Exercise 1:

~~~
---
title: "HW2 Exercise 1"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
date: "3/3/2020"
output: pdf_document
---

```{r, echo=FALSE}
# knitr::opts_chunk$set(fig.width=6, fig.height=4)
```



```{r, warning=FALSE}
library(readr)
library(ggfortify)
library(survival)
library(ggplot2)
```

```{r}
transplant <- read_table2("~/Documents/Columbia/STAT W5703/HW/HW2/transplant.txt",
   col_names = c("time", "type", "Indicator"), skip = 8)
```
### Problem 1
```{r}
# make the variable "type"" a factor
transplant$type <- factor(transplant$type)
censored_index <- which(transplant$Indicator==0)
plot(density(transplant[censored_index,]$time), main = "relapse time density plot for right-censored patients")
```

As we can see from the density plot, the right-censored patients, which are alive patients, do not have a fixed censoring time; hence, it is reasonable to assume that the right censoring in this dataset is random.

### Problem 2
```{r}
autoplot(survfit(Surv(time, Indicator)~type, data=transplant), conf.int=TRUE, col=c(1,3), main="Kaplan-Meier estimates for transplant data")
```
~~~

We can see from the above plot that there is a minor difference in the transplant survival rate between the two types. In general, Type 1 (allogeneic) transplant seems to be slightly more efficient since Type 2 (autologous) transplant has a large drop in survival after 55 months.

### Problem 3

```{r}
model_exp <- survreg(Surv(time, Indicator)~type, data=transplant, dist = "exponential")
summary(model_exp)
```

We can see from the output that type 2 (autologous), on average, has 0.325 lower survival than type 1 (allogeneic) transplant, and this difference is not significant as the associatd p-value is 0.25. This result agrees with my intuition from last point since the previous plot only showed type 2 (autologous) transplant has a slightly lower survival than type 1.

### Problem 4

From the summary table, we see that the likelihood ratio statistics gives a p-value of 0.25, indicating that the difference between type 1 and type 2 treatments is not significant. It also agrees with the insignificant coeeficient for type2. This conclusion depends on the exponential model assumption.

### Problem 5

```{r}
plot(survfit(Surv(time, Indicator)~type, data=transplant), conf.int=TRUE, col=c(1,3),
main="Exponential v.s. K-M fits")
x <- seq(0, 70, 1)
lines(x, 1-pexp(x, exp(-model_exp$coefficients[1])), col="darkred", lwd=2)
lines(x, 1-pexp(x, exp(-sum(model_exp$coefficients))), col="darkgreen", lwd=2)
legend("topright", legend = c("type 1", "type 2"),
      col=c("darkred", "darkgreen"), lwd = 2)
```

We observe that the exponential model does not fit very well. For both transplant type, we see some departures from the pointwise confidence intervals of the K-M estimates. Although the expoential model fits type 2 better, the fitted line of type 2 deviates from the confidence intervals at around time = 28.

### Problem 6

```{r}
model_wei <- survreg(Surv(time, Indicator)~type,data=transplant)
summary(model_wei)
```

~~~
```{r}
type1_index <- which(transplant$type==1)
fit.wei1<-survreg(Surv(time, Indicator)~type, data=transplant[type1_index,])
fit.wei2<-survreg(Surv(time, Indicator)~type, data=transplant[-type1_index,])
gamma1=1/exp(fit.wei1$scale)
gamma2=1/exp(fit.wei2$scale)

plot(survfit(Surv(time, Indicator)~type, data=transplant), conf.int=TRUE, col=c(1,3),
main="Weibull (Split data) v.s. K-M fits")
x <- seq(0, 70, 1)
lines(x, 1-pweibull(x, gamma1, exp(coef(fit.wei1)[1])), col="darkred",lwd=2)
lines(x, 1-pweibull(x, gamma2, exp(coef(fit.wei2)[1])), col="darkgreen",lwd=2)
```
~~~

When we split the data according to the transplant types as what we did in the lecture notes, we can see the model fits even worse; hence, we may want to use the whole dataset.

~~~
```{r}
plot(survfit(Surv(time, Indicator)~type, data=transplant), conf.int=TRUE, col=c(1,3),
main="Weibull (whole data) v.s. K-M fits")
x <- seq(from=0,to=70,by=1)
lines(x, 1-pexp(x, exp(-model_exp$coefficients[1])), col="darkred", lwd=2)
lines(x, 1-pexp(x, exp(-sum(model_exp$coefficients))), col="darkgreen", lwd=2)
lines(x, 1-pweibull(x, model_wei$scale, exp(coef(model_wei)[1])), col="purple")
lines(x, 1-pweibull(x, model_wei$scale, exp(sum(coef(model_wei)))), col="orange")
```
~~~

Still, using the whole dataset does not give a better fit. This result agrees with what we have from the weibull fit model as we can see the p-value of the model is 0.37, indicating that the weibull does not fit the model well and there is no improvement from the exponential model.

Exercise 2:

~~~
---
title: "STAT5703 HW2 Exercise 2"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(fig.width=12, fig.height=8, fig.path='Figs/', warning=FALSE,
message=FALSE)
~~~

~~~~
```

## Exercise 2.
#### Question 1.
```{r}
data <- read.table('scores.txt', header = TRUE)
```

```{r}
# Complete case analysis.
c1 <- cov(data,use="complete")
c1
```

```{r}
# Available case analysis.
c2 <- cov(data,use="pairwise")
c2
```

```{r}
# Mean imputation
data_mean=data
for(i in 1:ncol(data_mean)) {
  data_mean[ , i][is.na(data_mean[ , i])] <- mean(data_mean[ , i], na.rm = TRUE)
}
c3 <- cov(data_mean)
c3
```

```{r}
# Mean imputation with bootstrap
cov<-matrix(rep(0,25),ncol=5)
for(i in 1:400){
  sam<-sample(nrow(data),22,replace=TRUE)
  temp <- sapply(data[sam,], function(x) ifelse(is.na(x), mean(x, na.rm = TRUE), x))
  cov <- cov + cov(temp)
}
c4 <- cov/400
c4
```

```{r,warning=FALSE}
# The EM-algorithm
~~~~

```
library(Amelia)
Completed_data <- amelia(data,m=1,p2s=0)
c5 <- cov(Completed_data$imputations$imp1)
c5
```

Mean imputation and Mean imputation with the bootstrap have smaller covariance than others. Only EM-algorithm has a negative covariance of x1 and x2.

#### Question 2.

By delta method, we can get $\sqrt{n}(\hat \lambda_1-\lambda_1)\to N(0,2\lambda_1^2)$, therefore asymptotic normality of $\hat\lambda_1$ is :$\hat\lambda_1\to N(\lambda_1,\frac{2\lambda_1^2}{n})$, the confidence interval of $\lambda_1$ is:

$$[\frac{\hat\lambda_1}{1+ z_{1-\alpha/2}\sqrt\frac{2}{n}},\frac{\hat\lambda_1}{1- z_{1-\alpha/2}\sqrt\frac{2}{n}}]$$

Because $\lambda_1$ is the largest eigenvalue of the population covariance matrix, we can get $\hat\lambda_1$ from each method and the intervals of $\lambda_1$:

```{r}
get_interval <- function(lambda) {
  n=nrow(data)
  print(paste0('[',lambda/(1+sqrt(2/n)*qnorm(0.975)),', ',lambda/(1-sqrt(2/n)*qnorm(0.975)),']'))}
get_interval(max(eigen(c1)$value))
get_interval(max(eigen(c2)$value))
get_interval(max(eigen(c3)$value))
get_interval(max(eigen(c4)$value))
get_interval(max(eigen(c5)$value))
```

Complete case analysis and available case analysis give us a higher covariance than Mean imputation but also have larger confidence intervals because our data only has few complete records.

The EM-algorithm generates a smaller range of confidence interval than Complete case and available case but larger than mean imputation(with bootstrap or not)

Therefore Mean imputation with the bootstrap might be a good method to handle missing data for this particular scores data.

#### Question 3.

```{r}
library(SMPracticals)
cov(mathmarks)
get_interval(max(eigen(cov(mathmarks))$value))
```

Using EM-algorithm generates a cloest confidence interval of $\lambda_1$ from the full data.

Therefore the EM-algorithm might be the best method to fill in the missing data in this case which is not consistent with the result we thought at question2, because the data size in questions before is really small.

#### Question 4.

partially observed vectors:
$$X_i=\begin{bmatrix}
X_{io} \\
X_{im} \\
\end{bmatrix}$$
we have that,
$$\mu^{(k)}=\begin{bmatrix}
\mu_{io}^{(k)} \\
\mu_{im}^{(k)} \\
\end{bmatrix}
,
\Sigma^{(k)}=\begin{bmatrix}
\Sigma_{ioo}^{(k)}&\Sigma_{iom}^{(k)} \\
\Sigma_{imo}^{(k)}&\Sigma_{imm}^{(k)} \\
\end{bmatrix}$$
Then, for E-step:
Because of
$$E(X_i|X_io)=\begin{bmatrix}
X_{io} \\
E(X_{im}|X_{io}) \\
\end{bmatrix}
$$
$$
E(X_iX_i^T|X_{io})=\begin{bmatrix}
X_{io}X_{io}^T&X_{io}E(X_{im}^T|X_{io}) \\
E(X_{im}|X_{io})X_{io}^T&E(X_{im}X_{im}^T|X_{io}) \\
\end{bmatrix}\\$$
where from the propertites of multivariate normal distribution,
$$
E(X_{im}|X_{io})=\mu_{im}^{(k)}+\Sigma_{imo}^{(k)}(\Sigma_{ioo}^{(k)})^{-1}(X_{io}-\mu_{io}^{(k)})\\$$
$$E(X_{im}X_{im}^T|X_{io})=Cov(X_{im}|X_{io})+E(X_{im}|X_{io})E(X_{im}|X_{io})^T\\=(\Sigma_{imm}^{(k)}-\Sigma_{imo}^{(k)}(\Sigma_{ioo}^{(k)})^{-1}\Sigma_{iom}^{(k)})+E(X_{im}|X_{io})E(X_{im}|X_{io})^T$$
Then, for M-step:
$$\mu^{(k+1)}:\frac{1}{n}\sum_{i=1}^nE(X_i|X_{io})=0,\
\Sigma^{(k+1)}:\frac{1}{n}\sum_{i=1}^nE(X_iX_i^T|X_{io})-\mu^{(k+1)}\mu^{(k+1)^T}=0$$
To simplify using the information above, we can get:
$$\mu^{(k+1)}:\sum_{i=1}^n(\hat{X_i}-\mu)=0,\

\Sigma^{(k+1)}:\sum_{i=1}^n(\Sigma-(\hat{X_i}-\mu)(\hat{X_i}-\mu)^T-C_i^{(k)})=0$$

Exercise 3:

~~~
---
title: "Ex3_bx2168_hl3339_wf2255"
author: "Banruo Xie;Wen Fan;Hanjun Li"
date: "3/7/2020"
output:
  pdf_document: default
  html_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r message = F}
library(dplyr)
library(lubridate)
```
~~~

(1)

$a_1$ means the probability of rainy day given the previous day is rainy day

$a_2$ means the probability of no rain day given the previous day is rainy day

$a_3$ means the probability of rainy day given the previous day is no rain day

$a_4$ means the probability of no rain day given the previous day is no rain day

(2)

Let $X_i$ represent whether the ith day is rainy.

By Bayesian formula:

$P(X_n=0)=$
$P(X_n=0|X_{n-1}=0)P(X_{n-1}=0)+P(X_n=0|X_{n-1}=1)P(X_{n-1}=1)$
$= a_1P(X_n=0)+a_3(1-P(X_n=0))$

Therefore, $P(X_n=0)=\frac{a_3}{1-a_1+a_3}$

```{r}
data <- read.csv('./CentralPark.csv', header = T)
data$DATE <- as.POSIXct(strptime(as.character(data$DATE), "%m/%d/%y"))
data <- data %>% mutate(is_rain = if_else(PRCP>=1.5,T,F))
data$month <- month(data$DATE)
data$will_rain <- append(data$is_rain,c(NA))[2:(length(data$is_rain)+1)]
print(c((nrow(data %>% filter(month == 7, is_rain, will_rain)))/
         nrow(data %>% filter(month == 7, is_rain)),
       (nrow(data %>% filter(month == 7, is_rain, !will_rain)))/
         nrow(data %>% filter(month == 7, is_rain)),
       (nrow(data %>% filter(month == 7, !is_rain, will_rain)))/
         nrow(data %>% filter(month == 7, !is_rain)),
       (nrow(data %>% filter(month == 7, !is_rain, !will_rain)))/
         nrow(data %>% filter(month == 7, !is_rain))))
```

(4)

Hypothesis test: $H_0: p_{00} = p_{11}$, $H_1:p_{00} \neq p_{11}$

$p_{00}$ is the probability of rainy day given the previous day is rainy day

$p_{11}$ is the probability of no rain day given the previous day is no rain day

Since $p_{00}$ and $p_{11}$ are independent, therefore,
$\hat p_{00}\xrightarrow[\infty]{D}N(\hat p_{00},\frac{\hat p_{00}\left(1-\hat p_{00}\right)}{n_0})$

$\hat p_{11}\xrightarrow[\infty]{D}N(\hat p_{11},\frac{\hat p_{11}\left(1-\hat p_{11}\right)}{n_1})$

$\hat p_{00}-\hat p_{11}\xrightarrow[\infty]{D}N(0,\frac{\hat p_{00}\left(1-\hat p_{00}\right)}{n_0}-\frac{\hat p_{11}\left(1-\hat p_{11}\right)}{n_1})$

```{r}
a1 = (nrow(data %>% filter(month == 7, is_rain, will_rain)))/
  nrow(data %>% filter(month == 7, is_rain))
a4 = (nrow(data %>% filter(month == 7, !is_rain, !will_rain)))/
  nrow(data %>% filter(month == 7, !is_rain))
print(pnorm((a1-a4)/sqrt(a1*(1-a1)/nrow(data %>% filter(month == 7, is_rain))
                    +a4*(1-a4)/nrow(data %>% filter(month == 7, !is_rain)))))
```

Therefore, we reject $H_0$

(5)

```{r}
data$will_rain2 <- append(data$will_rain,c(NA))[2:(length(data$will_rain)+1)]
```

$H_0$: Higher model chain can not improve.
$H_1$: Higher model chain does improve.

Using likelihood ratio test:

$\begin{aligned}\Lambda_{n} &=2\left\{\ell(\hat{\mathbf{P}})_{\text {second order}}-\ell(\hat{\mathbf{P}})_{\text {first order}}\right\}=2\left\{\sum_{r=1}^{S} \sum_{s=1}^{S} \sum_{t=1}^{S} n_{r s t} \log \hat{p}_{r s t}-\sum_{s=1}^{S} \sum_{t=1}^{S}n_{. st} \log \hat{p}_{st}\right\} \\&=2\left\{\sum_{r=1}^{S} \sum_{s=1}^{S} \sum_{t=1}^{S} n_{r s t} \log \hat{p}_{r s t}-\sum_{r=1}^{S} \sum_{s=1}^{S} \sum_{t=1}^{S} n_{r s t} \log \hat{p}_{st}\right\}=2 \sum_{r=1}^{S} \sum_{s=1}^{S} \sum_{t=1}^{S} n_{r s t} \log \left(\frac{\hat{p}_{r s t}}{\hat{p}_{s t}}\right)\end{aligned}$

By asymptotic theory,
$\Lambda_{n} \frac{\mathcal{D}}{n \rightarrow \infty} \chi_{(S-1)^{2}}^{2}$
```{r}
p00 <- (nrow(data %>% filter(month == 7, is_rain, will_rain)))/
  nrow(data %>% filter(month == 7, is_rain))
p01 <- (nrow(data %>% filter(month == 7, is_rain, !will_rain)))/
  nrow(data %>% filter(month == 7, is_rain))
p10 <- (nrow(data %>% filter(month == 7, !is_rain, will_rain)))/
  nrow(data %>% filter(month == 7, !is_rain))
p11 <- (nrow(data %>% filter(month == 7, !is_rain, !will_rain)))/
  nrow(data %>% filter(month == 7, !is_rain))

r000 <- nrow(data %>% filter(month == 7, is_rain, will_rain, will_rain2))
r001 <- nrow(data %>% filter(month == 7, is_rain, will_rain, !will_rain2))
r010 <- nrow(data %>% filter(month == 7, is_rain, !will_rain, will_rain2))
r011 <- nrow(data %>% filter(month == 7, is_rain, !will_rain, !will_rain2))
r100 <- nrow(data %>% filter(month == 7, !is_rain, will_rain, will_rain2))
r101 <- nrow(data %>% filter(month == 7, !is_rain, will_rain, !will_rain2))
r110 <- nrow(data %>% filter(month == 7, !is_rain, !will_rain, will_rain2))
r111 <- nrow(data %>% filter(month == 7, !is_rain, !will_rain, !will_rain2))

p000 <- r000/(r000 + r001)
p001 <- r001/(r000 + r001)
p010 <- r010/(r010 + r011)
p011 <- r011/(r010 + r011)
```

```
p100 <- r100/(r100 + r101)
p101 <- r101/(r100 + r101)
p110 <- r110/(r110 + r111)
p111 <- r111/(r110 + r111)

result <- 2* (r000*log(p000/p00) + r001*log(p001/p01) + r010*log(p010/p10)
           + r011*log(p011/p11) + r100*log(p100/p00)
           + r101*log(p101/p01) + r110*log(p110/p10)
           + r111*log(p111/p11))


pchisq(result,2)
```

Therefore, we fail to reject $H_0$, higher model chain does not improve fit of the data.

Exercise 4:

---
title: "STAT5703 HW2 Exercise 4"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
output: pdf_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(fig.width=12, fig.height=8, fig.path='Figs/', warning=FALSE,
message=FALSE)
```

## Exercise 4.

#### Question 1.

The independent random variables N is in multinomial distribution, the joint distribution is

$$
P_{\theta}(N_{A}, N_{C}, N_{G}, N_{T})=\frac{n!}{N_{A} ! N_{C} ! N_{G} ! N_{T} !} p_{A}^{N_{A}} \cdot p_{C}^{N_{C}} \cdot p_{G}^{N_{G}} \cdot P_{T}^{N_T}
$$

#### Question 2.

log likelihood:

$$
L_{\theta}=\log P_{\theta} = \log(n!) - \sum_{x \in\{A, C, G, T\}}{N_x!}+\sum_{x \in\{A, C, G, T\}} N_{x} \log p_{x}
$$

$$

\begin{aligned}
\frac{d L_{\theta}}{d \theta} &=\sum_{x} N_{x} \frac{d \log(p_{x})}{d \theta} \\
&=N_{A} \frac{-1}{1-\theta} + N_{C} \frac{1-2 \theta}{\theta-\theta^{2}}+ N_{G} \frac{2 \theta-3 \theta^{2}}{\theta^{2}-\theta^{3}} + N_{T} \frac{ 3 \theta^{2}}{\theta^{3}} =0
\end{aligned}
$$

$$
-N_{A}+N_{C} \frac{1-2 \theta}{\theta}+ N_{G} \frac{2 -3 \theta}{\theta} + N_{T} \frac{ 3 (1-\theta)}{\theta} =0\\$$
$$
-N_{A} \theta +N_{c}(1-2\theta) +N_G(2-3 \theta)+3 N_{T}(1-\theta)=0\\$$
$$
\theta\left(N_{A}+2 N_{C} + 3N_G+ 3 N_{T}\right)=N_{C}+2 N_{G}+3 N_{T}\\
$$
$$
\hat\theta = \frac{N_{C}+2 N_{G}+3 N_{T}}{N_{A}+2 N_{C} + N_G+ 3 N_{T}}\\
$$

#### Question 3.

In this case we have$\hat\theta \to N(\theta, \frac{1}{nI(\theta)})$, where $I(\theta)$ is the Fisher Information.

$$
\begin{aligned}
I(\theta) &= -E[\frac{d^{2} L_{\theta}}{d \theta^{2}}]\\
 &= -E[\frac{2\theta a-a-b\theta}{\theta^2(1-\theta)^2}]\\
&=n \cdot \frac{1+\theta + \theta^2}{\theta (1-\theta)}
\end{aligned}
$$
where $a = N_{C}+2 N_{G}+3 N_{T}$, $b=N_{A}+2 N_{C} + N_G+ 3 N_{T}$
$E[a]=n(\theta+\theta^2+\theta^3)$, $E[b]=n(1+\theta+\theta^2)$
Therefore, the asymptotic distribution is $N(\theta, \frac{\theta (1-\theta)}{n(1+\theta + \theta^2)})$

#### Question 4.
We want $E[T]=\theta = n(a_A(1-\theta)+a_C(\theta-\theta^2)+a_G(\theta^2-\theta^3)+a_T(\theta^3))$
Therefore$a_A = 0$, $a_C = 1/n$, $a_G = 1/n$, $a_T = 1/n$

#### Question 5.

$$
Var[T] = Var[\frac{N_C + N_T + N_G}{n}] = Var[1 - \frac{N_A}{n}]=\frac{Var[N_A]}{n^2}=\frac{\theta(1-\theta)}{n}
$$
$$
efficiency(T, \hat{\theta})=\frac{Var[T]}{Var[\hat\theta]}=1+\theta+\theta^2
$$

#### Question 6.

log-likelihood if $p_i$ doesn't depend on $\theta$, and using Lagrange:

$$
Lagrange_{p_x;\lambda}=\sum_{x \in\{A, C, G, T\}} N_{x} \log p_{x}-\lambda(\sum_{x \in\{A, C, G, T\}}p_x-1) \\
$$

$$
\frac{Lagrange_{p_x;\lambda}}{\partial p_{x}}=\frac{N_{x}}{p_{x}}-\lambda=0
$$
By solving it, we get,

$$
p_{x}=N_x/\lambda\\
\sum_{x \in\{A, C, G, T\}}p_{x}=\sum_{x \in\{A, C, G, T\}}\frac{N_x}{\lambda}=1\\
$$
Therefore $$\lambda=n\\$$
$$
\hat p_x=\frac{N_x}{n}, \forall x \in \{A,C,G, T\}
$$
Compare with $p_x$ depends on $\theta$ :
$$\hat p_A = 1-\hat \theta, \hat p_C = \hat\theta-\hat \theta^2, \hat p_G = \hat\theta^2-\hat \theta^3, \hat p_T = \hat \theta^3$$
Both are unbiased estimator, but $p_A$ depends on $\theta$ needs observed occurrences for all bases, $p_A$ not depends on $\theta$ noly need one of them but has 2 more free parameters.

#### Question 7.

The likelihood ratio test for testing the hypothesis:$P=P(\theta)$,

$$
\Lambda = 2 \sum_{x \in\{A, C, G, T\}} N_x\log \frac{\hat p_x}{\hat p_x(\theta)}=2 \sum_{x \in\{A, C, G, T\}} N_x\log \frac{N_x}{n \hat p_x(\theta)} \sim \chi_{2}

$$

Exercise 5:

~~~
---
title: "HW2 Exercise 5"
author: "Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)"
date: "3/7/2020"
output:
  pdf_document: default
  html_document:
    df_print: paged
---

```{r, warning=FALSE}
library(readr)
library(dplyr)
transplant <- read_table2("~/Documents/Columbia/STAT W5703/HW/HW2/transplant.txt",
    col_names = c("time", "type", "Indicator"), skip = 8)

# preprocessing
transplant$type <- factor(transplant$type)
type1_index <- which(transplant$type==1)
type2_index <- which(transplant$type==2)
death_index <- which(transplant$Indicator==1)
```


### Problem 1
Suppose the relapse rates for both treatment group are the same. i.e. $P_A = P_B = p$. We also
see that the total number of death $n_d^{(k)}$ is independent of the number of at risk groups
$n_A^{(k)}$ and $n_B^{(k)}$, so
\begin{align*}
P(y^{(k)}=m|n_A^{(k)}=m_A,n_B^{(k)}=m_B,n_d^{(k)}=m_d) &= \frac{P(y^{(k)}=m,
n_A^{(k)}=m_A,n_B^{(k)}=m_B,n_d^{(k)}=m_d)}{P(n_A^{(k)}=m_A,n_B^{(k)}=m_B,n_d^{(k)}=m_
d)}\\
&= \frac{P(y^{(k)}=m,n_d^{(k)}=m_d|
n_A^{(k)}=m_A,n_B^{(k)}=m_B)P(n_A^{(k)}=m_A,n_B^{(k)}=m_B)}{P(n_d^{(k)}=m_d|
P(n_A^{(k)}=m_A,n_B^{(k)}=m_B)}\\
&= \frac{\binom{m_A}{m} p^m(1-p)^{m_A-m}\binom{m_B}{m_d-m} p^{m_d-m}(1-p)^{m_B-
m_d+m}}{\binom{m_A+m_B}{m_d}p^{m_d}(1-p)^{m_A+m_B-m_d}}\\
&= \frac{\binom{m_A}{m}\binom{m_B}{m_d-m}}{\binom{m_A+m_B}{m_d}}
\end{align*}
Hence, it is a $HyperGeometric(n_{A}^{(k)}+n_{B}^{(k)}, n_{A}^{(k)}, n_{d}^{(k)})$
~~~

### Problem 2
From previous problem, we have the conditonal probability $$P(y^{(k)}=m|n_A^{(k)}=m_A,n_B^{(k)}=m_B,n_d^{(k)}=m_d)=\frac{\binom{n_A^{(k)}}{y^{(k)}} \binom{n_B^{(k)}}{n_d^{(k)}-y^{(k)}}}{\binom{n^{(k)}}{n_d^{(k)}}}$$
Using the formulas for the expected value and variance of a hypergeometric distribution: given $h(x; N,n,k)$, $$E[X]=\frac{n*k}{N}$$ $$Var(X)=\frac{n*k*(N-k)*(N-n)}{N^2*(N-1)}$$

so $$E^{(k)}=\frac{n_A^{(k)}n_d^{(k)}}{n^{(k)}}$$ $$V^{(k)}=\frac{n_A^{(k)}n_d^{(k)}(n^{(k)}-n_A^{(k)})(n^{(k)}-n_d^{(k)})}{(n^{(k)})^2(n^{(k)}-1)}=\frac{n_A^{(k)}n_d^{(k)}n_B^{(k)}n_s^{(k)}}{(n^{(k)})^2(n^{(k)}-1)}$$
Hence, shown.

### Problem 3
We have
\begin{align*}
var(y^{(k)}-E^{(k)}) &= var(E[y^{(k)}-E^{(k)}| n_{A}^{(k)}, n_{B}^{(k)}, n_{d}^{(k)}])+E[var(y^{(k)}-E^{(k)}|n_{A}^{(k)}, n_{B}^{(k)}, n_{d}^{(k)})]\\
&= var(E[y^{(k)}|n_{A}^{(k)}, n_{B}^{(k)}, n_{d}^{(k)}]-E^{(k)})+E[V^{(k)}]\\
&= 0 + E[V^{(k)}]\\
&= E[V^{(k)}]
\end{align*}

### Problem 4
We have $$var[\sum_{k=1}^K(y^{k}-E^{(k)})]=\sum_{k=1}^{K} var[y^{(k)}-E^{(k)}]+2\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}cov[y^{(i)}-E^{(i)},y^{(j)}-E^{(j)}]$$
Let the condition $(n_{A}^{(i)}, n_{B}^{(i)}, n_{d}^{(i)},n_{A}^{(j)}, n_{B}^{(j)}, n_{d}^{(j)})$ be $C$.
By the Law of Total Variance,
\begin{align*}
cov[y^{(i)}-E^{(i)},y^{(j)}-E^{(j)}] &= cov[E[y^{(i)}-E^{(i)}| C],E[y^{(j)}-E^{(j)}|C]]+E[cov[y^{(i)}-E^{(i)},y^{(j)}-E^{(j)}|C]]\\
\end{align*}
We have $E[y^{(i)}-E^{(i)}|C]=E[y^{(j)}-E^{(j)}| C]=0$ and $cov[y^{(i)}-E^{(i)},y^{(j)}-E^{(j)}| C]=cov[y^{(i)},y^{(j)}| C]$ from problem 3, and $y^{(i)}$, $y^{(j)}$ are two independent hypergeometric variables; hence, $$cov[y^{(i)}-E^{(i)},y^{(j)}-E^{(j)}| C]=cov[y^{(i)},y^{(j)}| C]=0$$. As a result,
\begin{align*}
cov[y^{(i)}-E^{(i)},y^{(j)}-E^{(j)}] &= cov[E[y^{(i)}-E^{(i)}| C],E[y^{(j)}-E^{(j)}|C]]+E[cov[y^{(i)}-E^{(i)},y^{(j)}-E^{(j)}|C]]\\
&= 0+0 = 0
\end{align*}
and $$var[\sum_{k=1}^K(y^{k}-E^{(k)})]=\sum_{k=1}^{K} var[y^{(k)}-E^{(k)}]$$

### Problem 5

In our dataset, for each k, we have 1 death if the data is not censored, so $n_d^{k}=1$, $\forall k$

```{r}
n <- nrow(transplant)
k <- length(death_index)
na <- length(type1_index)
nb <- length(type2_index)
nd <- 1
df <- transplant[order(transplant$time),] %>% filter(Indicator==1)
y = numeric(k)
E = numeric(k)
Var = numeric(k)
for (i in seq_len(k)){
  E[i] <- na*nd/(na+nb)
  Var[i] <- na*nb*nd*(na+nb-nd)/((na+nb)^2*(na+nb-1))
  if(df[i,]$type==1){
    y[i] <- 1
    na <- na - 1
  } else {
    y[i] <- 0
    nb <- nb - 1
  }
}

Z <- sum(y-E)/sqrt(sum(Var))
pnorm(Z)
```

Using the logrank_test from "coin" library, we got the same conclusion:

```{r, message=FALSE}
library(coin)
logrank_test(Surv(time, Indicator)~type, data=transplant, distribution = "asymptotic")
```

Since the resulting p-value is insignificant, we do not have enough evidence to reject H_0.