# ST516 Final Project

ST516

2021-11-11

## Instructions

For this project, you must work on your own to answer all of the questions. 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 Markdown file(s) that we could execute to replicate your results (it is your responsibility to verify that your code is well-documented and completely self-contained).

There are two components of the project: a simulation study and some data analysis. You must complete both components.

## Project Tasks

### Task 1: Simulation Study

For this part of the project you will need to perform a simulation study to investigate the properties of four sample statistics: the mean, the 25th percentile, the minimum, and a 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. For the purpose of this question, you should act as if this dataset **is the population(s)** of interest.

The two files: `yrbss_2007.rds` and `yrbss_2017.rds`, contain the respondents from 2007 and 2017 respectively. You can read these files and create the corresponding R data frames `yrbss_2007` and `yrbss_2017`, by downloading them, moving them to your working directory, and running:

```
yrbss_2007 <- readRDS("yrbss_2007.rds")
yrbss_2017 <- readRDS("yrbss_2017.rds")
```

You can find a description of the variables at the bottom of this document.

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 2017. 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, `quantile(x, prob = 0.25)` will give you the $25^{th}$ percentile of the values in `x`.

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 2017 and 2007**, 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.**

Make sure you comment on the center, spread and shape for the sampling distribution of each statistic as the sample size increases. You should also contrast the behaviour of the sample statistics.

Here are some statements, that are sometimes true, you might find useful to reflect on in your summary:

- The sampling distribution of an estimate is always centered around the true parameter value.
- The precision of an estimate (the standard deviation of the sampling distibution) decreases according to the square root of the sample size.
- As the sample size increases, the sampling distribution of an estimate always approaches a Normal distribution.

**Task 2: Data Analysis**

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

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

1. How has the BMI of high-school students changed between 2007 and 2017? Are high-schoolers getting more overweight?
2. In 2017, are 12th graders more or less likely than 9th graders to be "physically active at least 60 minutes per day on 5 or more days"?
3. How much sleep do highschoolers get?

**Description of variables**

| Column | Corresponding survey question |
|---|---|
| `year` | 4-digit year of survey – 1991, 1993, etc. |
| `bmi` | Body mass index (BMI) |
| `age` | How old are you? |
| `sex` | What is your sex? |
| `grade` | In what grade are you? |
| `race4` | 4-level variable from race and ethnicity questions |
| `q9` | During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol? |
| `q32` | During the past 30 days, on how many days did you smoke cigarettes? |
| `q77` | During the past 7 days, how many glasses of milk did you drink? |
| `q88` | On an average school night, how many hours of sleep do you get? |
| `qn79` | `TRUE` if the respondent answered 5, 6 or 7 days to the question: During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day |

You can read more about the questions asked at the CDC website in the Combined Datasets Users Guide