Exercise ark #3.
Ratio and regression estimation

Oğuz–Alper, Melike & Pekarskaya, Tatsiana, Statistics Norway

October 21, 2020

Exercise 1

(R code available) The data set agsrs.csv contains information on the number of farms in 1987 for the SRS of $n = 300$ counties from the population of the $N = 3\,078$ counties in the United States. In 1987, the United States had a total of 2 087 759 farms and the total number of acres devoted to farming was 964 470 625.

1. Plot the data.
2. Use ratio estimation to estimate the total number of acres devoted to farming in 1992, using the number of farms in 1987 as the auxiliary variable.
3. Repeat (2), using regression estimation.
4. Which method gives the most precision: ratio estimation with auxiliary variable $acres87$, ratio estimation with auxiliary variable $farms87$, or regression estimation with auxiliary variable $farms87$? Why? (Lohr, 2019, p.157)

Exercise 2

(R code available) The data file counties.csv contains information on land area, population, number of physicians, unemployment, and a number of other quantities for an SRS of 100 of the 3 141 counties in the United States (U.S. Census Bureau, 1994). The total land area for the United States is 3 536 278 square miles; 1993 population was estimated to be 255 077 536.

1. Draw a histogram of the number of physicians for the 100 counties.
2. Estimate the total number of physicians in the United States, along with its standard error, using $N\bar{y}$.
3. Plot the number of physicians vs. population for each county. Which method do you think is more appropriate for these data: ratio estimation or regression estimation?
4. Using the method you chose in (3), use the auxiliary variable population to estimate the total number of physicians in the United States, along with the standard error.
5. The "true" value for total number of physicians in the population is 532 638. Which method of estimation came closer?

6. Repeat parts (1)–(4) with y = farm population and x = land area. The "true" value for farm population is $t_y = 3\,871\,583$ (Lohr, 2019, pp.157-158)

Exercise 3

Suppose the population consists of 4 households with total income of 210, 350, 700 and 1260(in units of kr. 1000) for households 1, 2, 3, 4 respectively. The numbers of adults in the households are 1, 1, 2 and 3 for households 1, 2, 3, 4. We take a simple random sample of size 2 for estimating the total income t (= 2520). Consider the two estimators $\hat{t}_e = N\bar{y}_s$ and $\hat{t}_r = t_x \frac{\bar{y}_s}{\bar{x}_s}$ where the auxiliary variable x_i is the number of adults in a household i.

1. Find the expected value and variance (exact values) of the two estimators. Which one would you choose?

2. Compute the approximate 95% confidence intervals (assuming approximate normality) based on the two estimators for all 6 possible samples. What are the exact confidence levels for the two CI procedures based on $\hat{t}_e$ and $\hat{t}_r$? (Bjørnstad, 2018)

Exercise 4

(R code available) We are going to use a dataset pop_industry.csv which contains a population of 415 companies with a given NACE (business classification code). For each company there is registered information on turnover (y) and number of employees (x). We are going to estimate turnover of the population of these 415 companies by using different estimation methods.
As far as sampling plan is concerned: all units with more than 50 employees, should be surveyed. For the rest – should be applied simple random sampling. The size of the sample should be such that units from the group with > 50 employees and $<= 50$ employees together will be 25 companies.

1. Calculate estimates of total value of variable y and 95% CI with a help of ratio ($\hat{t}_r$) and expansion ($\hat{t}_e = N\bar{y}$) estimators. Make sampling 10 times and fill in the table below. Compare the estimators. Which one would you prefer?

# sample	$\hat{t}_r$	CI for $\hat{t}_r$	$\hat{t}_e$	CI for $\hat{t}_e$
1				
2				
...				
10				

2. By taking an SRS of 25 companies from the population we get an empiric standard error for ratio estimator equal to 730 and for expansion estimator – 1589(values are in thousands). Compare sample plans in (1) with SRS. (Bjørnstad, 2016, p.66)