Stockholm University

SAMPLE SURVEYS, ST306G. EXAM

Department of statistics
Edgar Bueno
2020–12–08

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.** SRSWR = Simple random sampling with replacement; **iii.** $U[0\,,1)$ = Uniform distribution between 0 and 1; **iv.** $\hat{t}_{ra}$ = ratio estimator; **v.** $\hat{t}_{ht}$ = Horvitz-Thompson estimator; **vi.** HT estimator = Horvitz-Thompson estimator.

Part one. Multiple choice

1. Which of the following sentences is **correct**:

   (a) In simple random sampling, the sampling unit and observation unit differ.

   (b) An SRSWR can be seen as selecting n independent samples of size one with elements selected with probabilities $1/N$.

   (c) Among two sampling strategies A and B, the one that is more precise shall be preferred.

   (d) The sample mean is an unbiased estimator for the population mean.

2. Which of the following is **correct** regarding nonresponse:

   (a) The bias due to nonresponse can be ignored only if the sampling design is SRS.

   (b) The main problem caused by nonresponse is potential bias.

   (c) Doubling the required sample size for a survey reduces the nonresponse bias approximately to half, i.e. $B(\hat{t}|2n) \approx 0.5\, B(\hat{t}|n)$.

   (d) The bias due to nonresponse can be ignored if the response rate is larger than 95%.

3. Which of the following is **correct** regarding SRS of size n from a population of size N:

   (a) The probability that the elements i and i' $(i \neq i')$ are simultaneously selected in the sample is n/N.

   (b) If coupled with the Horvitz-Thompson estimator, the resulting strategy is less efficient than SRSWR with the HT-estimator.

   (c) It is generally less cost-effective than cluster sampling.

   (d) If coupled with the ratio estimator, the resulting strategy is less efficient than SRSWR with the ratio estimator.

4. Which of the following is **not** correct about one-stage cluster sampling with HT estimator:

   (a) It usually has more precision per dollar spent than SRS with the HT estimator.

   (b) The inclusion probabilities are the same for all elements in the population.

   (c) It tends to be efficient when the variability between clusters is large.

   (d) The intraclass correlation coefficient is positive for most real-life populations.

5. Which of the following is **not** correct regarding SRS with the ratio estimator:

   (a) It satisfies $E(\hat{t}_{ra}) = E(t_x\, \hat{t}_y/\hat{t}_x) = E(t_x\, \hat{t}_y)/E(\hat{t}_x) = t_x\, t_y/t_x = t_y$.

   (b) Usually it offers a gain in precision compared to SRS with the HT-estimator.

   (c) Its variance can be approximated by Taylor series.

   (d) It is biased, although usually only slightly.

6. Which of the following is **not** correct regarding one-stage cluster sampling:

   (a) As the sampling units differ from the observation units, non-sampling errors are introduced in the estimation process.

   (b) The clusters can be selected by any sampling design.

   (c) If the clusters are of different sizes, they can be selected by SRS.

   (d) Analizing survey data from cluster sampling as if it was selected by SRS will usually underestimate the variances.

7. Which of the following is **not** correct regarding the strategy that couples SRS with the regression estimator:

   (a) It has a better chance of reducing nonrespose than the strategy that couples SRS with the ratio estimator.

   (b) It is biased, although usually only slightly.

   (c) It usually yields smaller variance than SRS with the Horvitz-Thompson estimator.

   (d) It usually yields smaller variance than SRS with the ratio estimator.

8. Which of the following is **not** correct regarding nonresponse:

   (a) Sources of nonresponse can be classified into survey content, methods of data collection and respondent characteristics.

   (b) Designed experiments can be used to discover causes of nonresponse before the survey is carried out.

   (c) A well-designed form for the respondent may reduce item nonresponse.

   (d) Telephone surveys usually yield larger response rates than in-person surveys.

Part two. Estimation

A statistician is developing an image recognition program. To this end, she has access to a pool of 378 283 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 eight 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 five 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 5.

The variable `type` in the dataset indicates the group, where 1 and 2 denote "Animals" and "Objects", respectively.

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 37 060 images that one of her colleagues is using in a current project. A simple random sample of 144 images was selected from this group. Although the statistician is not familiar with the images, her colleague suggests that the *degree of red* (variable `red` 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 red for the images in this group is 1 336 273. Use the ratio 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 fourth group consists of images that 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. The data for the fifth group is not available. However, the estimates for the whole population of images are available as shown in the row labeled “Total” in the following table. Using these estimates as well as those obtained for the other groups, estimate the total of the function of interest for the fifth group, as well as the corresponding CVe and a 95% confidence interval.

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						
Group 5	$\hat{t}_y$ — CVe						
	95% CI						
Total	$\hat{t}_y$ — CVe	203 476 249	2.74%	66 653 554	4.71%	136 822 695	3.87%
	95% CI	192 543 923	214 408 575	60 495 637	72 811 471	126 447 737	147 197 653