# Assignment 1 (survey sampling)

## Data

The data set *families* contains information about 43886 families living in the city of Cyberville. The city has four regions:

1. The Northern region has 10149 families.
2. The Eastern region has 10390 families.
3. The Southern region has 13457 families.
4. The Western region has 9890 families.

For every family, the following information is recorded:

- Family type
- Size of the family
- Number of children in the family
- Family income
- Region ( i.e 1 corresponds to North as described above)
- Education level of the head of the household.

In the following exercises you will try to learn about the families of Cyberville by using sampling. For every correct question you get 0.1 points.

## Exercises

**a.** Take a simple sample of 600 families. Estimate the following population parameters, calculate the estimated standard error of these estimates and form 95% confidence intervals:

i. The proportion of husband-wife family (family type = 1 in the data).

ii. The average number of children per family.

iii. The average number of persons per family.

**b.** Take 100 samples of size 400.

i. For each sample, find the average family income.

ii. Find the average and standard deviation of these 100 estimates and make a histogram of the estimates.

iii. Superimpose a plot of a normal density with that mean and standard deviation of the histogram and comment on how well it appears to fit.

iv. For each of the 100 samples, find a 95% confidence interval for the population average income. How many of those intervals actually contain the population target.

v. Take 100 samples of size 100. Compare the averages, standard deviations and histograms to those obtained for a sample of size 400 and explain how the theory of simple random sampling relates to the comparisons.

**c.** Stratify the families into four strata by region (North, East, South and West).

i. What are the sampling fractions for proportional allocation and optimal allocation?

ii. Allocate 500 observations proportionally to the four regions and estimate the average income from the stratified sample. Estimate the standard error and form a 95% confidence interval. Compare your results to the results of the simple random sample.