One-way ANOVA

Mengye Liu

2022/04/20

```
library(knitr)# making tables in rmd
library(dplyr)# group_by, summarise
```

Firstly, using data from problem 3.9 as an example.

```
#import the data
Strength = c(3129, 3000, 2865, 2890,3200, 3300, 2975,
             3150,2800, 2900, 2985, 3050, 2600, 2700, 2600, 2765)
Tech = factor(c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4))
data = data.frame(strength = Strength, tech = Tech)
head(data)
```

```
##   strength tech
## 1     3129    1
## 2     3000    1
## 3     2865    1
## 4     2890    1
## 5     3200    2
## 6     3300    2
```

Then we can get the anova table.

```
fit = aov(strength~tech, data = data)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## tech         3 489740  163247   12.73 0.000489 ***
## Residuals   12 153908   12826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Or we can use function **kable** in package **knitr** to make it as a table in your homework file.

```
opts <- options(knitr.kable.NA = "")
summ = summary(fit)
kable(summ[[1]])
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tech	3	489740.2	163246.73	12.72811	0.0004887
Residuals	12	153908.2	12825.69		

Notice: the ANOVA output of R here does not have SS_{total} line, in your homework you need to compute it additionally to get the full ANOVA table.

Also we can get the value in the anova table by hand or by formula.

```r
N = length(data$strength)# total amount of obs
t = length(levels(data$tech))
g = data %>%
  group_by(tech) %>%
  summarise(m = mean(strength), n = n(), .groups = 'drop')#m: yi._bar; n: n_i
s_overall = sum(data$strength)#y.. = sum(yij)
m_overall = mean(data$strength)#y.._bar
```

Compute SS_{total}, Sum of squares of Total line in ANOVA table.

```r
SST1 = var(data$strength)*(N-1)
SST2 = sum((data$strength - m_overall)^2)##sum(yij - y.._bar)^2
SST3 = sum(data$strength^2) - s_overall^2/N ##sum(yij)^2 - y..^2/N
SST4 = sum(data$strength^2) - N*m_overall^2 ##sum(yij)^2 - N* y.._bar^2
```

Compute $SS_{between}$

```r
SSB1 = sum(g$n*(g$m - m_overall)^2)##sum(n_i(yi._bar- y.._bar)^2)
SSB2 = sum(g$n* g$m^2) - s_overall^2/N ##sum(n_i*yi._bar^2) - y..^2/N   (sum(yi.^2/n_i) - y..^2/N)
SSB3 = sum(g$n* g$m^2) - N*m_overall^2
```

Then SSE = SST - SSB. We can have the ANOVA table as

```r
df = c(t-1, N-t, N-1)
SS = c(SSB1, SST1 - SSB1, SST1)
MS  = SS/df
MS = c(MS[-3],NA)
F_value = c(MS[1]/MS[2],NA,NA)
anova = data.frame(df,SS,MS,F_value)
rownames(anova) = c('Between treatment', 'Error', 'Total')
kable(anova)
```

	df	SS	MS	F_value
Between treatment	3	489740.2	163246.73	12.72811
Error	12	153908.2	12825.69	
Total	15	643648.4		