R Notebook

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1(a) The data has 4 columns and 30 rows.

```
my_url <- "http://ritsokiguess.site/STAC32/cholest.csv"
cholesterol <- read_csv(my_url)
```

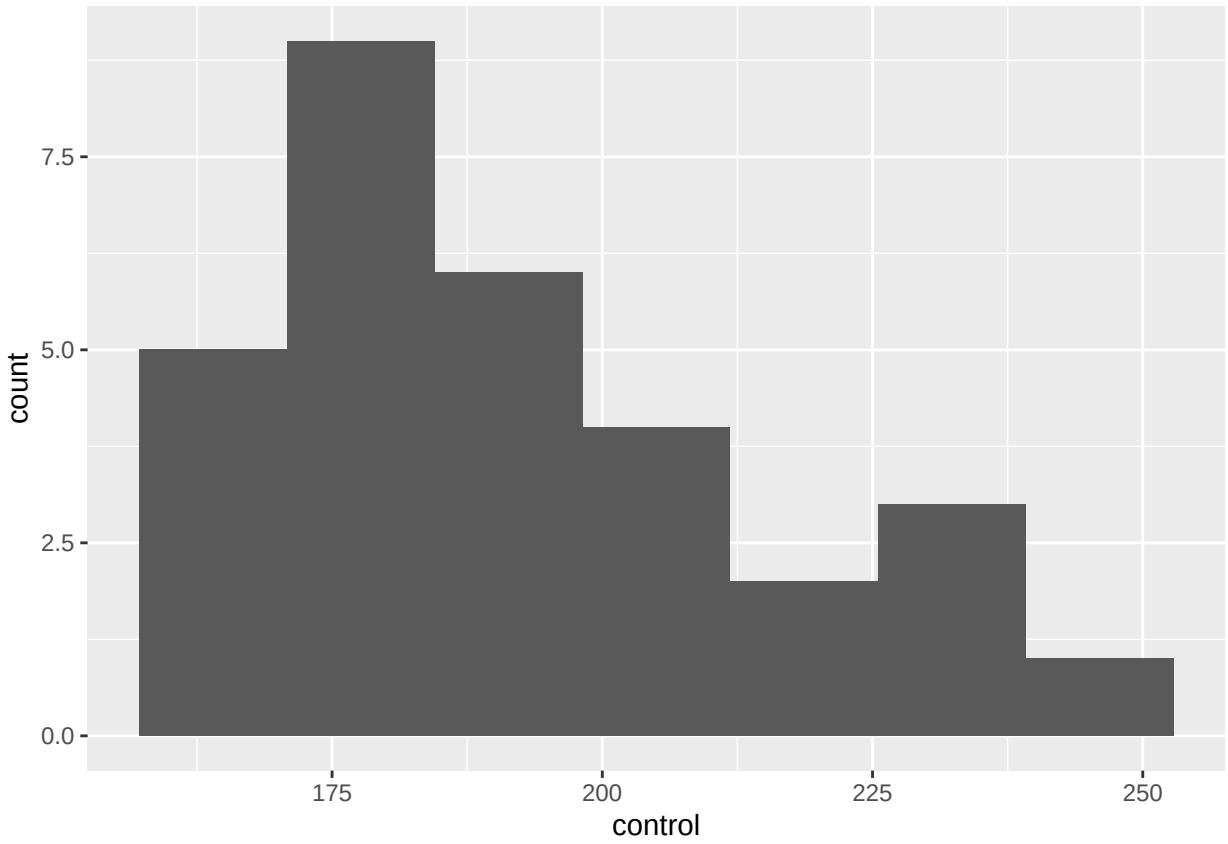
```
## Parsed with column specification:
## cols(
##   `2-Day` = col_double(),
##   `4-Day` = col_double(),
##   `14-Day` = col_double(),
##   control = col_double()
## )
```

```
cholesterol
```

```
## # A tibble: 30 x 4
##    `2-Day` `4-Day` `14-Day` control
##      <dbl>   <dbl>    <dbl>   <dbl>
##  1     270     218      156     196
##  2     236     234       NA     232
##  3     210     214      242     200
##  4     142     116       NA     242
##  5     280     200       NA     206
##  6     272     276      256     178
##  7     160     146      142     184
##  8     220     182      216     198
##  9     226     238      248     160
## 10     242     288       NA     182
## # ... with 20 more rows
```

1(b) There is only one quantitative variable(cholesterol levels of the control patients), and 0 categorical variable, so I choose geometric histogram. The shape is right skewed, with a long tail at the right, and the cholesterol level bunching on the left, cholesterol level with most patients is about 180

```
ggplot(cholesterol, aes(x = control)) + geom_histogram(bins = 7)
```



1(c) let mu be the mean level of cholesterol levels of the control patients H0: mu = 200 , Ha: mu < 200

```
t.test(cholesterol$control, mu = 200, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  cholesterol$control
## t = -1.6866, df = 29, p-value = 0.05121
## alternative hypothesis: true mean is less than 200
## 95 percent confidence interval:
##      -Inf 200.0512
## sample estimates:
## mean of x
##  193.1333
```

since p-value is 0.05121, sightly larger than 0.05, so we fail to reject null hypothesis and conclude that population mean cholesterol level is equal to 200. That is to say, at level of 0.05, we consider mean of cholesterol level of people in good health is 200.

1(d) the 95% confidence interval of mean value of level of cholesterol level of control patience is (184.8, 201.5) So by 95% confidence, population mean cholesterol would take between 184.8-201.5

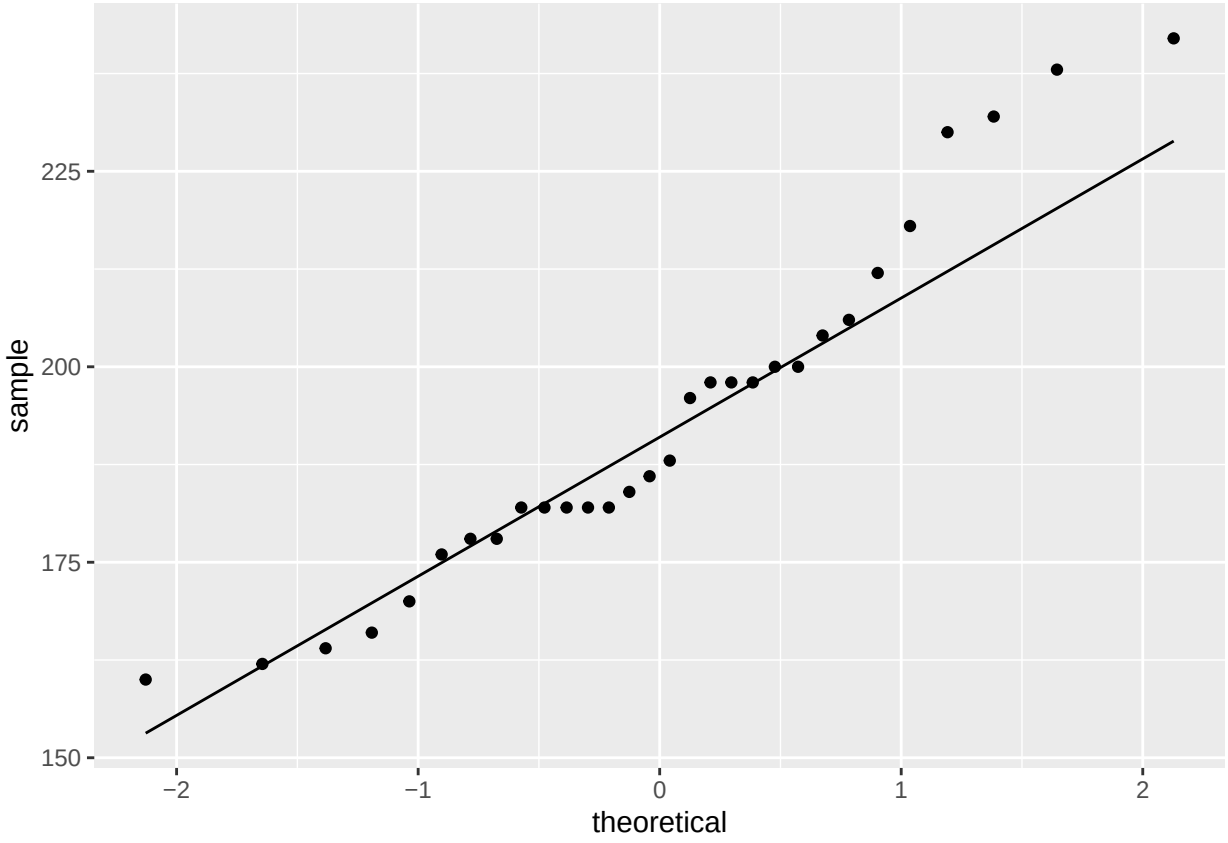
```
t.test(cholesterol$control)
```

```
##
##  One Sample t-test
##
## data:  cholesterol$control
## t = 47.436, df = 29, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  184.8064 201.4603
## sample estimates:
## mean of x
##  193.1333
```

1(e)

```
ggplot(cholesterol,aes(sample=control))+
stat_qq()+stat_qq_line()
```



Firstly, the graph of control patient in (b) is right skewed, which is not appropriate for t-test, t-test needs normal distributed sample. And by looking at the normal quantile plot, some points above 1 diverge from the straight line, which is a little bit violate normality assumption. Secondly, the sample size is 30, which is sufficiently large for central limit theorem. Although the source population is right skewed, we can still consider the distribution of sample means is approximately normally distributed. In conclusion, we can trust the the t procedure in this question.

2(a) the data has 2 columns and 21 rows

```
my_url <- "http://ritsokiguess.site/STAC32/anchoring.csv"
anchor <- read_csv(my_url)
```

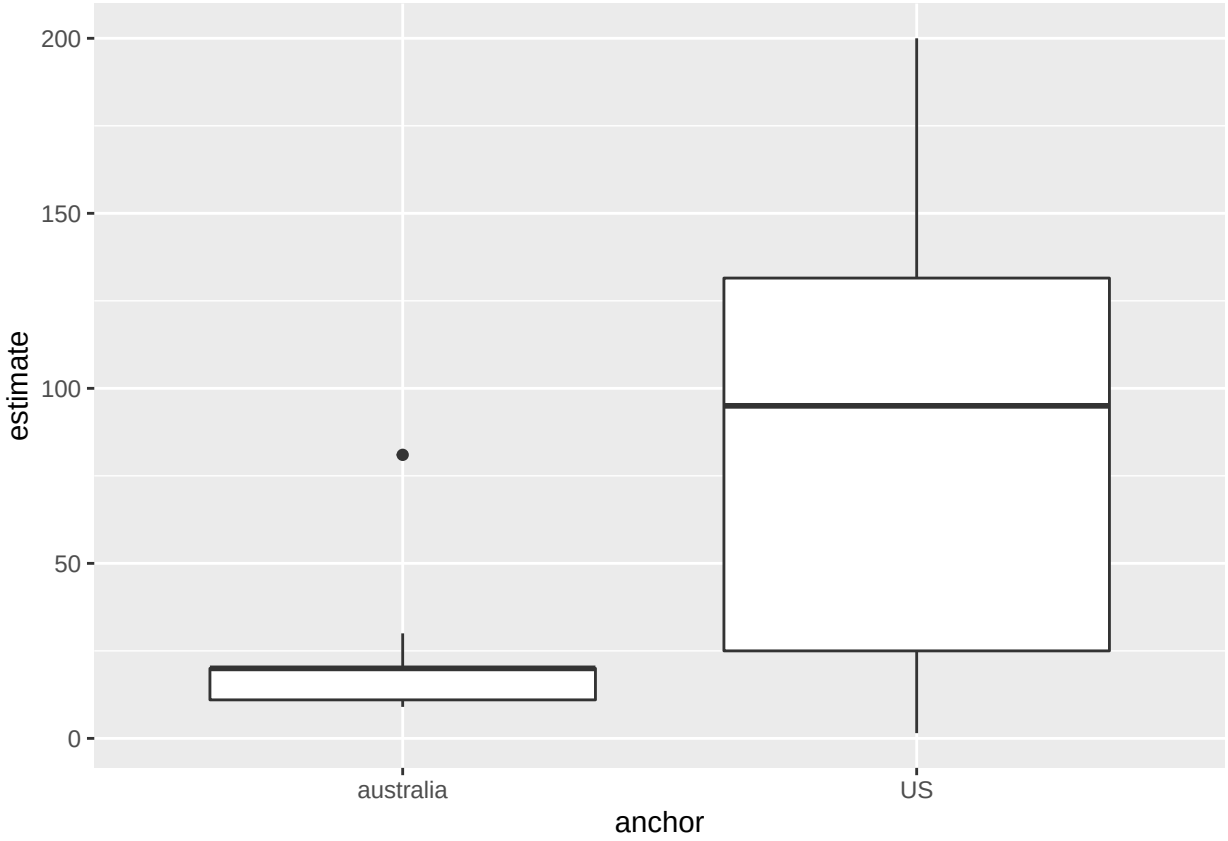
```
## Parsed with column specification:
## cols(
##   anchor = col_character(),
##   estimate = col_double()
## )
```

```
anchor
```

```
## # A tibble: 21 x 2
##    anchor estimate
##    <chr>     <dbl>
##  1 US         20
##  2 US         90
##  3 US          1.5
##  4 US        100
##  5 US        132
##  6 US        150
##  7 US        130
##  8 US         40
##  9 US        200
## 10 US         20
## # ... with 11 more rows
```

2(b) First,we need to compare the distribution of students' estimation in two different group, Second, we have 1 categorical variable and 1 quantitative variable here, so it is suitable for us to use 2 sided box plot.

```
ggplot(anchor, aes(x = anchor, y = estimate)) + geom_boxplot()
```



2(c) Choosing between welch t-test and pooled t-test, we need to consider whether two groups have same spread. Since the height of box plot australia is much smaller than that of us, estimate of australia has a much smaller variation than estimate of us. With a different variance, we should choose Welch-Satterthwaite t-test.

2(d) since we want to test if us > aus, in alphabetical order, we want to test if aus < us, so proper alternative is "less". H0: aus = us, Ha: aus < us

```r
t.test(estimate ~ anchor, data = anchor, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  estimate by anchor
## t = -3.0261, df = 10.558, p-value = 0.006019
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -26.63839
## sample estimates:
## mean in group australia        mean in group US
##                22.45455                88.35000
```

2(e) By looking at the p-value $= 0.006 < 0.05$, so we reject null hypothesis and conclude that aus $<$ us, which is to say, students given Australia as an anchor have smaller mean estimate than students given US as an anchor.