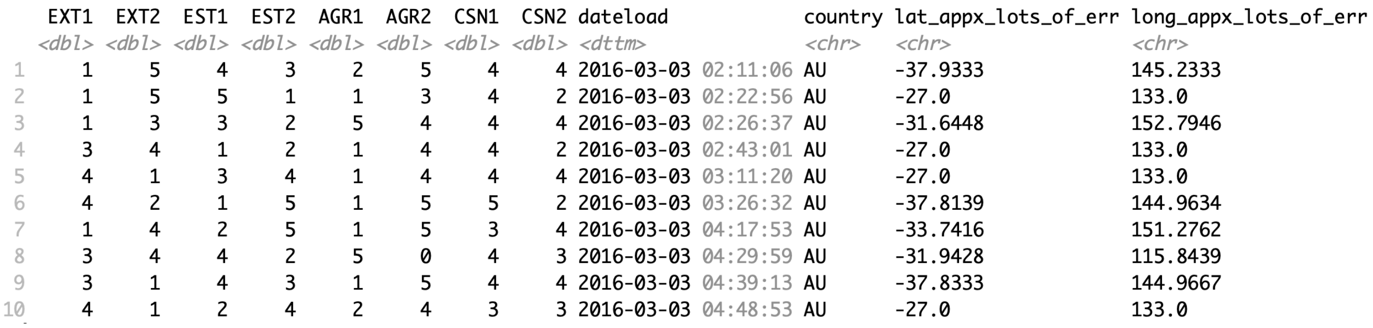
Midterm Project of MATH3426, Spring 2020

Dataset: *Big Five Personality Test*

The Big Five personality traits, also known as the five-factor model (FFM) and the OCEAN model, is a taxonomy, or grouping, for personality traits. When factor analysis (a statistical technique) is applied to personality survey data, some words used to describe aspects of personality are often applied to the same person. For example, someone described as conscientious is more likely to be described as "always prepared" rather than "messy". This theory is based therefore on the association between words but not on neuropsychological experiments. This theory uses descriptors of common language and therefore suggests five broad dimensions commonly used to describe the human personality and psyche.

This original dataset contains 1,015,342 questionnaire answers collected online by Open Psychometrics. Each participant needs to answer around 100 questions, Information about the participant is also recorded.

In our analysis, we use a smaller version of the original dataset which contains 335,429 questionnaire answers and each participant answered 8 questions. Below is a snapshot of the first 10 rows of our data.



```
    EXT1  EXT2  EST1  EST2  AGR1  AGR2  CSN1  CSN2 dateload            country lat_appx_lots_of_err long_appx_lots_of_err
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dttm>              <chr>   <chr>                <chr>
 1     1     5     4     3     2     5     4     4 2016-03-03 02:11:06 AU      -37.9333             145.2333
 2     1     5     5     1     1     3     4     2 2016-03-03 02:22:56 AU      -27.0                133.0
 3     1     3     3     2     5     4     4     4 2016-03-03 02:26:37 AU      -31.6448             152.7946
 4     3     4     1     2     1     4     4     2 2016-03-03 02:43:01 AU      -27.0                133.0
 5     4     1     3     4     1     4     4     4 2016-03-03 03:11:20 AU      -27.0                133.0
 6     4     2     1     5     1     5     5     2 2016-03-03 03:26:32 AU      -37.8139             144.9634
 7     1     4     2     5     1     5     3     4 2016-03-03 04:17:53 AU      -33.7416             151.2762
 8     3     4     4     2     5     0     4     3 2016-03-03 04:29:59 AU      -31.9428             115.8439
 9     3     1     4     3     1     5     4     4 2016-03-03 04:39:13 AU      -37.8333             144.9667
10     4     1     2     4     2     4     3     3 2016-03-03 04:48:53 AU      -27.0                133.0
```

Basically, this data frame has 335,429 rows and 12 columns. Each row represents the 1 participant. There are 8 questions answered by each participant, ‘EXT1’, ‘EXT2’, ‘EST1’, ‘EST2’, ‘AGR1’, ‘AGR2’, ‘CSN1’, ‘CSN2’. Each participant answer these questions by providing scale answers 1=Disagree, 3=Neutral, 5=Agree. The information about these questions are as follows.

EXT1 I am the life of the party.
EXT2 I don't talk a lot.
EST1 I get stressed out easily.
EST2 I am relaxed most of the time.
AGR1 I feel little concern for others.
AGR2 I am interested in people.
CSN1 I am always prepared.
CSN2 I leave my belongings around.

Some other information about the participant is also provided. These variables are
dateload The timestamp when the survey was started.
country The country, determined by technical information (NOT ASKED AS A QUESTION)
lat_appx_lots_of_err approximate latitude of user. determined by technical information, THIS IS NOT VERY ACCURATE. Read the article "How an internet mapping glitch turned a random Kansas farm into a digital hell" https://splinternews.com/how-an-internet-mapping-glitch-turned-a-random-kansas-f-1793856052 to learn about the perils of relying on this information
long_appx_lots_of_err approximate longitude of user

We regard the provided dataset as the population and apply sampling method to estimate population parameters.

In your R code file, you must begin with setting up the random seed by

> set.seed(put your student id number here)

Part I. Simple Random Sampling

We first estimate the population mean of variable "EXT1".

1. Find the population mean and population total of "EXT1".
2. Use simple random sampling with n=1500 to estimate the population mean and total of "EXT1" respectively. Construct 96% confidence interval for the population mean and total of "EXT1". What are the error bounds for these estimates?
3. Repeat step 2 for 100 times and calculate the relative frequencies that the constructed confidence intervals contain the population mean and total, respectively. Are the relative frequencies close to 96%?
4. Use the sample variance from step 2, calculate the the sample size lower bound for estimating the population mean of "EXT1" with an error bound 0.015, at the significance level 0.05.
5. Use the sample size lower bound from step 4, random sample this number of observations and calculate the error bound at the significance level 0.05, is it close to 0.015? Repeat this process 100 times and calculate the relative frequency that the error bound is smaller than 0.015.

We then estimate the population proportion of answer "5" for variable "EXT2".

6. For the variable "EXT2", what is the population proportion of answer "5"?
7. Use simple random sampling with n=5000 to estimate the population proportion of answer "5" for variable "EXT2". What is the confidence interval and error bound of your estimate at significance level 0.05 ?
8. Repeat step 7 for 100 times and calculate the relative frequencies that the confidence intervals contain the population proportion of answer "5" for the variable "EXT2".
9. Use the sample variance from step 7, calculate the sample size lower bound for estimating the population proportion of answer "5" of variable "EXT2" with an error bound of 0.02, at the significance level 0.04.

Part II. Stratified Simple Random Sampling

By looking at the country code of all participants in the provided dataset, the participants come from 6 different countries

"AU", "CA", "DE", "GB", "US", "PH"

In this part, we implement stratified SRS to estimate population parameters for the variables "EST1" and "EST2".

We first estimate the population mean of variable "EST1".

1. Find the population sizes of the 6 strata and population mean of "EST1".
2. To estimate the population mean of "EST1", use simple random sampling to sample 400 from 'AU', 600 from 'CA', 140 from 'DE', 660 from 'GB', 200 from 'PH' and 1200 from 'US'. Compute your point estimate and error bound at significance level 0.05. Construct the confidence interval.
3. Repeat step 2 for 100 times and calculate the relative frequencies that the constructed confidence intervals contain the population mean of "EST1". Is the relative frequency close to 95%?
4. Use the sample variances and allocation weights from step 2, calculate the the sample size lower bound for estimating the population mean of "EST1" with an error bound 0.02, at the significance level 0.05.
5. Use the sample size lower bound from step 4 and allocating weights as step 2, randomly sample these numbers of observations and calculate the error bound at the significance level 0.05, is it close to 0.02? Repeat this process 100 times and calculate the relative frequency that the error bound is smaller than 0.02.

We then estimate the population proportion of answer “5” for variable “EST2”.

6. Find population proportion (also for each stratum) of answer “5” for variable “EST2”.
7. Use stratified simple random sampling with n=5000 and Neyman allocation to estimate the population proportion of answer “5” for the variable “EST2”. What is the confidence interval and error bound of your estimate at significance level 0.05 ?
8. Repeat step 7 for 100 times and calculate the relative frequencies that the confidence intervals contain the population proportion of answer “5” for the variable “EST2”.
9. Use the sample variances from step 7 with Neyman allocation, calculate the sample size lower bound for estimating the population proportion of answer “5” of variable “EST2” with an error bound of 0.02, at the significance level 0.05.
10. Repeat step 9, but use proportional allocation.
11. Use the sample size obtained from step 10 and with proportional allocation, randomly sample this number of observations from the population by stratified SRS and calculate the error bound. Repeat the process 100 times and calculate the relative frequencies that the error bound is less than 0.02.

Part III. Systematic Random Sampling

Now, we apply systematic random sampling to estimate the population mean of variable “AGR1”.

We estimate the population mean of variable “AGR1”.

1. Find the population mean of variable “AGR1”.
2. Choose k=100 and do a systematic random sampling. Calculate the point estimate and its error bound at significance level 0.05. Use half-sample estimator to estimate the required variance.
3. Choose k=100 and do a systematic random sampling. Calculate the point estimate and its error bound at significance level 0.05. Use successive differences estimator to estimate the required variance.
4. Choose k=100 and do a systematic random sampling. Calculate the point estimate and its error bound at significance level 0.05. Use 10 repeated system samples to estimate the required variance.
5. Use step 4 to construct 95% confidence interval of population mean of “AGR1”. Repeat 100 times and calculate the relative frequency of those intervals containing the population mean.

Your final submission should include your reports of the above questions and your complete R codes.