PLSC 502: "Statistical Methods for Political Research"

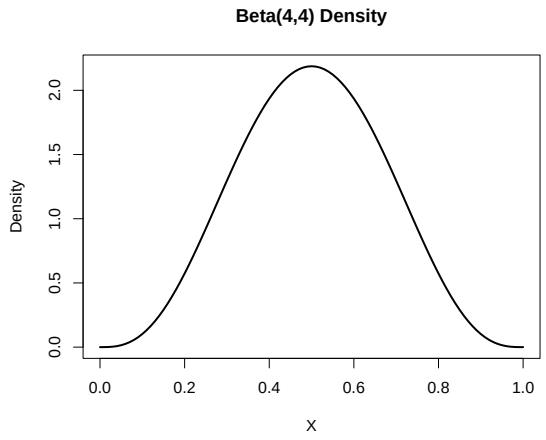
Exercise Five

October 21, 2022

Part I

We'll start with a simulation-based exercise. The beta distribution is a two-parameter probability density with support on the unit [0,1] interval. It's parameters are "shape" parameters that define the shape of the probability density; they are typically denoted a and b, and the distribution is sometimes written as $\mathcal{B}(a, b)$. Those two parameters a and b completely characterize the shape of the density of a beta-distributed variate.

For this part of the exercise, we're going to work with a beta distribution with $a = 4$ and $b = 4$ [so, $\mathcal{B}(4, 4)$]; that specific density is illustrated in the figure here:



The general goal of this part of the simulation is to show how Normal-based confidence intervals provide poor coverage when sample sizes are small. To that end, do the following:

1. Choose some different sample sizes; for example, you might consider $N \in \{6, 15, 40, 100, 500\}$. Make sure that at least a couple of them are pretty small.
2. For a given sample size, simulate samples of size N by drawing N values from a $\mathcal{B}(4, 4)$ distribution;[1] do this K times, where K is large.
3. For each of the K samples of size N, estimate $\hat{a}$ and $\hat{b}$, as well as $\hat{a}$'s and $\hat{b}$'s standard errors (that is, $\sigma_{\hat{a}}$ and $\sigma_{\hat{b}}$).[2]
4. Using the estimates $\hat{a}$ and $\hat{b}$ and their standard errors $\sigma_{\hat{a}}$ and $\sigma_{\hat{b}}$, construct K 95-percent two-tailed confidence intervals for $\hat{a}$ and $\hat{b}$, on the assumption that $\hat{a} \sim \mathcal{N}(a, \sigma^2_{\hat{a}})$ and $\hat{b} \sim \mathcal{N}(b, \sigma^2_{\hat{b}})$.
5. Repeat steps (2-4) for each of the values of N in (1).
6. Conclude by discussing:
   (a) ...how the confidence intervals around $\hat{a}$ and $\hat{b}$ change with N, and
   (b) ...how well the Normal-based intervals perform in terms of "coverage." Specifically, for each different value of N, does the empirical reality (the number of those K confidence intervals for a and b that contain the true values $a = 4$ and $b = 4$) match the theoretical expectation associated with a 95-percent confidence interval? Explain any differences you find.

[1] This can be done using the `rbeta` command with options `shape1=4` and `shape2=4`.

[2] This step is probably most easily done using the `fitdistr` command in the `MASS` package; using that command, you can specify (e.g.):

```
> foo <- fitdistr(X,"beta",start=list(shape1=1,shape2=1))
```

and the resulting object will have estimates $\hat{a}$ and $\hat{b}$ in `foo$estimate` and estimated standard errors in `foo$sd`.

Part II

The "real" data for this exercise are drawn from Washington University's *The American Panel Study* (TAPS), which collected data on a panel of respondents over several monthly waves from 2012-2017. (Here, we will be looking only at single-response questions, so the data are cross-sectional.) The data have $N \approx 1100$, and are available on the course github repository, in the "Exercises" folder, in a file named `PLSC502-2022-ExerciseFive.csv`. In addition to a respondent identifier variable (`WUSTLID`), the pool of independent variables in those data are:[3]

- Political party identification indicators – binary variables for `Democrat` (Democratic Party) and `GOP` (Republican Party), with independents serving as a reference category;
- `Ideology` – a seven-point Likert-type indicator variable, where higher values indicate greater political conservatism (right-wing) and lower values indicating greater progressivism (left-wing);
- `Education` – measured as a twelve-category ordinal variable with values ranging from 3 to 15, where the lowest value corresponds to a 5th-6th grade level of education and the highest reflects a doctoral degree;
- `Income` – a 15-category ordinal variable, where higher values indicate higher income levels (where each unit roughly corresponds to an increase of $10,000 in annual income);
- The respondent's age in years, as of 2016 (`Age2016`);
- `Female` – a binary indicator of sex, naturally-coded;
- Racial classifications – binary indicator variables for `White`, `Black`, and `Asian` identification (with "other" as the reference category);
- `FT.Communists` is the respondent's placement of "communists" on a 0-100 "feeling thermometer" scale;
- `InterviewDuration` records the number of seconds that each respondent took to complete the (on-line) survey;
- `HowManyKids` indicates the number of children (under the age of 18) that each respondent has in their household;
- Finally, the TAPS survey also asked a series of yes-no / binary-response questions related to respondents' specific behaviors and preferences, including:
    - Have you ever taken the shampoo and conditioner bottles from a hotel or motel? (`StealShampoo`; 0=no, 1=yes)
    - During the past year, have you ever run out of gas while driving a car or other vehicle? (`RunOutOfGas`; 0=no, 1=yes)
    - Have you ever looked directly at the sun to see an eclipse without using a filter? (`LookedAtEclipse`; 0=no, 1=yes)
    - Have you ever stolen a street sign? (`StolenStreetSign`; 0=no, 1=yes)
    - Would you rather be attacked by a big bear or a swarm of bees? (`BeesOrBear`; 0=bees, 1=bear)

Note that several of the variables in these data have missing values, some of them in substantial numbers.

[3] The data also contain a set of survey `weights` that reflect the sampling scheme; you can ignore those, for now.

Using these data, do the following:

1. Calculate and report the 95 percent confidence intervals for:
   (a) The mean of `Age2016`;
   (b) The mean of `FT.Communists`;
   (c) The mean of `InterviewDuration`;
   (d) The proportion of "bear" responses to `BeesOrBear`.
2. Plot the 80, 95, and 99-percent confidence intervals for each of the four variables in (1).
3. Discuss, in words, two of the four confidence intervals in (1). What do they *mean* in substantive terms?
4. Use both significance tests with $\alpha = 0.10$ and P-values to examine the following hypotheses:
   (a) $\overline{\texttt{Ideology}} = 5$.
   (b) $\overline{\texttt{HowManyKids}} = 0.5$.
   (c) $\overline{\texttt{Female}} = 0.5$.
   (d) $\overline{\texttt{StealShampoo}} = 0.9$.

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 Friday, October 28, 2022; submit your materials in electronic format – via e-mail attachment – to Tuba (`tzs5636@psu.edu`) *and* to me (`zorn@psu.edu`). This exercise is worth 50 possible points.