Homework #1 **20180047 김경수**

*rds 파일들은 다음과 같이 implement 하였다.
clinical<-readRDS("clinical.rds") *# Clinical annotation*
expression<-readRDS("expression.rds") *# Gene expression*
mutation<-readRDS("mutation.rds") *# Mutation*

Homework #1-1

1. Procedure & R code

사건 B에 대한 사건 A의 조건부 확률은 다음과 같다. $P(A|B) = P(A \cap B)/P(B)$.
표본공간 S를 모든 환자들로 설정하여, 위의 정의와 같이 구할 수도 있지만,
본 레포트에서는 표본공간 S*를 특정 subtype을 가지는 환자들로 설정하여,
n(특정 subtype을 가진 환자들 중 5년보다 오래 생존한 환자들의 수) / n(S*) 로 직접적으로 구하였다.

(1) 먼저 subtype이 Normal-like인 clinical data들을 dplyr::filter 함수를 이용하여 제거한다.
R code : `clinical_filter<-dplyr::filter(clinical, subtype!="Normal-like")`
(2) unique 함수를 이용하여 분석해야 할 subtype들을 확인한다.
R code : `subtype<-unique(clinical_filter$subtype)`

NA를 제외한 각각의 subtype에 대해서 아래의 (3)~(5)과정을 각각 수행한다.
R code : `key <- " subtype ' s name "`
(3) dplyr::filter 함수를 이용하여 subtype이 key인 clinical data들만 뽑아낸다.
R code : `clinical_filter_key<-dplyr::filter(clinical_filter, subtype==key)`
(4) length와 which 함수를 이용하여 각 subtype별, 5년보다 오래 생존한 환자의 수와, 총 환자의 수와, 그로 인해 계산된 5-year survival probability를 변수에 저장한다,

```
R code : survived_patients<-length(which(clinical_filter_key$survival_time>1825))
         total_patients<-length(which(clinical_filter_key$survival_time>=0))
         survival_rate<-survived_patients/total_patients
```

(5) round 함수를 이용하여 소수점 셋째 자리까지 나타내며, print와 paste 함수를 이용하여 interface에 표시한다.

```
R code : print(paste("This is the result which subtype is ", key, ".  Total number
of the survived patients are ", survived_patients, ". The number of all patients are
", total_patients, ". In conclusion, 5-year survival probability of breast cancer
patients which subtype is ", key, "is ", round(survival_rate, 3), "."))
```

2. Results

각 subtype별 환자들의 5-year survial probability는 다음과 같다.

Table 1. 각 subtype별 환자들의 5-year survival probability

Subtype	Luminal A	Basal-like	Luminal B	HER2-enriched
survived patients	59 people	28 people	24 people	8 people
Total patients	196 people	83 people	104 people	45 people
Probability	0.301	0.337	0.231	0.178

Homework #1-2

1. Procedure & R code

Normal-like를 제외하고 총 4가지 subtype이 존재한다.(Luminal A, Basal-like, Luminal B, HER2-enriched). 각각의 subtype들에 대해 다음의 과정을 수행한다.
R code : `key<-"subtype ' s name "`

사건 B에 대한 사건 A의 조건부 확률은 다음과 같다. $P(A|B) = P(A \cap B)/P(B)$.
표본공간 S를 모든 환자들로 설정하여, 위의 정의와 같이 구할 수도 있지만,
본 레포트에서는 표본공간 S*를 특정 subtype을 가지는 환자들로 설정하여,
n(특정 subtype을 가진 환자들 중 특정 mutation을 가지고 있는 환자들의 수) / n(S*) 로 직접적으로 구하였다.

(1) 특정 subtype의 clinical data와 mutation data를 dplyr::filter, which 함수를 이용하여 얻는다.

```
R code : clinical_key<-dplyr::filter(clinical,subtype==key)
         sample_id_filter<-clinical_key$sample_id
         mutation_key<-dplyr::filter(mutation, sample_id %in% sample_id_filter)
```

(2) Top 10 highly driving mutation을 구해야 하므로, 다음과 같이 변수를 설정한다.
gene_key_raw : mutation vector(Hugo_Symbol).
gene_key : 중복을 허용하지 않은 mutation vector.
cnt : 각 mutation이 일어난 횟수를 기록하기 위한 vector.

```
R code : gene_key_raw<-(mutation_key$Hugo_Symbol)
         gene_key<-unique(gene_key_raw)
         cnt<-rep(0,length(gene_key))
```

(3) cnt에 gene_key vector에 있는 mutation이 일어난 횟수를 for문을 이용하여 기록한다.

R code :

```
for(i in 1:length(gene_key_raw)){
     for(j in 1 : length(gene_key)){
         if(gene_key_raw[i]==gene_key[j]){
             cnt[j]<-cnt[j]+1
             break
         }
     }
}
```

(4) for문과 max, which.max 함수를 이용하여 top 10 highly driving mutation을 구하고, 이를 console 창에 표시한다.(round 함수를 이용하여 소수 셋째자리까지 표현)

R code :

```
print(paste("This is the result which subtype is ",key,". "))
for(i in 1:10){
          temp<-gene_key[which.max(cnt)]
          print(paste("Total number of patients is ",length(sample_id_filter),". The number of patients with ",temp," mutation is ",max(cnt),". In conclusion, conditional probability is ",round(max(cnt)/length(sample_id_filter),3),"."))
          cnt[which.max(cnt)]<-0
      }
```

2. Results

각 subtype별 환자들의 5-year survial probability는 다음과 같다.

Table 1. Luminal A 환자들의 Top 10 highly driving mutation

Mutation	Patients (people)	Total patients (people)	Probability
TP53	75	196	0.383
PIK3CA	75		0.383
TTN	50		0.255
MUC12	37		0.189
GATA3	32		0.163
MUC16	27		0.138
MAP3K1	27		0.138
MUC4	24		0.122
CSMD2	20		0.102
UBR4	19		0.097

Table 2. Basal-like 환자들의 Top 10 highly driving mutation

Mutation	Patients (people)	Total patients (people)	Probability
TTN	54	83	0.651
TP53	32		0.386
PIK3CA	22		0.265
DST	17		0.205
MUC16	17		0.205
RYR2	12		0.145
XIRP2	11		0.133
CMYA5	11		0.133
AHNAK2	11		0.133
SYNE2	10		0.12

Table 3. Luminal B 환자들의 Top 10 highly driving mutation

Mutation	Patients (people)	Total patients (people)	Probability
TTN	57	104	0.548
PIK3CA	55		0.529
TP53	37		0.356
MUC16	33		0.317
AHNAK	23		0.221
GPR112	19		0.183
SPEN	18		0.173
MUC17	15		0.144
MAP3K1	13		0.125
MLL3	13		0.125

Table 4. HER2-enriched 환자들의 Top 10 highly driving mutation

Mutation	Patients (people)	Total patients (people)	Probability
PIK3CA	25	45	0.556
MAP3K1	12		0.267
TP53	10		0.222
TTN	9		0.2
GATA3	8		0.178
FLNA	7		0.156
SYNE2	7		0.156
CDH1	6		0.133
MUC4	6		0.133
ADNP	6		0.133