UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Exam in: STK4600 Statistical methods for social sciences
Day of exam: Thursday, December 12, 2013
Exam hours: 14.30 – 18.30
This examination paper consists of 3 pages.
Appendices: None
Permitted materials: Lecture notes, student's own notes and approved calculator

Make sure that your copy of this examination paper
is complete before answering.

Exercise 1
We shall draw a sample of physicians (medical doctors) from the register of the " Den norske lægeforening" to assess the total use of antibiotics in 1999. The register consisted of 14841 physicians . The physicians are in various occupational categories as shown in table 1 below.

Table 1

Position/ occupation	Hospital	Private practise	Academic/ administrative	Unknown	Total
Number of physicicans	7771	4861	854	1355	14841

a) Explain why it is sensible to take a stratified sample after occupational category. Under what conditions will proportional allocation give minimum variance for the stratified estimator?

b) Assume you shall select a stratified simple random sample of $n = 1000$, with strata being the occupational groups in table 1 above. How would you allocate the sample units to the different strata to obtain proportional allocation? What are the inclusion probabilities in this case?

The variable of interest for a physician, y_i, is the total amount of antibiotics discharged during one year, measured in units of 100 " daily doses". We make the following guesses regarding the population stratum totals t_h and stratum variances σ_h^2 :

Table 2

Position/ occupation	Hospital	Private practise	Academic/ administrative	Unknown	Total
Totals t_h	100 000	118 000	2000	16 000	236 000
Variances σ_h^2	25	100	1	25	

c) Compute the variance and standard error of the stratified estimator, with proportional allocation, for the total t and the population mean t/N, assuming the values in the table above are correct.

Continued on page 2

d) From table 2, compute the stratum means $\mu_h = t_h / N_h$, and use these values together with the values of σ_h^2 to compute the total population variance σ^2. Assume the sample is a simple random sample of size n = 1000, without stratification. Find the variance of the sample mean based estimator $N\bar{y}_s$. How large must the sample be for this estimator to be as good as the stratified estimator in part c) based on a sample of n = 1000?

e) Assume again that the values in table 2 are correct. Determine the allocation of the 1000 physicians in the sample such that the stratified estimator for total amount of antibiotics has minimum variance.

f) All the prescriptions discharged by the physicians in the sample can be regarded as a sample from the population of all prescriptions discharged in this year. What kind of sample of prescriptions is this and is it self-weighting for the following sample design for the physicians

I. Proportional allocation of the physicians

II. Optimal allocation of the physicians as in part d)

A stratified random sample of 1000 physicians with proportional allocation gives the following results:

Table 3

Stratum	Hospital	Private practise	Academic/ administrative	Unknown
sample mean $\bar{y}_h$	11.8	22.10	3.00	13.10
stratum sample variance s_h^2	4.3^2	7.1^2	0.5^2	4.0^2

g) Derive an estimate for the mean antibiotics use per physician and find the standard error of the estimate.

h) Give a 95% confidence interval for the mean antibiotics use per physician and give an interpretation of the interval.

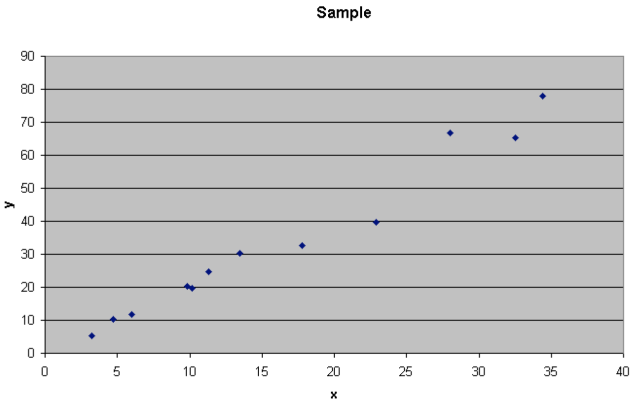
Exercise 2

Suppose, for a certain population U, with N = 234 units, we intend to estimate a total $t = \sum_{i=1}^{N} y_i$, and we take a simple random sample of size n = 12 from U. Assume that no information on an auxiliary variable is available at the time of sampling. By the time we have the data in hand and are ready to estimate T, the information on a variable x has become available: Its population mean is $\bar{x} = 18.2$ and the sample data on x as well as y is available:

Sample unit	1	2	3	4	5	6	7	8	9	10	11	12
x	3.2	4.7	6.0	9.8	10.2	11.3	13.5	17.8	22.9	28.0	32.5	34.4
y	5.3	10.2	11.8	20.3	19.7	24.7	30.3	32.6	39.6	66.7	65.4	78.0

We shall consider the y-values as realized values of random variables Y. The scatter plot is shown below

Continued on page 3



Consider the following 3 predictors of the total T:

Expansion estimator: $\hat{T}_{exp} = N \cdot \overline{Y}_s$, with $\overline{Y}_s = \sum_{i\in s} Y_i \ / n$

Ratio estimator: $\hat{T}_R = t_x \cdot \dfrac{\overline{Y}_s}{\overline{x}_s}$ where $\overline{x}_s = \sum_{i\in s} x_i / n$

Regression estimator: $\hat{T}_{reg} = N \cdot \overline{Y}_s + N\hat{\beta}_2(\overline{x} - \overline{x}_s)$, where $\hat{\beta}_2 = \dfrac{\sum_{i\in s}(x_i - \overline{x}_s)Y_i}{\sum_{i\in s}(x_i - \overline{x}_s)^2}$.

a) Each estimator above can be expressed on the form $\hat{T} = \sum_{i\in s} Y_i + \sum_{i\notin s} \hat{Y}_i$ where $\hat{Y}_i$ is a prediction for the value y_i. Derive $\hat{Y}_i$ for the three estimators.

b) For which population models are these three estimators the BLU predictor? Which of these models would you use for these data?

c) Compute the estimates for the total T from these 3 estimators. Which of these estimates do you think is closest to T? Here you can use that $\sum_{i\in s} x_i Y_i = 9387.16$, $\sum_{i\in s} x_i^2 = 4443.21$.

d) Suppose we have missing data on the y-values for units 1, 9 and 12 in the sample, but we know the x-values for these units. Suggest ways to impute the y-values for these units based on the three different models in part b).

END