ADVANCED DATA ANALYSIS

HW1

Fan Yang
UNI: fy2232
01/31/2018

Problem 1

(a)

The reject region given is $S \geq 16$, which contradicts with the two-sided alternative hypothesis. So I do in 2 ways as follow.

1)

Assume $H_a : \eta > 0$

Because any one observation is equally likely to be above or below the population median η, the number of $X_i \geq \eta = 0$ will have a binomial distribution with mean = 0.5.

$$\begin{aligned} 1 - \alpha &= Pr(S \geq 16 | H_0) \\ &= \sum_{i=16}^{25} \binom{25}{i} \times (\frac{1}{2})^{25} \\ &= 0.11476 \\ \alpha &= 0.88524 \end{aligned}$$

Therefore, the level of the test is 0.88524.

2)

Assume reject region is $S \geq 16$ and $S \leq 9$

Because any one observation is equally likely to be above or below the population median η, the number of $X_i \geq \eta = 0$ will have a binomial distribution with mean = 0.5.

$$\begin{aligned} 1 - \alpha &= Pr(S \geq 16 \text{ and } S \leq 9 | H_0) \\ &= 2 \times \sum_{i=16}^{25} \binom{25}{i} \times (\frac{1}{2})^{25} \\ &= 0.22952 \\ \alpha &= 0.770477 \end{aligned}$$

Therefore, the level of the test is 0.770477.

(b)

$$
\begin{aligned}
Pr(X_i > \eta_0) &= Pr(X_i > 0) \\
&= 1 - Pr(\frac{X_i - 0.5}{1} \leq -0.5) \\
&= 1 - Pr(Z \leq -0.5)
\end{aligned}
$$

where Z follows $N(0,1)$.

$$
= 0.6915
$$

So S follows $Bin(25, 0.6915)$.

$$
\begin{aligned}
\text{power} &= Pr(\text{reject } H_0|H_1) \\
&= Pr(S \geq 16|H_1) \\
&= \sum_{i=16}^{25} \binom{25}{i} \times (0.6915)^i \times (1 - 0.6915)^{25-i} \\
&= 0.78355
\end{aligned}
$$

Therefore, the power of the test is 0.78355.

Problem 2

(a)

In [1]:
```
pretest = c(30,28,31,26,20,30,34,15,28,20,
            30,29,31,29,34,20,26,25,31,29)
posttest = c(20,30,32,30,16,25,31,18,33,25,
            32,22,34,32,32,27,28,29,32,32)
diff = pretest-posttest
knitr::kable(cbind(pretest,posttest,diff))
print("mean of pretest-posttest is")
mean(diff)
print("standard deviation of pretest-posttest is")
sd(diff)
```

pretest	posttest	diff
30	20	10
28	30	-2
31	32	-1
26	30	-4
20	16	4
30	25	5
34	31	3
15	18	-3
28	33	-5
20	25	-5
30	32	-2
29	22	7
31	34	-3
29	32	-3
34	32	2
20	27	-7
26	28	-2
25	29	-4
31	32	-1
29	32	-3

```
[1] "mean of pretest-posttest is"
```

-0.7

```
[1] "standard deviation of pretest-posttest is"
```

4.43787526211646

test statistis is defined as

$$\begin{aligned} t^* &= \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \\ &= \frac{-0.7 - 0}{4.4379/\sqrt{20}} \\ &= -0.7054 \end{aligned}$$

$$\text{while } t_{n-1}(\alpha/2) = 2.093 > |t^*| = 0.7054$$

$$\begin{aligned} \text{p-value} &= Pr(t > |t^*|) \\ &= 0.48912 \end{aligned}$$

Therefore, we fail to reject H_0.

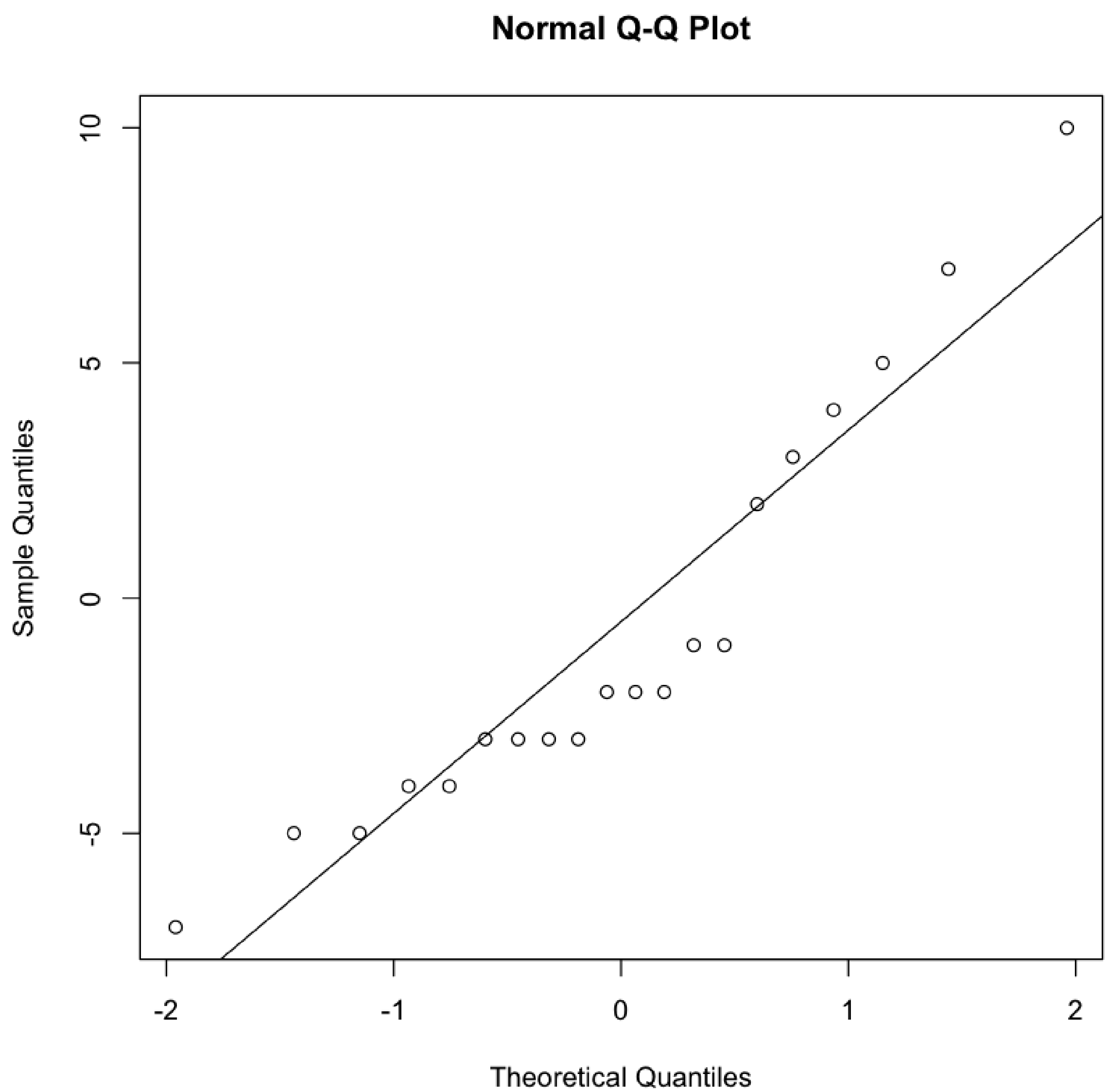
In [32]:
```
t.test(diff, mu=0)
```

```
        One Sample t-test

data:  diff
t = -0.7054, df = 19, p-value = 0.4891
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.77699  1.37699
sample estimates:
mean of x
     -0.7
```

We need to assume that pretest-posttest follows normal distribution.

In [2]:
```
qqnorm(diff)
qqline(diff)
```



We can conclude that the difference is approximately follows normal distribution.

(b)

The $100(1-\alpha)\%$ confidence interval is

$$\bar{X} \pm t_{n-1}(\alpha/2)s/\sqrt{n}$$

which is

$$-0.7 \pm 2.093 \times 4.4379/\sqrt{20}$$
$$[-2.77699, 1.37698]$$

In [44]:
```
-0.7+ 2.093*4.4379 / sqrt(20)
-0.7- 2.093*4.4379 / sqrt(20)
```

1.37697726398858

-2.77697726398858

(c)

In [46]:
```
signdiff = diff / abs(diff)
sum(signdiff>0)
```

6

The test statistic $T^* = \sum I(X_i > 0) = 6$
and $T \sim Bin(n, 0.5)$
$|T - n/2| = |6 - 10| = 4$

$$1 - \alpha = Pr(T > T')$$
$$= \sum_{i=T'}^{20} \binom{20}{i} 0.5^{20}$$

when $T' = 7,\ Pr(T > 7) = 0.94234$ and when $T' = 6,\ Pr(T > 6) = 0.979305$
when $T' = 13,\ Pr(T < 13) = 0.94234$ and when $T' = 14,\ Pr(T < 14) = 0.979305$
Let's calculate the p-value
p-value= $2min(P(T \leq 6), P(T \geq 6)) = 0.115318$, which is less than α
Therefore, we fail to reject H_0.

(d)

In [65]:
```
library(BSDA)
SIGN.test(diff, md=0,,alternative="two.sided",conf.level=0.95)
```

```
        One-sample Sign-Test

data:  diff
s = 6, p-value = 0.1153
alternative hypothesis: true median is not equal to 0
95 percent confidence interval:
 -3.000000  1.650588
sample estimates:
median of x
         -2

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.8847     -3 -1.0000
Interpolated CI       0.9500     -3  1.6506
Upper Achieved CI     0.9586     -3  2.0000
```

So the 95% confidence interval for η is $[-3.000000, 1.650588]$

Problem 3

```
In [5]: Active = c(9.00,9.50,9.75,10.00,13.00,9.50)
Noexe = c(11.50,12.00,9.00,11.50,13.25,13.00)
knitr::kable(cbind(Active,Noexe))
print("mean of Active is")
mean(Active)
print("standard deviation of Active is")
sd(Active)
print("mean of Noexe is")
mean(Noexe)
print("standard deviation of Noexe is")
sd(Noexe)
```

```
| Active| Noexe|
|------:|-----:|
|   9.00| 11.50|
|   9.50| 12.00|
|   9.75|  9.00|
|  10.00| 11.50|
|  13.00| 13.25|
|   9.50| 13.00|
```

```
[1] "mean of Active is"
```

10.125

```
[1] "standard deviation of Active is"
```

1.44697961284878

```
[1] "mean of Noexe is"
```

11.7083333333333

```
[1] "standard deviation of Noexe is"
```

1.52000548244625

two sample t-test

Denote μ_1 as the mean of the Active group and μ_2 as the mean of none-execise groups.
$H_0:\ \mu_1 = \mu_2; \quad H_1:\ \mu_1 \neq \mu_2$

test statistis is defined as

$$t^* = \frac{\bar{Y}_1 - \bar{Y}_2}{SE(\bar{Y}_1 - \bar{Y}_2)}$$

$$\text{where } SE(\bar{Y}_1 - \bar{Y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \text{ with } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 1.4839$$

$$t^* = \frac{10.125 - 11.708}{1.4839 \times \sqrt{\frac{1}{6} + \frac{1}{6}}}$$

$$= -1.84806$$

$$\text{while } t_{n_1+n_2-2}(\alpha/2) = 2.228 > |t^*| = 1.84806$$

$$\text{p-value} = Pr(t > |t^*|)$$

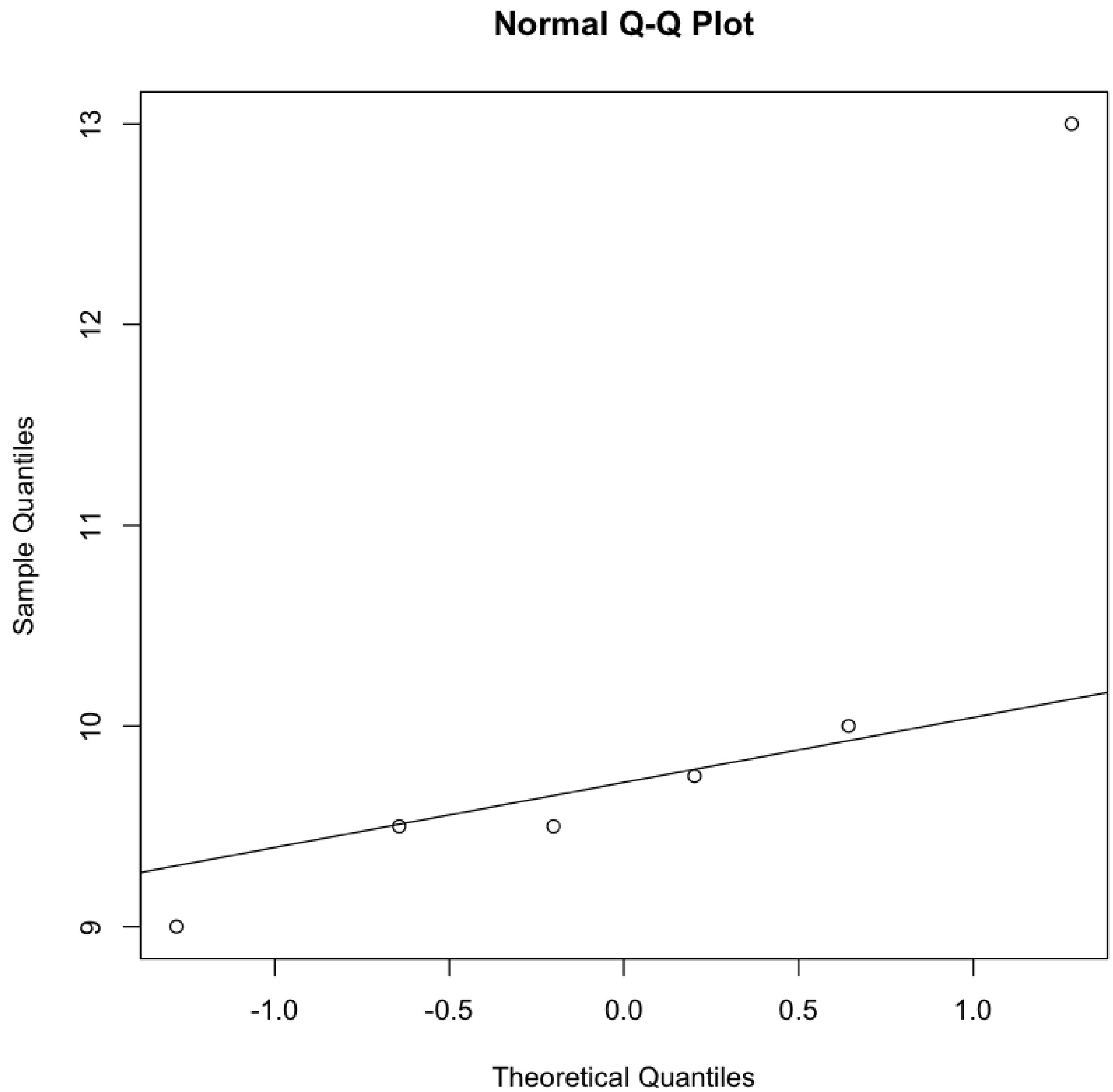
$$= 0.09434$$

Therefore, we fail to reject H_0.

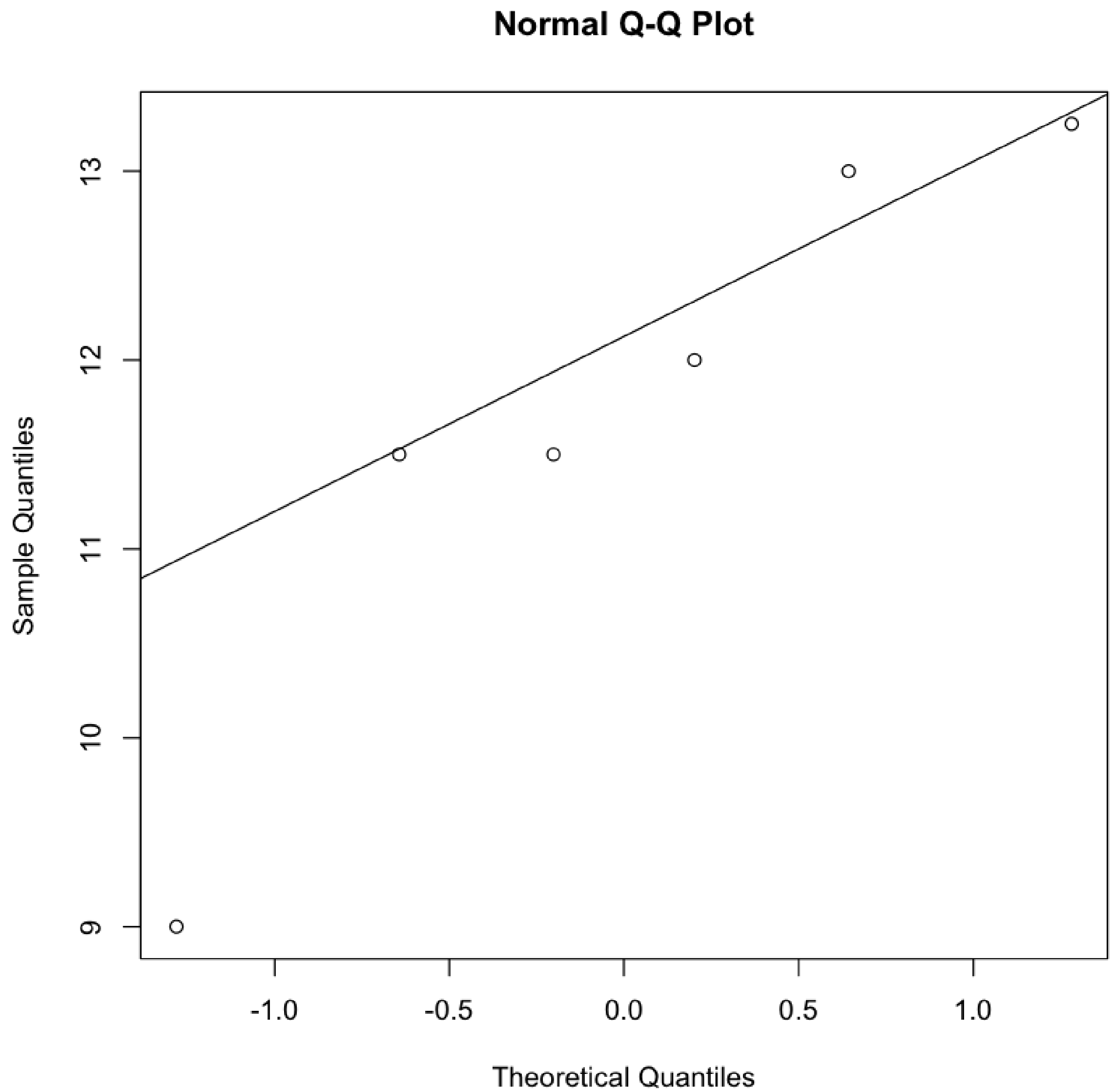
Assumption. In order to use this test we need to assume that X follows normal distribution. And the two group should have the same variance σ^2. Now we use qqplot to check.

In [6]:

```
qqnorm(Active)
qqline(Active)
```



In [7]:
```
qqnorm(Noexe)
qqline(Noexe)
```



According to the above qqplot, we can assume the sample follows normal distribution.

In [8]:
```
t.test(Active,Noexe,var.equal=T)
```

```
        Two Sample t-test

data:  Active and Noexe
t = -1.8481, df = 10, p-value = 0.09434
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.492301  0.325634
sample estimates:
mean of x mean of y 
 10.12500  11.70833
```

(Wilcoxon) Mann-Whitney two sample procedure

In [134]:
```
wilcox.test(Active,Noexe)
```

```
Warning message in wilcox.test.default(Active, Noexe):
“cannot compute exact p-value with ties”

	Wilcoxon rank sum test with continuity correction

data:  Active and Noexe
W = 9, p-value = 0.1705
alternative hypothesis: true location shift is not equal to 0
```

The Sign test do not need to assume normal sample or sysmtric distribution, we only need the sample to be indenpendent.