Tutorial 6:

Introduction to Data Science

Stratified Sampling

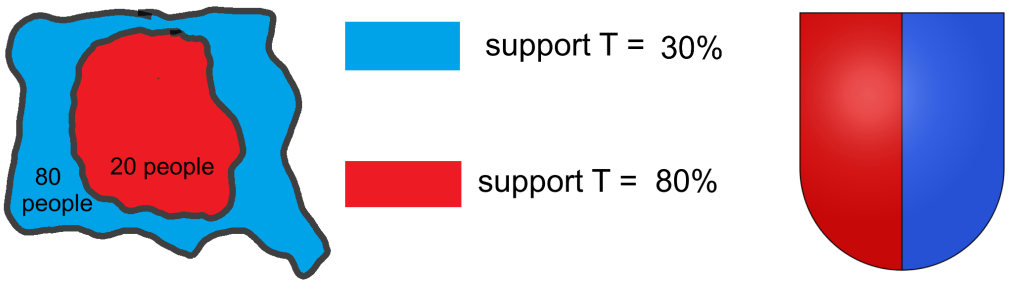
Ernst C. Wit

October 21, 2024

Exercises

1. **Voting in Blue-Red Land.**
Context: Blue-Red Land has a population of 100 voters for an upcoming election, where they have the choice between 2 candidates: T and H. The 100 voters live in 2 counties:

   - Blue County: size = 80
     - 24 support T, 56 support H
   - Red County: size = 20
     - 16 support T, 4 support H



A pollster is interested in predicting the outcome of the elections, i.e., she wants to estimate p = true fraction of T voters. She proposes to take the following stratified sample:

- Sample **5 voters** from Red County,

$$X_r = \text{number of T supporters among 5 in Red County}$$

- Sample **5 voters** from Blue County,

$$X_b = \text{number of T supporters among 5 in Blue County}$$

Tasks:

(a) What is the value of the parameter that the pollster tries to estimate?

(b) She defines a **stratified** estimator:

$$\hat{p}_{\text{strat1}} = 0.5 \times \frac{X_r}{5} + 0.5 \times \frac{X_b}{5}.$$

Is the estimator unbiased?

(c) The pollster get some advise to adjust the estimator to make it unbiased. She gets two suggestions:

$$\begin{aligned} \hat{p}_{\text{strat2a}} &= c \times \hat{p}_{\text{strat1}} \\ \hat{p}_{\text{strat2b}} &= w \times \frac{X_r}{5} + (1-w) \times \frac{X_b}{5}. \end{aligned}$$

i. Find the values c and w to make the estimators unbiased.

ii. Why is $\hat{p}_{\text{strat2a}}$ a invalid estimator, and why is $\hat{p}_{\text{strat2b}}$ a valid estimator?

(d) In order to see how good the estimator is, determine the standard deviation of $\hat{p}_{\text{strat2b}}$.

(e) Compare the value of the standard deviation with that of the SRS estimator and the stratified estimator from the lecture, i.e.,

$$\begin{aligned} SD[\hat{p}_{\text{simple}}] &= 0.15 \\ SD[\hat{p}_{\text{strat}}] &= 0.14 \end{aligned}$$

(f) **Voter Poll Simulation.** Start by creating a country consisting of two counties as described above with the number of T and H supporters as indicated. Then repeat 10,000 times the following sampling schemes:

i. **Simple Random Sampling.** Sample 10 individuals from the entire country with replacement, record X the number supporting T, and calculate

$$\hat{p}_{\text{simple}} = \frac{X}{10}$$

ii. **Proportional Stratified Random Sampling.** Sample 2 individuals from the Red County and 8 individuals from the Blue County with replacement, record X_r and X_b the number supporting T in the two samples, respectively, and calculate

$$\hat{p}_{\text{strat}} = \frac{X_r + X_b}{10}$$

iii. **Weighted Stratified Random Sampling.** Sample 5 individuals from the Red County and 5 individuals from the Blue County with replacement, record X_r and X_b the number supporting T in the two samples, respectively, and calculate

$$\hat{p}_{\text{strat2b}} = w \times \frac{X_r}{5} + (1-w) \times \frac{X_b}{5},$$

with the w you found in c(i).

Calculate empirically the mean and standard deviation for each of the three estimators.

(g) What is the approximate distribution of $\hat{p}_{\text{strat2b}}$? [Hint: Make use of the Central Limit Theorem 2 times and use the fact that the sum of two normal distributions is again normally distributed.]

(h) Make a histogram of the values for $\hat{p}_{\text{strat2b}}$ you found in question (f) and superimpose the density of this approximate distribution from (g). How good is the approximation?