Assignment2025

SID 540798904

1.

(a)

Let μ = the population mean of test A - the population mean of test B; which means that if μ > 0, the test A is better than test B (new better than old)

Given that we want to determine whether the newer is better than older, we can make hypotheses that :

$$H0 : \mu <= 0 \qquad H1 : \quad \mu > 0$$

(b)

Given that we got 2 groups of sample from 2 category, and we don't know any statistics from the population, we need to stimulate the population mean by the sample. The t-test can do it, so we can use T-test or T-statistics to evaluate the hypotheses. Also, the T-statistics is suitable for the small number of sample. Given that we need to judge whether μ <= 0, we need to use the one side T-statistics.

(c)

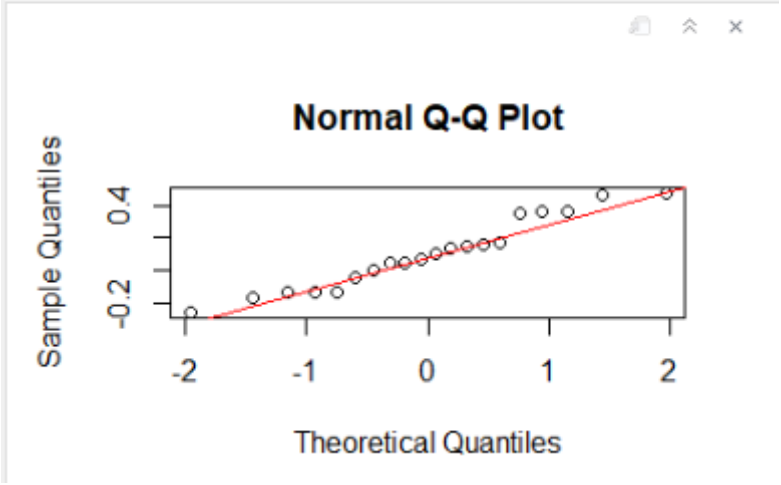
In T-statistics, we need to plot the qq plot (Quantile-Quantile Plot). If the distribution is near a line, the μ in population is near a normal distribution. Then we can use T-statistics for μ.

In R: we can use

```
diff = test.A - test.B
qqnorm(diff)
qqline(diff, col = "red")
```

The result is

```r
diff = test.A - test.B
qqnorm(diff)
qqline(diff, col = "red")
```



Which the distribution is near a line, so it can use T-statistics.

(d)

The formula of T-statistics

$$T = \frac{\bar{X} - \mu_0}{\widehat{SE}_0(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

Where u0 = 0, so we can calculate the mean of μ and the standard deviation of μ, then calculate the T-score, where due to the hypothesis, we need to calculate P(T > t_diff).

Using R :

mean_diff = mean(diff)
sd_diff = sd(diff)
t_diff = (mean_diff - 0) / (sd_diff/ sqrt(20))
t_diff

Which result in 2.096674.

```r
mean_diff = mean(diff)
sd_diff = sd(diff)
t_diff = (mean_diff - 0) / (sd_diff/ sqrt(20))
t_diff
```

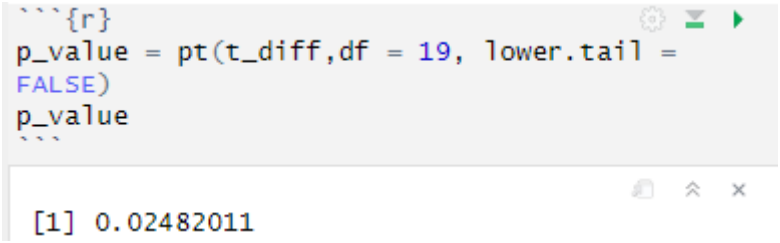
```
[1] 2.096674
```

P-value :

p_value = pt(t_diff,df = 19, lower.tail = FALSE)

p_value

Which result in 0.02482011.



In this case, we need to test μ <= 0, so it is one side T-test. After calculating the T-score, using pt() to stimulate the population and the freedom is n – 1 which is 19 and due to the one side T-test, we use the lower.tail = False to get the p-value, rather than * 2.

(e)

if use 5%, which means that 0.05 significance level and 95% confidence interval. The p-value 0.02482001is lower than the given significance level, so we reject H0 and conclude that the new tire significantly improves breaking deceleration.

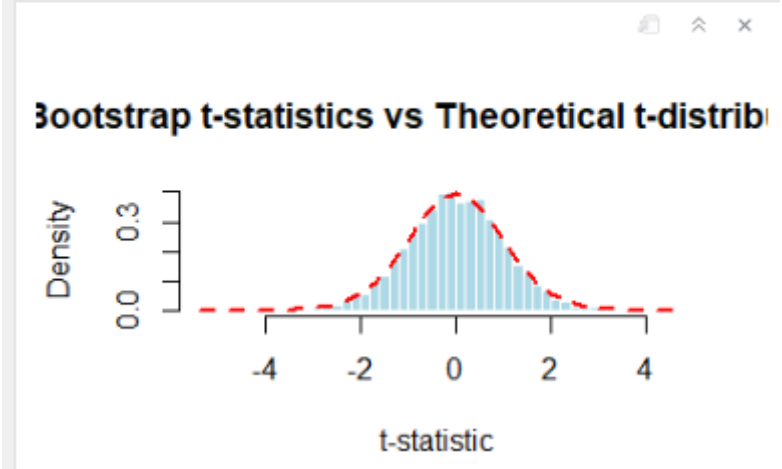
(f)

Using the following code

```
set.seed(1)
time = 10000
n = length(diff)
T.stats.sim <- numeric(time)

for (i in 1 : time) {
  samp = sample(diff, size = n, replace = TRUE)
  T.stats.sim[i] = (mean(samp) - mean(diff)) / (sd(samp) / sqrt(n))
}
hist(T.stats.sim, breaks = 50, prob = TRUE,
     main = "Bootstrap t-statistics vs Theoretical t-distribution",
     xlab = "t-statistic", col = "lightblue", border = "white")

curve(dt(x, df = n - 1), add = TRUE, lty = 2, col = "red", lwd = 2)
```

Result :

```r
set.seed(1)
time = 10000
n = length(diff)
T.stats.sim <- numeric(time)

for (i in 1 : time) {
  samp = sample(diff, size = n, replace = TRUE)
  T.stats.sim[i] = (mean(samp) - mean(diff)) /
(sd(samp) / sqrt(n))
}
hist(T.stats.sim, breaks = 50, prob = TRUE,
     main = "Bootstrap t-statistics vs
Theoretical t-distribution",
     xlab = "t-statistic", col = "lightblue",
border = "white")

curve(dt(x, df = n - 1), add = TRUE, lty = 2,
col = "red", lwd = 2)
```



Using 10000 times replacemental draws to stimulate the t distribution graph, and draw the t distribution curve on this graph, we can see that it is quite close. Then, we calculate the percentage of the T score of the given sample at, then calculate the P-value.

(g)

In R:

mean(T.stats.sim > t_diff)

Result : 0.0199

```r
mean(T.stats.sim > t_diff)
```

```
[1] 0.0199
```

The code find the percentage of data which is higher than the T score, and in this case, it is the P-value. The P-value is 0.0199, which is quite closed to the P-value using T-statistics, still smaller than the 5%, so we can also reject the H0.

2.

(a)

Let xa = mean of population in online tutoring, xb mean of population in in-person tutoring. So when it is no effect, xa_bar should equal xb_bar.

H0 : xa = xb H1 : xa != xb

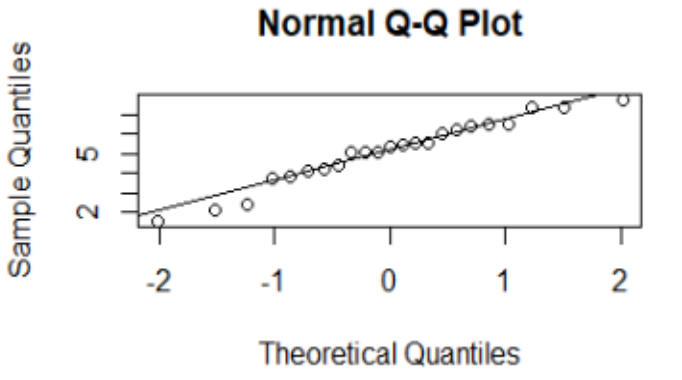
(b)

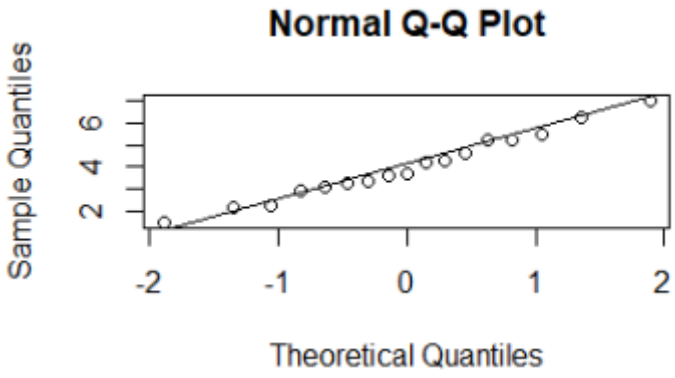
We need to check the 2 sample is followed by normal shaped. Using QQ Plot can do that.

In R:

```
qqnorm(group.A)
qqline(group.A)
qqnorm(group.B)
qqline(group.B)
```

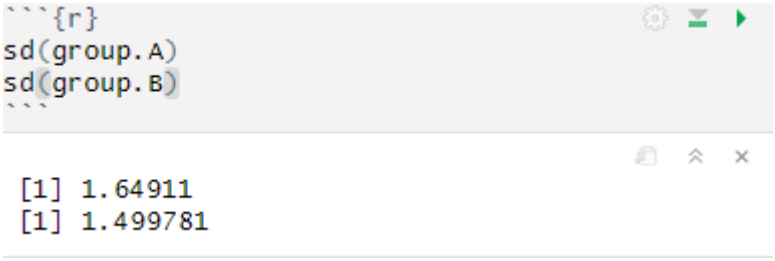
Result





We can see that this 2 sample all fit in a line, so we can say that it follow a normal shape.

Also, we can check the standard deviation



Which is similar.

(c)

We can use the following formula to calculate the T score. The standard deviation of groupA and B is similar, so we can use classic T-test.

First, the total formula is :

$$T = \frac{xa_{bar} - xb_{bar}}{\sigma p_{bar} * \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

In which xa_bar, xb_bar means the sample means of 2 sample set, σpbar means the weight average standard deviation of 2 sample set, m and n is the set size of 2 sample set.

The σp_bar can be calculated by that:

$$\sigma p_{bar} = \sqrt{\frac{(m-1)\sigma xa_{bar}^2 + (n-1)\sigma xb_{bar}^2}{m+n-2}}$$

Where σxa_bar is the standard deviation of sample a and σxb_bar is the standard deviation of sample b.

We can use R to calculate the σp_{bar} at first, which is

```
m = length(group.A)
n = length(group.B)
sd_a = sd(group.A)
sd_b = sd(group.B)
mean_a = mean(group.A)
mean_b = mean(group.B)
sigma_p_bar = round(sqrt( (((m - 1) * sd_a^2) + ((n - 1) * sd_b^2)) / (m + n - 2) ), 3)
sigma_p_bar
```

Result:

```r
m = length(group.A)
n = length(group.B)
sd_a = sd(group.A)
sd_b = sd(group.B)
mean_a = mean(group.A)
mean_b = mean(group.B)
sigma_p_bar = round(sqrt( ((m - 1) * sd_a^2 +
(n - 1) * sd_b^2) / (m + n - 2) ), 3)
sigma_p_bar
```

```
[1] 1.588
```

Then we can calculate the T score

t_score = round((mean_a - mean_b) / (sigma_p_bar * sqrt(1/m + 1/n)),3)

t_score

Result:

```r
t_score = round((mean_a - mean_b) /
(sigma_p_bar * sqrt(1/m + 1/n)),3)
t_score
```

```
[1] 2.165
```

(d)

This case is a 2 side T test because we want to determine whether the 2 sample set have a similar means. Given that the critical region of rejection at the 5%, we need to find the 97.5% of t distribution, we can use qt() to do that.

qt(0.975, df = m + n - 2)

Result:

```
[1] 2.024394
```

We can see that the t_score is out of the confident interval area (-2.024394, 2.024394), so we should reject H0 and accept H1, which means that there is effect on this 2 in-person tutoring.

Or we can caculate P-value by

p_value_2 = 2 * pt(t_score,df = m + n - 2,lower.tail = FALSE)

p_value_2

Which result to :

```r
p_value_2 = 2 * pt(t_score,df = m + n -
2,lower.tail = FALSE)
p_value_2
```

```
[1] 0.03673386
```

Which is smaller than the given 5% level of significance. So we reject the H0.

(e)

Using code below to procedure the Welch Two Sample t-test

t.test(group.A, group.B, var.equal = FALSE)

Result:

```r
t.test(group.A, group.B, var.equal = FALSE)
```

```
        Welch Two Sample t-test

data:  group.A and group.B
t = 2.1964, df = 36.294, p-value =
0.03453
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 0.0845232 2.1143003
sample estimates:
mean of x mean of y
 5.100000  4.000588
```

We can see that the T score is quite close, and df is a little smaller than the m+n-2 = 38. P-value is 0.03453, and quite close to the result of classical two-sample t-test, still smaller than 5% so we can also reject H0, accept H1.

3.

(a)

H0 : device preference is independent of age.

H1 : device preference is not independent of age.

If we reject H0, then we have H1, which means that device preference is associated with age.

(b)

Expect e_ij = row_total_i * column_total_j / total

	Laptop	Desktop	Tablet	Total
Under 18	11.54	10.15	8.31	30
18-29	11.54	10.15	8.31	30
30-49	15.39	13.54	11.08	40
50+	11.54	10.15	8.31	30
Total	50	44	36	130

(c)

The chi-squared test should satisfy :

- Sample size n is large.
- All the categories have large probabilities.

The n is 130 in this situation and all the categories Ej is bigger than 5, so according to lecture content, it is a good practice to use chi-squared test.

(d)

We need to calculate T and then use chi-squared distribution about that.

$$T = \frac{(\mathrm{O}_1 - E_1)^2}{\mathrm{E}_1} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

In R, we can use that

- First input all the data

```
under_18 = c(12,6,12)
one8_29 = c(14,10,6)
three0_49 = c(16,12,12)
five0_ = c(8,16,6)
Oij = rbind(under_18, one8_29, three0_49, five0_)
rownames(Oij) = c("Under 18", "18–29", "30–49", "50+")
colnames(Oij) = c("Laptop","Desktop","Tablet")

# or easily use that
Eij = matrix(c(
   11.54, 10.15, 8.31,
   11.54, 10.15, 8.31,
   15.38, 13.54, 11.08,
   11.54, 10.15, 8.31
), nrow = 4, byrow = TRUE)
```

rownames(Eij) = c("Under 18", "18–29", "30–49", "50+")

colnames(Eij) = c("Laptop", "Desktop", "Tablet")

- Then calculate the T

  T = sum((Oij - Eij)^2 / Eij)

  T

Result

```r
under_18 = c(12,6,12)
one8_29 = c(14,10,6)
three0_49 = c(16,12,12)
five0_ = c(8,16,6)
Oij = rbind(under_18, one8_29, three0_49,
five0_)

rownames(Oij) = c("Under 18", "18-29", "30-49",
"50+")
colnames(Oij) = c("Laptop","Desktop","Tablet")

# or easily use that
Eij = matrix(c(
  11.54, 10.15, 8.31,    # Under 18
  11.54, 10.15, 8.31,    # 18-29
  15.38, 13.54, 11.08,   # 30-49
  11.54, 10.15, 8.31     # 50+
), nrow = 4, byrow = TRUE)

rownames(Eij) = c("Under 18", "18-29", "30-49",
"50+")
colnames(Eij) = c("Laptop", "Desktop",
"Tablet")
T = sum((Oij - Eij)^2 / Eij)
T
```

```
[1] 9.898673
```

(e)

First we need to calculate the df or freedoom degree, which is

In R:

df = (nrow(Oij) - 1) * (ncol(Oij) - 1)

p_value = pchisq(T, df, lower.tail = FALSE)

p_value

Result:

```r
df = (nrow(Oij) - 1) * (ncol(Oij) - 1)
p_value = pchisq(T, df, lower.tail = FALSE)
p_value
```

```
[1] 0.1289845
```

Which is bigger than 5% significance, so we can accpet H0, which means that the device preference is independent of age.