# Assignment 1

Statistical Thinking 2020

Due 6pm Sunday 20 September 2020.

## Instructions

This assignment is a group assignment. Only one submission for each group is required. Your Group Number, and all group member names and ID numbers must be stated in the YAML section of the Rmarkdown file and on any other files submitted.

Follow instructions provided in each section below. Write your answers in the RMarkdown file so that it compiles to produce the desired results. Insert your response to (or code chunk for) each question part in the space immediately following the relevant question.

Once all tasks are completed, you will need to upload **two (2)** separate files to the **Assignment 1 2020** link on Moodle for your group. These files will have names such as:

1. **GroupNumber_A1.Rmd**
2. **GroupNumber_A1.pdf** (obtained by first rendering to .html and then printing to .pdf)

Show and evaluate all code chunks in your submission file using the 'echo=TRUE' and 'eval=TRUE' chunk options, and format the output so that it does not run off the page when printed. This can be achieved by setting the **global** *knitr* option to **echo=TRUE**. You can also suppress all other messages and warnings - as per the following command (to be included in the first code chunk of your .Rmd file):

---

```
knitr::opts_chunk$set(echo = TRUE, eval = TRUE, warning = FALSE,
    message = FALSE, error = FALSE, tidy.opts = list(width.cutoff = 60),
    tidy = TRUE)
```

---

Anything that is not part of your submission should not be included in the .Rmd, even if it is not evaluated or does not appear in the rendered file.

## Assessment marks

There are 100 marks available in total for this assignment. Your final mark will be based on the completeness and clarity of responses. If you do not submit both the .Rmd and corresponding .pdf (from .html) document, or if your .Rmd file does not render as submitted, then 20 marks will automatically be deducted from the total mark.

## General comments

- All Groups are to do all questions in this Assignment, regardless of your unit code.
- Refer to the **R Markdown Quick Reference** available under the **RStudio** *Help* menu (located at the very top of the **RStudio** environment) for help with formatting your report. Note the section on *LaTeX Equations* for how to insert mathematical symbols into your document. A useful resource for the LaTeX math symbols can be found at https://www.caam.rice.edu/~heinken/latex/symbols.pdf.
- You may discuss your ideas about the Assignment with your classmates, including on the Discussion Forum. Groups may even work together, however every group must complete their own submission which should accurately reflect the work and efforts of all students in your Group. Do not send questions about the Assignment to either the Chief Examiner or directly to your tutors.
- Experience suggests it is a poor strategy for a group to allocate questions to individuals and then just to present them together as if they were each undertaken collectively by the group. Not only do we see the marks suffer from this strategy, but as the semester continues the groups that work well together also learn more from each other and the unit overall. The best strategy is to have each group member attempt each question on their own, with the group meeting together after a few days to review and decide whether the questions can be answered directly, or if additional work is required. Be sure to leave enough time to combine your efforts and review the final report before submitting the assignment.
- Organise your submission by Question, with the individual question parts labelled indicating both the Question number and part (e.g. "Q1.a"). Separate each question using a sub-sub-section heading (three hashtags) in your document (e.g., ### Question 1) to ensure adequate separation between questions.

## The Dataset

This Assignment uses a dataset named *NHANES* that is available from the **NHANES R** package. Review the available **R Help** for further information about the dataset. Note the **Disclaimer** to consider our use of this dataset for educational purposes only.

Follow the steps shown in the code chunks below to "clean" the data for use before answering the Assignment questions. The code shown here is also available in the **A1.R** script file available on Moodle.

1. Load the *NHANES* data library, from the *NHANES* package, keeping only the rows with a distinct case IDs as shown in the code chunk below. The package library *tidyverse* is required, while the *broom* and *kableExtra* packages may be of use.
2. Select only the data relating to adults, aged 18 and over, and variables *Gender*, *Age*, *HomeOwn*, *BPSysAve*, *BPSys2* and *BPSys3*. Save these variables in a tibble named *dt*.
3. Drop rows containing missing values in the variables contained in *dt*.[1]
4. Convert the variables *Age*, *BPSysAve*, *BPSys2* and *BPSys3* from integer type to numerical.

---

```
options(digits = 6)
library(tidyverse)
library(broom)
library(kableExtra)
# install.packages('NHANES')
```

[1] Overall this data has 2.9% of values missing, mostly from the *BPSys* measurements. The *visdat* and *naniar* packages are particularly helpful for understanding patterns of missing values in a dataset.

```r
library(NHANES)
dt <- NHANES %>% distinct(ID, .keep_all = TRUE)
dt <- dt %>% filter(Age >= 18) %>% dplyr::select(Gender, Age,
    HomeOwn, BPSysAve, BPSys2, BPSys3)
dt <- dt %>% drop_na()
dt1 <- dt %>% mutate(Age = as.numeric(Age), BPSysAve = as.numeric(BPSysAve),
    BPSys2 = as.numeric(BPSys2), BPSys3 = as.numeric(BPSys3))
```

---

Be sure to understand the variables to be used and the general context that produced the dataset before you start to answer the questions.

**Question 1 [50 marks]**

Consider the two readings of systolic blood pressure selected for the assignment, named *BPSys2* and *BPSys3*.

a. [7 marks] Produce a single plot that displays the sample distributions of the two variables, *BPSys2* and *BPSys3*. Explain why the plot is relevant to the comparison of the two variables. Detail the important elements of the plot in your report, and comment on any features apparent from the plot of particular interest.

b. [6 marks] Produce a plot to appropriately display the sample information regarding the distribution of the difference variable $Diff = BPSys3 - BPSys2$. Detail the important elements of the plot in your report, and comment on any features apparent from the plot of particular interest.

c. [7 marks] Produce a selection of suitable summary statistics for each of the relevant variables from parts a. and b. Describe each statistic produced and explain its relevance.

The NHANES study developers are considering removal of one of these two blood pressure measurements from a future NHANES study, which would help to reduce the cost of the survey, even if only by a relatively small amount. They want to know what is the difference between the average systolic blood pressure of respondents, as determined by these two measurements.

d. [10 marks] Use a Bootstrap-based approach to produce a 95% confidence interval for the average difference in systolic blood pressure of respondents, as measured by *BPSys2* and *BPSys3*. Report your interval and explain how it was obtained. Include a plot of the empirical Bootstrap sample density in your discussion, explain what it represents as well as how the plot relates to the interval produced.

e. [10 marks] Use a CLT-based approach to produce a 95% confidence interval for the average difference in systolic blood pressure measurements, corresponding to part d. Report this alternative interval, and compare it to the one obtained using the Bootstrap method in part d. Explain the relative benefits of *each* approach used to produce the competing confidence intervals.

f. [10 marks] Explain why the measures *BPSys2* and *BPSys3* are not independent and why it is important to take the dependence in to account. What would be the result for each of parts d. and e. if it were to be assumed that the two populations (for the two measurements) were independent?

**Question 2 [50 marks]**

We are now interested in exploring the difference between the average systolic blood pressure, as measured by *BPSysAve*, for people who own their own home (*HomeOwn*="Own") versus people who are renting (*HomeOwn*="Rent").

a. [4 marks] Produce a single plot that displays the sample distribution of the *BPSysAve* variable, for each of the two *HomeOwn* groups of interest, i.e. the "Owners" and the "Renters". Detail the important elements of the plot in your report, and comment if anything seems of particular interest. Explain the steps undertaken to produce the plot.

b. [4 marks] Produce a selection of summary statistics for *BPSysAve*, for each of the two *HomeOwn* groups of interest. Comment on anything interesting you find in these summaries.

c. [8 marks] Estimate the average difference in *BPSysAve* for "Owners" relative to "Renters", and using a CLT-based approach, report a 95% confidence interval for this difference and report on the corresponding outcome of the formal two-sided hypothesis test. In your report, detail the form of the null and alternative hypotheses, and explain how you reached the conclusion of the test.

d. [10 marks] Consider the **R** code provided in the code chunk below (and available in the **R** script file named **A1.R**, noted in the Introduction). *Explain* what the code does, and how it is used to test the relevant hypotheses detailed in part c.

---

```
# Code chunk for Q2 part d.

dt2 <- dt %>% filter(HomeOwn != "Other")
n  # student to add
R  # student to add

# student to add student to add

Rdt2 <- dt2

for (r in 1:R) {
    Rdt2 <- Rdt2 %>% mutate(BPSysAve = sample(dt2$BPSysAve, n,
        replace = FALSE))
    Rdt2S <- Rdt2 %>% group_by(HomeOwn) %>% summarise(mean = mean(BPSysAve))
    RDiff[r] <- Rdt2S %>% summarise(Diff = mean[1] - mean[2])
}

# student to add additional lines
```

---

e. [12 marks] Supplement the code provided and produce a two-sided test based on the sampling distribution described in part d. Discuss the rationale for the test, carefully explain how to determine the outcome of the test, and report the corresponding "strength of evidence" that results.

f. [12 marks] If you restrict the cases to only include men aged between 35 and 44 (inclusive) the strength of evidence changes. Applying the same method as in part e., show the test results, report the "strength of evidence" for the conclusion, and discuss any apparent reasons for the difference in the results compared to part e.