HW 2

309554005 黃睿宇

Problem 1

(a)

```
pizza <- read.table(file="pizza2.txt", header=TRUE)
a <- pizza[pizza[,"heat"]=="Coal", "rating"]
sum(a) / length(a)
```

```
## [1] 4.688824
```

```
summation  = 0
for (i in a){
  summation = summation + sum(i - mean(a))^2
}
(summation / (length(a)-1)) ^ 0.5
```

```
## [1] 0.4867479
```

average rating of the pizzas baked by coal = 4.688824
standard deviation of the ratings baked by coal = 0.4867479

```
a <- pizza[pizza[,"heat"]=="Wood", "rating"]
sum(a) / length(a)
```

```
## [1] 3.8764
```

```
summation  = 0
for (i in a){
  summation = summation + sum(i - mean(a))^2
}
(summation / (length(a)-1)) ^ 0.5
```

```
## [1] 1.537248
```

average rating of the pizzas baked by wood = 3.8764
standard deviation of the ratings baked by wood = 1.537248

```r
a <- pizza[pizza[,"heat"]=="Gas", "rating"]
sum(a) / length(a)
```

```
## [1] 2.961013
```

```r
summation  = 0
for (i in a){
  summation = summation + sum(i - mean(a))^2
}
(summation / (length(a)-1)) ^ 0.5
```

```
## [1] 1.817251
```

average rating of the pizzas baked by gas = 2.961013
standard deviation of the ratings baked by gas = 1.817251

The average rating of the pizzas baked by coal is the largest, and the rating is the most concentrated. On the contrary, the average rating of the pizzas baked by gas is the smallest, and the rating is the most dispersed.

(b)

```r
model <- lm(rating ~ heat, data=pizza)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: rating
##            Df Sum Sq Mean Sq F value    Pr(>F)
## heat        2  58.04  29.022  9.8749 8.184e-05 ***
## Residuals 197 578.98   2.939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value $= 8.184 \times 10^{-5} < 0.05$, we could conclude that there is significant difference among variables in heat source. Through F value, we could also infer that the difference within each heat source might be larger than that among different heat sources.

(c)

```r
summary(model)
```

```
##
## Call:
## lm(formula = rating ~ heat, data = pizza)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -3.506 -1.715  0.379  1.562  2.039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6888     0.4158  11.277  < 2e-16 ***
## heatGas      -1.7278     0.4376  -3.948 0.000109 ***
## heatWood     -0.8124     0.5389  -1.507 0.133289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.714 on 197 degrees of freedom
## Multiple R-squared:  0.09112,    Adjusted R-squared:  0.08189
## F-statistic: 9.875 on 2 and 197 DF,  p-value: 8.184e-05
```

For coal: estimated coefficient $= 4.6888$, p-value $< 2 \times 10^{-16}$
For gas: estimated coefficient $= -1.7278$, p-value $= 0.000109$
For wood: estimated coefficient $= -0.8124$, p-value $= 0.133289$

(d)
In univariate analysis, we could only know the information of each variable. While in ANOVA, we could know whether the averages in different heat sources are equal or not. And from the result obtained in (a) is similar as (c), we could know that the variation of coal is the largest among the three.

Problem 2

```r
model1 <- lm(rating ~ heat + area + cost, data=pizza)
summary(model1)
```

```
##
## Call:
## lm(formula = rating ~ heat + area + cost, data = pizza)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98864 -0.52516  0.00599  0.51428  1.92332
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.72260    0.34461   2.097  0.03731 *
## heatGas         -1.59555    0.20526  -7.773 4.52e-13 ***
## heatWood        -0.45753    0.26056  -1.756  0.08069 .
## areaEVillage     4.17970    0.24628  16.971  < 2e-16 ***
## areaLES          2.37294    0.26106   9.089  < 2e-16 ***
## areaLittleItaly  0.78700    0.25268   3.115  0.00212 **
## areaSoHo         3.65362    0.24498  14.914  < 2e-16 ***
## cost             0.43865    0.06613   6.633 3.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7957 on 192 degrees of freedom
## Multiple R-squared:  0.8092, Adjusted R-squared:  0.8022
## F-statistic: 116.3 on 7 and 192 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(rating ~ heat_re + area + cost, data=pizza)
summary(model2)
```

```
##
## Call:
## lm(formula = rating ~ heat_re + area + cost, data = pizza)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97759 -0.51011 -0.02969  0.52497  2.15583
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.96212    0.31668   3.038  0.00271 **
## heat_re         -0.87601    0.09242  -9.479  < 2e-16 ***
## areaEVillage     4.10646    0.24378  16.845  < 2e-16 ***
## areaLES          2.26091    0.25405   8.900 4.08e-16 ***
## areaLittleItaly  0.69163    0.24774   2.792  0.00577 **
## areaSoHo         3.54383    0.23768  14.910  < 2e-16 ***
## cost             0.44911    0.06618   6.786 1.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7997 on 193 degrees of freedom
## Multiple R-squared:  0.8062, Adjusted R-squared:  0.8002
## F-statistic: 133.8 on 6 and 193 DF,  p-value: < 2.2e-16
```

If we use "heat_re" as the predictor variables, it will have the weight error (the weight of gas is two times larger than that of wood).

```
model1 <- lm(rating ~ heat + area + cost, data=pizza)
predict(model1, data.frame(heat="Coal", area="LittleItaly", cost=2.5))
```

```
##        1
## 2.606232
```

```
model2 <- lm(rating ~ heat_re + area + cost, data=pizza)
predict(model2, data.frame(heat_re=0, area="LittleItaly", cost=2.5))
```

```
##        1
## 2.776521
```

prediction of model a = 2.606232
prediction of model b = 2.776521

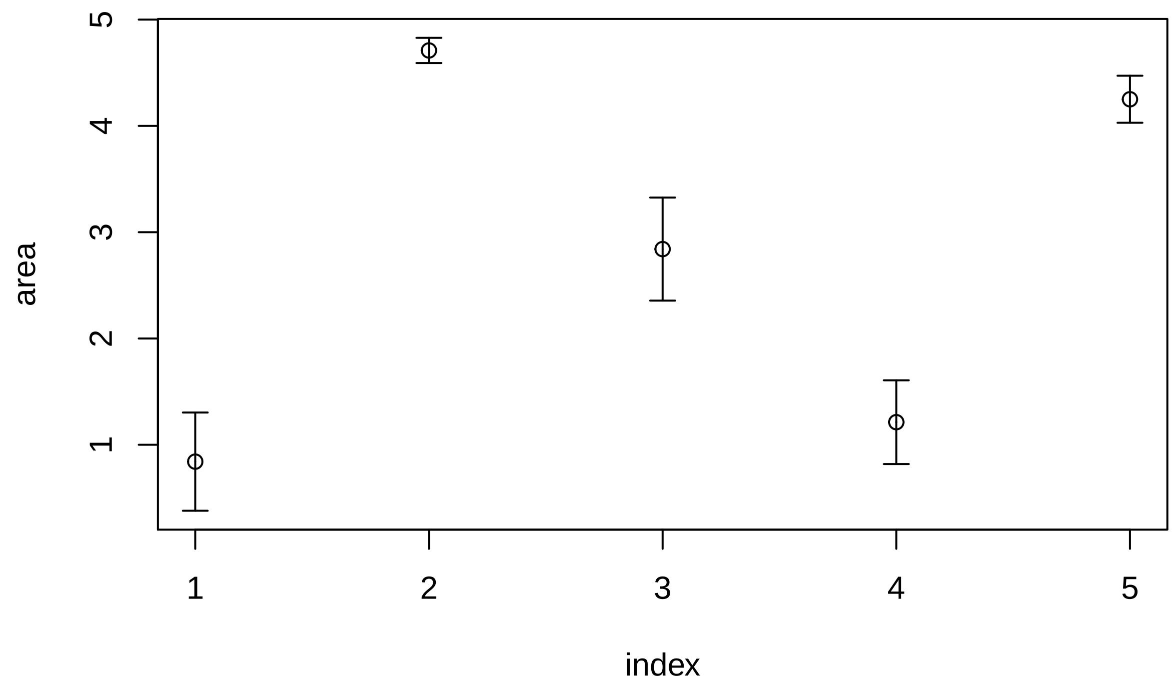
Problem 3

```r
x1 <- pizza[pizza[,"area"]=="Chinatown", "rating"]
m1 <- mean(x1)
U1 <- m1 + qnorm(0.975) * sd(x1) / sqrt(length(x1))
L1 <- m1 - qnorm(0.975) * sd(x1) / sqrt(length(x1))

x2 <- pizza[pizza[,"area"]=="EVillage", "rating"]
m2 <- mean(x2)
U2 <- m2 + qnorm(0.975) * sd(x2) / sqrt(length(x2))
L2 <- m2 - qnorm(0.975) * sd(x2) / sqrt(length(x2))

x3 <- pizza[pizza[,"area"]=="LES", "rating"]
m3 <- mean(x3)
U3 <- m3 + qnorm(0.975) * sd(x3) / sqrt(length(x3))
L3 <- m3 - qnorm(0.975) * sd(x3) / sqrt(length(x3))

x4 <- pizza[pizza[,"area"]=="LittleItaly", "rating"]
m4 <- mean(x4)
U4 <- m4 + qnorm(0.975) * sd(x4) / sqrt(length(x4))
L4 <- m4 - qnorm(0.975) * sd(x4) / sqrt(length(x4))

x5 <- pizza[pizza[,"area"]=="SoHo", "rating"]
m5 <- mean(x5)
U5 <- m5 + qnorm(0.975) * sd(x5) / sqrt(length(x5))
L5 <- m5 - qnorm(0.975) * sd(x5) / sqrt(length(x5))

y <- c(m1, m2, m3, m4, m5)
U <- c(U1, U2, U3, U4, U5)
L <- c(L1, L2, L3, L4, L5)

require(plotrix)
```

```
## Loading required package: plotrix
```

```
## Warning: package 'plotrix' was built under R version 4.0.3
```

```r
plotCI(1:5, y, ui=U, li=L, xlab="index", ylab="area")
```



We could see that the largest mean rating is in EVillage, while its confidence interval is the smallest. The smallest mean rating is in Chinatown, while its confidence interval is the largest.