

B. Exercises

You should be prepared to present the exercises and at least an attempt to solve them. A complete solution is not required but you should at least have read, understood and attempted to solve them before going to the lecture.

For February 7, 2012

11. A population consists of 7 000 businesses and one intends to draw a sample of 500 enterprises. One intends to stratify after size or industry, but does not know what is best. From the frame and previous studies the following numbers (exact) and guesstimated values.

After industry

Industry	1	2	3	4
Numbers	2000	3000	1500	500
Likely mean value	50	30	100	400
Likely standard dev.	200	100	700	2000

After size

Number of employed	1-10	11-50	51-100	101 -
Number	3000	3500	450	50
Likely mean value	10	27	200	6500
Likely standard dev.	7	20	100	4000

- Derive the optimal sample sizes in both cases. (The costs per company are assumed to be equal regardless of size).
- Compute also the expected variance of the total estimate and decide which stratification that is best.
- Compute the design effects in both cases by comparing the expected variance with that of a simple random sample of the same size. (deff is defined as the ratio of the variances, with that of SRS in the denominator)

12. A non anonymous course evaluation was done on the first course in statistics with 190 registered students. In total only 57 answers were handed in. One of the questions, Y_i , dealt with total working time on the course. At the time of the analysis the result at the first test, X_i , is also known for all 190. Of the 190 registered students only 121 got a result. (the remaining 69 did not take the test at the first opportunity) and 46 of these 121 had also replied to the evaluation.

a) Mean working time for the 46 who did the test was 32 h/week and their average test result was 67 points. The mean value for all 121 who made the test was 58. A linear regression between the result and working time gave the regression line $Y_i = 12,0 + 0,30 \cdot X_i + \varepsilon_i$, where the residual variance is estimated to 64. Make a regression estimate of the mean working time for all doing the test. (Assume that the sample can be considered to be a SRS from all taking the test) and estimate the standard deviation.

b) Of the 69 students without a test result 11 replied. Their observed mean working time was 16,8 h/week and the sample variance was 169. Estimate the mean working time among all not taking the test and the standard error of the estimate. (Assume that the sample can be viewed as SRS).

c) Estimate the average working time for all students and the standard error.

13. There are 5 900 establishments in a certain industry in a business register (2 500 with less than 10 employees, 2 000 with 11- 50 employees, 1 000 with 51 – 250 employees, 300 with 251 – 1000 employees and 100 with more than 1000 employees. One intends to draw a sample in order to study a variable whose variances in the five strata are judged to be 10, 50, 1000, 12 000 and 100 000.

a) Determine the stratum sizes in a sample of 25 establishments using optimal allocation.

b) Do the sampling and give the numbers of the selected establishments. (The establishments are numbered from 1 to 5 900 after size in increasing order. (Tables of random numbers can be found in many statistical textbooks e.g. Lohr. You can also find them using a computer program)

c) The observed means in the five strata are 2.24 Mkr, 25.7 Mkr, 243.2 Mkr, 792.2 Mkr resp 3 920,4 Mkr. Observed mean square errors ($\sum_{sh} (y_i - \bar{y})^2 / n_h$) are 15.72, 43.12, 842.31, 15787.12 and 84002.98 Mkr². Estimate the total and the standard error of this estimate.