UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Exam in: STK4600 Statistical methods for social sciences.
Day of exam: Thursday June 9, 2016
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

Assume the tax authorities (Skatteetaten) want to audit income tax returns by taking a stratified simple random sample of 200 returns from a certain district in Norway of a total of 10 000 returns. The strata are groups of returns showing similar adjusted gross incomes. In the population of 10000 returns the strata are:

Stratum	Size of stratum	Sample	Observed proportions of frauds
1: under kr. 300 000	5000		0.030
2: 300 000 – 700 000	4000		0.050
3: Over 700 000	1000		0.200
		200	

a) The main task is to discover frauds. Why do you think it is a good thing to stratify according to gross incomes instead of taking a simple random sample? A proportional allocation is decided. How many tax returns should be selected from the different strata?
b) Assume we use proportional allocation and get the data in the table. Estimate the proportion of frauds in the district, calculate the standard error and the 95% confidence interval.
c) What is the statistical interpretation of the 95% confidence interval calculated in part b. How do you interpret the 95% coverage?
d) Suppose it is expected that the proportion of frauds in the 3 strata are around 1, 5, and 15 percent respectively. In estimating the total number of frauds how should the allocation be in order to minimize the variance of the stratified estimator for the proportion of frauds in the district?
e) Suppose we find that the observed proportions of frauds with optimal allocation are 0.034, 0.069 and 0.171. Derive the estimate, standard error and 95% confidence interval.
f) Assume the true proportions of frauds in the 3 strata are 0.01, 0.05 and 0.15 for strata 1, 2 and 3 respectively. Derive the true variances for the stratified estimator of the total proportion of frauds for optimal allocation and proportional allocation.

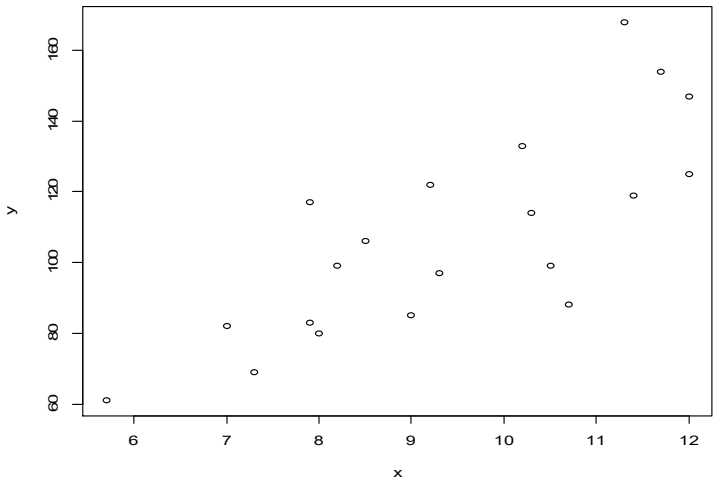
Continued on page 2

Exercise 2

Foresters want to estimate the average age of trees in a stand. Determining age is cumbersome because one needs to count the tree rings on a core taken from the tree. In general, though, the older the tree, the larger the diameter, and diameter is easy to measure. The foresters measure the diameter in all 1132 trees and find that the population mean is 10.3. They randomly select 20 trees for age measurement with the following results:

Tree no.	Diameter, x	Age, y		Tree no.	Diameter, x	Age, y
1	12.0	125		11	5.7	61
2	11.4	119		12	8.0	80
3	7.9	83		13	10.3	114
4	9.0	85		14	12.0	147
5	10.5	99		15	9.2	122
6	7.9	117		16	8.5	106
7	7.3	69		17	7.0	82
8	10.2	133		18	10.7	88
9	11.7	154		19	9.3	97
10	11.3	168		20	8.2	99

A scatter plot of the data:



a) Consider the following two possible population models for uncorrelated Y_i's :

Model 1: $E(Y_i) = \beta x_i$ and $Var(Y_i) = \sigma^2 x_i$.
Model 2: $E(Y_i) = \beta x_i$ and $Var(Y_i) = \sigma^2$

Looking at the scatter plot, which of the two models do you think describes the data best?

We have the following values of various statistics that you can use in answering the questions in this exercise:

Continued on page 3

- Mean of x in the sample is 9.4
- Mean of y in the sample is 107.4
- The estimated σ^2 in model 1 is equal to 32.30
- The estimated σ^2 in model 2 is equal to 321.85
- The sum of x^2 in the sample is 1832.63
- The sample variance of y equals 821.52
- The sum of $(y \cdot x)$ is 20980.4
- The sum of the squares of (y/x) equals 2656.8

b) Estimate the mean age based on model 1using the BLU estimator. Compute the model-based standard error and an approximate 95% model-based confidence interval.

c) Estimate the mean age based on model 2 using the BLU estimator. Compute the model-based standard error and an approximate 95% model-based confidence interval.

d) Under what condition is the sample mean based estimator model-unbiased under both models 1 and 2. Is the sample mean based estimator then the BLU estimator under any of the two models? Give a reason for your answer.

e) Compare the estimates and standard errors in b) and c) with the sample-mean based estimate and its model-based standard error under the model 2 without any x-variable, that is, $E(Y_i) = \beta$. What can you conclude regarding the value of including x as an auxiliary variable in the estimation?

Assume now that we want to estimate the population mean $\overline{R}$ of the ratio age to diameter, $R = Y/x$.

f) Under what model for Y is the BLU estimator for $\overline{R}$ given by

$$\hat{\overline{R}} = \frac{1}{n}\sum_{i \in s}(Y_i / x_i) \ .$$

g) Under the model you found in part f), derive the model-based standard error of the estimator in part f) and compute the values of SE and the 95% confidence intervals for $\overline{R}$.

h) When will $\overline{R} = \overline{Y} / \overline{x}$? Use the estimates in parts b) and c) to estimate $\overline{Y} / \overline{x}$. Compare these estimates and 95% confidence intervals with the results in part f).