November 12, 2025

Stat 440 HW 5 Assignment, due Wednesday 11/26/25

Instructions: HW Set 5 consists of 6 problems worth a total of 85 points, due 11/25/25, 11:59pm on ELMS. (OK to hand in by 11:59pm 11/26.) 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 each) Thompson, Ch.12, # 1, 3 and 5; and Ch. 13 # 1.

(2.) (*20 points*) Problem #2 on the Midterm described a simple random sample from a population, as follows. The adult population of a small city consists of consists of 50,000 people, of whom 18,000 are registered Democrats, 20,000 registered Republicans, and 12,000 are Independents. A Simple Random Sample of size 600 of the whole adult population is taken to find the proportion of people favoring candidate Jones versus candidate Smith in a two-person race for mayor. In the survey, the numbers of people in each registration category and their favored candidates are given in the following table:

	Democrat	Republican	Independent
Jones	100	105	125
Smith	90	120	60

Use these data

(a) to provide a 95% confidence interval **post-stratification** estimator for the proportion of the whole population favoring candidate Jones, and

(b) to compare that estimator and its variance with the one based on a multiple-regression estimator for the average of $y_i = I_{[i \text{ favors Jones}]}$ with predictor variables $\mathbf{x}_i = (I_{[i \text{ is Democrat}]}, I_{[i \text{ is Independent}]})$.

You could either derive and use formulas for the regression estimator in part (b), or else you could create a dataset of 600 units with filled-in columns according to the Table data and use (unweighted) linear-regression software on that dataset.

The next problem continues the use of the stratified dataset from problem # (3) from HW4. Follow the steps given there to read the data-frame "cty.unemp" with `readRDS` from the dataset `ctyunemp.rds` in the Scripts directory of the course web-page, and to group the data into 4 strata based on quartiles of the variable `pop2010`. Our objective now is to create a ratio estimator and confidence interval for the national unemployment rate (for counties of population less than 10^6) in 2017 from columns `unemployed_2017` and `civilian_labor_force_2017` of a stratified random sample of the data from your data-frame.

(3.) (*25 points*) Consider the data `ctyunemp.rds` in the Scripts directory of the course web-page. Perform the following computational steps, for which you may use the function you wrote for problem (3) of Homework Set 4 to draw stratified random samples from these data.

(a). Create a column `stratlab` of labels with values 1 to 4 for the quartiles of `pop2010`. (Stratum 1 is the set of entries with the 25% lowest values of 2010 county populations, etc.)

(b) Draw a stratified simple random sample with respective stratum sample sizes 80, 40, 50, 100 from the 3139 counties in your dataset, and create stratfied SRS estimates $\hat{t}_y$ of the total number unemployed and $\hat{t}_x$ of the total number in the civilian labor-force in the whole set of counties in the dataset.

(c). The ratio $\hat{t}_y/\hat{t}_x$ is a ratio estimator for the unemployment rate in the whole set of counties in the dataset. Find the estimator and an estimator for its variance, based on the approximate method given in class and in the book (Chapter 7) for variances of ratio-estimators based on general survey weights.

(d). By drawing a large number of independent stratified SRS samples with the same stratum sample sizes 80, 40, 50, 100 as in parts (b)–(c), verify by simulation that the sample variance of the ratio estimators you compute from those samples is close to the variance you estimated from a single sample in part (c).