Lung Cap(cc) t.test

Module3: Hypothesis testing

ALY6010

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi

Date: Nov 20th, 2022

[Introduction]

In the real world, a researcher estimates mean of population by samples. In this module, I do t-test in R, with t.test() operation. I use dataset about lung capacity which is lung volume of someone. I will extract sample about lung capacity. By comparing critical value and p-value & alpha, I will decide to reject or not to reject null hypothesis. Furthermore, I will subset lung capacity sample from age<11 and smokers. I decide alpha as 0.01 to make more accurate test.

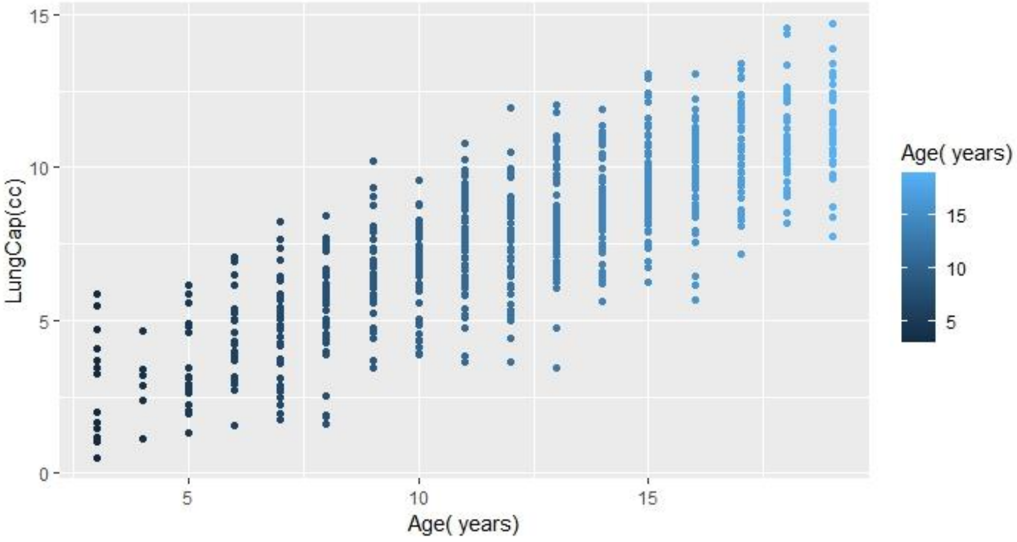
[Analysis of dataset]

Table1. pysco::describe of LungCap.xls

X	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
LungCap(cc)	1	725	7.86	2.66	8.0	7.94	2.71	0.51	14.68	14.17	-0.23	-0.33	0.10
Age(years)	2	725	12.33	4.00	13.00	12.45	4.45	3.00	19.00	16.00	-0.26	-0.71	0.15
Heigh(inches)	3	725	64.84	7.20	65.4	65.04	7.71	35.30	36.50	36.50	-0.23	-0.51	0.27

- ✓ The data's observations are 725 and There are 6 variables. I decide LungCap as target data, and there are ages, height, smoke or not, gender, caesarean or not. Mean of LungCap is 7.86 and SD is 2.66. Min and Max of LungCap are 0.51 and 14.68. There is Age which can categorize data. I think height is highly affected by age, so I will analyze age only. Age is from 3 to 19. All age is below 20.

[Jitter plot of LungCap by Age]



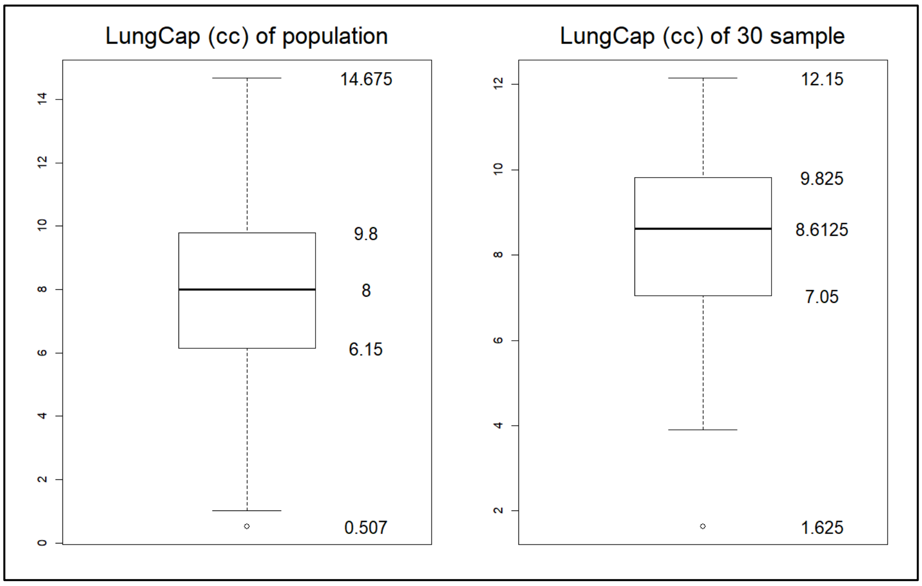
[Analysis Part 1 – Comparing Critical Value]

```
[1]   8.425   4.775   9.025   7.325 11.800   1.625   7.825   3.900
      12.150   4.900   5.550   7.200   9.375   8.800
[15] 11.400   9.800   4.525 10.600   8.375   6.800 10.550   9.650
      9.825   9.500 10.350   9.700 10.000   7.700
[29] 7.050   7.700
```

- ✓ Extraced 30 Sample from whole data to do t.test

[Boxplot]

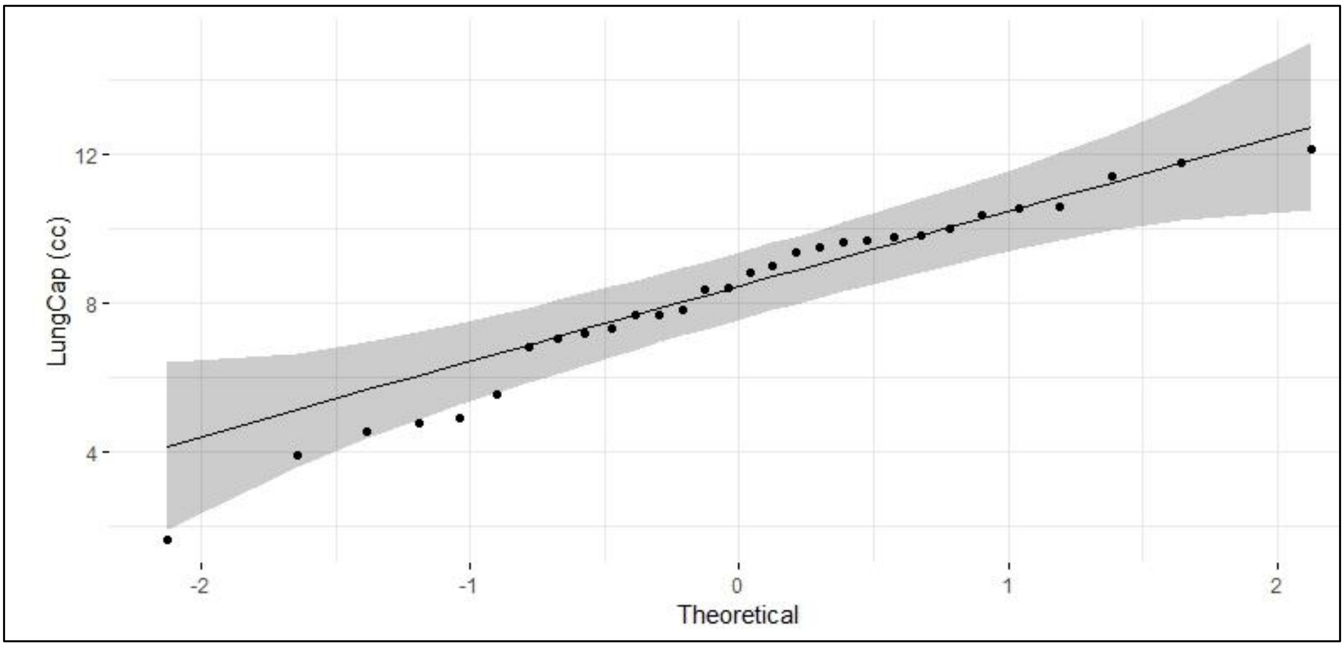
Boxplot1. LungCap (cc) of Population vs 30 Sample



- ✓ Plot1: 30 Sample(right) is extracted from population by sample() operation. n ≥ 30, so we can do t.test, but cross-check with qqplot.

[qqplot]

qqplot1. LungCap (cc) 30 Sample qqplot for normality check



- ✓ Plot2: Visual inspection of the data normality using Q-Q plots (quantile-quantile plots) (STHDA, 2022).

[Find t Critical Values in R]

<table>
<tr><td colspan="2">1. State the null and alternative hypothesis
null hypothesis: true mean is equal to 7.86
alternative hypothesis: true mean is not equal to 7.86

2. set a significance level of alpha as 0.01
qt(p=.01/2, df=29, lower.tail=F)
[1] 2.756386
-2.756386 < t < 2.756386

a <- mean(Lung$`LungCap(cc)`)
s <- sd(LungCap)
n <- 30-1
xbar <- mean(LungCap)
t <- (xbar-a)/(s/sqrt(n))

t= 0.73793</td></tr>
<tr><td>**Left-tail**</td><td>**Right-tail**</td></tr>
<tr><td>> qt(p=.01, df=29, lower.tail=T)
[1] -2.462021

-2.462021 < t (0.73793)</td><td>> qt(p=.01, df=29, lower.tail=F)
[1] 2.462021

t (0.73793) < 2.462021</td></tr>
</table>

[interpretation with critical value]

1. (two-tail test) Null hypothesis is not rejected since -2.756386 < t < 2.756386
2. (left-tail test) Null hypothesis is not rejected since -2.462021 < t (0.73793)
3. (right-tail test) Null hypothesis is not rejected since t (0.73793) < 2.462021
4. df is the degrees of freedom (df=29)
5. t-value interval is 2.756 (two-tail), 2.46 (right-tail), -2.46(left-tail)
6. sample estimates is mean value of the sample: 8.206
7. In every test, null hypothesis is not rejected, so there is no significant difference between population mean and sample mean.
   + Use n ≥ 30 data sample qqplot to see namality

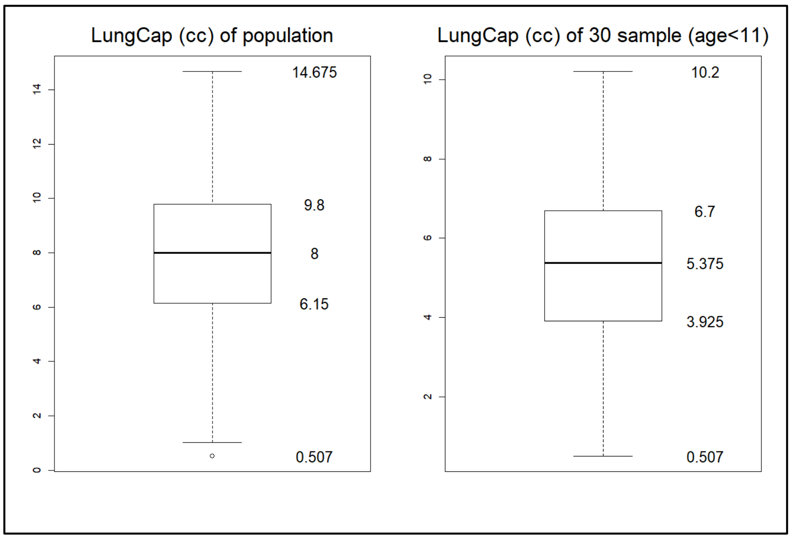
[Analysis Part 2-1 – Comparing P-value with t.test()]

```
[1] 6.225   6.450   6.700   1.775   2.550   8.425   4.425
[8] 4.825   6.950   5.375   3.975   6.000   6.175   6.950
[15] 3.825   5.775   6.125   6.300   6.575   5.025   6.125
[22] 6.950   2.250   7.975   6.850   3.100   5.375   5.950
[29] 6.100   3.925
```

- ✓ Extracted Sample which is age < 11

[Boxplot]

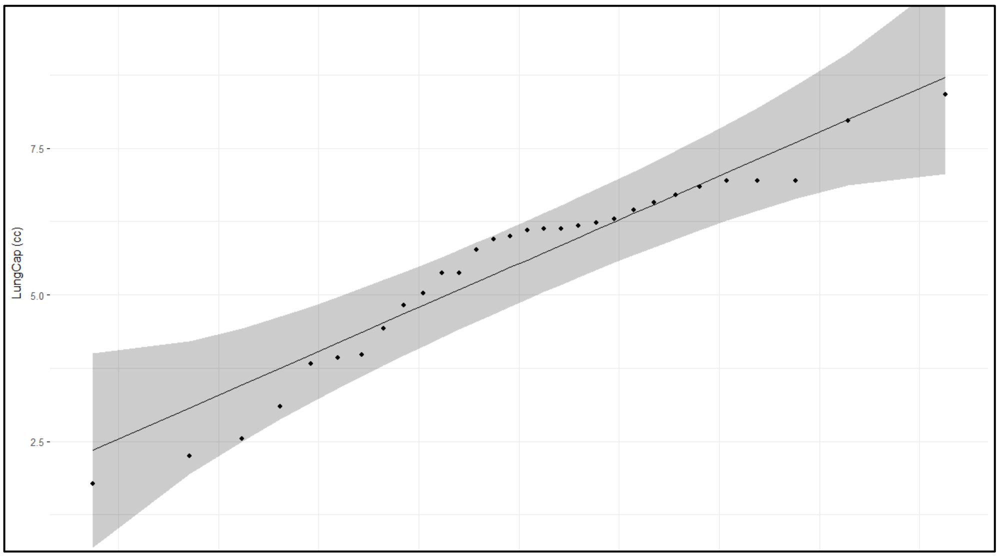
Boxplot2. LungCap (cc) of population vs age<11



- ✓ Plot2: comparing age < 11 & n=30 sample with population

[qqplot]

qqplot2. LungCap (cc) of 30 sample of age<11 group



- ✓ Plot2: Visual inspection of the data normality using Q-Q plots (quantile-quantile plots). (STHDA, 2022).

[Shapiro-Wilk normality test and t.test()]

```
> Shapiro-Wilk normality test
                    data:   LungCap.age30
              W = 0.93788, p-value = 0.07975

> t.test(LungCap.age30, mu = 7.86)
                    One Sample t-test
                    data:   LungCap.age30
            t = -7.88, df = 29, p-value = 1.087e-08
         null hypothesis: true mean is equal to 7.86
    alternative hypothesis: true mean is not equal to 7.86
               95 percent confidence interval:
                     4.888517 6.113150
                     sample estimates:
                        mean of x
                         5.500833
```

Left tail	Right tail
> t.test(LungCap.age30, mu = 7.86, + alternative = "less") One Sample t-test data: LungCap.age30 t = -7.88, df = 29, p-value = 5.433e-09 null hypothesis: true mean is equal to 7.86 alternative hypothesis: true mean is less than 7.86 95 percent confidence interval: -Inf 6.009531 sample estimates: mean of x 5.500833	> t.test(LungCap.age30, mu = 7.86, + alternative = "greater") One Sample t-test data: LungCap.age30 t = -7.88, df = 29, p-value = 1 null hypothesis: true mean is equal to 7.86 alternative hypothesis: true mean is greater than 7.86 95 percent confidence interval: 4.992136 Inf sample estimates: mean of x 5.500833

[interpretation with p-value]

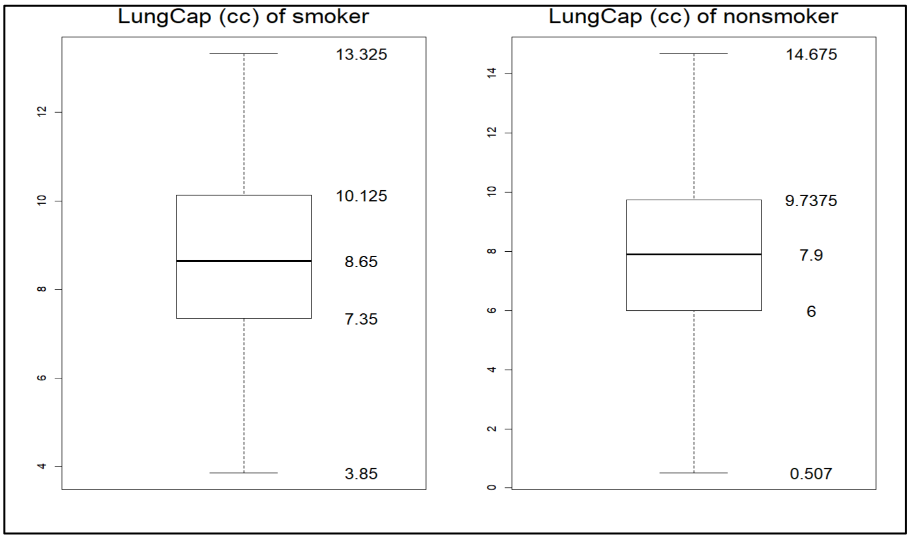
1. (two-tail test) reject null hypothesis, since 1.087e-08 < 0.01.
2. (left-tail test) reject null hypothesis, since 5.433e-09 < 0.01.
3. (right-tail test) null hypothesis is not rejected since 1 > 0.01.
4. df is the degrees of freedom (df=29)
5. p-value is the significance level of the t-test (p-value, 1.087e-08, 5.433e-09, 1)
6. sample estimate is mean value of the sample: 5.5008
7. two-tail test and left-tail test reject null hypothesis, so there is significant difference between population mean and sample mean.
8. right-tail test p-value is bigger than alpha, so null hypothesis is not rejected. It means that we can not find the evidence that population and sample mean is different.

   + Use n ≥ 30 data sample qqplot to see namality and Shapiro-Wilk normality test p-value is 0.2142 (bigger than 0.05)

[Analysis Part 2-2 – Comparing P-value with t.test()]

[Boxplot]

Boxplot3. smoker vs non-smoker of LungCap (cc)



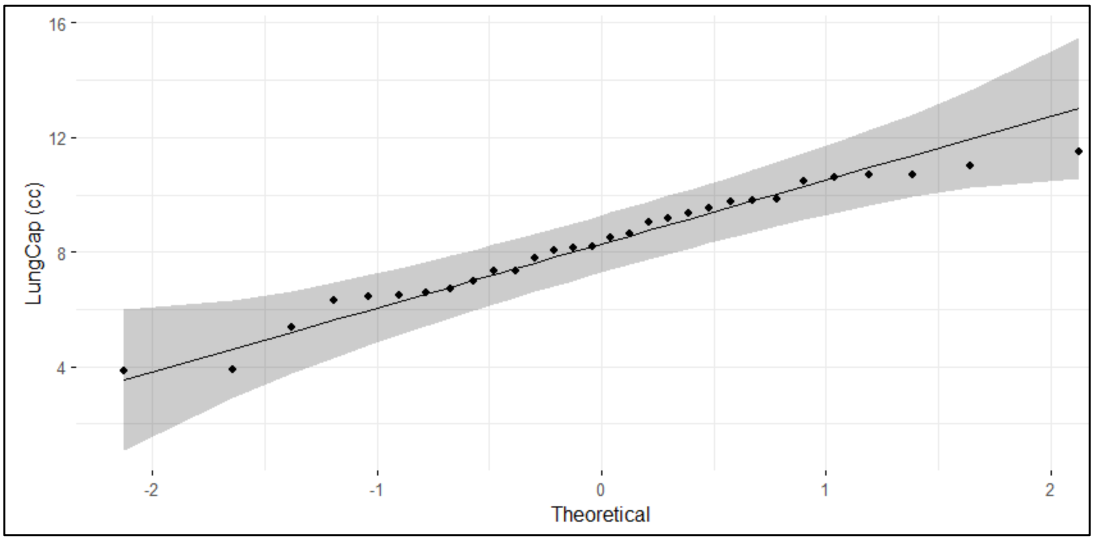
✓ Plot3: comparing all the smokers and nonsmokers, and decide to compare smokers sample with population, because it’s far from population mean.

[Extracted Sample from smoker group]

```
[1] 8.500    7.800    6.450    6.475    8.200    10.625    9.750    10.450
     7.000 11.025    6.700    9.350    7.350    9.550
[15] 3.900 10.700    5.375 10.700    8.650    3.850    6.575    9.850
      9.175    9.050    8.125 11.500    6.325    8.075
[29]   7.350    9.800
```

[qqplot]

qqplot3. LungCap (cc) of 30 smokers



✓

[Shapiro-Wilk normality test & t.test()]

```
> shapiro.test(LungCap.smk)
Shapiro-Wilk normality test
data:   LungCap.smk
W = 0.9614, p-value = 0.3362

> t.test(LungCap.smk, mu = 7.86)
One Sample t-test
data:   LungCap.smk
t = 1.1253, df = 29, p-value = 0.2697
null hypothesis: true mean is equal to 7.86
alternative hypothesis: true mean is not equal to 7.86
95 percent confidence interval:
7.521429 9.026905
sample estimates:
mean of x
8.274167
```

Left-tail	**Right-tail**
> t.test(LungCap.smk, mu = 7.86, + alternative = "less") One Sample t-test data: LungCap.smk t = 1.1253, df = 29, p-value = 0.8652 null hypothesis: true mean is equal to 7.86 alternative hypothesis: true mean is less than 7.86 95 percent confidence interval: -Inf 8.899523 sample estimates: mean of x 8.274167	> t.test(LungCap.smk, mu = 7.86, + alternative = "greater") One Sample t-test data: LungCap.smk t = 1.1253, df = 29, p-value = 0.1348 null hypothesis: true mean is equal to 7.86 alternative hypothesis: true mean is greater than 7.86 95 percent confidence interval: 7.64881 Inf sample estimates: mean of x 8.274167

[interpretation with p-value]

1. (in two-tail test) Null hypothesis is not rejected since 0.2697 > 0.01.
2. (in left-tail test) Null hypothesis is not rejected since 0.8652 > 0.01.
3. (in right-tail test) Null hypothesis is not rejected since 0.1348 > 0.01.
4. df is the degrees of freedom (df=29)
5. p-value is the significance level of the t-test (p-value = 0.03986, 0.9801, 0.01993)
6. sample estimates is mean value of the sample: 8.2741
   + Use n ≥ 30 data sample qqplot to see namality and Shapiro-Wilk normality test p-value is 0.3362 (bigger than 0.05)

[Conclusion]

There was a big difference in lungcap according to age, and samples extracted based on age < 11 showed a significant difference from the mean. There was no significant difference according to smoker.

In part 1, I extracted a sample from LungCap data and performed t.test. Initially, 30 random samples were taken. Their t-value is within critical value range, verifying the null hypothesis.

In part 2, I extract additional samples based on age and smoke. As shown in the jitter plot, age had a great effect on Lungcap, I compare lungcap by age. I choose 30 random samples from age <11 group. The null hypothesis was rejected in the two-tail and left-tail. Based on whether smoke or not, the mean appeared higher in yes group. I extract 30 samples from smokers. The results showed that all p-values were greater than 0.01, so the null hypothesis was not rejected.

[References]

RADHAKRISHNA. (2019). Lung capacity. kaggle. Retrieved from https://www.kaggle.com/datasets/radhakrishna4/lung-capacity

Kelly Black. (2015). Calculating p values. R Tutorial. Retrieved from https://www.cyclismo.org/tutorial/R/pValues.html

Zach (2020, August 6). How to find t critical values in r. STATOLOGY. Retrieved from https://www.statology.org/t-critical-value-r/

STHDA. (n.d.). One-sample t-test in r. Retrieved from http://www.sthda.com/english/wiki/one-sample-t-test-in-r

Zach (2020, October 22). How to select random samples in r (with examples). STATOLOGY. Retrieved from https://www.statology.org/random-sample-in-r/

finnstats. (2021, June 14). Reading data from excel files (xls,xlsx,csv) into r-quick guide. R bloggers. Retrieved from https://www.r-bloggers.com/2021/06/reading-data-from-excel-files-xlsxlsxcsv-into-r-quick-guide/

Zhang, Z. & Wang, L. (2022). Advanced statistics using r. Retrieved from https://advstats.psychstat.org/book/hypothesis/index.php

DataFlair. (n.d.). Introduction to hypothesis testing in r. Retrieved from https://data-flair.training/blogs/hypothesis-testing-in-r/

[Appendix: Code]

```
install.packages("devtools", "ggpubr", "readxl", "psych")
library("easypackages")
libraries("readxl", "ggpubr", "psych")
setwd("C:\\Users\\14083\\Desktop\\6010\\Module3\\dataset from TA")

#Import dataset using read.csv()
Lung <- read_excel("LungCap.xls")
headtail(Lung,5)
describe(Lung)

Lung$`LungCap(cc)`
LungCap <- sample(x=Lung$`LungCap(cc)`, size=30)
LungCap
boxplot(LungCap, ylab = "LungCap(cc)", xlab = FALSE)

#see jitter plot
ggplot(Lung, aes(x=`Age( years)`, y=`LungCap(cc)`, color=`Age( years)`)) + geom_point()
ggplot(Lung, aes(x=Smoke, y=`LungCap(cc)`, color=Smoke)) + geom_point()

#boxplot of popluation
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(Lung$`LungCap(cc)`, col='white')
fivenum(Lung$`LungCap(cc)`)
text(y=fivenum(Lung$`LungCap(cc)`), labels=fivenum(Lung$`LungCap(cc)`), x=1.35,
cex=1.5)
mtext("LungCap (cc) of population", side=3, line=1, cex=2)

#boxplot of Sample
opar <- par(no.readonly = TRUE)
par(fig=c(0.5, 1, 0, 1), new=TRUE)
boxplot(LungCap, col='white')
fivenum(LungCap)
text(y=fivenum(LungCap), labels=fivenum(LungCap), x=1.35, cex=1.5)
mtext("LungCap (cc) of 30 sample", side=3, line=1, cex=2)
par(opar)

#ggqqplot without ggplot2
ggqqplot(LungCap, ylab = "LungCap (cc)",
            ggtheme = theme_minimal())

#find critical value in R
qt(p=.01, df=29, lower.tail=T)
qt(p=.01, df=29, lower.tail=F)
qt(p=.01/2, df=29, lower.tail=F)
```

```
a <- mean(Lung$`LungCap(cc)`)
s <- sd(LungCap)
n <- 30-1
xbar <- mean(LungCap)
t <- (xbar-a)/(s/sqrt(n))
t

#Sample of age
headtail(Lung,5)
table(Lung$`Age( years)`)
Lu.age.sub <- subset(Lung, `Age( years)` < 11,
                     select = `LungCap(cc)`:`Age( years)`)
Lu.age.sub
LungCap.age30 <- sample(x=Lu.age.sub$`LungCap(cc)`, size=35)
LungCap.age30

#boxplot of popluation
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(Lung$`LungCap(cc)`, col='white')
fivenum(Lung$`LungCap(cc)`)
text(y=fivenum(Lung$`LungCap(cc)`), labels=fivenum(Lung$`LungCap(cc)`), x=1.35,
cex=1.5)
mtext("LungCap (cc) of population", side=3, line=1, cex=2)

#boxplot of Sample
opar <- par(no.readonly = TRUE)
par(fig=c(0.5, 1, 0, 1), new=TRUE)
boxplot(Lu.age.sub$`LungCap(cc)`, col='white')
fivenum(Lu.age.sub$`LungCap(cc)`)
text(y=fivenum(Lu.age.sub$`LungCap(cc)`), labels=fivenum(Lu.age.sub$`LungCap(cc)`),
x=1.35, cex=1.5)
mtext("LungCap (cc) of 30 sample (age<11)", side=3, line=1, cex=2)
par(opar)

#ggqqplot without ggplot2
ggqqplot(LungCap.age30, ylab = "LungCap (cc)",
         ggtheme = theme_minimal())

shapiro.test(LungCap.age30)
# One-sample t-test
t.test(LungCap.age30, mu = 7.86)

#left tail
t.test(LungCap.age30, mu = 7.86,
       alternative = "less")
```

```
#right tail
t.test(LungCap.age30, mu = 7.86,
        alternative = "greater")
headtail(Lung,5)
table(Lung$Smoke)

Lu.smk.yes <- subset(Lung, Smoke == "yes",
                      select = c("LungCap(cc)","Smoke"))
Lu.smk.no <- subset(Lung, Smoke == "no",
                      select = c("LungCap(cc)","Smoke"))

#boxplot of popluation
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(Lu.smk.yes$`LungCap(cc)`, col='white')
fivenum(Lu.smk.yes$`LungCap(cc)`)
text(y=fivenum(Lu.smk.yes$`LungCap(cc)`), labels=fivenum(Lu.smk.yes$`LungCap(cc)`),
x=1.35, cex=1.5)
mtext("LungCap (cc) of smoker", side=3, line=1, cex=2)

#boxplot of Sample
opar <- par(no.readonly = TRUE)
par(fig=c(0.5, 1, 0, 1), new=TRUE)
boxplot(Lu.smk.no$`LungCap(cc)`, col='white')
fivenum(Lu.smk.no$`LungCap(cc)`)
text(y=fivenum(Lu.smk.no$`LungCap(cc)`), labels=fivenum(Lu.smk.no$`LungCap(cc)`),
x=1.35, cex=1.5)
mtext("LungCap (cc) of nonsmoker", side=3, line=1, cex=2)
par(opar)

Lu.smk.yes
LungCap.smk <- sample(x=Lu.smk.yes$`LungCap(cc)`, size=30)
LungCap.smk

#ggqqplot without ggplot2
ggqqplot(LungCap.smk, ylab = "LungCap (cc)",
          ggtheme = theme_minimal())

shapiro.test(LungCap.smk)
# One-sample t-test
t.test(LungCap.smk, mu = 7.86)
#left tail
t.test(LungCap.smk, mu = 7.86,
        alternative = "less")
#right tail
t.test(LungCap.smk, mu = 7.86,
        alternative = "greater")
```