Write an R script to solve the problems prefixed with an R. In Word (or any other word processor), answer all the problems prefixed with a W, without using R. Then convert the word processed file (not the R script) into pdf format. Submit both files, named according to the convention given in Homework 0, attached to a single email.

Write an R script for the following.

Information on this data set from its source, statcrunch.com: “This data set contains information on how the two comic book companies have faired [sic] at the box office. Each movie from both Marvel and DC is listed with name of the film and release date. The Domestic and Foreign gross of each film is provided in millions of dollars along with the total Worldwide gross. The Adjusted column modifies this total for inflation.”

You are interested in the question: *Do DC and Marvel blockbuster movies gross on average the same amounts?*

To address this question, you consider the random variable that is the adjusted gross income of a randomly selected movie (made by DC or Marvel) from the population represented. As an aside, you might think about exactly what “the population represented” is in this case — that is not an easy question!

R1. Put the csv file comicBookMovies.csv (with that filename) into your working directory so that it can be read into R. Read that csv file into a data frame called movieData in R.

R2. Make simultaneous density plots of DC and Marvel adjusted gross revenues in the samples contained in this data set. Your plots should have a legend, and they should have an appropriate label on the x axis. (You should notice a prominent right skew in these plots. That, along with various other things, indicates that we should be hesitant to conduct a 2-sample t tests on the data in this form. For now, we will proceed to do so. Later we will discuss such issues in greater depth.)

R3. Compute the sample standard deviations of the adjusted gross revenues in this data set for DC and Marvel separately.

R4. Based on the density plots and the standard deviations, it is reasonable to assume that the adjusted gross income random variable has equal variances within the DC and Marvel groups. Conduct a 2-sample equal variances t test and compute a 2-sample equal variances t confidence interval for the mean DC adjusted revenues minus the mean Marvel adjusted revenues.

Answer the following, but don’t include any R code in your write-up.

The description of this data set from statcrunch.com is as follows. “Results of an experiment to test whether directed reading activities in the classroom help elementary school students improve aspects of their reading ability. A treatment class of 21 third-grade students participated in these activities for eight weeks, and a control class of 23 third-graders followed the same curriculum without the activities. After the eight-week period, students in both classes
took a Degree of Reading Power (DRP) test which measures the aspects of reading ability that the treatment is designed to improve. **Reference:** Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics. **Original source:** Schmitt, Maribeth C., The Effects on an Elaborated Directed Reading Activity on the Metacomprehension Skills of Third Graders, Ph.D. dissertation, Purdue University, 1987.

<table>
<thead>
<tr>
<th>Column</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>class</td>
<td>treatment, control</td>
<td>which class the student was in</td>
</tr>
<tr>
<td>score</td>
<td>points</td>
<td>student’s score on Degree of Reading Power (DRP) test</td>
</tr>
</tbody>
</table>

We are interested in the question: **Do directed reading activities in the classroom help elementary school students improve aspects of their reading ability?**

We design a study that uses available data from this data set. We define $S$ (short for `score`) to be the random variable whose value is the score of a randomly selected student from the population represented. Also, we define $\mu_1$ to be the random variable mean of $S$ within the untreated (control) population, and $\mu_2$ to be the random variable mean of $S$ within the treatment (directed reading activity) population.

An individual in this study is an elementary school student from among the population represented.

**W1.** To what population does statistical inference validly extend in this study? As usual, if the results don’t extend validly beyond the individuals in the study, then you should address that question of where the variability is: what would vary if the study were repeated over and over? (If there weren’t a source of variability, it wouldn’t be sensible to draw statistical inferences.)

**W2.** In the population represented, do the results carry strong causative weight? (That is, might the results of this study, if positive, be used in an argument that directed reading instruction causes higher DRP scores?) Why or why not?

**W3.** Below is a simultaneous density plot of the DRP scores within the two classes.

![DRP Score Density Plot](image-url)
Is there anything unusual here worth investigating? If so, how would you investigate it?

W4. You enter the following code into R, and get the following output:

```R
> sd(subset(readingData, class == "control")$score)
[1] 17.14873
> sd(subset(readingData, class == "treatment")$score)
[1] 11.00736
```

What does this say about DRP scores? (Be precise: for example, are these random variable standard deviations or sample standard deviations?)

W5. Based on the plots and the above calculations, should you assume equal variances when conducting \( t \) statistical inference techniques here?

W6. You enter the following code into R, and get the following output:

```R
> t.test(score ~ class, data = readingData)

Welch Two Sample t-test
data: score by class
t = -2.3109, df = 37.855, p-value = 0.02638
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -18.67588  -1.23302
sample estimates:
mean in group control mean in group treatment
  41.52174     51.47619
```

In terms of \( S, \mu_1, \) and \( \mu_2 \) defined above, what are the null and alternative hypotheses of the hypothesis test conducted here?

W7. In terms of \( S, \mu_1, \) and \( \mu_2 \) defined above, what parameter is being estimated in the 95% confidence interval computed here?

W8. What is the point estimate for this parameter, obtained from the output above?

W9. What is a 95% confidence interval for this parameter, obtained from the output above?

W10. Obtained from the output above, for the test computed here:

(a) What is the value of the test statistic?

(b) How many degrees of freedom does the test statistic’s \( t \) distribution under the null hypothesis have?
(c) What is the $p$-value?

W11. Interpret the $p$ value, using the usual significance level of 0.05. That is, state whether or not you found statistically significant evidence. Be sure to do so not in generic terms about “the null hypothesis”, but instead in terms of the specific null hypothesis that was tested here.