Lab 6 (80 pts.): Two Sample *T* and Matched Pairs T Procedure Objectives: Confidence interval and significance tests for two samples.

A (35 points) Food Intake and Weight Gain (Data Set: ex07-36wtgain.txt)

If we increase our food intake, we generally gain weight. Nutrition scientists can calculate the amount of weight gain that would be associated with a given increase in calories. In one study, 16 non-obese adults, aged 25 to 36 years, were fed 1000 calories per day in excess of the calories needed to maintain a stable body weight. The subjects maintained this diet for 8 weeks, so they consumed a total of 56,000 extra calories. According to theory, 3500 extra calories will translate into a weight gain of 1 pound. Therefore, we expect each of these subjects to gain 56,000/3500 = 16 pounds (lb). Here are the weights before and after the 8-week period expressed in kilograms (kg):

Subject	1	2	3	4	5	6	7	8
Weight before	55.7	54.9	59.6	62.3	74.2	75.6	70.7	53.3
Weight after	61.7	58.8	66.0	66.2	79.0	82.3	74.3	59.3
Subject	9	10	11	12	13	14	15	16
Weight before	73.3	63.4	68.1	73.7	91.7	55.9	61.7	57.8
Weight after	79.1	66.0	73.4	76.9	93.1	63.0	68.2	60.3

Solution:

```
> wtgain = read.table(file="ex07-36wtgain.txt",header=T)
> wtgain
> attach (wtgain)
```

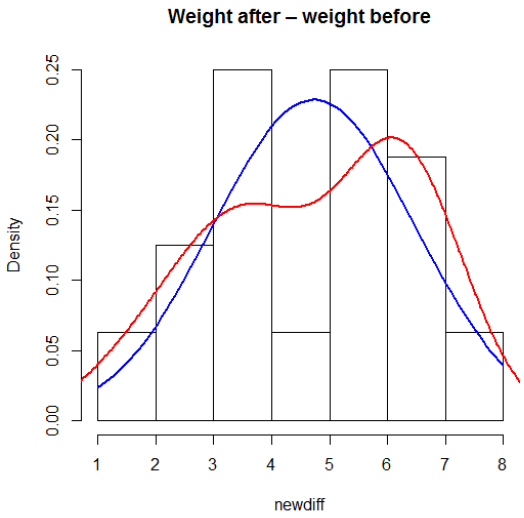
1. (5 pts.) Should you use two sample t or matched pairs t procedure to analyze the data? Please explain your answer without referring to the format of the data.

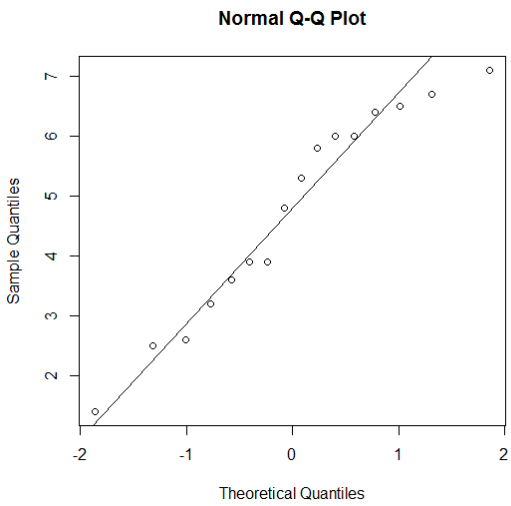
Solution:
This is a before/after situation where the original weight of the person would confound the difference caused by the experiment itself.

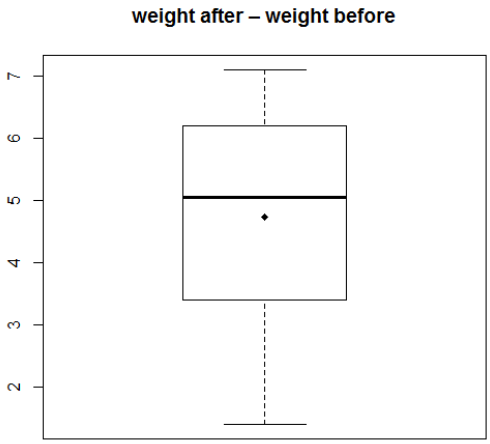
2. (5 pts.) Do you think these data are Normally distributed? Use graphical methods to examine the appropriate distribution. Write a short summary of your findings.

Solution:

```
> gain = wta - wtb

> hist(gain, freq = F, main = "weight after - weight before")
> curve(dnorm(x,mean=mean(gain),sd=sd(gain)),col="blue",lwd=2,
+   add=T)
> lines(density(gain), col = "red", lwd = 2)

> qqnorm(gain)
> qqline(gain)

> boxplot(gain, main = "weight after - weight before")
> points(mean(gain), pch = 18)
```







Even though the histogram slightly skewed and definitely not normal, the two curves look approximately the same. In addition, the boxplot looks roughly symmetrical with no outliers and the QQ plot looks roughly normal. Therefore I would conclude that the distribution is roughly normal. I would conclude that with a sample size of 16, this distribution is normal enough to perform the t-test.
roughly symmetric.

3. (10 pts) Find the 95% confidence interval of the weight gain and interpret your result.

Solution:
The same code is used for part 3 and part 4.

```
> t.test(wta,wtb,conf.level=0.95,mu=16/2.2, alternative="two.sided",
+   paired = T)

        Paired t-test

data:  wta and wtb
t = -5.8233, df = 15, p-value = 3.355e-05
alternative hypothesis: true difference in means is not equal to 7.272727
95 percent confidence interval:
 3.801008 5.661492
sample estimates:
mean of the differences
```

```
4.73125
```

The output for this part is highlighted in green in the output.

The 95% confidence interval is (3.801008, 5.661492).

We are 95% confident that the population mean weight gain when fed 1000 extra calories per day is between 3.801008 and 5.661492 kg.

4. (10 pts) Test the null hypothesis that the mean weight gain is 16 lb. (Hint: first convert pounds to kg. Because there are 2.2 kg per pound, you will need to divide the mean weight gain by 2.2.) Be sure to specify the null and alternative hypotheses, the test statistic with degrees of freedom, and the *P*-value. What do you conclude?

Solution:

The output for this part is highlighted in blue in the output.

$Note: \frac{16}{2.2} = 7.2727$

Step 0: Definition of the terms

μ_D is the population mean weight gain.

Step 1: State the hypotheses

H_0: $\mu_D = 7.273$
H_a: $\mu_D \neq 7.273$

Step 2: Find the *Test Statistic*.

$t_t = -5.8233$

Step 3: Find the *p-value, report DF:*

DF = 15
P-value = 3.355e-5

Step 4: Conclusion:

$\alpha = 0.05$
Since 3.355e-5<0.05, we reject H_0

The data does provide strong evidence (P-value = 3.355e-5) to the claim that the population mean of weight gain is different from 2.2 kg (7.273 kg).

5. (5 pts) Compare the answers of 3 and 4. Are they the same or different? What are the practical consequences of the results?

Solution:
They are saying the same thing because the confidence interval, (3.801008, 5.661492) does not contain 7.273. Therefore, we should reject H_0.

B (45 points) House Prices (Data Set: houseprice.txt – web site)

How much more would you expect to pay for a home that has four bedrooms than for a home that has three? Here are some data for West Lafayette, Indiana. These are the asking prices (in dollars) that the owners have set for their homes.

Four-bedroom homes					
149,900	169,900	175,000	189,000	206,900	225,000
249,900	289,900	320,000	339,900	399,900	429,900
320,000	269,900				

Three-bedroom homes					
79,500	82,000	89,999	90,000	99,900	100,000
106,900	113,900	115,000	117,500	122,900	129,900
139,900	145,000	149,000	150,000	157,900	164,900
189,900	219,900	260,000	274,900	295,000	

Solution:

```
> houseprice = read.table(file = "houseprice.txt", header = T)
> houseprice
```

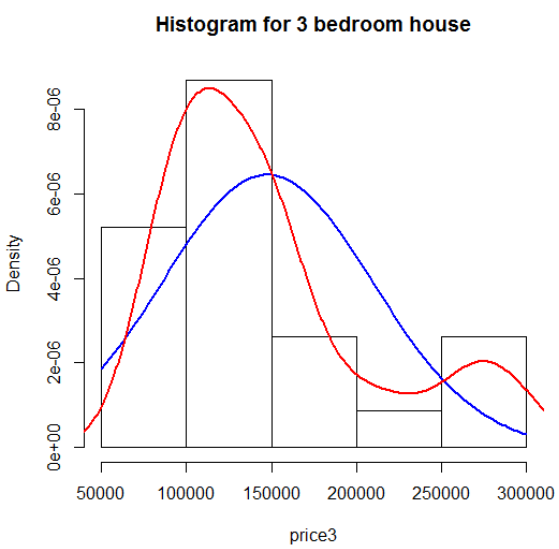
1. (5 pts.) Should you use two sample t or matched pairs t procedure to analyze the data? Please explain your answer without referring to the format of the data.

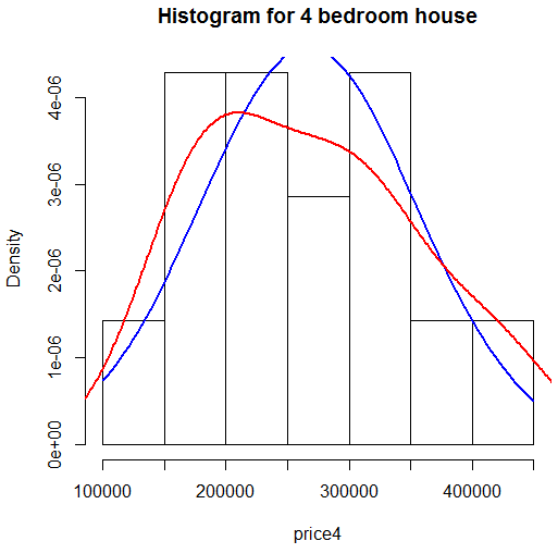
Solution:
Since we are looking at different sizes of houses with no other information provided like neighborhood, there are no confounding variables so we should use two sample t procedure. State that the number of four-bedroom homes is not the same as the number of three-bedroom homes or that the data is not matched will be marked wrong.

2. (5 pts) Do you think these data are **normally** distributed? Use graphical methods to examine the appropriate distributions. Write a short summary of your findings.

Solution:

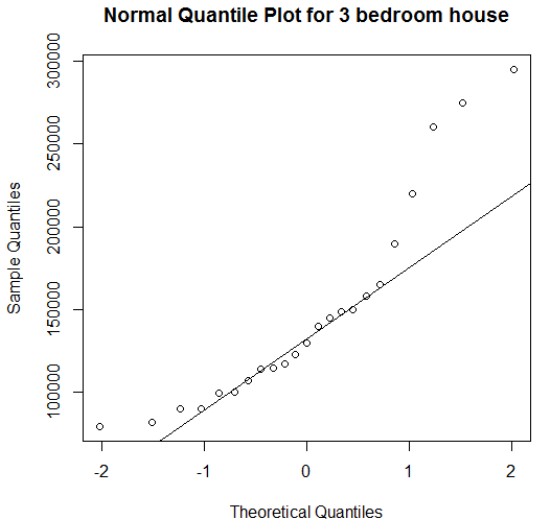
```
> price3 <- subset(houseprice, Bedroom == "3")
> price4 <- subset(houseprice, Bedroom == "4")
>
> hist(price3$Price, freq = FALSE, main="Histogram for 3 bedroom house")
> curve(dnorm(x, mean=mean(price3$Price), sd=sd(price3$Price)), col="blue",
   lwd=2, add=TRUE)
> lines(density(price3$Price),col = "red", lwd=2)
>
> hist(price4$Price, freq = FALSE, main="Histogram for 4 bedroom house")
> curve(dnorm(x, mean=mean(price4$Price), sd=sd(price4$Price)), col="blue",
   lwd=2, add=TRUE)
> lines(density(price4$Price),col = "red", lwd=2)
>
> boxplot(price3$Price, main = "3 Bedroom House")
> points(mean(price3$Price))
>
> boxplot(price4$Price, main = "4 Bedroom House")
> points(mean(price4$Price))
>
> qqnorm (price3$Price, main = "Normal Quantile Plot for 3 bedroom house")
> qqline (price3$Price)
>
> qqnorm (price4$Price, main = "Normal Quantile Plot for 4 bedroom house")
> qqline (price4$Price)
```

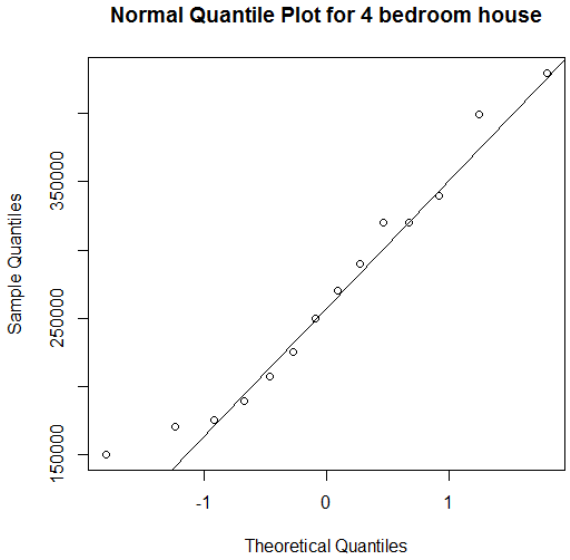

Histogram for 3 bedroom house
Density
0e+00
2e-06
4e-06
6e-06
8e-06
50000
100000
150000
200000
250000
300000
price3


Histogram for 4 bedroom house
Density
0e+00
1e-06
2e-06
3e-06
4e-06
100000
200000
300000
400000
price4


3 Bedroom House
100000
150000
200000
250000
300000


4 Bedroom House
150000
250000
350000





From the histograms, boxplots and QQplots, both of the distributions are right skewed. In fact, the data for the 3-bedroom homes has outliers at higher values.

3. (5 pts) These data are not SRSs from a population. Give a justification for use of the two-sample *t* procedures in this case.

Solution:
Since the data is from SRSs, and $n_1 + n_2 = 37$ and both distributions are right skewed (have similar distributions), the t-procedure is appropriate.

4. (5 pts) Would you consider using a one-sided alternative for this analysis? Explain why or why not.

Solution:
It is possible to consider a one sided test. Since usually a 4 bedroom home is more expensive than 3 bedroom home. However, this is not necessarily true since other factors affect the price of homes. In the one-sided case, the alternative would be H_a: $\mu_4 - \mu_3 > 0$.

5. (10 pts) Follow the four-step procedure, test the null hypothesis that the mean asking prices for the two sets of homes are equal versus the two-sided alternative. Give the test statistic with degrees of freedom, the *P*-value, and your conclusion. You may assume a 0.01 significance level in your discussion.

Solution:

There are two ways to do this. I will give you the code and output for each.

Method 1:

```
> t.test(price3$Price, price4$Price, conf.level = 0.99, alternative =
   "two.sided")

        Welch Two Sample t-test

data:  price3$Price and price4$Price
t = -4.4753, df = 20.976, p-value = 0.0002091
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -194674.07  -43789.99
sample estimates:
mean of x mean of y
 147560.8  266792.9
```

Method 2:

```
> t.test(houseprice$Price ~ houseprice$Bedroom, conf.level = 0.99, alternative
   = "two.sided")

        Welch Two Sample t-test

data:  houseprice$Price by houseprice$Bedroom
t = -4.4753, df = 20.976, p-value = 0.0002091
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -194674.07  -43789.99
sample estimates:
mean in group 3 mean in group 4
       147560.8        266792.9
```

The output for this part is highlighted in blue in the output.

Step 0: Definition of the terms

μ_4 = the population mean price of 4-bedroom homes in West Lafayette.
μ_3 = the population mean price of 3-bedroom homes in West Lafayette.

Step 1: State the hypotheses

H_0: $\mu_4 - \mu_3 = 0$
H_a: $\mu_4 - \mu_3 \neq 0$

Step 2: Find the *Test Statistic*.

t_t = - 4.4753

Step 3: Find the *p-value, report DF:*

DF = 20.976
P-value = 0.0002091

Step 4: Conclusion:

$\alpha = 0.01$
Since $0.0002091 < 0.01$, we should reject H_0.
The data provides strong evidence (P-value = 0.0002091) to the claim that population mean of price of 4 bedroom and 3 bedroom are different.

6. (5 pts) Give a 99% confidence interval for the difference in mean asking prices.

Solution:

The output for this part is highlighted in green in the output.

The 99% confidence interval for the difference in mean asking prices is (-194674.07, -43789.99).

7. (5 pts) Interpret your 99% confidence interval.

Solution:
We are 99% confident that the difference in mean asking prices for 4-bedroom homes and 3-bedroom homes in West Lafayette is between (-194674.07, -43789.99).

8. (5 pts) Compare the answers of 5 and 6. Are they saying the same thing? What are the practical consequences of the results?

Solution:
They are the same. The confidence interval does not contain 0. So we reject null hypothesis.