Briefly discuss how your sampled data compare to the population as a whole, particularly with respect to the `DateOfBirth`, `Active`, `PartyID`, `Female`, and `Primary` variables.

3. Repeat part 2 many[2] times, and compare the sampling distribution of the means and variances of those variables to their theoretical quantities.

4. Draw many cluster samples of $N_c = 10$ clusters, using precincts as your primary sampling unit (cluster). Again, briefly discuss the similarities and differences between the values of the data in the clustered samples and those in the population for the five variables mentioned above.

5. Draw many stratified random samples with a sample size equal to 1% of the population. Stratify on `PartyID`, sampling such that half of your sample are Republicans and the other half Democrats (that is, oversample from the two major parties, and undersample (to zero) from those identifying with "Other" parties). Again, discuss briefly how your sample differs from the population on `DateOfBirth`, `Active`, `Female`, and `Primary`.

6. Finally, illustrate and briefly describe the empirical sampling distribution of the sample mean and variance of the proportion of Republican voters where the (simple random) sample size is $N = 20$, and again when the sample size is $N = 800$.

As usual, use plots, words, or combinations thereof to complete this exercise. Submit your answers **in PDF format**. In addition to your answers, please include a copy of all computer code used to conduct your simulations, generate your figures, etc. This can be in any form – a separate `.R` or `.do` file, an appendix in the PDF, or as a `.Rmd` or similar format containing both content and code. This homework exercise is due by 11:59 p.m. ET on Tuesday, October 17, 2023; submit your materials in electronic format – via e-mail attachment – to Nathan (`nam@psu.edu`) *and* to me (`zorn@psu.edu`). This exercise is worth 50 possible points.

[2] I'll leave it to you to decide what "many" is.