Assignment #2

Stat 332: Sampling and Experimental Design (Spring 2024)

Professor: Christian Boudreau — Due at 11:59 pm on Monday Jun 17, 2024

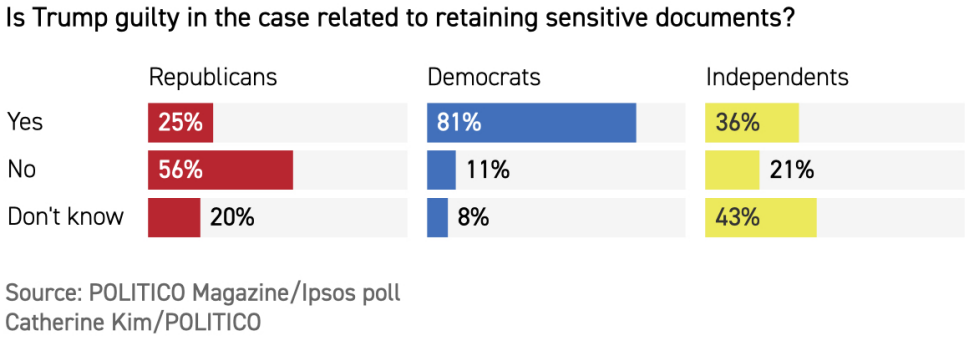
Instructions:

- Teamwork is allowed and encouraged; however, everyone must hand in their own assignment showing that they understood what they wrote. You are not allowed to collaborate with anyone who is not a STAT 332 classmate.
- Justify all your answers.
- One solution per question; if you provide multiple solutions, you will receive 1/2 the points of the inferior solution.
- Write legibly; if the TA or I have to guess at your answers, it's unlikely to be in your favour.
- Make sure you include your `R` code and relevant output for Q#4.
- Crowdmark:
  - You should soon received an automated email from Crowdmark saying that you can now submit/upload your solution. If you have not received the email by noon on Wednesday, please start by checking your spam or junk mail folder. If the email is not there, email me. Note that, even if you did not get the email, you should still be able to submit/upload your assignment by simply going to Crowdmark, and signing into your account.
  - To ensure smooth and efficient grading, please make sure to upload your solution separately for each question.
  - Crowdmark allows you to resubmit anytime before the due date; once the due date has passed, your submission is locked in and you will thus not be able to resubmit.
  - Late submissions will be accepted, but are subject to a 5% per hour penalty*; click here for more information on Crowdmark and late submissions.
  - Technical difficulties with Crowdmark:
    - Consult Crowdmark Help
    - Watch this short video about submitting an assignment on Crowdmark
    - E-mail me
    - If tragedy strikes and you cannot upload your assignment to Crowdmark, email it to me, so I have proof of when you completed your assignment (otherwise, the above mentioned lateness penalty will apply).
- Since I do not want the reduced lateness penalty to interfere with me posting these solutions to assignment #2 before the midterm, I'm making the following changes. There will be no 48 hours reduced lateness penalty. However, the 5% reduced lateness penalty for an additional 24 hours remains. Again, please recall that you have to email me (at least) 12 hours before the original due date to take advantage of this reduced penalty. To compensate for the fact that they are no 48 hours reduced lateness penalty, I am allowing for a 0% reduced lateness penalty if you submit your assignment within 3 hours past the deadline. Though you will not get penalized (i.e., 0% lateness penalty), this counts towards your one time reduced lateness penalty.
- See course outline for more information on assignments, and how much each of them contributes to your overall grade.

*This is the default penalty, and is applied when there are no extenuating circumstances.

Question #1 (5 points):

A recent POLITICO Magazine/Ipsos poll asked 1,005 adults (i.e., ages 18 or older) Americans if they believe that Trump is guilty in the case related to retaining sensitive documents. As expected, results vary greatly depending on the polytical affiliation; see table below.



(a) Using the fact that (as of Jan 2024) 26% of adult Americans are Republicans, 28% are Democrats and 46% are Independents, estimate the overall proportion of Americans who believe that Trump is guilty [†]

(b) Provide a 95% CI for the estimate obtained in part (a)

Question #2 (10 points):

A practitioner makes a mistake in applying the optimal/Neyman allocation, and uses

$$n_h = n \frac{W_h S_{U_h}^2}{\sum_{h=1}^{H} W_h S_{U_h}^2}$$

instead of the correct formula given by

$$n_{h,o} = n \frac{W_h S_{U_h}}{\sum_{h=1}^{H} W_h S_{U_h}}$$

(a) Let $V_{\text{STSRS,e}}(\bar{y}_\pi)$ be the variance of $\bar{y}_\pi$ under this erroneous allocation (where "e" in the subscript stands for erroneous). Find the relative loss of precision due to the use of error; i.e., find a formula for

$$\frac{V_{\text{STSRS,e}}(\bar{y}_\pi) - V_{\text{STSRS,o}}(\bar{y}_\pi)}{V_{\text{STSRS,o}}(\bar{y}_\pi)}$$

where $V_{\text{STSRS,o}}(\bar{y}_\pi)$ is the variance of $\bar{y}_\pi$ under the optimal/Neyman allocation, and is given by

$$V_{\text{STSRS,o}}(\bar{y}_\pi) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h S_{U_h}\right)^2 - \frac{1}{N}\sum_{h=1}^{H} W_h S_{U_h}^2$$

Note: to get full marks, you must simplify your answer; in particular, your answer should not contain any double sums.

[†] We obviously know that he is guilty of a lot more crimes that jus the sensitive documents case.

(b) Show that the relative loss of precision is only about 10% for a very large population (i.e., $N \to \infty$) consisting of 3 strata where both the stratum relative sizes and the stratum standard deviations are in the ratio 1:2:3; i.e., $W_2 = 2W_1$ and $W_3 = 3W_1$, and $S_{U_2} = 2S_{U_1}$ and $S_{U_3} = 3S_{U_1}$.

Question #3 (10 points)**:**

In class, we saw how to use `survey package` to compute $\bar{y}_\pi$ and $\hat{t}_\pi$ under SRS and STSRS. We also saw how to use use that package to compute $\widehat{V}_{\text{SRS}}(\bar{y}_\pi)$, $\widehat{V}_{\text{SRS}}(\hat{t}_\pi)$, $\widehat{V}_{\text{STSRS}}(\bar{y}_\pi)$ and $\widehat{V}_{\text{STSRS}}(\hat{t}_\pi)$. However, the `survey package` must be able to compute $\bar{y}_\pi$ and $\hat{t}_\pi$, as well as their corresponding variance estimators, under more sampling designs that just SRS and STSRS. Hence, the `survey package` uses more general formulas than the ones given in class. However, these more general formulas simplify to the ones given in class, and this question is concerned with proving this.

The formula used by the `survey package` for estimating a population mean $\bar{y}_U$ is given by

$$\bar{y}_{\text{svy}} = \frac{1}{w_{\bullet\bullet}} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}\, y_{hi}$$

where

$$w_{\bullet\bullet} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}$$

The variances estimator of $\bar{y}_{\text{svy}}$ is given by

$$\widehat{V}_{\text{svy}}(\bar{y}_{\text{svy}}) = \sum_{h=1}^{H} \frac{n_h(1 - n_h/N_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi} - \bar{e}_{h\bullet})^2$$

where

$$e_{hi} = \frac{w_{hi}}{w_{\bullet\bullet}}(y_{hi} - \bar{y}_{\text{svy}}) \qquad \text{and} \qquad \bar{e}_{h\bullet} = \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi}$$

Using the above formulas and the ones given in class show that:

(a) $\bar{y}_{\text{svy}}$ simplifies to $\bar{y}_S$ under SRS

(b) $\widehat{V}_{\text{svy}}(\bar{y}_{\text{svy}})$ simplifies to $\widehat{V}_{\text{SRS}}(\bar{y}_S)$ under SRS

(c) $\bar{y}_{\text{svy}}$ simplifies to $\bar{y}_\pi$ under STSRS

(d) $\widehat{V}_{\text{svy}}(\bar{y}_{\text{svy}})$ simplifies to $\widehat{V}_{\text{STSRS}}(\bar{y}_\pi)$ under STSRS

Question #4 (15 points)**:**

This question is concerned with the Survey of Youth in Custody (SYC). This survey was conducted by the U.S. Department of Justice amongst juveniles and young adults in longterm state-operated juvenile institutions by the end of 1987. Out of the 23,655 juveniles/young adults in custody, a sample of 2,621 was selected using a stratified sampling design (see variable `stratum`). These individuals were then interviewed about family background, previous criminal history, drug and alcohol use, etc. More information about the survey can be found here.

Since this survey uses a fairly complex stratified sampling design, the statistician(s) in charge computed sampling weights (see variable `finalwt`) for people to use when analysing SYC data. You will thus have to use the `weights` argument of the `svydesign` function instead of the `probs` argument that we have used thus far; more information about the arguments of the `svydesign` function can be found here.

File `syc.txt`, which contains the dataset, is available on the STAT 332 LEARN website. The complete list of available variables and their description is given at the top of file `syc.txt`, but this question is concerned with the 8 variables given in the table below. Missing values should be handled the same way as in Assignment #1.

Variable	Description
`age`	age (99 = missing)
`race`	race (1 = white, 2 = black, 3 = Asian/Pacific Islander, 4 = Native, 5 = other, 9 = missing)
`sex`	gender (1 = male, 2 = female, 9 = missing)
`numarr`	# of prior arrests (99 = missing)
`prviol`	previously arrested for violent crime (1 = yes, 0 = no, 9 = missing)
`everdrug`	ever used illegal drugs (1 = yes, 0 = no, 9 = missing)
`stratum`	stratum number (labelled 1, ..., 16)
`finalwt`	sampling weight

Stratum sizes

h	N_h	h	N_h	h	N_h	h	N_h
1	2724	5	3504	9	624	13	744
2	3192	6	376	10	520	14	847
3	4107	7	56	11	672	15	824
4	2705	8	528	12	384	16	1848

Using the `survey package` with weight variable `finalwt` and properly accounting that SYC used a stratified sampling design, give a point estimate and corresponding 95% CI for each of the following:

(a) Estimate the average age of juveniles/young adults in custody.

(b) Estimate the mean number of prior arrests.

(c) Estimate the proportion of juveniles/young adults in custody that have used illegal drugs.

(d) Estimate the proportion of juveniles/young adults in custody that were previously arrested for a violent crime.

(e) Estimate the proportion of juveniles/young adults in custody that have used illegal drugs and were previously arrested for a violent crime.

(f) Estimate the proportion of males amongst juveniles/young adults in custody.

(g) Estimate the proportion of African Americans amongst juveniles/young adults in custody.

(h) Estimate the proportion of African American males amongst juveniles/young adults in custody.

(i) From census data, we know that African-American males are about 7.6% of the youth population in the U.S. Using your answer to part (h), what can you conclude about racial bias when it comes to youth in custody?