# Lab 9 Homework: Errors in Inference

## Background

In this lab, we consider the data set provided to be the entire population of interest. Because we are considering the data set to be the entire population of interest, we know the true population distribution of the provided variables. We select random samples from this data, and for each sample we perform estimation (with a confidence interval) and testing (with a hypothesis test). Given that we know the true parameter values, we can assess the overall performance of the confidence intervals and hypothesis tests.

## The Data: Youth Risk Behavior Surveillance System (YRBSS)

Every two years the CDC conducts national surveys in schools to monitor and assess the six largest contributors to youth morbidity and mortality. These contributors include not only health risks such as high body mass index, but also risky behaviors such as tobacco and alcohol use, as well as drunk driving and failure to use seatbelts. In 2013, 47 states participated in this school-based survey, yielding 13,583 respondents and 213 variables. A subset of this data set which has no missing data for 16 selected variables is provided in the file `yrbss2013.csv`.

**For this section, you will need to import the `yrbss2013.csv` data set, as well as submit the files contained within `TestingFunctions.R`.**

## 1. Explore inferential results when we repeatedly sample from `weight_kg`

### (a) Examine the population distribution of `weight_kg`.

- Describe the shape of the population distribution.
- What is the true population mean of `weight_kg`?
- What is the true population standard deviation of `weight_kg`?

### (b) Consider taking samples of size $n = 300$ from this population. Are sampling distribution assumptions satisfied for valid inference? Why or why not?

### (c) Fill in the blanks and circle the correct word choice. Consider inference at the $\alpha = 0.05$ level of significance.

We are testing $H_0$ : ____ versus $H_a$ : ____. In the hypothesis test, we run the risk of committing a (**Type I** / Type II) error because in reality the null hypothesis is actually (**true** / false). The targeted (**Type I** / Type II) error rate is ____ and the targeted confidence interval coverage is ____. Because sampling distribution assumptions (**are** / are not) satisfied, we expect the observed Type I error rate and confidence interval coverage from simulation results to (**equal** / not equal) the targeted levels.

### (d) Write and execute code to perform inference for 100 samples of size $n = 20$ from the `weight_kg` variable at the $\alpha = 0.05$ level of significance. Explore the hypothesis test results.

- Describe the shape of the distribution of the sample means:
- Describe the shape of the distribution of the t test statistics:
- Describe the shape of the distribution of the p-values:
- What is the percent of samples that commit an error in hypothesis test results?

### (e) Write and execute code to plot the confidence intervals. Explore the confidence interval results.

What percent of samples have confidence intervals that capture the true parameter value? Describe what else you notice about the confidence intervals. (Do they appear to be of the same length? Do they appear to be randomly scattered about the true mean? Do they appear to provide reasonable bounds for the mean?)