Statistical Inference Concepts Review Assignment

We now revisit and further strengthen our understanding of basic statistical inference, confidence intervals, and hypothesis testing. You will use the functions `t.test` (for numeric data) and `prop.test` (for data taking two categorical values) to perform hypothesis tests and construct confidence intervals.

In each part, you will first need to identify whether the data is *numeric* or *nominal*. Numeric data consists of numbers, which can be added, subtracted, etc. Nominal (or *categorical*) data consists of *categories* that do not have any inherent order to them. Examples include eye color (green/blue/brown), species (dog/cat), and self-identified race (black/white/…). The `t.test` function is used primarily with numeric data, such as comparing whether the mean outcome in two groups is the same or different, while the `prop.test` function is used to test whether proportions (such as the proportion of men with blue eyes and the proportion of women with blue eyes) are the same or different.

For the purpose of this assignment, regard a P-value less than 0.05 as being "statistically significant" in terms of the evidence against the null hypothesis.

Problem 1

Consider the `TitanicSurvival` dataset, which contains data on the passengers on the titanic, including their class and whether they survived. Before proceeding, make sure you have the package `carData` installed by running the command

```
library(carData)
```

If this line does not run, then you will need to run the command `install.packages("carData")` in the console and try again. Next, we load the data

```
data(TitanicSurvival)
knitr::kable(head(Titanic))
```

Class	Sex	Age	Survived	Freq
1st	Male	Child	No	0
2nd	Male	Child	No	0
3rd	Male	Child	No	35
Crew	Male	Child	No	0
1st	Female	Child	No	0
2nd	Female	Child	No	0
3rd	Female	Child	No	17
Crew	Female	Child	No	0
1st	Male	Adult	No	118
2nd	Male	Adult	No	154
3rd	Male	Adult	No	387
Crew	Male	Adult	No	670
1st	Female	Adult	No	4
2nd	Female	Adult	No	13
3rd	Female	Adult	No	89
Crew	Female	Adult	No	3
1st	Male	Child	Yes	5
2nd	Male	Child	Yes	11
3rd	Male	Child	Yes	13
Crew	Male	Child	Yes	0
1st	Female	Child	Yes	1
2nd	Female	Child	Yes	13
3rd	Female	Child	Yes	14
Crew	Female	Child	Yes	0
1st	Male	Adult	Yes	57
2nd	Male	Adult	Yes	14
3rd	Male	Adult	Yes	75
Crew	Male	Adult	Yes	192
1st	Female	Adult	Yes	140
2nd	Female	Adult	Yes	80
3rd	Female	Adult	Yes	76
Crew	Female	Adult	Yes	20

We will focus on the variables `survived` and `sex`:

```
survived <- TitanicSurvival$survived
sex <- TitanicSurvival$sex
```

a. **Define a reasonable sampling population. What limitations will there be if we try to generalize our conclusions beyond this population?**

   Answer: A reasonable sampling population includes a sample size usually greater than or equal to 30. A sample of female and males in the titanic would give a reasonable sample size but out conclusions would only be pertained to this sample rather than a generalization.

b. Are the pair of variables `sex` and `survived` (a) both nominal, (b) both numeric, (c) nominal and numeric, or (d) numeric and nominal?

   Answer: (a) both nominal

c. **Form a hypothesis about whether men or women are more likely to survive the crashing of the titanic.** No correct answer here, this will just determine which alternative hypothesis you will use later. The null hypothesis will be that the survival proportion in the population would be the same. Possible alternatives are (a) males and females survive at different rates, (b) males are more likely to survive than females, and (c) females are more likely to survive than males.

   Answer: hypothesis males and females survive at different rates

d. **What proportion of females survived the disaster? What about males?**

```
print(table(sex, survived))
```

```
        survived
sex       no yes
  female 127 339
  male   682 161
```

```
prop.table(table(sex, survived))
```

```
        survived
sex               no        yes
  female 0.09702063 0.25897632
  male   0.52100840 0.12299465
```

   Answer: About 25% were females who survived the disaster while about 12% people were males who survived the disaster.

e. **Use the `prop.test` function to perform a hypothesis test on the proportions of male and female survivors, using the alternative you chose in part (c). What is the P-value associated with this test? Do we have evidence to reject the null hypothesis that survival rates are the same for males and females in favor of your alternative?**

```
sex_survived_table <- table(sex, survived)
print(sex_survived_table)
```

```
        survived
sex       no yes
  female 127 339
  male   682 161
```

```
## Your code here
prop.test(sex_survived_table)
```

```

	2-sample test for equality of proportions with continuity c

data:  sex_survived_table
X-squared = 363.62, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5865065 -0.4864599
sample estimates:
   prop 1    prop 2 
0.2725322 0.8090154 
```

   Answer: p-value < 2.2e-16 which is less than .05. This tells us that we have statistically sufficient evidence to reject the null hypothesis and accept the alternative hypothesis (On average, men and women have different survival rates)

f. **Construct a 95% confidence interval for the difference in the proportions of male and female survivors.**

```
## Your code here
prop.test(sex_survived_table)$conf.int
```

```
[1] -0.5865065 -0.4864599
attr(,"conf.level")
[1] 0.95
```

g. A limitation of both of the tests (`t.test` and `prop.test`) used in this homework assignment is that the different outcomes are *independent* of one another, e.g., whether Miss. Elisabeth Walton Allen survives has no bearing on whether Master. Hudson Trevor Allison survives. **Does this assumption seem to be reasonable for this data? Why or why not?**

   Answer:yes because whether a woman survivers has no bearing on whether a man survives thus both sample groups are independent of each other

Problem 2

For our second task, we'll analyze the `iris` dataset available in R. For background on this data, run the command `?iris` in your console. This dataset gives measurements in centimeters of the variables sepal length and width and petal length and width, for 50 flowers from each of three species of iris flowers.

Our scientific question is: do different flowers (virginica and versicolor) have different petal lengths on average? First, let's get the data for `species` and `petal_length` for these two variables:

```
data("iris")
iris <- subset(iris, Species != "setosa")
species <- iris$Species
petal_length <- iris$Petal.Length
```

a. **Define a reasonable sampling population. What limitations will there be if we try to generalize our conclusions beyond this population?**

   Answer: sample size greater than 30 and since two kinds of flowers are different it must be sampled independently from each other. Again, our conclusions would only represent the sample rather than to generalize other flowers.

b. **Are the pair of variables `species` and `petal_length` (a) both nominal, (b) both numeric, (c) nominal and numeric, or (d) numeric and nominal?**

   Answer:(c) nominal and numeric, 'petal length' is numeric while 'species' is nominal

c. **Perform a hypothesis test comparing the average petal lengths of versicolors to virginicas, under the alternative that the petal lengths are not the same, using the `t.test` function. What is the P-value associated with this test? Do we have evidence to reject the null hypothesis that the petal lengths are the same?**

```
## Your code here
t.test(petal_length~species, alternative="two.sided")
```

```

	Welch Two Sample t-test

data:  petal_length by species
t = -12.604, df = 95.57, p-value < 2.2e-16
alternative hypothesis: true difference in means between group
95 percent confidence interval:
 -1.49549 -1.08851
sample estimates:
mean in group versicolor  mean in group virginica 
                   4.260                    5.552 
```

   Answer: p-value < 2.2e-16 which is less than the alpha value of .5 this shows that we can sufficient evidence to reject the null hypothesis and accept the alternate hypothesis

d. **Construct a 99.5% confidence interval for the difference in average petal lengths between virginicas and versicolors (the confidence level can be changed from the default in `t.test` by specifying the `conf.level` argument; see `?t.test`). What can you conclude from this interval?**

```
## Your code here
t.test(petal_length~species, alternative="two.sided", conf.lev
```

```

	Welch Two Sample t-test

data:  petal_length by species
t = -12.604, df = 95.57, p-value < 2.2e-16
alternative hypothesis: true difference in means between group
99.5 percent confidence interval:
 -1.5865763 -0.9974237
sample estimates:
mean in group versicolor  mean in group virginica 
                   4.260                    5.552 
```

   Answer: Difference in average petal length between virginica and versicolor is between -1.5866 and -.9974 with 99.5% confidence.

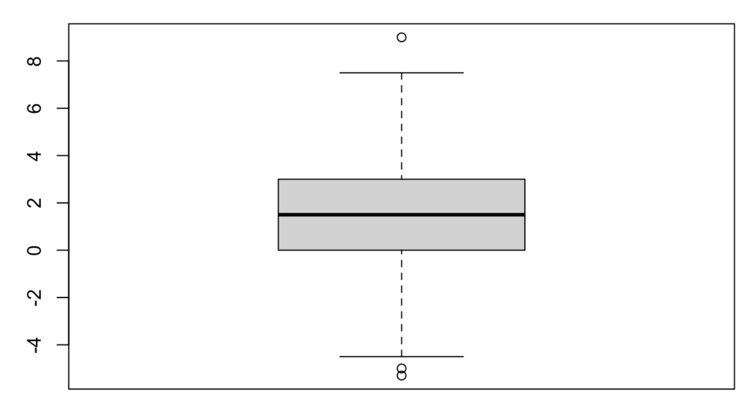
Problem 3

The last question we will be interested in is: do parents tend to, on average, be taller or shorter than their children? We will focus on a famous dataset consisting of this height data. We will look only at the oldest child in each family and restrict attention to males.

```
height_file <-
  "https://raw.githubusercontent.com/data-8/materials-fa17/master/
heights <- subset(read.csv(height_file), childNum == 1 & gender ==
```

Our interest will be in the *difference* between the father's and son's heights:

```
son_minus_father <- heights$childHeight - heights$father
boxplot(son_minus_father)
```



a. **Give a guess as to *why* (scientifically) father's and son's might not have the same height on average. On the basis of this, what issues might we run into if we try to generalize the results from the sampling population to other populations?**

   Answer: based on the boxplot above, we are able to see that the distribution of the difference of height between father and son is normal and thus shows they might not have the same height on average

b. **Perform a hypothesis test to test whether the difference in height is zero on average, under the alternative that it is not zero. What is the P-value associated with this test? Do we have evidence to reject the null hypothesis that the heights are the same?**

```
## Your code here
t.test(son_minus_father)
```

```

	One Sample t-test

data:  son_minus_father
t = 7.1139, df = 178, p-value = 2.653e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.9797514 1.7319805
sample estimates:
mean of x 
 1.355866 
```

   Answer: based on the information above, we are able to see that our pvalue of 2.653e-11 is less than the alpha value of .05 which means that there is statistically sufficient evidence to reject the null hypothesis and accept the alternate hypothesis (heights are not the same)

c. **Construct a 99.5% confidence interval for the difference in average height. What can you conclude from this interval?**

```
## Your code here
t.test(son_minus_father, conf.level = .995)
```

```

	One Sample t-test

data:  son_minus_father
t = 7.1139, df = 178, p-value = 2.653e-11
alternative hypothesis: true mean is not equal to 0
99.5 percent confidence interval:
 0.814112 1.897620
sample estimates:
mean of x 
 1.355866 
```

   Answer: Assuming no bias, results give an approximate 99.5% confidence interval for the difference in average height of (.8141, 1.8976).