Stockholm University

SAMPLING AND ESTIMATION, ST720A. EXAM

Department of statistics
Edgar Bueno
2021–12–03

General Instructions:

i. The exam should be solved individually. **ii.** The exam is divided into two parts. The first part (first hour) consists of twelve multiple choice questions. In the second part (last four hours) you should estimate the indicated parameters using the provided sample data. **iii.** Write your anonymity code at the top of each page; **iv.** At the end of the first part, you should hand-in pages 3 and 4; **v.** At the end of the second part, you should hand-in pages 5 and 6; **vi.** You should submit also another file showing how you obtained the estimates (see below). For example, `R` code, Excel file or handwritten notes.

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

i. This part is closed book. The only aids allowed are blank paper, pen and a dictionary. **ii.** The number of points granted in this part is given by $\max(0\,, 4(a-3))$, where a is the number of right answers, for a maximum of 36 points. **iii.** Please mark *clearly* your chosen option. **iv.** Marking two or more options in the same question will invalidate the results for that question.

Second part, Estimation. In this part 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.

i. This part is open book. You are free to use also your own notes, lecture notes or any other material that you consider appropriate. **ii.** This part is computer based. You will be assigned a computer in the computer room. You can bring the files you consider appropriate. **iii.** The data required for this part can be downloaded from http://tenta.stat.su.se/tenta.html. **iv.** Once you have retrieved the data, access to internet will not be allowed any longer during the exam. **v.** Once you have handed in your exam to the invigilator go to http://tenta.stat.su.se/tenta.html again and upload a file showing how you obtained the estimates. **vi.** Points in this part are granted according to the following table (for a maximum of 64).

	Total			Domain 1			Domain 2		
Exercise	Point	CVe	CI	Point	CVe	CI	Point	CVe	CI
1	3	3	2	1	1	2	1	1	2
2	3	3	2	1	1	2	1	1	2
3	3	3	2	1	1	2	1	1	2
4	3	3	2	1	1	2	1	1	2

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

Part one. Multiple choice

1. Which of the following sentences is **not** correct regarding nonsampling errors:

   (a) One type of nonsampling error is called 'undercoverage'.

   (b) One type of nonsampling error is called 'goodness-of-fit of an estimator'.

   (c) If the sample size is large, the nonsampling errors are in general more serious than the sampling errors.

   (d) It is more difficult to estimate or assess the size of nonsampling errors than the size of sampling errors.

2. Which of the following sentences is **not** correct:

   (a) The term undercoverage refers to units that belong to the target population but are not included in the frame population.

   (b) If a survey contains sensitive questions, the estimates may suffer from measurement errors.

   (c) Nonresponse leads to nonresponse variance, which is a type of nonsampling error.

   (d) The error due to a respondent declaring having voted for a party for which he did not vote, is called measurement error.

3. Which of the following sentences is **not** correct:

   (a) A census is a survey in which the aim is to measure the entire population.

   (b) Register-based statistics use administrative registers to produce statistics.

   (c) The resulting statistical error due to the frame not including all the units in the target population is called sampling error.

   (d) A census may suffer from nonsampling errors.

4. Which of the following is **not correct** regarding a fixed-size sampling of clusters in the estimation of totals using the π estimator:

   (a) If y_k is constant, a simple random sample of clusters will have a variance equal to zero.

   (b) If $\pi_{Ii} \propto t_i$, the variance would be equal to zero.

   (c) If $\bar{y}_{U_i} = \bar{y}_U$ and $\pi_{Ii} \propto N_i$, the variance would be equal to zero.

   (d) If $\bar{y}_{U_i} \propto 1/N_i$ and π_{Ii} is constant, the variance would be equal to zero.

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

   (a) Nonprobability sampling usually yields nearly unbiased estimates.

   (b) The number of parameters of interest may be large, even in the thousands.

   (c) Usually, nonresponse errors are negligible in practice.

   (d) Carrying out a census implies that the estimates are error free.

6. Which of the following sentences is **correct** regarding stratified sampling:

   (a) Selecting all the elements from a stratum generates a nonsampling error known as "overcoverage".

   (b) Sampling should be carried out independently in each stratum.

   (c) If the main aim of a survey is domain estimation, then the strata may overlap each other.

   (d) If is always more efficient than simple random sampling.

7. Which of the following **is a reason** for implementing cluster sampling in a survey:

   (a) Data will be collected through telephone interviews in a national survey.

   (b) It allows for an efficient supervision of the field work.

   (c) A reliable sampling frame of elements is available.

   (d) Cluster sampling usually yields a small design effect.

8. Which of the following is **correct** regarding one-stage cluster sampling:

   (a) Every population element in the selected clusters is observed.

   (b) The clusters may overlap with each other.

   (c) The sample of clusters must be selected by a fixed-size sampling design.

   (d) Some elements may not belong to any cluster.

9. Under which of the following sampling designs is the sample mean an unbiased estimator of the population mean:

   (a) Under one-stage cluster sampling.

   (b) Under Bernoulli sampling.

   (c) Under Pareto πps sampling.

   (d) Under SI sampling.

10. Which of the following is **correct** regarding estimation of the finite population covariance between y and z, $S_{yz,U} = \frac{1}{N-1}\left(t_{yz} - \frac{t_y t_z}{N}\right)$:

    (a) If N is unknown, inserting π estimators in place of the unknown totals (including N) yields an unbiased estimator.

    (b) If N is known, inserting π estimators in place of the unknown totals yields an unbiased estimator.

    (c) If N is known, estimators other than the one in (b) have found to be unbiased.

    (d) Even if N is known, no unbiased estimators are available.

11. From a design-based approach, the coefficients $B_1, \cdots, B_q$ of a regression of y in terms of $z_1, \cdots, z_q$ can be interpreted as:

    (a) The most likely coefficients, given the observations, in a model of the type $y = \beta_1 z_1 + \cdots + \beta_q z_q + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.

    (b) The best linear unbiased estimates of the coefficients in a model of the type $y = \beta_1 z_1 + \cdots + \beta_q z_q + \epsilon$.

    (c) The coefficients of the best fitting plane of the type $B_1 z_1 + \cdots + B_q z_q$ to the y values.

    (d) None of the above.

12. Which of the following is **correct** with respect to the estimator of the median $\hat{M} = \hat{F}^{-1}(0.5)$ under a probability sampling design:

    (a) It is unbiased for the population median $M = F^{-1}(0.5)$.

    (b) An unbiased estimator of its variance is available.

    (c) The confidence interval is centered around the point estimate.

    (d) None of the above.

Part two. Estimation

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

1. A beverage company is planning to launch a new product in a city. In order to determine the potential market of the product they carry out a survey with the aim of estimating the total sales of a similar product, Toro Rosso, during the previous year as follows. A Pareto πps sample of size 50 from the 2370 grocery stores in the city is drawn with inclusion probabilities proportional to the number of employees of the store (variable `employees`). The observed sample can be found in the file `Sample1.csv` or in the sheet *Sample1* of the Excel file `Beverage.xlsx`. It is known that the total number of employees is 138 768. Use the sample data to estimate the total sales of Toro Rosso (variable `sales` in SEK) in the city using the π estimator. Estimate also the total sales by type of store (variable `type`, 1=chain store, 2=other). Estimate also the CV and a 95% confidence interval. (Fill in the values in the table below.)

	Total	Chain stores	Other
$\hat{t}_y$ — CVe			
95% CI			

2. The same company selected a Poisson sample of expected size equal to 200 from the grocery stores in another city with inclusion probabilities proportional to the number of employees of the store (variable `employees`). The observed sample can be found in the file `Sample2.csv` or in the sheet *Sample2* of the Excel file `Beverage.xlsx`. Use the sample data to estimate the total sales of Toro Rosso (variable `sales`) in the city using the ratio estimator with `employees` raised to the power 1.2 ($employees^{1.2}$) as the auxiliary variable. It is known that the total number of employees is 113 407 and the total number of $employees^{1.2}$ is 272 404. Estimate also the total sales by type of store (variable `type`, 1=chain store, 2=other). Each estimate must be accompanied by the estimated coefficient of variation and a 95% confidence interval. (Fill in the values in the table below.)

	Total	Chain stores	Other
$\hat{t}_y$ — CVe			
95% CI			

3. In a third city, the company selected a stratified sample from the population of grocery stores as follows. The stores were divided into four strata with respect to the number of employees: the 1107 stores with less than 14 employees make up the first stratum, the 545 stores having between 14 and 49 employees make up the second stratum, the 343 stores having between 50 and 107 employees make up the third stratum, the remaining 135 stores make up the fourth stratum. A simple random sample of 26 stores was drawn from the first stratum, 39 from the second, 39 from the third and 48 from the fourth stratum. The observed sample can be found in the file `Sample3.csv` or in the sheet *Sample3* of the Excel file `Beverage.xlsx`. Use the selected sample to estimate the ratio of the total sales (variable `sales`) over total number of employees (variable `employees`). Estimate also the ratio by type (variable `type`, 1=chain store, 2=other). Each estimate must be accompanied by the estimated coefficient of variation and a 95% confidence interval. (Fill in the values in the table below.)

	Total	Chain stores	Other
$\hat{R}$ — CVe			
95% CI			

4. The beverage company has drawn a Bernoulli sample with parameter $\lambda = 0.08197$ from the grocery stores in a fourth city. The observed sample can be found in the file `Sample4.csv` or in the sheet *Sample4* of the Excel file `Beverage.xlsx`. Use this sample to estimate the average sales per store in the city. Estimate also the average by type (variable `type`, 1=chain store, 2=other). Each estimate must be accompanied by the estimated coefficient of variation and a 95% confidence interval. (Fill in the values in the table below.)

	Total	Chain stores	Other
$\hat{\bar{y}}$ — CVe			
95% CI			