October 22, 2025

Stat 440 HW 4 Assignment, due Wednesday 11/5/25

Instructions: HW Set 4 consists of 4 problems worth a total of 80 points, due 11/05/25, 11:59pm on ELMS. Your answers to the submitted HW problems should show complete reasoning and (wherever you used it) computer code, along with page and line references in the Thompson book or other book source (or my web-page Handouts or slides) for where the topic is covered.

(1.) (10 points) Thompson, Ch.11, #11.1.

(2.) (10 points) Thompson, Ch.11, #11.2.

Download the data-frame "cty.unemp" by reading with `readRDS` from the dataset `ctyunemp.rds` in the Scripts directory of the course web-page. See the short code-description at the end of this assignment document to see how I extracted it from the `usdata` R-package, keeping only the 3139 county values with non-missing values in the columns `pop2010, civiilian_labor_force_2017, unemployed_2017, some_college_2017`. In the next problem, the goal is to estimate the national unemployment rate. We will assume that the dataset total of civilian labor force in 2017 of 160,582,000 is correct, so the unemployment rate will be estimated as the estimated total unemployed (from stratified SRS samples) divided by that number. Here y_i in county i is the number of civilian unemployed in 2017 in that county; the target of estimation is t_y (total civilian unemployed nationally in 2017); and the estimates $\hat{t}_y$ will come from stratified random samples.

(3.) (45 points) (a) Write an R function to draw a Stratified SRS sample of 130 counties stratified by quartiles of 2010 population, respectively with 10,20,40,60 sample counties drawn from quartiles 1, 2, 3, 4 of pop2010, and estimate both the total number of civilian unemployed (in all counties) from the stratSRS sample data, and also the variance $var(\hat{t}_y)$ of that estimated number unemployed. Apply your function 1000 times and compare the variance $\widehat{\text{var}}(\hat{t}_y)$ you estimated at first to the variance of the 1000 new $\hat{t}_y$ estimates from the sample sampling design. Also compare with the average of the variance estimates $\widehat{\text{var}}(\hat{t}_y)$ you get from your 1000 new samples (all `stratSRS` of size 130). Also make sure that your estimates of the national unemployment rates from all your samples are reasonable. (The national rate was 4.36% : what proportion of your $\hat{t}_y$ estimates fall within $\pm 1.96 \cdot \{\widehat{\text{var}}(\hat{t}_y)\}^{1/2}$ of the correct value $t_y = 160,582,000 * 0.04355 = 6993346$?)

(b) Now use the theory of optimal stratum allocation in the chapter, based on the stratumwise variances of y_i calculated on the whole dataset and the same 4 strata of `pop2010 quartiles`, to calculate what would be the optimal stratum sizes to have used in your stratified SRS sample of size 130, from the viewpoint of getting an estimator of t_y with the smallest poassible variance ? And also find out what that optimized variance is.

(c), (d) Using R code similar to what you wrote in part (a), repeat all the steps in (a)–(b) with total-estimates for the same outcome variable y_i, now using strata defined as the counties in the four quartiles of the county median percentages of persons with at least "some college" (I think the ACS definition is at least 1 year of junior- or 4-year college).

(e) Which of these four stratified-sample designs give the smallest variance for stratified-SRS-sampling to estimate t_y ? Can you provide some explanation why this best-performing design was the winner ?

(4). (15 points) Thompson, Ch.8, #8.1(a)–(c).

Steps Used to Extract Data

```
> library(usdata)
##  after having installed the package    usdata  of ACS  data
##   which contains data-frames  "county" and "county_complete"
> cty.unemp = cbind.data.frame(county_complete[,c("civilian_labor_force_2017",
     "unemployed_2017","some_college_2016")], county$pop2010)
     !is.na(county_complete$civilian_labor_force_2017),]
             ## now nothing missing, 3139 counties
> c(ty.tot = sum(cty.unemp$civi), unemp.rate =
             sum(cty.unemp$unemp)/sum(cty.unemp$civi))
      ty.tot   unemp.rate
1.605824e+08 4.355128e-02

> dim(cty.unemp)
[1] 3139    4
> names(cty.unemp)
[1] "civilian_labor_force_2017" "unemployed_2017"
[3] "some_college_2016"         "county$pop2010"

>  saveRDS(cty.unemp, "ctyunemp.rds")
```