**Assignment Survey data Analysis – Traffic Data**

**Questions? Stuck? E-mail:** [P.lugtig@uu.nl](mailto:P.lugtig@uu.nl)

The final assignment involves doing survey data analysis for statistical inferences, and preparing a presentation where you:

* Summarize the research topic
* Present the (statistical) findings following the research questions as stated in the assignment
* Aim for a discussion on the topic

The grade will take these 3 elements into account and the R-file displaying how the computations were carried out. You will work in a group.

The analysis need to be carried out in R using the survey package. Output are to be reported in a brief scientific way and should include a description of the results and of the required computations. Hand in (by email), the short scientific report of the survey analysis, the R-code and the presentation. Different sampling strategies (e.g., SRS, Stratified, Cluster sampling) need to be described, in detail. Include descriptions of your frame and target and population, intended sample size, and method of sample selection. Discuss potential non-sampling, sampling, and non-response errors.

* The scientific report should be maximally 10 pages (Fontsize 12, single spaced)
* The presentation should take maximally 12 minutes, after which there will be 8 minutes for questions and discussion.
* There is **no limit** on your R code. Do include comments in your code to help your teachers understand what you did.

--------

**The data and assignment**

You will work with a population dataset that contains all incoming and outgoing car traffic into the city of Utrecht of two weeks in January and February 2017 (23 January – 5 Feb). You can find the data in the files:

“traffic.csv.zip” (unzip the file and open the csv in R)

The file is openly available from <https://utrecht.dataplatform.nl/data?page=7&res_format%5B0%5D=CSV>

Note that the data in the files are slightly different. Some pre-processing has been done (such as the addition of English variable labels). Value labels can be found in the word-document “Legenda Utrecht-hb.docx”. Also, data from multiple days have been merged into 1 dataset.

The datafile contains 9951859 records, so will take some time to load into R. Each record in the file contains one observation of a vehicle passing by a camera. Cameras are placed at all the main roads in Utrecht (so they offer a representative picture of “traffic” in the city. It could well be that individual cars have multiple observations in the dataset (e.g. if they drive around the city), but due to privacy reasons, we don’t know whether this is the case. The file contains the following variables:

* Country: country of origin of the car passing by (‘Netherlands’, or ‘other’)
* Class: Type of vehicle (see file “legenda(…)” for more info)
* Camera: camera no. (in total there are 99 cameras at the main roads of Utrecht)
* Date: the data in format “ddmmyyyy”
* Time: Local time in Utrecht measures in hours [range 00-24]
* Noise: registered noise of the vehicle passing by measured in Decibel
* Noiseunits: The city council is interested not only knowing how many vehicles pass by, but also how much noise they produce. A convenient method to do this is by using a variables that weights every vehicle by the amount of noise they make. On average vehicles have a noise unit of 1. Vehicles producing twice as much noise as the average have a noise unit level of 1.1 (as with Db, the scale is logarithmic)

Utrecht city council is interested in estimating the **total amount of traffic** and **noise** in the city of Utrecht permanently, but doing this every day of the year is cumbersome. They would at most like to sample 50000 cases over a two-week period in one-hour periods of observation. So, if they do observations at any one camera, they would like to so for at least 1 full hour. Utrecht city council can turn cameras on or off for period of 1 hour; data will then be streamed automatically to a database

Your task in this assignment is to set up a sampling plan based on the population data for the two weeks you have available. Focus on the variable “ noiseunits” for estimating the total amount of traffic over the 2-week period using the population data. Remember that vehicles on average have a value of 1 on this variable, so this is good proxy for the number of vehicles. It could of course be that traffic is very different in say summer than when compared to February, but lets for now assume that the data in the file “ traffic” contain the actual population data.

Take the following steps in designing a sampling plan

1. Screen the data. Note that because the dataset is large, you may want to screen samples of the data. Screen for **outliers**, **missings**, etc.Perhaps you want to recode variables. Report shortly on your findings, and how you solved any issues
2. What would happen if you would select **50000** cases from **one random day** out of **14** to estimate the total “ noiseunits” over the 2-week period? Please estimate the **survey total**, as well as the **standard error**
3. What would happen if we would **randomly select one particular Camera**, and sample **50000** cases? What would by the estimate of the **population mean**, and **standard error**?
4. What would happen if we would **randomly select one particular hour of the day**, and sample **50000** cases? What would by the estimate of the **population mean, and standard error**?

All three variables (day, time and Camera) can be thought of as potential variables to stratify on. We can oversample from particular times of the day, days, locations, or combinations of these strata.

1. Investigate the variation in “noiseunits” within and between both strata in the population dataset. What is your conclusion about the necessity to stratify if the goal is to estimate the total noiseunits with the least amount of bias, not caring about precision?

Tip: the function **aggregate** may be helpful

1. (related to Q5.) And what if the goal was to estimate the population total for noiseunits both with the least amount of bias and the highest precision? Please to not actually design the sample, or evaluate the precision, but explain how you would stratify the sample. Note that there are many possible strata, too many if we want to sample approximately 50000 cases . A good strategy to decide what strata are useful would be
2. Inspect the between-strata variance of Time,day and Location, including interactions.
3. Group the data into a useful number of strata (e.g. 5-10).
4. Use the design that you have explained in the previous question and estimate the population total and precision, again with a sample of 50000 as a maximum. Please use an estimator that in **theory is unbiased**, so that you know that the population total is unbiased. Evaluate the precision by drawing 1 sample only, and then estimate the standard error.
5. Imagine that the council would not have a fixed budget, but would rather have a level of precision as determined by the **Coefficient of Variation of .005**. Under the sampling design you specified under Q6-7, how large should the sample be to achieve this? Use the same strata as in Q6-7.