Homework #2 **20180047 김경수**

*rds 파일들은 다음과 같이 implement 하였고, subtype이 NA인 데이터들을 다음과 같이 배제시켰다.

```
clinical<-readRDS("clinical.rds") # Clinical annotation
expression<-readRDS("expression.rds") # Gene expression
mutation<-readRDS("mutation.rds") # Mutation

sample_id_filter<-intersect(colnames(expression),
intersect(unique(mutation$sample_id),
clinical$sample_id[which(!is.na(clinical$subtype))]))
clinical_filter<-dplyr::filter(clinical,sample_id %in% sample_id_filter)
mutation_filter<-dplyr::filter(mutation, sample_id %in% sample_id_filter)
expression_filter<-expression[,sample_id_filter]
```

Homework #2-1

1. Procedure & R code

이론적 배경은 다음과 같다.
각 stage별 환자의 수가 각각 1단계부터 3단계까지 79, 254, 86이고, 모분산 또한 불명확하므로, Welch's T-test(one-sided)를 통해 단계의 증가와 환자의 생존시간 감소와의 상관관계에 대해 파악하였다.
Let i < j. (i, j = 1,2,3, i != j)
H0 : stage i의 환자의 생존시간 <= stage j의 환자의 생존시간
HA : stage I의 환자의 생존시간 > stage j의 환자의 생존시간
if(p-value < 0.5) -> reject H0, else accept H0

(1) boxplot 함수를 이용하여 유방암의 stage와 survival time의 상관관계에 대한 그래프를 그린다.

```
R code : boxplot(survival_time~stage,data=clinical_filter,xlab="Clinical
Stage",ylab="Survival Time (days)")
```

(2) t.test(built-in function)를 이용하여, p-value값을 얻음으로써 단계의 증가와 환자의 생존시간 감소와의 상관관계에 대해 파악한다. (H0, HA, 그리고 H0를 reject, accept하는 기준 등 자세한 내용은 앞에서 서술하였다.)

```
R code :
t.test(clinical_filter$survival_time[which(clinical_filter$stage=="stage i")],
clinical_filter$survival_time[which(clinical_filter$stage=="stage ii")],
'greater')$p.value

t.test(clinical_filter$survival_time[which(clinical_filter$stage=="stage i")],
```

```
clinical_filter$survival_time[which(clinical_filter$stage=="stage iii")],
'greater')$p.value

t.test(clinical_filter$survival_time[which(clinical_filter$stage=="stage ii")],
clinical_filter$survival_time[which(clinical_filter$stage=="stage iii")],
'greater')$p.value
```

2. Results

유방암의 stage(i, ii, iii)와 survival time의 상관관계의 그래프는 다음과 같다.

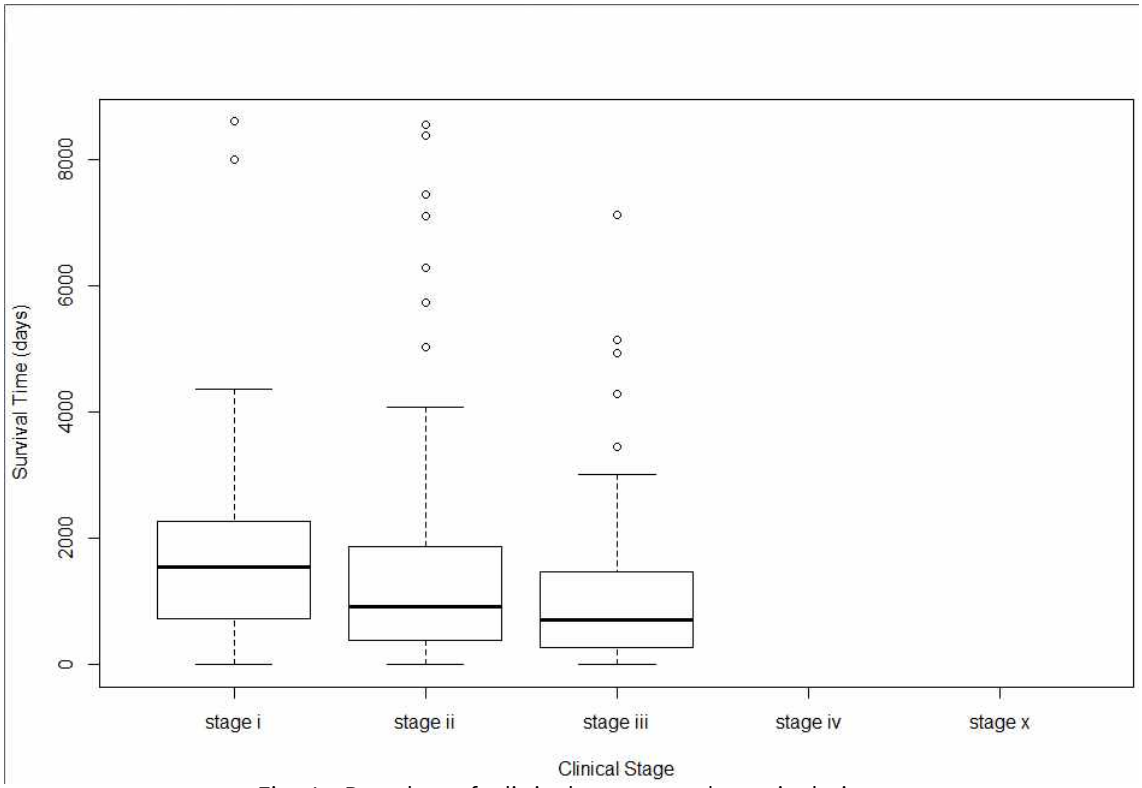

Fig 1. Boxplot of clinical stage and survival time

또한 Welch's T-test의 p-value와 그로 인한 H0의 채택여부는 다음과 같다. 이를 바탕으로 Stage2와 Stage3의 경우를 제외하고는 significance level 0.05에서 stage가 증가할수록 survival time이 감소한다는 결론을 내릴 수 있다.

Table 1. Results of Welch's T-test

	Stage 1 vs Stage 2	Stage 1 vs Stage 3	Stage 2 vs Stage 3
p-value	0.0075	0.0014	0.1151
Reject H0?	yes	yes	no

Homework #2-2

1. Procedure & R code

이론적 배경은 다음과 같다.
각 stage별 expression data를 추출하였으며, stage별 환자의 수가 각각 1단계부터 3단계까지 79, 254, 86이고, 각각의 gene에 대한 expression data들의 모분산 또한 불명확하므로, 각각의 gene에 대해 Welch's T-test(two-sided)를 적용하였다. 이를 통해 주어진 gene이 (differentially expressed gene)DEG인지 확인하였다.
Let i < j. (i, j = 1,2,3, i != j)
H0 : 주어진 gene의 stage i에서의 expression = 주어진 gene의 stage j에서의 expression
HA : 주어진 gene의 stage i에서의 expression != 주어진 gene의 stage j에서의 expression
if(p-value < 0.5) -> reject H0, else accept H0

또한 각 gene에 대해 DEG analysis를 수행하는데, 반복 수행 횟수에 비례하여 false positive가 증가한다. 따라서 이 비율의 통제가 필요하므로, multiple testing error correction을 진행한다.(FDR correction)

(1) Expression data에서 각 stage에 속하는 환자의 expression data를 which function을 이용하여 뽑아낸다.
R code :

```
stage1<-as.character(clinical_filter$sample_id[which(clinical_filter$stage=="stage i")])
stage2<-as.character(clinical_filter$sample_id[which(clinical_filter$stage=="stage ii")])
stage3<-as.character(clinical_filter$sample_id[which(clinical_filter$stage=="stage iii")])
expression.stage3<-expression_filter[,stage3]
expression.stage2<-expression_filter[,stage2]
expression.stage1<-expression_filter[,stage1]
```

(2) 각 gene에 대해 위에서 명시한 t-test를 진행하기 위해, sapply function을 이용하여 각 gene의 p-value 값들을 저장한다.
R code :

```
pvalues<-sapply(c(1:dim(expression_filter)[1]),
                FUN=function(k){
                  pval<-t.test(expression.stage1[k,],expression.stage2[k,])$p.value
                  return(pval)
                })
```

(3) FDR correction을 진행한 후, FDR correction을 진행하기 전/후의 DEG의 수를 length function을 이용하여 비교한다.
R code :

```
pvalues.adj<-p.adjust(pvalues,method = "fdr")
length(pvalues[which(pvalues<0.05)])
length(pvalues.adj[which(pvalues.adj<0.05)])
```

(5) Stage2와 stage3, stage1과 stage3에 대해서도 위의 과정을 반복한다.

2. Results

각 stage의 조합에 대해서, significance level 0.05일 시 DEG의 개수는 다음과 같다.

Table 2. Results of Welch's T-test

	Stage 1 vs Stage 2	Stage 1 vs Stage 3	Stage 2 vs Stage 3
Number of DEGs (before FDR correction)	1542	2042	1915
Number of DEGs (after FDR correction)	1	19	8

Homework #2-3

1. Procedure & R code

이론적 배경은 다음과 같다.
우리가 원하는 것은 significance level 0.05에서 각각의 subtype이 변화하면 survival time에 significant difference를 주는지 여부를 확인하는 것이다. ANOVA test를 적용하여 이를 해결하였다.

H0 : expectation of survival time is equal regardless of subtype.
HA : At least one of the expectation of survial time of subtype is different with each other.
if(Pr(>F) < 0.05) -> reject H0, else accept H0

만약 H0가 기각되면 post-hoc analysis를 진행하여야 한다. Accept될 경우 진행하지 않아도 된다.

(1) Subtype이 NA, Normal-like인 data를 제거한 후, boxplot을 그린다.
R code : clinical_filter<-dplyr::filter(clinical, subtype!="NA")
clinical_filter<-dplyr::filter(clinical_filter, subtype!="Normal-like")
boxplot(survival_time~subtype,data=clinical_filter,xlab="Subtype",ylab="Survival Time (days)")

(2) 위에서 서술한 ANOVA test를 진행하고 F value와 Pr(> F)를 확인한다.(H0가 accpet 되었으므로 post-hoc analysis는 진행하지 않는다.)

R code : `anova.result <- aov(survival_time~subtype,data=clinical_filter)`

```
summary_ANOVA <- summary(anova.result)
summary_ANOVA
```

2. Results

ANOVA test의 결과는 다음과 같다. F value는 1.485이고 Pr(> F)는 0.218이다. 따라서 significance level 0.05에서 subtype의 변화에 따른 survival time에 대한 significant difference는 없다고 결론 내릴 수 있다.

Table 3. Results of ANOVA test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
subtype	3	8486086	2828695	1.485	0.218
Residuals	424	807871675	1905358		