R Assignment 3: Intro and Assignment

The assignment will be accepted for marking until the due date given on CourseSpaces. The dropbox will close at that point, and late assignments will not be accepted after that point. **You must try to submit your assignment early, if possible**.

Your assignment should be submitted electronically through the CourseSpaces page. Please submit in **pdf format** a typed document. Non-pdf file will not be accepted by CourseSpaces. If you don't know how to export a Word or OpenOffice document to a pdf, please come to Office hours and we will show you how.

At the top of the first page of your assignment, type the following information in the upper left-hand corner:

Last Name, First Name
Student Number
STAT 260 Assignment 3
Instructor: Chi Kou

Instructions: In this assignment, we are creating confidence intervals and testing hypotheses for a single mean or two means.

The two commands we will use are as follows:

- For a single mean: t.test(x, alternative, mu, conf.level)
- For two means: t.test(x, y, alternative, mu, conf.level, var.equal)

Using t.test for confidence intervals for a single mean: If we are using the t.test command to create a confidence interval, we only need to specify x, our vector of observations, and **conf.level**, our desired confidence level.

Example: Suppose we have a set of five steel bolts, and we measure their weights in grams:

$$12.3, 12.5, 12.7, 11.5, 15.3$$

We would like to create a 97% confidence interval for μ, the mean bolt weight. First, we create a vector containing our data. (see the introduction to assignment 1 for review, if needed.) bolt.weight <- c(12.3, 12.5, 12.7, 12.1, 12.6)

Next, we call for my 97% confidence interval.

t.test(bolt.weight, conf.level = 0.97)

The output is as follows:

```
One Sample t-test

data: bolt.weight
t = 115.5025, df = 4, p-value = 3.37e-08
alternative hypothesis: true mean is not equal to 0
97 percent confidence interval:
  12.08483  12.79517
sample estimates:
mean of x
        12.44
```

Since we are creating a confidence interval, only the second half of the output is relevant to us. The 97% confidence interval is $(12.08483, 12.79517)$, and the sample mean $\bar{x} = 12.44$.

Using t.test for hypothesis tests for a single mean: If we are using the t.test command to test a hypothesis, we must specify x, our vector of observations, mu, our hypothesized value for μ, and **alternative**, the form of our alternative hypothesis.

For **alternative**, we either specify that **alternative = "two.sided"**, or **alternative = "less"**, or **alternative = "greater"**.

Example: Suppose we want to test the alternative hypothesis that the true mean of the bolts is less than 13. That is, we want to test the hypotheses $H_0 : \mu = 13$, $H_1 : \mu < 13$.

We have already created a vector containing the data, so now we call for a hypothesis test.

```
t.test(bolt.weight, mu = 13, alternative = "less")
```

The output is as follows:

```
One Sample t-test

data: bolt.weight
t = -5.1995, df = 4, p-value = 0.003259
alternative hypothesis: true mean is less than 13
95 percent confidence interval:
         -Inf 12.66961
sample estimates:
mean of x
        12.44
```

For a hypothesis test, the first half of the output is of interest.

- **t = -5.1995** tells us the observed value of the test statistic.
- **df = 4** tells us how many degrees of freedom were used for the t-distribution.
- **p-value = 0.003259** tells us the p-value for our hypothesis test.

Important: The next line, **alternative hypothesis: true mean is less than 13**, is just a statement of the alternative hypothesis being tested. R is **not** telling you what your conclusion should be; that is up to you to determine.

Here, since the p-value is less than 0.01, we conclude there is **very strong** evidence against the null hypothesis.

Using t.test for confidence intervals and/or hypothesis tests for two means:

For two means, we use the same **t.test** command, specifying x and y, the names of the two vectors containing the data from the first and second sample, respectively. We must also specify **var.equal=TRUE** (if we are using pooled procedures) or **var.equal=FALSE** (if we are using unpooled procedures).

Example: The following are the observed lifespans of two random samples of the lifespans (in years) of USB drives made by two manufacturers.

Company 1: 4.1, 4.3, 5.1, 5.2, 5.4

Company 2: 6.1, 4.3, 7.9, 2.8, 5.1, 4.7

Let μ_1, μ_2 be the true mean lifespan of USB drives made by Company 1, Company 2 respectively. Construct a 99% confidence interval for $\mu_1 - \mu_2$.

First, we create the vectors, and find their standard deviations:

```
> company.1 <- c(4.1, 4.3, 5.1, 5.2, 5.4)
> company.2 <- c(6.1, 4.3, 7.9, 2.8, 5.1, 4.7)
> sd(company.1)
[1] 0.580517
> sd(company.2)
[1] 1.727136
```

We then calculate the larger standard deviation divided by the smaller standard deviation:

```
> sd(company.2)/sd(company.1)
[1] 2.975169
```

Since this ratio is larger than 1.4, we will use **unpooled** procedures (i.e. we will indicate **var.equal=FALSE**). If the ratio had been less than or equal to 1.4, we would use pooled procedures, and would indicate **var.equal = TRUE**.

The following is my command and output to construct the 99% confidence interval

```
> t.test(company.1, company.2, conf.level=0.99, var.equal=FALSE)

            Welch Two Sample t-test

data: company.1 and company.2
t = -0.43919, df = 6.3028, p-value = 0.6752
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -3.061548   2.401548 sample
estimates:
mean of x mean of y
       4.82         5.15
```

The 99% confidence interval is (-3.061548, 2.401548).

Example: For the USB data, test the hypotheses $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 \neq 0$.

As with the previous example, we will be using unpooled procedures, so we will indicate **var.equal=FALSE**. We will also need to specify that mu = 0 (i.e. our hypothesized parameter value in the null hypothesis is 0), and that **alternative="two.sided"**.

The following is my command and output:

```
> t.test(company.1, company.2, mu=0, alternative = "two.sided", var.equal=FALSE)

            Welch Two Sample t-test

data: company.1 and company.2
t = -0.43919, df = 6.3028, p-value = 0.6752
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -2.147353   1.487353
sample estimates:
mean of x mean of y
     4.82        5.15
```

The observed value of my test statistic is $t_{obs} = -0.43919$, and the p-value was calculated using t with 6.3028 degrees of freedom (Note that R can use non-integer degrees of freedom). The p-value is 0.6752.

We would conclude that there is little to no evidence against H_0; the data are consistent with the two means being equal.

Assignment: For each of the following questions, carry out all calculations using your R. Calculations done by any other method will not be awarded marks.

For each question, copy and paste the commands and the output into a word processing document.

1. In a soup factory, we take a random sample of 8 cans of tomato soup, and measure their sodium content (in mg). The following are our observations.

   510 520 515 516 517 519 522 510

   a) **(1 mark)** Find a 96% confidence interval for the true mean sodium content. Also include your command and output.

   b) **(1 mark)** Using your confidence interval, decide if 515 is a reasonable estimate for μ, the true mean sodium content.

2. A company which makes concrete slabs are testing the cube compressive strength (in N/mm^2) of their slabs. They take a random sample of 5 slabs and measure their cube compressive strength. The following are their observations.

   35.1 34.4 35.8 36.1 35.7

   (a) **(1 mark)** Give the command and output to test the alternative hypothesis that μ, the true mean cube compressive strength is greater than 35 N/mm^2.

   (b) **(1 mark)** What is the observed value of the test statistic?

   (c) **(1 mark)** What is the p-value for our test?

   (d) **(1 mark)** Suppose we are testing at a significance level of $\alpha = 0.01$, what would the conclusion be?

3. Nutritional researchers are comparing the sodium levels of two different brands of canned black beans. They randomly selected 5 cans from each of the two brands. The following are the observed sodium levels (in mg).

   Brand 1: 580, 592, 588, 589, 583

   Brand 2: 579, 582, 577, 591, 581

   Let μ_1, μ_2 be the true mean sodium level for cans of black beans made by Brand 1, Brand 2 respectively.

   (a) **(1 mark)** Using R, calculate and compare the standard deviations of the two samples. Indicate whether you should use pooled procedures or unpooled procedures for these data.

   (b) **(1 mark)** Give the command and output to test the hypotheses $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 \neq 0$.

   (c) **(1 mark)** What is the p-value for our test?

   (d) **(1 mark)** What is the strength of evidence against H_0?

4. Suppose you are interested in the effect of an experimental drug on blood pressure. Blood pressures in mmHg are measured before and after treatment from a random sample of 15 participants. The following data result:

Pre	134	103	116	113	124	120	128	122	123	108	134	108	111	125	134
Post	134	106	110	115	122	126	130	118	125	110	138	111	115	125	130

a) **(1 mark)** Find a 95% confidence interval for the true mean of the difference in blood pressures (Pre – Post). Include your command and output.

b) **(1 mark)** Give the command and output to test if there is any effect of the experimental drug on blood pressure.

c) **(1 mark)** What is the observed value of the test statistic?

d) **(1 mark)** What is the distribution of the test statistic (with parameter if any) under H_0?

e) **(1 mark)** What is the p-value for our test?

Instruction for Question 4:

The command for paired test is:

```
t.test(x1, x2, mu =, alternative =, conf.level = , paired = TRUE)
```

where x1 and x2 are the two vectors of equal size and $\bar{x}_D = x1 - x2$ (in this order); all other parameters are similar to those for two independent sample hypothesis tests.

To save time, you can copy the following data into R:

```
x1 <- c(134,103,116,113,124,120,128,122,123,108,134,108,111,125,134)   # for Pre
x2 <- c(134,106,110,115,122,126,130,118,125,110,138,111,115,125,130)   # for Post
```