SAMPLE SURVEYS, ST306G. EXAM
Department of statistics
Edgar Bueno
2020–11–02

General Instructions:

- Read carefully the enclosed instructions for exam submission. There you find all the necesssary information about submission, anonymous code, etc.

- For questions about the **content** of the exam, contact the course coordinator on email edgar.bueno@stat.su.se. Incoming e-mail questions are answered continuously during the exam.

- **Practical** help is only available during the **first hour** of the exam by email expedition@stat.su.se.

- If you, despite the instructions, have problems submitting the exam, email the exam to tenta@stat.su.se. However this is only done in exceptional cases.

- If the course coordinator needs to send out information to all students during the exam, this is done to your registered email address. Check your email during the exam.

- The exam should be solved individually.

- The exam is divided into two parts. The first part consists of eight multiple choice questions. In the second part you should estimate the indicated parameters using the provided sample data.

- You should submit pages 3 (solutions to the first part) and 5 (solutions to the second part).

- You should submit also another file showing how you obtained the estimates. For example, R code, Excel file or handwritten notes.

First part, Multiple choice. This part consists of eight multiple choice questions, each with four options and *one single correct answer.*

- The number of points granted in this part is given by $\max(0\,,\frac{25}{6}(a-2))$, where a is the number of right answers, for a maximum of 25 points.

- Please mark *clearly* your chosen option.

- Marking two or more options in the same question will invalidate the results for that question.

Second part, Estimation. In this part a sample is provided and you are asked to estimate several parameters. Each point estimate must be accompanied by the CVe (estimated coefficient of variation) and a 95% confidence interval.

Each correct estimate grants 1.25 points, for a maximum of 75 points.

You are free to use the "tool" you consider appropriate for obtaining the estimates (e.g. using R, SAS, Excel, calculator, etc.).

Grading criteria: Grading of the exam is according to the following table:

Points	0—10	11—50	51—60	61—70	71—80	81—90	91—100
Grade	F	Fx	E	D	C	B	A

Note: The following notation/abbreviations will be used: **i.** SRS = Simple random sampling without replacement; **ii.** $U[0\,,1)$ = Uniform distribution between 0 and 1; **iii.** $\hat{t}_{ra}$ = ratio estimator; **iv.** $\hat{t}_{ht}$ = Horvitz-Thompson estimator.

Part one. Multiple choice

1. Which of the following is **correct** regarding stratified SRS if the aim is to reduce the variance of the Horvitz-Thompson estimator:
   (a) If $S_{y,U_1} = S_{y,U_2}$, then the same sample size should be used in both strata.
   (b) If $S_{y,U_1} > S_{y,U_2}$, then a larger sample size should be used in the first stratum.
   (c) If $S_{y,U_1} < S_{y,U_2}$, then a larger sample size should be used in the first stratum.
   (d) If $S_{y,U_1} = S_{y,U_2}$, then a larger sample size should be used in the larger stratum.

2. A university conducts a survey about work environment. They want estimates for each department. Which of the following is **correct**:
   (a) Each department is a domain of study.
   (b) Each department is a stratum.
   (c) Obtaining results for men is not possible as this subset overlaps with the department.
   (d) Obtaining results for men is possible only if a stratum including all men in the population is created.

3. Which of the following sentences is **correct** regarding stratification:
   (a) Selecting all the elements from a stratum generates a nonsampling error known as "overcoverage".
   (b) If strata are well defined, it is a very efficient sampling design.
   (c) Sampling should be carried out independently in each stratum.
   (d) If the main aim of a survey is domain estimation, then the strata may overlap each other.

4. Which of the following sentences is **not** correct:
   (a) Nonresponse leads to nonresponse variance, which is a type of nonsampling error.
   (b) The term undercoverage refers to units that belong to the target population but are not included in the frame population.
   (c) The error due to a respondent declaring having voted for a party for which he did not vote, is called measurement error.
   (d) If a survey contains sensitive questions, the estimates may suffer from measurement errors.

5. Which of the following is **not** correct regarding SRS of size n from a population of size N:
   (a) It is unbiased for the estimation of totals.
   (b) The probability that the elements i and i' ($i \neq i'$) are simultaneously selected in the sample is $n(n-1)/(N(N-1))$.
   (c) It assigns a probability of $1/\binom{N}{n}$ to every sample of n different elements.
   (d) Let υ_i ($i = 1, \cdots, N$) be independent realizations from $U[0\,, 1)$. Selecting the n elements with largest υ values yields an SRS sample.

6. Which of the following is **not** a one-stage cluster sampling design:
   (a) From a list of teachers, select one among the first twenty in the list by SRS, then select every 20th teacher until the end of the list.
   (b) First, select all schools; then select a simple random sample of teachers in each school.
   (c) First, select a systematic sample of schools; then select all teachers in the selected schools.
   (d) First, select an SRS of schools; then select all teachers in the selected schools

7. Which of the following sentences is **not** correct regarding one-stage cluster sampling:
   (a) It is generally more cost-effective than SRS.
   (b) It is often used when a sampling frame of the elements in the population is not available.
   (c) The inclusion probability of an element coincides with the inclusion probability of the cluster it belongs to.
   (d) When the intraclass correlation coefficient is large, it will be more efficient than SRS.

8. Which of the following is **not** correct about estimation of totals assisted by an auxiliary variable fairly proportional to the study variable:
   (a) $\hat{t}_{ra}$ is expected to have a smaller variance than $\hat{t}_{ht}$.
   (b) $\hat{t}_{ra}$ is expected to have a smaller bias than $\hat{t}_{ht}$.
   (c) $\hat{t}_{ra}$ is expected to yield a shorter 95% confidence interval than $\hat{t}_{ht}$.
   (d) $\hat{t}_{ra}$ is expected to have a smaller mean squared error than $\hat{t}_{ht}$.

Part two. Estimation

A statistician is developing an image recognition program. To this end, she has access to a pool of 444 259 images. For each image, she needs to compute the value of a function and then find the total of such values. She is also interested in obtaining the total by type of image (Animals or Objects).

Computing the desired value for each image takes around one minute. Therefore, obtaining the totals of interest would take more than ten months. For this reason, the statistician has decided to select a sample of images and estimate the totals of interest based on the values observed in the sample. Due to varying availability of information regarding the images she decided to divide the pool of images into four groups and select independent samples in each group.

For each group, you should provide **i.** a point estimate of the total by type of image (Animals or Objects) and the overall total, **ii.** the corresponding estimated coefficient of variation (CVe) of the estimators, and, **iii.** a 95% confidence interval. Fill in your estimates in the table on page 4.

Note: The *csv* files have headers, columns are separated by commas and the decimal indicator is the point.

1. The first group consists of a set of 154 043 images that one of her colleagues is using in a current project. A simple random sample of 45 images was selected from this group. Although the statistician is not familiar with the images, her colleague suggests that the *degree of blue* (variable `blue` in the dataset) that he has found for each image may be well correlated to the function of interest. He says that the total degree of blue for the images in this group is 7 705 881. Use the regression estimator for obtaining the desired estimates. The observed data can be found in the file `group1.csv` or in the sheet *group1* of the Excel file `image.xlsx`.

2. The second group consists of a set of 88 414 images that her research assistant has been working with. A simple random sample of 132 images was selected from this group. Her assistant has classified the images into three groups according to the size in kilobytes of each image (variable `size` in the dataset, where 1, 2 and 3 denote "small", "medium" and "large", respectively). The assistant tells her that there are 17 147 images in the category "small", 53 022 in the category "medium" and 18 245 in the category "large". Use the poststratified estimator for obtaining the desired estimates. The observed data can be found in the file `group2.csv` or in the sheet *group2* of the Excel file `image.xlsx`.

3. The third group consists of a set of images that are available at four different sources on the internet (variable `source` in the dataset). Downloading the whole set of images would require a long time, therefore she decided to only download a simple random sample of images from each source as follows. 75 images out of 15 169 were downloaded from the first source; 100 out of 22 753 were downloaded from the second source; 125 out of 27 303 were downloaded from the third source; and 15 out of 3792 were downloaded from the fourth source. Use the Horvitz–Thompson estimator for obtaining the desired estimates. The observed data can be found in the file `group3.csv` or in the sheet *group3* of the Excel file `image.xlsx`.

4. The remaining images have not been digitized, they are in printed format and stored in 1661 folders. The researcher has decided to select a simple random sample of 7 folders, scan all the images in the selected folders, calculate the value of the function and register the type for each scanned image. Use the Horvitz–Thompson estimator for obtaining the desired estimates. The observed data can be found in the file `group4.csv` or in the sheet *group4* of the Excel file `image.xlsx`.

5. Estimate the total of the function of interest for the population of images (by type). Provide, also, the corresponding CVe and a 95% confidence interval for the total value of the function.

Group		Total	Animals	Objects
Group 1	$\hat{t}_y$ — CVe			
	95% CI			
Group 2	$\hat{t}_y$ — CVe			
	95% CI			
Group 3	$\hat{t}_y$ — CVe			
	95% CI			
Group 4	$\hat{t}_y$ — CVe			
	95% CI			
Total	$\hat{t}_y$ — CVe			
	95% CI			