# Instructions

For this project, you must work on your own to answer all of the questions. You may interact by e-mail with the course TA and/or the course instructor, but you may not interact with any of your fellow students.

There are two deadlines:

1. **Deadline 1: due end of week 9** This is a progress report. Please see the details of this submission posted as the homework for week 9.
2. **Deadline 2: due end of week 10** This is the completed project. The primary deliverable is a pdf report, but you should also submit an R script that we can execute to replicate your results (it is your responsibility to verify that your code is well-documented and completely self-contained).

# Project Questions

1. **Simulation Study** Your assignment for this part of the project is to perform a large simulation study to investigate the properties of four sample statistics: the mean, 25th pecentile, minimum and difference in medians.

   We are providing you with observational data on body mass index (BMI) from the Youth Risk Behavior Surveillance System (YRBSS), a large survey of high school students in the United States of America. The two files: `yrbss_2003.rds` and `yrbss_2013.rds`, contain the respondents from 2003 and 2013 respectively. You can read these files and create the corresponding R data frames `yrbss_2003` and `yrbss_2013`, by downloading them to your working directory, and running:

   ```
   yrbss_2003 <- readRDS("yrbss_2003.rds")
   yrbss_2013 <- readRDS("yrbss_2013.rds")
   ```

   You can find a description of the variables in `yrbss_codebook.xlsx`.

   You should act as if the dataset is the *population(s)* of interest.

   (a) Using repeated samples of size $n = 10, 100, 1000$ from the `bmi` variable, describe the sampling distribution of the sample mean of BMI in 2013. Include at least one plot to help describe your results. Report the means and standard deviations of the sampling distributions, and describe how they change with increasing sample size.

   (b) Repeat the simulation in part (a), but this time use the $25^{th}$ percentile as the sample statistic. In R, the `quantile()` function will give you sample quantiles of a sample of data.

   (c) Repeat the simulation in part (a), but this time use the sample minimum as the sample statistic.

   (d) Describe the sampling distribution of the difference in the sample median BMI between 2013 and 2003, by using repeated samples of size $(n_1 = 5, n_2 = 5)$, $(n_1 = 10, n_2 = 10)$ and $(n_1 = 100, n_2 = 100)$. Report the means and standard deviations of the sampling distributions, and describe how they change with the different sample sizes.

   (e) Summarize your results.

2. **Data Analysis** For this part of your assignment your task is to analyze the data to answer the questions of interest. Your solution must include a non-technical summary of your findings.

   Using the same data as the Simulation Study, but now treating the YRBSS as a sample from the population of all USA high-school students:

   - How has the BMI of high-school students changed between 2003 and 2013? Are high-schoolers getting more overweight?
   - Are male high-schoolers more likely to smoke than female high-schoolers in 2013?
   - How much TV do highschoolers watch?