

# Stats in R assignment

*Your name here*

*Due March 24*

Load any packages you need to complete your work here

```
library(tidyverse)
```

## 1. Load NHANES

Load in the nhanes dataset and use a few functions of your choice to investigate the dataset.

NHANES is the National Health and Nutrition Examination Survey (NHANES) program at the CDC. You can read a lot more about NHANES on the CDC's website or Wikipedia. NHANES is a research program designed to assess the health and nutritional status of adults and children in the United States. The survey is one of the only to combine both survey questions and physical examinations. It began in the 1960s and since 1999 examines a nationally representative sample of about 5,000 people each year.

## 2. Descriptive statistics

- A. Create a new dataset that filters nhanes for Adults (Age  $\geq 18$ )
- B. What is the mean and standard deviation of Pulse for adults in nhanes?
- C. How many missing values are there for Pulse in the adults dataset?

## 3. Confidence Intervals

- A. Use `set.seed(324)` and then `filter()` and `sample_n()` to create a new dataset that is a random sample of 20 adults, defined as Age  $\geq 18$ . Take a look at the Pulse column for this new dataset.
- B. Use the adults dataset from #2 to calculate a 95% confidence interval for Pulse and provide an interpretation of what it means
- C. Add an argument to the `t.test` function to change the confidence level and calculate a 99% confidence interval for these data. What happens as we increase the confidence level?
- D. Instead of using the sample of adults we created in #2, create a 95% confidence interval starting with all adults in nhanes (3707 rows). What happens to the 95% CI as sample size increases?

## 3.

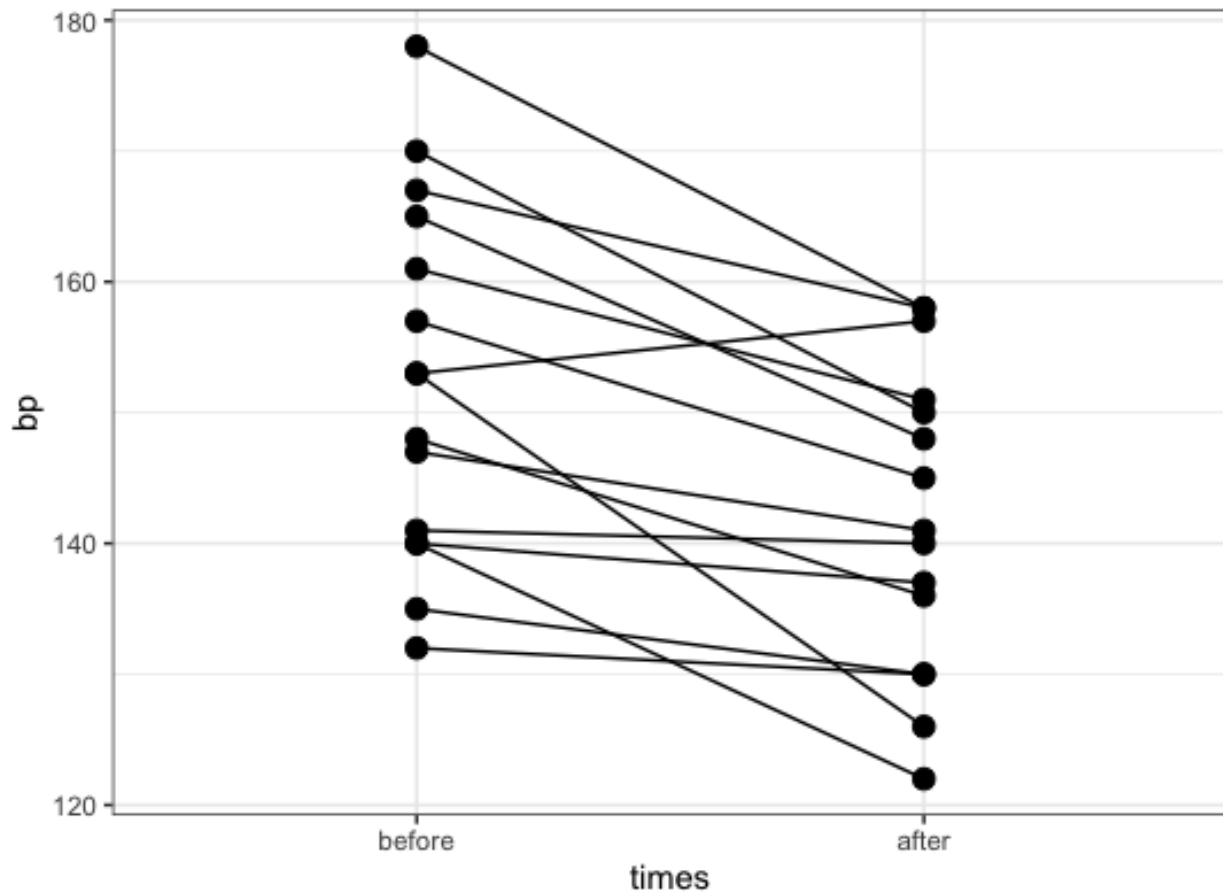
The drug atenolol is used to reduce blood pressure. A group of 15 patients with high blood pressure report the following systolic pressures (measured in mm Hg). Let's say we were interested in testing for a difference in blood pressure before and after treatment.

A. Read in the `atenolol.csv` datafile and take a look to make sure it loaded correctly.

B. Using the `ggplot2` library, create a plot showing each patients' blood pressure before and after drug treatment. Hint: use `geom_point()` and `geom_line(aes(group = id))` to connect the before points with the after.

```
# first refactor times so that "before" comes first
aten <- aten %>%
  mutate(times = factor(times, levels = c("before", "after")))
```

Here is how your plot should look:



- C. Think through the assumptions of a t-test to decide which type of t-test to run.
- D. Run the test you selected in B and interpret what you have found

#### 4. Chi square

- A. Use the `nhanes adults` dataset (created in #2A) to create an `xtabs()` object that stores a cross tabulation of Race and Diabetes
- B. Use `prop.table()` to calculate the proportion of each Race that has Diabetes
- C. Run a chi square contingency table test (test of independence) for these data
- D. Interpret what you have found