Lab01 - t tests

Your Name Here

Goals

Your goal is to get some practice working with data sets in R. The idea is to take the example code I provided for analyzing the mean body temperature and adapt it for analyzing the mean difference in the volumes of the left hippocampus between the twins who are unaffected and affected by schizophrenia.

Part 1: Reminder of R Markdown

Recall that all R code has to go in between the lines that start with "`{r}" and end with "`".

There are three main ways to run R code. First, whenever you knit the document, all chunks will be run in a "fresh" R session.

However, as you're going along you will also want to run commands in a working session so that you can check that your code runs without having to knit the whole document. To do that, you can run individual code chunks by clicking the green "Play" arrow at the top right corner of the chunk.

You can also select individual lines of code you want to run and choose "Run... Run Selected Line(s)" from the menu at the top of the editor window.

Try this with the R code chunk below:

```
2 + 2
```

```
## [1] 4
```

```
log(10)
```

```
## [1] 2.302585
```

Part 2: Loading Packages

R comes with a decent amount of built-in functionality, but to do anything useful you will need to load *packages* that contain additional functionality. You load packages with the `library` command. For this lab, we will need 3 packages that add extra functionality to R: readr, mosaic, dplyr, and ggplot2. Calls to load these packages are in the code chunk below. Go ahead and run this code chunk now.

```
library(readr)
library(dplyr)
library(ggplot2)
library(mosaic)
```

Part 3: Read in the data and take a first look

The following R code reads in the data set from the twins study, and stores it in a data frame called twins. Think of a data frame as R's way of representing a spreadsheet - it's the most common way of storing a data set in R. The original spreadsheet is stored on my website in a file called a csv file (csv stands for comma separated values). To read in csv files, we can use a function called `read_csv`. We have to give the data frame a name so that we can work with it later. Here I have called it creativity. The arrow pointing to the left `<-` says to store the results of `read_csv` in the object named on the left side.

Run the code below now to read in the data set. No need to modify this code.

```
twins <- read_csv("http://www.evanlray.com/data/sleuth3/ex0202_twins_schizophrenia.csv")
```

```
## Parsed with column specification:
## cols(
##   Unaffected = col_double(),
##   Affected = col_double()
## )
```

```
twins <- twins %>% mutate(
  difference = Unaffected - Affected
)
```

Use the `head` function to look at the first 6 rows of the data set (which is called `twins`) .

```
head(twins)
```

```
## # A tibble: 6 x 3
##   Unaffected Affected difference
##        <dbl>    <dbl>      <dbl>
## 1       1.94     1.27      0.67
## 2       1.44     1.63     -0.190
## 3       1.56     1.47      0.09
## 4       1.58     1.39      0.19
## 5       2.06     1.93      0.13
## 6       1.66     1.26      0.400
```

Use the `dim` function to find the number of rows (observational units) and columns (variables) in the data set:

```
dim(twins)
```

```
## [1] 15  3
```

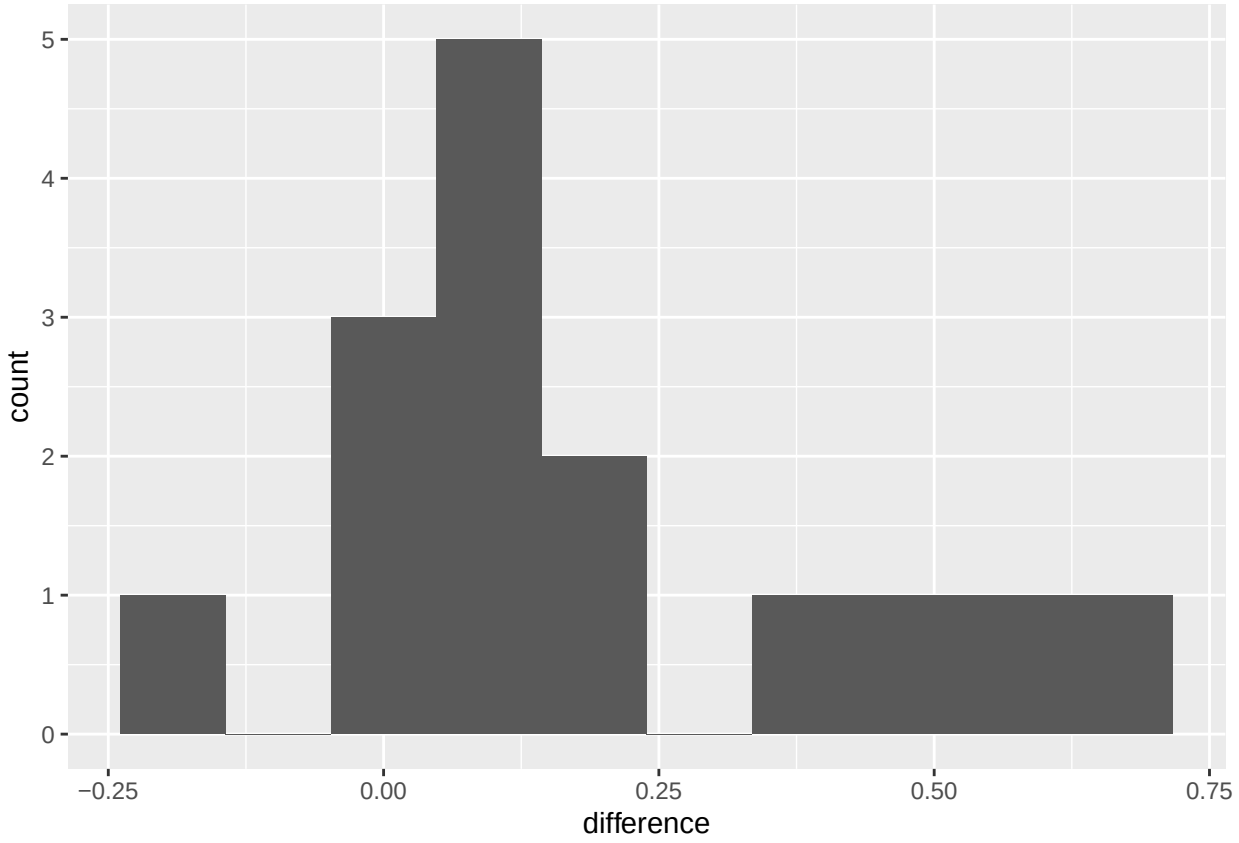
Calculate the mean, median, and standard deviation of differences in volume of the left hippocampus. I have provided an outline of the code to use. You need to provide the name of the data frame at the beginning of the first line, and the name of the variable from that data frame to summarize for the mean, median, and sd functions.

```
 twins %>%
  summarize(
    mean_difference = mean(difference),
    median_difference = median(difference),
    sd_difference = sd(difference)
  )
```

```
## # A tibble: 1 x 3
##   mean_difference median_difference sd_difference
##             <dbl>             <dbl>         <dbl>
## 1           0.199              0.11         0.238
```

Create a histogram of differences in volume of the left hippocampus (unaffected - affected). I have provided an outline of the code to use. You need to specify the name of the data frame, the variable from that data frame to use for the x axis of the plot, and a number of bins to use for the histogram.

```
ggplot(data = twins, mapping = aes(x = difference)) +
  geom_histogram(bins = 10)
```



Part 4: t test and confidence interval

Set up and run the necessary R code below to conduct a t test and find a confidence interval for the average difference.

Notation:

- δ: mean difference in volume of the left hippocampus (unaffected-affected)

Hypotheses:

- $H_0 : \delta = 0$

- $H_A : \delta \neq 0$

Conduct test:

```
t.test(~difference, mu=0, data=twins)
```

```
##
##  One Sample t-test
##
## data:  difference
## t = 3.2289, df = 14, p-value = 0.006062
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.0667041 0.3306292
## sample estimates:
## mean of x
## 0.1986667
```

The test statistic is $t = 3.2289$ on 14 degrees of freedom. The corresponding p-value is 0.006, which is strong evidence that the mean difference in the volume of the left hippocampus is greater than 0. In other words, there is strong evidence that the mean volume of the left hippocampus of the unaffected twin is larger than the mean volume of the left hippocampus of the affected twin. We don't know how these twins were selected for the study (and this is an observational study), so the conclusions we can draw are limited.

Confidence interval: We are 95% confident that the true mean difference in the volume of the hippocampus is between 0.067 and 0.331. In 95% of samples, confidence intervals constructed in this way would contain the true mean difference.