SAMPLING AND ESTIMATION, ST720A. EXAM

Department of statistics
Edgar Bueno
2021–10–29

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.** You should hand-in pages 2 to 5 (solutions to the first and second parts). **v.** You should submit (see below) 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 twelve multiple choice questions, each with four options and *one single correct answer*.

i. This part is closed book. **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	2	2	2	2	2	2
2	3	3	2	2	2	2	2	2	2
3	3	3	2	2	2	2	2	2	2
4	2	—	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:

   (a) A probability sampling design is one where $\pi_k > 0$ for all $k \in U$.

   (b) A sampling design is a probability distribution on the set of 2^N subsets of U.

   (c) An equal probability sampling design is one where $p(s)$ is constant for all samples s.

   (d) A measurable sampling design is one where $\pi_{kl} > 0$ for all $k, l \in U$.

2. Which of the following is **not** a frame imperfection:

   (a) Undercoverage.

   (b) Overcoverage.

   (c) Coding errors.

   (d) Duplicate listings.

3. Which of the following sentences is **not** correct regarding a design with a fixed size n:

   (a) $\sum_U \pi_k = n$.

   (b) $\sum_U \sum_U \pi_{kl} = n^2$.

   (c) $\sum_{l \in U} \pi_{kl} = (n-1)\pi_k$.

   (d) $\sum_U \sum_U \Delta_{kl} = 0$

4. Let $0 < \lambda_k \leq 1$ for all $k \in U$. Which of the following is **not** correct:

   (a) Under Systematic πps(λ) sampling, the estimator $\sum_s y_k/\lambda_k$ is unbiased for the total.

   (b) Under Poisson sampling $PO(\lambda)$, the estimator $\sum_s y_k/\lambda_k$ is unbiased for the total $\sum_U y_k$.

   (c) Under Pareto πps(λ) sampling, the estimator $\sum_s y_k/\lambda_k$ is unbiased for the total $\sum_U y_k$.

   (d) $\sum_U \lambda_k \leq N$.

5. Which of the following is **not** correct about systematic sampling in "its basic form":

   (a) It can be seen as a particular case of cluster sampling.

   (b) It is a measurable sampling design.

   (c) $V_{SY}(\hat{t}_\pi) < V_{SI}(\hat{t}_\pi)$ when the intraclass correlation coefficient is small.

   (d) It is a random size sampling design.

6. Which of the following is **not** correct about Poisson sampling, $PO(\lambda)$:

   (a) It is a random size sampling design.

   (b) If the λ values are proportional to the y values, then $V_{PO}(\hat{t}_\pi) = 0$.

   (c) The second order inclusion probabilities are given by $\pi_{kl} = \lambda_k \lambda_l$.

   (d) It is a strict πps sampling design.

7. Which of the following is **not** correct regarding simple random sampling of clusters, SIC:

   (a) It is a fixed size sampling design.

   (b) If the cluster totals t_i are constant, the strategy $(SIC, \hat{t}_\pi)$ is an efficient choice.

   (c) It is generally less costly than SI.

   (d) It is generally true that $V_{SIC}(\hat{t}_\pi) > V_{SI}(\hat{t}_\pi)$.

8. Which of the following is **not necessarily** correct regarding the π-estimator under a probability sampling design:

   (a) It is unbiased for the total $\sum_U y_k$.

   (b) Its variance is given by $\sum_U \sum_U \Delta_{kl} y_k y_l / (\pi_k \pi_l)$.

   (c) If $\pi_{kl} > 0$ for all $k, l \in U$, its variance is unbiasedly estimated by $\sum_s \sum_s (\Delta_{kl}/\pi_{kl}) y_k y_l / (\pi_k \pi_l)$.

   (d) If $\pi_{kl} > 0$ for all $k, l \in U$, its variance is unbiasedly estimated by $-(1/2) \sum_s \sum_s (\Delta_{kl}/\pi_{kl})(y_k/\pi_k - y_l/\pi_l)^2$.

9. Consider the estimators of the population mean $\hat{\bar{y}}_{U\pi} = \hat{t}_{y\pi}/N$ and $\tilde{y}_s = \hat{t}_{y\pi}/\hat{N}$ with $\hat{N} = \sum_s 1/\pi_k$. Which of the following is **correct**:

   (a) If the population size N is known, $\hat{\bar{y}}_{U\pi}$ must be preferred over $\tilde{y}_s$.

   (b) If the population size N is unknown, $\hat{\bar{y}}_{U\pi}$ must be preferred over $\tilde{y}_s$.

   (c) Often, we have $V_p(\tilde{y}_s) < V_p(\hat{\bar{y}}_{U\pi})$.

   (d) Both estimators are unbiased for $\bar{y}_U$.

10. Which of the following is **correct** regarding the difference estimator under a probability sampling design:

    (a) It is unbiased for the total $\sum_U y_k$.

    (b) It is unbiased if and only if all proxy values y_k^0 are positive.

    (c) If the proxy values y_k^0 largely differ from the y values, its bias will be large.

    (d) If the proxy values y_k^0 are fairly proportional to the y values, its variance will be small.

11. Which of the following is **correct** about the design effect:

    (a) It is the ratio of the variance of a sampling strategy compared to the variance of the π estimator under SI.

    (b) It is the ratio of the variance of a sampling strategy using SI compared to the variance of the π estimator under SI.

    (c) It is the ratio of the variance of a sampling strategy using the π estimator compared to the variance of the π estimator under SI.

    (d) None of the above.

12. Which of the following is **correct** regarding STSI if the aim is to reduce the variance of the π estimator:

    (a) If $S_{y,U_1} > S_{y,U_2}$, then a larger sample size should be used in the first stratum.

    (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 the same sample size should be used in both strata.

    (d) If $S_{y,U_1} = S_{y,U_2}$, then a larger sample size should be used in the larger stratum.

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 real estate company has selected a Bernoulli sample of expected size equal to 100 from the 74 225 properties in a city. The observed sample can be found in the file `Sample1.csv` or in the sheet *Sample1* of the Excel file `REstate.xlsx`. Use the sample data to estimate the total real estate value (variable `value`, in SEK) of the properties in the city using the π estimator. Estimate also the total value by type of property (variable `type`, 1=apartment, 2=other). Estimate also the CV and a 95% confidence interval. (Fill in the values in the table below.)

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

2. The same company selected a Poisson sample of expected size equal to 200 from the properties in another city with inclusion probabilities proportional to the square root of the area of the property (the variable `area` indicates the area of each property in the sample in square meters). The observed sample can be found in the file `Sample2.csv` or in the sheet *Sample2* of the Excel file REstate.xlsx. It is known that the total area of the apartments is 11 478 961, the total area of the other type of properties is 2 471 918, the total of the square root of the area of the apartments is 1 269 350 and the total of the square root of the area of the other type of properties is 288 488. Use the sample data to estimate the total real estate value (variable `value`) of the properties in the city using the ratio estimator with area as the auxiliary variable. Estimate also the total value by type of property (variable `type`, 1=apartment, 2=other). Estimate also the CV and a 95% confidence interval. (Fill in the values in the table below.)

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

3. In a third city, the company selected a Pareto πps sample of size 50 with λ values proportional to the square root of the area of the property (the variable `area` indicates the area of each property in the sample in square meters). The observed sample can be found in the file `Sample3.csv` or in the sheet *Sample3* of the Excel file `REstate.xlsx`. It is known that the total of the square root of the area is 866 370. Use the sample data to estimate the average value per square meter i.e. the ratio of `value` over `area` for the whole city and by type of property (variable `type`, 1=apartment, 2=other). Estimate also the CV and a 95% confidence interval. (Fill in the values in the table below.)

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

4. The real estate company has selected a simple random sample of 150 out of the 61 195 properties in a fourth city. The observed sample can be found in the file `Sample4.csv` or in the sheet *Sample4* of the Excel file `REstate.xlsx`. Use this sample to estimate the median of the value for the whole city. Find also a 95% confidence interval. (Fill in the values in the table below).

	Total
$\hat{\breve{y}}$	
95% CI	