Homework 3

Ruoyuan Qian

2019/10/26

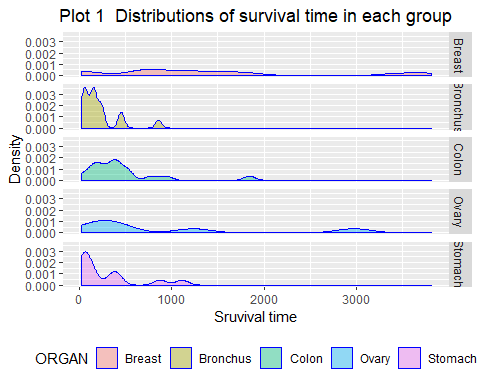
# Problem 2

## a)

surv = read\_csv(".\\data\\SurvCancer.csv")

## Parsed with column specification:  
## cols(  
## SURVIVAL = col\_double(),  
## ORGAN = col\_character()  
## )

ggplot(surv, aes( x = SURVIVAL,   
 fill = ORGAN,group=ORGAN)) +   
 geom\_density(alpha = .4, adjust = .5, color = "blue") +   
 facet\_grid(ORGAN ~. )+  
 theme(legend.position = "bottom")+  
 labs(x = "Sruvival time",  
 y = "Density",  
 title = "Plot 1 Distributions of survival time in each group"  
 )+  
 theme(plot.title = element\_text(hjust = 0.5))



surv %>%   
 group\_by(ORGAN) %>%   
 mutate( median = median(SURVIVAL),  
 IQR = paste(quantile(SURVIVAL)[2],"-",  
 quantile(SURVIVAL)[4])) %>%   
 select(-SURVIVAL) %>%   
 distinct() %>%   
 knitr::kable(caption = "Table 1 Discriptive statistics of each group",  
 format = "html")

Table 1 Discriptive statistics of each group

ORGAN

median

IQR

Stomach

124

46 - 396

Bronchus

155

72 - 245

Colon

372

189 - 519

Ovary

406

239.75 - 1039.5

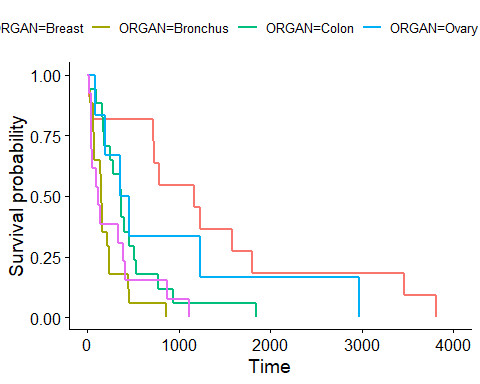
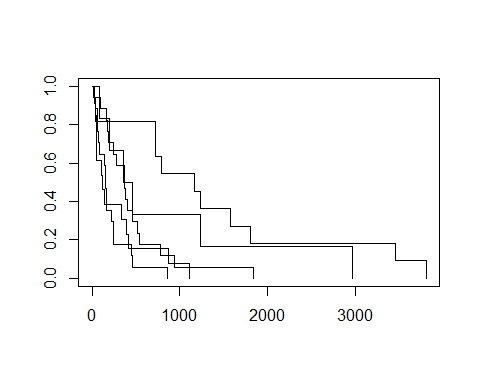
Breast

1166

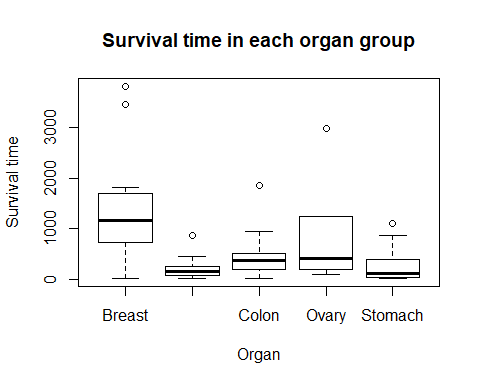
723 - 1692.5

surv\_p =  
surv %>%   
 mutate(status = 1)

fit<-survfit(Surv(SURVIVAL,status)~ORGAN,data=surv\_p)  
plot(fit)   
  
  
ggsurvplot(fit)



boxplot(SURVIVAL~ORGAN,data=surv\_p,main="Survival time in each organ group",xlab="Organ", ylab="Survival time")



## b)

Calculated by R

res1<-aov(SURVIVAL~factor(ORGAN), data=surv)  
  
summary(res1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(ORGAN) 4 11535761 2883940 6.433 0.000229 \*\*\*  
## Residuals 59 26448144 448274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

qf(0.99,4,59)

## [1] 3.654961

Calculated by hand

sum((surv[,1])^2)

## [1] 57955866

(sum(surv[,1]))^2

## [1] 1278205504

n = surv %>%   
 pull(ORGAN) %>%   
 length()  
  
k = surv %>%   
 distinct(ORGAN) %>%   
 pull(ORGAN) %>%   
 length()  
  
total\_ss = sum((surv[,1])^2)-(sum(surv[,1]))^2/n  
  
sum\_y2 = (sum(surv[,1]))^2/n  
  
between\_ss =   
surv %>%   
 group\_by(ORGAN) %>%   
 mutate(n = n(),  
 mean = mean(SURVIVAL)) %>%   
 distinct(mean,n) %>%   
 ungroup(ORGAN) %>%   
 mutate(mean^2\*n,  
 sum(mean^2\*n),  
 between\_ss = sum(mean^2\*n)-sum\_y2) %>%   
 select(between\_ss) %>%   
 distinct() %>%   
 pull(between\_ss)   
  
whithin\_ss = total\_ss - between\_ss  
  
MSB = between\_ss/(k-1)  
MSE = whithin\_ss/(n-k)  
  
F\_stat = MSB/MSE

## c)

pairwise.t.test(surv$SURVIVAL, surv$ORGAN, p.adj='bonferroni')

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: surv$SURVIVAL and surv$ORGAN   
##   
## Breast Bronchus Colon Ovary   
## Bronchus 0.00025 - - -   
## Colon 0.00608 1.00000 - -   
## Ovary 1.00000 0.38575 1.00000 -   
## Stomach 0.00153 1.00000 1.00000 0.75283  
##   
## P value adjustment method: bonferroni

# For Tukey, we need to use another function with an object created by aov()  
TukeyHSD(res1,conf.level = 0.99)

## Tukey multiple comparisons of means  
## 99% family-wise confidence level  
##   
## Fit: aov(formula = SURVIVAL ~ factor(ORGAN), data = surv)  
##   
## $`factor(ORGAN)`  
## diff lwr upr p adj  
## Bronchus-Breast -1184.32086 -2067.6073 -301.03446 0.0002385  
## Colon-Breast -938.49733 -1821.7837 -55.21093 0.0053072  
## Ovary-Breast -511.57576 -1670.0752 646.92367 0.5630900  
## Stomach-Breast -1109.90909 -2045.0583 -174.75983 0.0013962  
## Colon-Bronchus 245.82353 -537.1262 1028.77324 0.8208402  
## Ovary-Bronchus 672.74510 -411.1997 1756.68989 0.2271084  
## Stomach-Bronchus 74.41176 -766.6111 915.43467 0.9981461  
## Ovary-Colon 426.92157 -657.0232 1510.86636 0.6659115  
## Stomach-Colon -171.41176 -1012.4347 669.61114 0.9568289  
## Stomach-Ovary -598.33333 -1724.9413 528.27467 0.3772923

#install.packages("DescTools")  
library(DescTools)

##   
## Attaching package: 'DescTools'

## The following objects are masked from 'package:arsenal':  
##   
## %nin%, N

DunnettTest(surv$SURVIVAL, factor(surv$ORGAN))

##   
## Dunnett's test for comparing several treatments with a control :   
## 95% family-wise confidence level  
##   
## $Breast  
## diff lwr.ci upr.ci pval   
## Bronchus-Breast -1184.3209 -1832.455 -536.1867 9.1e-05 \*\*\*  
## Colon-Breast -938.4973 -1586.632 -290.3632 0.00221 \*\*   
## Ovary-Breast -511.5758 -1361.655 338.5031 0.36692   
## Stomach-Breast -1109.9091 -1796.099 -423.7192 0.00056 \*\*\*  
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

qt(1-0.001/2,59)

## [1] 3.46321

Stomach =   
surv %>%   
 filter(ORGAN=="Stomach")  
  
Colon =   
surv %>%   
 filter(ORGAN=="Colon")  
  
Ovary =   
surv %>%   
 filter(ORGAN=="Ovary")  
  
Bronchus =   
surv %>%   
 filter(ORGAN=="Bronchus")  
  
Breast =   
surv %>%   
 filter(ORGAN=="Breast")  
  
var.test(Stomach$SURVIVAL,Bronchus$SURVIVAL, alternative = "two.sided")

##   
## F test to compare two variances  
##   
## data: Stomach$SURVIVAL and Bronchus$SURVIVAL  
## F = 2.7232, num df = 12, denom df = 16, p-value = 0.06373  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.9425856 8.5821575  
## sample estimates:  
## ratio of variances   
## 2.723175

f\_s = var(Stomach$SURVIVAL)/var(Bronchus$SURVIVAL)  
  
mean\_sto = sum(Stomach %>% pull(SURVIVAL))/(Stomach %>% pull(SURVIVAL) %>% length())  
  
var\_sto =   
 sum((Stomach %>% pull(SURVIVAL)-mean\_sto)^2)/(Stomach %>% pull(SURVIVAL) %>% length()-1)  
  
mean\_bro = sum(Bronchus %>% pull(SURVIVAL))/(Bronchus %>% pull(SURVIVAL) %>% length())  
  
var\_bro =   
 sum((Bronchus %>% pull(SURVIVAL)-mean\_bro)^2)/(Bronchus %>% pull(SURVIVAL) %>% length()-1)  
  
n\_sto = (Stomach %>% pull(SURVIVAL) %>% length())  
  
n\_bro = (Bronchus %>% pull(SURVIVAL) %>% length())  
  
f\_a = qf(0.975,n\_sto-1,n\_bro-1)  
  
f\_s < f\_a

## [1] TRUE

s\_2 = ((n\_sto-1)\*var\_sto+(n\_bro-1)\*var\_bro)/(n\_sto+n\_bro-2)  
  
t\_s = (mean\_sto-mean\_bro)/sqrt(s\_2\*(1/n\_sto+1/n\_bro))  
  
  
p\_value = 2\*(1-(pt(t\_s,n-k)))  
p\_value

## [1] 0.4683403

t.test(Stomach$SURVIVAL,Bronchus$SURVIVAL,  
 paired=F,var.equal=T,  
 alternative = "two.sided", conf.level = 0.99)

##   
## Two Sample t-test  
##   
## data: Stomach$SURVIVAL and Bronchus$SURVIVAL  
## t = 0.7299, df = 28, p-value = 0.4715  
## alternative hypothesis: true difference in means is not equal to 0  
## 99 percent confidence interval:  
## -207.2975 356.1211  
## sample estimates:  
## mean of x mean of y   
## 286.0000 211.5882

t.test(Ovary$SURVIVAL,Bronchus$SURVIVAL,  
 paired=F,var.equal=T,  
 alternative = "two.sided", conf.level = 0.99)

##   
## Two Sample t-test  
##   
## data: Ovary$SURVIVAL and Bronchus$SURVIVAL  
## t = 2.5009, df = 21, p-value = 0.02074  
## alternative hypothesis: true difference in means is not equal to 0  
## 99 percent confidence interval:  
## -88.89153 1434.38173  
## sample estimates:  
## mean of x mean of y   
## 884.3333 211.5882

p\_value = 2\*(pt(2.5,21))  
p\_value

## [1] 1.979222

## d)

x <- c(2.9, 3.0, 2.5, 2.6, 3.2) # normal subjects  
y <- c(3.8, 2.7, 4.0, 2.4) # with obstructive airway disease  
z <- c(2.8, 3.4, 3.7, 2.2, 2.0) # with asbestosis  
  
#DunnettTest(list(x, y, z))  
  
## Equivalently,  
x <- c(x, y, z)  
g <- factor(rep(1:3, c(5, 4, 5)),  
 labels = c("Normal subjects",  
 "Subjects with obstructive airway disease",  
 "Subjects with asbestosis"))  
  
#DunnettTest(x, g)  
  
## Formula interface  
#boxplot(Ozone ~ Month, data = airquality)  
#DunnettTest(Ozone ~ Month, data = airquality)  
  
#DunnettTest(Ozone ~ Month, data = airquality, control="8", conf.level=0.9)

## d)

Stomach =  
 surv %>%   
 filter(ORGAN == "Stomach") %>%   
 pull(SURVIVAL)  
  
Bronchus =  
 surv %>%   
 filter(ORGAN == "Bronchus") %>%   
 pull(SURVIVAL)  
  
Colon =  
 surv %>%   
 filter(ORGAN == "Colon") %>%   
 pull(SURVIVAL)  
  
Ovary =  
 surv %>%   
 filter(ORGAN == "Ovary") %>%   
 pull(SURVIVAL)  
  
Breast =  
 surv %>%   
 filter(ORGAN == "Breast") %>%   
 pull(SURVIVAL)

#surv %>%   
# ggplot(aes(SURVIVAL,..density..,  
# fill = ORGAN))+  
# geom\_histogram()+  
# geom\_line(stat = 'density',size = 1)+  
# facet\_grid(ORGAN ~. )+  
# theme(legend.position = "bottom")+  
# labs(  
# title = "Density plot of organ group")

Stomach\_his =   
surv %>%   
filter(ORGAN=="Stomach") %>%   
 ggplot(aes(SURVIVAL,..density..))+  
 geom\_histogram(binwidth = 140)+  
 geom\_line(stat = 'density',size = 1)+  
 labs(x = "Stomach")  
  
Stomach\_qq =   
ggqqplot(Stomach,main="Q-Q plot for intervention group")

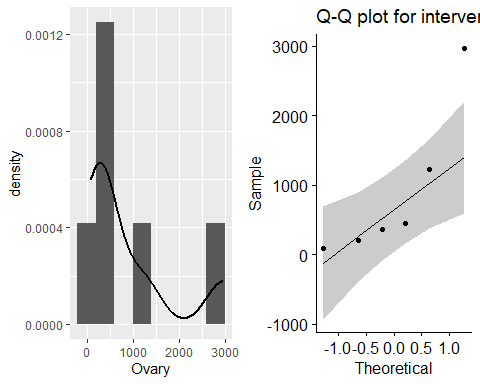
Bronchus\_his =   
surv %>%   
filter(ORGAN=="Bronchus") %>%   
 ggplot(aes(SURVIVAL,..density..))+  
 geom\_histogram(binwidth = 130)+  
 geom\_line(stat = 'density',size = 1)+  
 labs(x = "Bronchus")  
  
Bronchus\_qq =   
ggqqplot(Bronchus,main="Q-Q plot for intervention group")

Colon

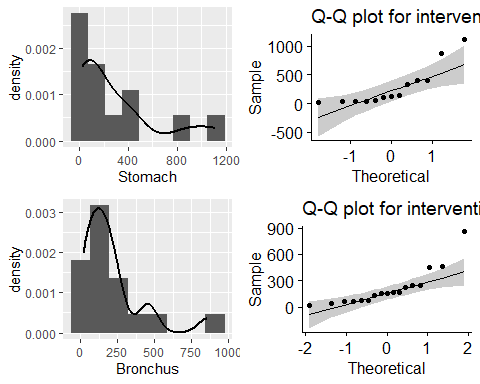
Colon\_his =   
surv %>%   
filter(ORGAN=="Colon") %>%   
 ggplot(aes(SURVIVAL,..density..))+  
 geom\_histogram(binwidth = 140)+  
 geom\_line(stat = 'density',size = 1)+  
 labs(x = "Colon")  
  
Colon\_qq =   
ggqqplot(Colon,main="Q-Q plot for intervention group")

Ovary

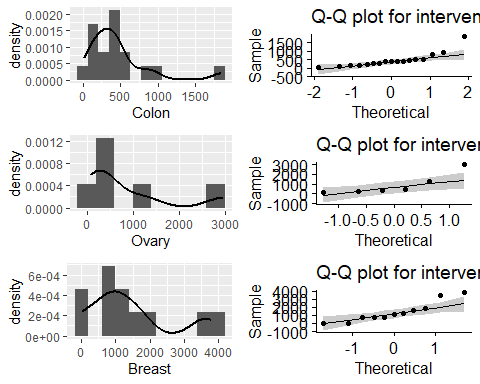
Ovary\_his =   
surv %>%   
filter(ORGAN=="Ovary") %>%   
 ggplot(aes(SURVIVAL,..density..))+  
 geom\_histogram(binwidth = 400)+  
 geom\_line(stat = 'density',size = 1)+  
 labs(x = "Ovary")  
  
Ovary\_qq =   
ggqqplot(Ovary,main="Q-Q plot for intervention group")  
  
multiplot(Ovary\_his,Ovary\_qq, cols=2)

 Breast

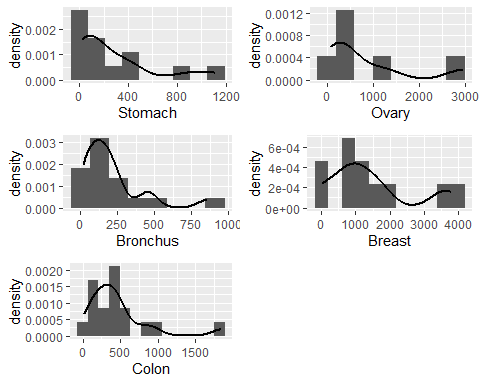
Breast\_his =   
surv %>%   
filter(ORGAN=="Breast") %>%   
 ggplot(aes(SURVIVAL,..density..))+  
 geom\_histogram(binwidth = 400)+  
 geom\_line(stat = 'density',size = 1)+  
 labs(x = "Breast")  
  
Breast\_qq =   
ggqqplot(Breast,main="Q-Q plot for intervention group")  
  
  
  
multiplot(Stomach\_his,Bronchus\_his,  
 Stomach\_qq,Bronchus\_qq,   
 cols=2)



multiplot(  
 Colon\_his,Ovary\_his,Breast\_his,  
 Colon\_qq,Ovary\_qq, Breast\_qq,  
 cols=2)



multiplot(Stomach\_his,Bronchus\_his,  
 Colon\_his,Ovary\_his,Breast\_his,  
 cols=2)



# Perform Shapiro-Wilk test  
shapiro.test(Stomach)

##   
## Shapiro-Wilk normality test  
##   
## data: Stomach  
## W = 0.75473, p-value = 0.002075

shapiro.test(Bronchus)

##   
## Shapiro-Wilk normality test  
##   
## data: Bronchus  
## W = 0.76596, p-value = 0.0007186

shapiro.test(Colon)

##   
## Shapiro-Wilk normality test  
##   
## data: Colon  
## W = 0.76056, p-value = 0.0006134

shapiro.test(Ovary)

##   
## Shapiro-Wilk normality test  
##   
## data: Ovary  
## W = 0.76688, p-value = 0.029

shapiro.test(Breast)

##   
## Shapiro-Wilk normality test  
##   
## data: Breast  
## W = 0.86857, p-value = 0.07431

### i)

Transformation can be done to tranaform the data into normality.

Krusak-Wallis test can be used when the normality is questionable.

### ii)

kruskal.test(surv$SURVIVAL~surv$ORGAN)

##   
## Kruskal-Wallis rank sum test  
##   
## data: surv$SURVIVAL by surv$ORGAN  
## Kruskal-Wallis chi-squared = 14.954, df = 4, p-value = 0.004798

res1<-aov(SURVIVAL~factor(ORGAN), data=surv)  
  
summary(res1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(ORGAN) 4 11535761 2883940 6.433 0.000229 \*\*\*  
## Residuals 59 26448144 448274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Problem 3

## a)

group 1: placebo and iron group 2: zinc and zinc+ iron

treatment = c("placebo","iron","zinc","zonc+iron")  
n = c(56,54,54,55)  
mean = c(1.1,1.4,0.7,0.8)  
se = c(0.2,0.2,0.1,0.1)  
  
data = tibble(treatment,n,mean,se)  
  
data\_nz =   
 data %>%   
 filter(treatment %in% c("placebo","iron"))  
  
data\_z =   
 data %>%   
 filter(treatment %in% c("zinc","zonc+iron"))

mean\_nz =  
 data\_nz %>%   
 mutate(new\_mean = sum(n \* mean)/sum(n)) %>%   
 select(new\_mean) %>%   
 distinct %>%   
 pull(new\_mean)  
  
mean\_z =  
 data\_z %>%   
 mutate(new\_mean = sum(n \* mean)/sum(n)) %>%   
 select(new\_mean) %>%   
 distinct %>%   
 pull(new\_mean)  
  
se\_nz =   
 data\_nz %>%   
 mutate(new\_se = sqrt(sum(se^2))) %>%   
 select(new\_se) %>%   
 distinct %>%   
 pull(new\_se)  
  
se\_z =   
 data\_z %>%   
 mutate(new\_se = sqrt(sum(se^2))) %>%   
 select(new\_se) %>%   
 distinct %>%   
 pull(new\_se)  
  
  
  
d = mean\_nz - mean\_z  
d\_se = sqrt(se\_nz^2+se\_z^2)

## b)

sd

data\_sd =   
data %>%   
 mutate(sd = se^2\*n) %>%   
 as.matrix()

sd\_pla = as.numeric(data\_sd[1,5])  
sd\_iro = as.numeric(data\_sd[2,5])  
  
n\_pla = as.numeric(data\_sd[1,2])  
n\_iro = as.numeric(data\_sd[2,2])  
  
   
# calculate the test-statistics  
f\_s = sd\_pla^2/sd\_iro^2  
  
# calculate F-value when type 1 error = .05  
f\_a = qf(0.975,n\_pla-1,n\_iro-1)  
  
f\_s < f\_a

## [1] TRUE

Sd is equal between placebo and iron.

sd\_z = as.numeric(data\_sd[3,5])  
sd\_zi = as.numeric(data\_sd[4,5])  
  
n\_z = as.numeric(data\_sd[3,2])  
n\_zi = as.numeric(data\_sd[4,2])  
  
   
# calculate the test-statistics  
f\_s = sd\_z^2/sd\_zi^2  
  
# calculate F-value when type 1 error = .05  
f\_a = qf(0.975,n\_z-1,n\_zi-1)  
  
f\_s < f\_a

## [1] TRUE

Sd is equal between zinc and zinc+iron.