Final assignment

The final assignment involves some reading on a specific topic, 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 groups of three and may choose your own partner(s). All students must present part of the slides (prepare everything together but decide beforehand who will do which part of the presentation).

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. Different sampling strategies (SRS, Stratified, Cluster sampling) need to be described. 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.

Description

The 2010 National Hospital Ambulatory Medical Care Survey (NHAMCS) is a national sample survey concerning visits to hospital outpatient and emergency departments, conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention (http://www.nber.org/data/national-hospital-ambulatory-medical-care-survey.html).

The survey is a component of the National Health Care Surveys, which measure health care utilization across a variety of health care providers. The micro-data file “nhamcsed2010-short” concerns the emergency department records with a limited number of variables. The file contains a total of n=34,936 visits by patients. The following variable information were stored.

Variable Name	Description
Month of visit [VMONTH]	Month of visit 01-12: January-December
Day of week of visit [VDAYR]	1=Sunday 2=Monday 3=Tuesday 4=Wednesday 5=Thursday 6=Friday 7=Saturday
Age patient [AGE]	0=Under 1 year, 1-99, 100=100 years and over
Arrival Time [ARRTIME]	-9=Missing, 00.00-23.59

Waiting Time [WAITTIME]	In minutes. Time of ED arrival, and time seen by MD/DO/PA/NP (Doctor of Medicine, Doctor of Osteopathy, Physician Assistant, and Nurse Practitioner) -9=missing -7=Not applicable (Not seen by MD/DO/PA/NP)
Length of Visit [LOV]	-9=Missing
Residence Patient [RESIDNCE]	-9 = Missing -8 = Unknown 1 = Private residence 2 = Nursing home 3 = Homeless 4 = Other
SEX	1 = Female 2 = Male
ETHNICITY [ETHUN]	-9= Blank 1 = Hispanic or Latino 2 = Not Hispanic or Latino
Race [RACEUN]	-9 = Missing 1 = White 2 = Black/African American 3 = Asian 4 = Native Hawaiian/Other Pacific Islander 5 = American Indian/Alaska Native 6 = More than one race reported
Arrival by Ambulance [ARREMS]	-9 = Blank -8 = Unknown 1 = Yes 2 = No
Expected source of payment [PAYPRIV]	0 = No 1 = Yes
Expected source of payment medicare [PAYMCARE]	0 = No 1 = Yes
Patient visit weight [PATWT]	Survey weight
Region [REGION] (Based on actual location of the hospital.)	1 = Northeast 2 = Midwest 3 = South 4 = West
[CSTRATM] Clustered PSU Stratum Marker	20108201-40400000 (note: content is masked)
[CPSUM] Clustered PSU Marker	5-100331 (note: content is masked)

The patient visit weight can be used to produce national estimates from the sample data. The data reflect a sample of all patient visits. When aggregating the patient visit weights, the n=34,936 sample records for 2010 represent the grand total of N=129,843,377 estimated visits made by all patients to Emergency Departments (EDs) in the United States. There are missing values in the recorded waiting times.

1. a. The variable CPSUM represents geographic primary sampling units (clusters), a variable CSTRATM represents relevant strata (e.g., emergency service areas within emergency departments or outpatient departments), variable PATWT represents an additional patient visit weight variable.

    a. Explain this multistage sampling design. Do the PATWT weights sum up to N, and what does this mean?

b. Define the svydesign() function with clusters, strata, and (additional) weights.
c. When considering gender differences in emergency visits, is it more likely for men than woman to visit an emergency department, explain?
d. Is there a mean difference between males and females in weights (PATWT), what do you conclude from this?

2. a. Estimate the average waiting time in the population under the specified svydesign. What do you conclude?

b. Exclude the cluster and strata information, (redefine the svydesign), how this change your result in 2a, explain?

c. Draw an SRS sample to estimate the average waiting time. Estimate the average waiting time, and compare the result(s) with the result(s) of 2a and 2b.

d. Draw a stratified sample to investigate the variability in waiting times over regions (Northeast, Midwest, South, West). Is there variability in waiting times over regions?

3. a. Explain that a ratio estimator can be used to improve the estimation of the average waiting time.
b. Describe potential variables for ratio estimation, and choose one to compute a ratio estimate of the average weighting time.
c. Give a 95% confidence interval for your preferable estimate of the average waiting time, and explain the result.

4. Investigate differences in waiting times among ethnic groups by fitting a linear regression model. Compare estimated results of a stratified sampling design and an unequal probability sampling design.

5. Assume unequal probability sampling, can you identify the important predictor variables of waiting times of patients visiting an ED? What can be concluded?