SURVMETH / SURV 625
Applied Sampling
Winter / Spring 2025

Homework 4

1. The following data were collected from a sample of $n = 10$ clusters that was selected from a large population (*assume that the sampling fractions are negligible*):

i	1	2	3	4	5	6	7	8	9	10	Totals
$t_{y,i}$	5	1	2	4	2	2	3	3	4	6	32
$t_{x,i}$	13	11	7	11	6	11	5	11	9	10	94

a) Compute the ratio mean $r = y/x$, where $y = \sum_i t_{y,i}$ is the total outcome and $x = \sum_i t_{x,i}$ is the realized sample size, and its standard error. Note that this is an example of simple random sampling of unequal-sized clusters. Use the ultimate cluster idea for variance estimation purposes (i.e., we don't really care how many stages of cluster sampling led to the realized sample sizes in each cluster; we assume a one-stage selection of ultimate clusters, where all units were sampled within them).

b) The mean is actually the proportion of individuals with a particular attitude (meaning that the *Y* variable is a binary indicator of whether a person has that attitude). Given this information, compute the simple random sampling variance, design effect, and *roh*. (*Hint:* Remember that when computing the design effect for these designs, the average sample size per cluster should be used.)

c) Estimate the variance if the sample size were tripled by tripling the number of primary stage cluster selections from 10 to 30.

d) Estimate the sampling variance if the sample size were tripled by tripling the subsampling rate in each cluster.

e) Compute the coefficient of variation of the denominator based on the current design [from part (a)]. Remember to account for the cluster sampling design in your calculation. Is the Taylor series approximation adequate?

2. The following are cluster totals from five strata, with two primary stage selections per stratum, for a binary variable named "total cholesterol greater than 200" ($t_{y,h,i}$):

h	i	$t_{y,h,i}$	$t_{x,h,i}$
1	1	16	23
	2	15	25
2	1	9	17
	2	5	15
3	1	8	20
	2	10	21
4	1	6	16
	2	10	19
5	1	10	12
	2	7	16

a) Compute an estimate of the proportion with total cholesterol greater than 200, and its standard error (you can ignore the finite population corrections again in this case). Make sure that you are carefully accounting for this specific type of cluster sampling design in your variance estimation.

b) Give a 95% confidence interval for the proportion, making sure to use the correct degrees of freedom according to this design.

c) Compute the design effect and *roh* for the proportion in (a).

d) Estimate the standard error expected if the sample size were doubled by doubling the number of primary selections from two to four in each stratum.

e) Compute the coefficient of variation of the denominator. Remember to account for the stratified cluster sampling design in your calculation. Is the Taylor series approximation adequate?