hw6

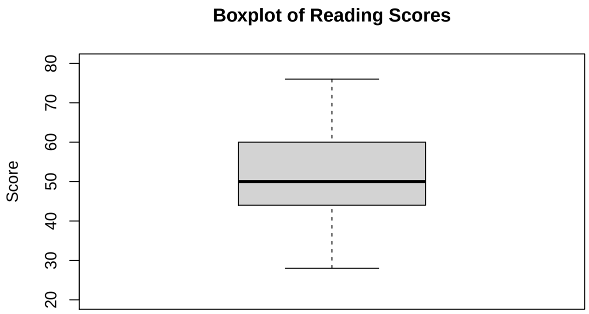
Kyu Park

2021 2 21

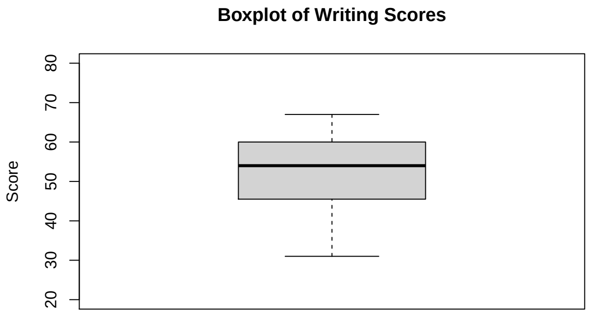
1.

a)

```
boxplot(hsb2$read, main = "Boxplot of Reading Scores", ylim = c(20, 80), ylab = "Score")
```

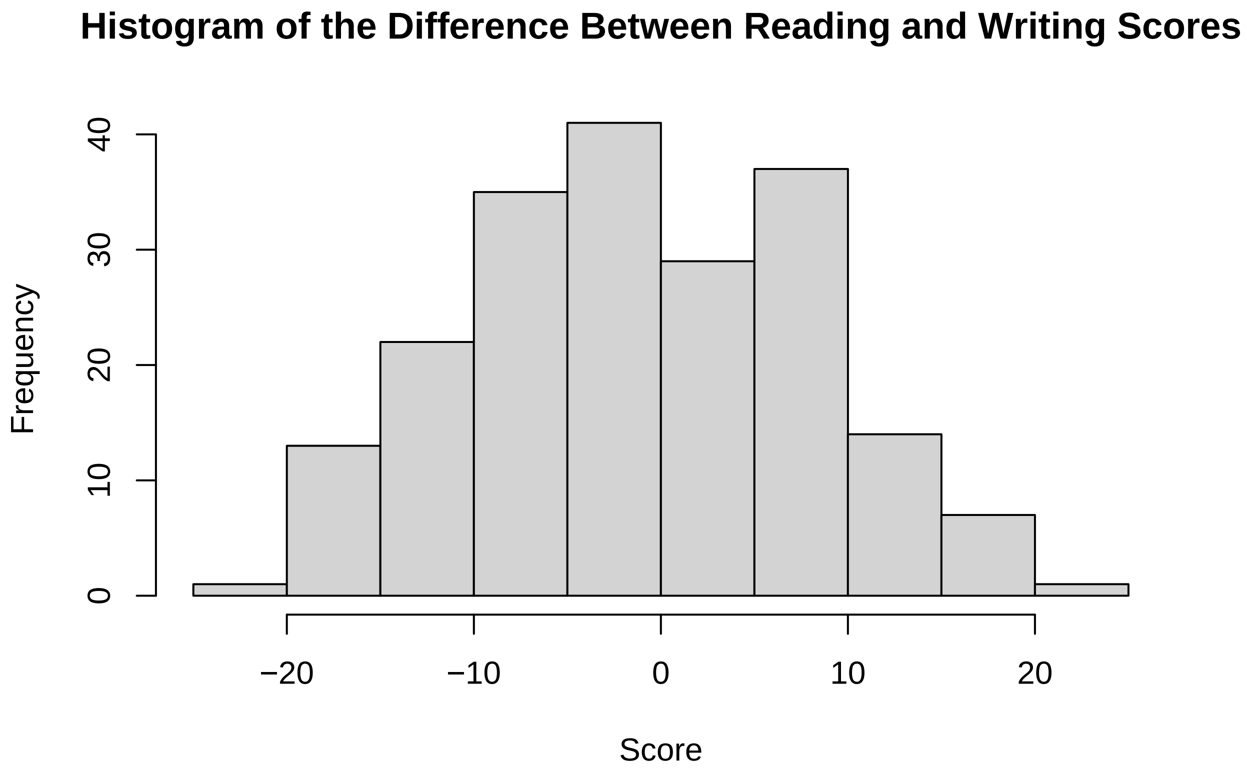


```
boxplot(hsb2$write, main = "Boxplot of Writing Scores", ylim = c(20, 80), ylab = "Score")
```



b)

```
hist(hsb2$read - hsb2$write, breaks = 10,
     main = "Histogram of the Difference Between Reading and Writing Scores",
     xlab = "Score")
```



c) The distribution of the difference between reading and writing scores are approximately normally distributed, centered at around 0. There are only a few that has extreme differences in reading and writing scores. This indicates that there is no clear difference in the average reading and writing scores.

d) No, the distribution of the difference between reading and writing scores demonstrated that individual students are likely to have similar reading and writing scores. This indicates that the reading and writing scores are dependent, since a student with a higher reading score is likely to receive a higher writing score, and likewise a student with a lower reading score is likely to receive a lower writing score.

e) Let null hypothesis be that there is no evident difference between reading and writing scores, $h_o : s_r = s_w$, and alternative hypothesis that there is an evident difference between reading and writing scores, $h_a : s_r \neq s_w$

f)

```
t.test(hsb2$read, hsb2$write, paired = TRUE)
```

```
##
##  Paired t-test
```

```
##
## data:  hsb2$read and hsb2$write
## t = -0.86731, df = 199, p-value = 0.3868
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7841424  0.6941424
## sample estimates:
## mean of the differences
##                  -0.545
```

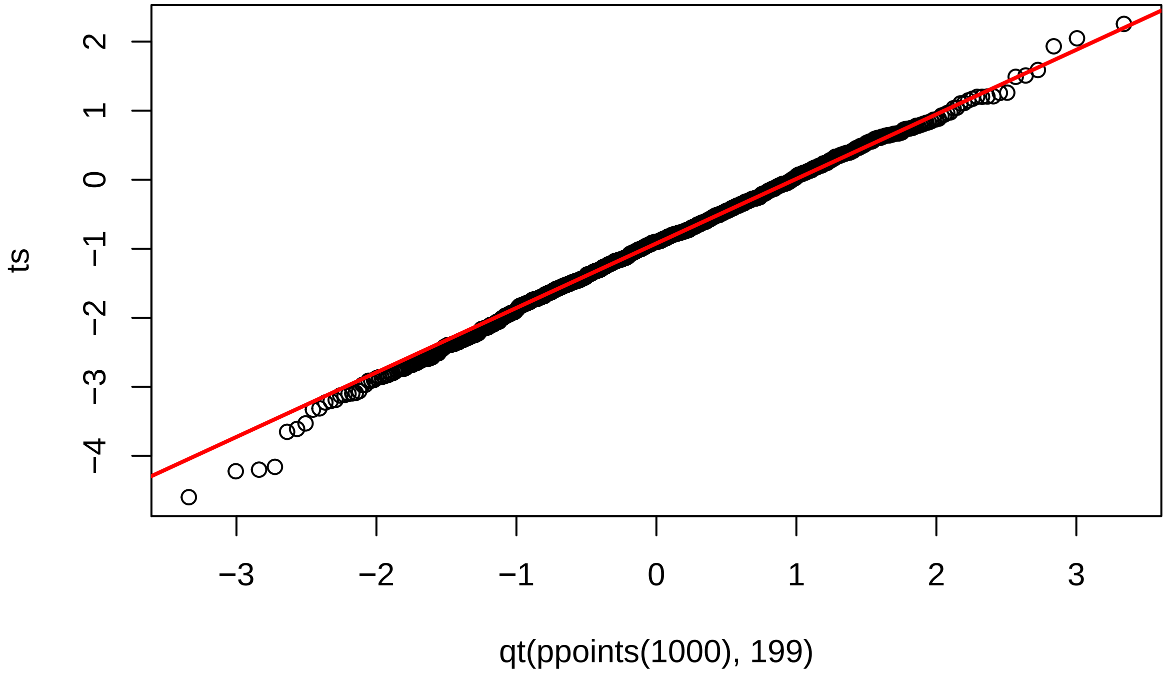
$p = 0.3868$, and since $p > 0.05$, accept the null hypothesis; there is no evident difference between reading and writing scores.

g) Assume that the sample was randomly selected and well represents the population, the distribution of the difference between reading and writing score is approximately normal, that the size of n=200 is sufficiently large enough, and the standard deviations of both reading and writing score are sufficiently similar.

h) i, ii, iii)

```
ts = c()
for (i in 1:1000)
{
  samples = sample(hsb2, 200, replace = TRUE)
  ts[i] = t.test(samples$read, samples$write, paired = TRUE, alternative = "two.sided")$statistic
}

qqplot(qt(ppoints(1000), 199),ts)
qqline(ts, col = "red", lwd = 2)
```



The sample t-statistics well follows the theoretical t-distribution, as shown in the graph. Degree of freedom of 199 was used.

2.

```
data2 = yrbss_samp
```

a)

```
hispanics = subset(data2, data2$hispanic == "hispanic")
nonhispanics = subset(data2, data2$hispanic == "not")
t.test(hispanics$weight, nonhispanics$weight)$conf.int
```

```
## [1] -3.270026  9.870416
## attr(,"conf.level")
## [1] 0.95
```

The 95% confidence interval for the difference of weight between hispanics and non-hispanics is (-3.270026 9.870416).

b) Assume that the sample was randomly selected and well represents the population, the distribution of the difference between hispanic and non-hispanic weight is approximately normal, that the sizes of both samples are sufficiently large enough, and the standard deviations of samples are sufficiently similar.

Let null hypothesis be that there is no evident difference in weight between hispanics and non-hispanics, $h_o : w_h = w_n$, and alternative hypothesis be that there is evident difference in weight between hispanics and non-hispanics, $h_a : w_h \neq w_n$.

```r
t.test(hispanics$weight, nonhispanics$weight)
```

```
##
##  Welch Two Sample t-test
##
## data:  hispanics$weight and nonhispanics$weight
## t = 1.0159, df = 39.091, p-value = 0.3159
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.270026  9.870416
## sample estimates:
## mean of x mean of y
##  71.46227  68.16208
```

p = 0.3159

Using pooled variance,

```r
t.test(hispanics$weight, nonhispanics$weight, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  hispanics$weight and nonhispanics$weight
## t = 0.92939, df = 97, p-value = 0.355
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.747405 10.347794
## sample estimates:
## mean of x mean of y
##  71.46227  68.16208
```

p = 0.355

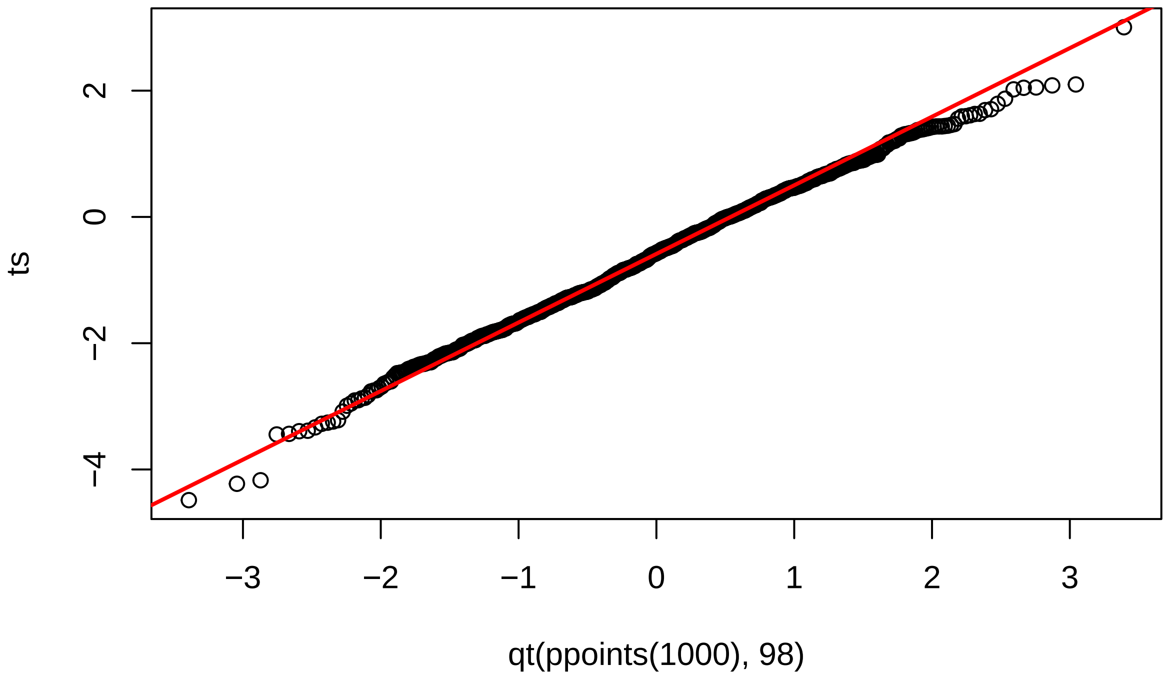
In 95% confidence level, both p = 0.3159 and p = 0.355 are higher than 0.05. Therefore, accept the null hypothesis; there is not enough evidence to say that there is evident difference in weight between hispanics and non-hispanics.

c) The sample was randomly selected and well represents the population, the distributions of the difference between hispanic and non-hispanic weight are approximately normal, and the sizes of both samples are sufficiently large enough, and the standard deviations of samples are sufficiently similar. All the assumptions made are valid.

d)

```r
ts = c()
for (i in 1:1000)
  {
  samples = sample(yrbss, 100, replace = TRUE)
  hispanicsample = subset(samples, samples$hispanic =="hispanic")
  nonhispanicsample = subset(samples, samples$hispanic == "not")
  ts[i] = t.test(hispanicsample$weight, nonhispanicsample$weight)$statistic
}

qqplot(qt(ppoints(1000), 98),ts)
qqline(ts, col = "red", lwd = 2)
```



The sample t-statistics well follows the theoretical t-distribution, as shown in the graph. Degree of freedom of 98 was used.

3.

a) Let null hypothesis be that the treatment group has the same survival rate with the control group, $h_o : p_c = p_t$, and alternative hypothesis that the control group has the survival rate less than the treatment group, $h_a : p_c < p_t$

```
prop.test(c(4,34), c(24, 69), correct = FALSE, alternative = "less")
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c4 out of 2434 out of 69
## X-squared = 7.8354, df = 1, p-value = 0.002562
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.1665322
## sample estimates:
##    prop 1    prop 2
## 0.1666667 0.4927536
```

P value is 0.002562, and less than 0.05 at 95% confidence level, therefore reject the null hypothesis; the control group has the survival rate less than the treatment group, indicating that the treatment is effective.

b) Normal approximation is legitimate if np and $n(1-p)$ are greater than or equal to 10. This means that all the elements in the table must be greater than 10. However, since not all the elements in the table are greater than 5, the normal approximation is not legitimate to be used here.

c) For the same reason above, normal approximation cannot be used for the confidence interval as well. Because there is not enough amount of sample, where the number of alive people in control group is less than 10, the normal approximation is not justified.

4.

a) Not paired, Intel's stock and Southwest's stocks are two different entities, and therefore independent variables. There is no apparent relationship between those two.

b) Paired, since the before-after effects of the same selected students are evaluated, therefore are paired.

c) Not paired, since two groups of different students are used to compare. Each group will not affect each other, therefore are not paired.

5.

a) Chi square test or 2-proportion test could be used. They both should provide apporoximately the same p value. For both 2-proportion Z test and chi square test, null hypothesis would be that there is no difference between the proportion of autism in no vitamin and vitamin group, $h_o : p_v = p_n$, and the alternative hypothesis would be that there is difference between the proportion of autism in no vitamin and vitamin group, $h_a : p_v \neq p_n$. 2-proportion test will provide Z score, which could be converted into p in normal distribution. Chi square test will provide chi square, which also could be converted into p with known degree of freedom in chi square distribution.

b) By hand 2-proportion test follows:

$$p = \frac{X_1 + X_2}{N_1 + N_2} = \frac{111 + 143}{181 + 302} = 0.52588$$

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{N_1} + \frac{1}{N_2})}} = \frac{0.61326 - 0.47351}{\sqrt{0.52588(1 - 0.52588)(\frac{1}{181} + \frac{1}{302})}} = 2.97737$$

$$p = 2P(Z < 2.97737)$$

```
2*pnorm(2.97737, lower.tail=FALSE)
```

```
## [1] 0.002907329
```

p = 0.002907329.

By hand chi-square follows:

$$E_{11} = \frac{(181)(254)}{483} = 95.18426, E_{12} = \frac{(181)(229)}{483} = 85.81573, E_{21} = \frac{(302)(254)}{483} = 158.81573, E_{22} = \frac{(302)(229)}{483} = 143.18426$$

$$\chi^2 = \frac{(111 - 95.18426)^2}{95.18426} + \frac{(70 - 85.81573)^2}{85.81573} + \frac{(143 - 158.81573)^2}{158.81573} + \frac{(159 - 143.18426)^2}{143.18426} = 8.86472$$

$df = 1$,

```
pchisq(8.86472, 1, lower.tail = FALSE)
```

```
## [1] 0.002907348
```

and $P(\chi_1^2 \geq X^2) = 0.002907348$.

Approximately, the chi-sqare test and 2-proportion z test give same p value of 0.0029.This is lower than 0.05 at 95% confidence interval level; null hypothesis is rejected and there is significant amount of evidence that the proportion of autism between no vitamin group and vitamin group may differ.

c) No, since the title may mislead the readers. The statistical conclusion gained from the tests above is that at 95% confidence level, there is significant amount of evidence that the proportions of autism between no vitamin and vitamin group are different. This does not necessarily mean that the vitamin may ward off autism. A better title could be "Vitamin May Potentially Alter the Probability of Autism"