Reliability Analysis-HW2

ID : 111024517 Name : 鄭家豪

due on 03/22

Problem 1

In the test, 25 controls were put on test and run until failure or until n=30 thousand cycles had been accumulated. Failures occurred at t=5, 21, and 28 thousand cycles. The other 22 controls did not fail by the end of the test.

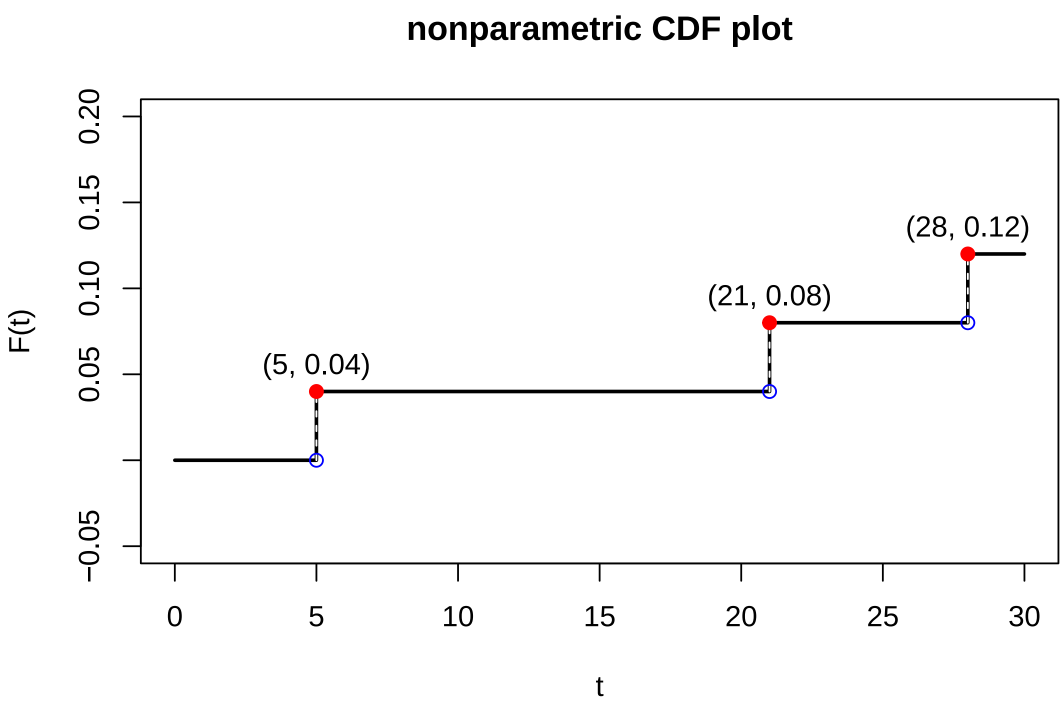
(a)

The empirical F(t) is

$$\hat{F}(5) = 1/25; \hat{F}(21) = 2/25; \hat{F}(28) = 3/25$$

And the plot:

```
x <- c(0,5,21,28,30)
y <- c(0,1/25,2/25,3/25,3/25)
plot(x, y, type = "s", xlab = "t", ylab = "F(t)",lwd=2,
     xlim = c(0,30),ylim=c(-0.05,0.2),
     main="nonparametric CDF plot")
segments(x0 = x[-c(1,5)],y0 = y[-c(4,5)],
         x1 = x[-c(1,5)],y1 = y[-c(1,5)],lty = 2,col = "white")
points(x[-c(1,5)], y[-c(1,5)], pch = 19, col = "red")
text(x=x[-c(1,5)], y=y[-c(1,5)],
     labels = paste("(", x[-c(1,5)], ", ", y[-c(1,5)], ")", sep = ""),
     pos = 3)
points(x[-c(1,5)], y[-c(4,5)], pch = 1, col = "blue")
```



(b)

For a fixed t,$X = n\hat{F}(t) \sim \text{Bin}(n, F(t))$. The conservative confidence interval I choose is Clopper-Pearson interval:$[\text{b}_{\alpha/2,x,n-x+1}, \text{b}_{1-\alpha/2,x+1,n-x}]$, where b is the quantile of Beta distribution. The CI for desired probability is

```r
#wrong
CI_CP <- function(alpha,sample){
  lower <- qbeta(p = alpha/2,shape1 = 30*sample,shape2 = 30-30*sample+1)
  upper <- qbeta(p = 1-alpha/2,shape1 = 30*sample+1,shape2 = 30-30*sample)
  result <- cbind(lower,upper)
  colnames(result) <- c("lower","upper")
  rownames(result) <- c("t=5","t=21","t=28")
  return(result)
}
knitr::kable(CI_CP(0.05,y[-c(1,5)]),escape = FALSE)
```

```r
# correct
x = 3 ## total fail number = 3
CI_CP <- function(alpha,sample){
  n = 25
  lower <- qbeta(p = alpha/2,shape1 = sample,shape2 = n-sample+1)
  upper <- qbeta(p = 1-alpha/2,shape1 = sample+1,shape2 = n-sample)
```

```r
  result <- cbind(lower,upper)
  return(result)
}
CI_CP(0.05,3)
```

```
         lower     upper
[1,] 0.0254654 0.3121903
```

(c)

The 100(1-α)% CI using the Jeffrey method is $[\mathrm{b}_{\alpha/2,x+1/2,n-x+1/2}, \mathrm{b}_{1-\alpha/2,x+1/2,n-x+1/2}]$, where b is the quantile of Beta distribution.

```r
## wrong
CI_Jef <- function(alpha,sample){
  lower <- qbeta(p = alpha/2,shape1 = 30*sample+1/2,shape2 = 30-30*sample+1/2)
  upper <- qbeta(p = 1-alpha/2,shape1 = 30*sample+1/2,shape2 = 30-30*sample+1/2)
  result <- cbind(lower,upper)
  colnames(result) <- c("lower","upper")
  rownames(result) <- c("t=5","t=21","t=28")
  return(result)
}
knitr::kable(CI_Jef(0.05,y[-c(1,5)]),escape = FALSE)
```

```r
# correct
x = 3 ## number of failure = 0.12*25
CI_Jef <- function(alpha,sample){
  n=25
  lower <- qbeta(p = alpha/2,shape1 = sample+1/2,shape2 = n-sample+1/2)
  upper <- qbeta(p = 1-alpha/2,shape1 = sample+1/2,shape2 = n-sample+1/2)
  result <- cbind(lower,upper)
  return(result)
}
CI_Jef(0.05,3)
```

```
          lower     upper
[1,] 0.03498475 0.2867275
```

(d)

The 100(1-α)% CI using the Jeffrey method is $\hat{F}(t) \pm z_{\alpha/2} \times \sqrt{\dfrac{\hat{F}(t)(1-\hat{F}(t))}{n}}$

```r
# wrong
CI_Wald <- function(alpha,sample){
  lower <- sample-qnorm(1-alpha/2,0,1)*sqrt(sample*(1-sample)/30)
  upper <- sample+qnorm(1-alpha/2,0,1)*sqrt(sample*(1-sample)/30)
  result <- cbind(lower,upper)
  colnames(result) <- c("lower","upper")
  rownames(result) <- c("t=5","t=21","t=28")
  return(result)
}
knitr::kable(CI_Wald(0.05,y[-c(1,5)]),escape = FALSE)
```

```r
# correct
CI_Wald <- function(alpha,Fhat){
  n=25
  l <- Fhat-qnorm(1-alpha/2,0,1)*sqrt(Fhat*(1-Fhat)/n)
  lower <- ifelse(l<0,0,l)
  upper <- Fhat+qnorm(1-alpha/2,0,1)*sqrt(Fhat*(1-Fhat)/n)
  result <- cbind(lower,upper)
  return(result)
}
CI_Wald(0.05,0.12)
```

```
     lower     upper
[1,]     0 0.2473826
```

(e)

why?

```r
# correct
CI_above <- rbind(CI_CP(0.05,3),CI_Jef(0.05,3),CI_Wald(0.05,0.12))
summary <- data.frame("Method"=c("Clopper-Pearson","Jeffrey","Wald"),
                      "lower"=CI_above[,1],"upper"=CI_above[,2],
                      "length"=CI_above[,2]-CI_above[,1])
knitr::kable(summary)
```

Method	lower	upper	length
Clopper-Pearson	0.0254654	0.3121903	0.2867249
Jeffrey	0.0349848	0.2867275	0.2517428
Wald	0.0000000	0.2473826	0.2473826

The Wald interval for binomial proportion is derived based on the asymptotic distribution of MLE.When the sample size is small or the sample proportion is close to 0 or 1,the statistical inference(like coverage probability) may not be good. For the CI derived from Clopper-Pearson or Jeffrey's method,they construct the exact confidence interval.So,in such cases,$\hat{F}(5) = 0.04$ and $\hat{F}(21) = 0.08$ are both close to 0,the Clopper-Pearson or Jeffrey's method would be preferred over the Wald interval for constructing a confidence interval for the binomial proportion.

(f)

(wrong) ~~Let X be the failure time of units,$X \sim \text{Exp}(\lambda = 2.3)$, the density $f(x) = 2.3\exp(-2.3x), x > 0$. $P(X \leq 365 \times 10) = \int_0^{3650} f(x)dx \approx 1$,which means that the probability of devices would fail in 10 years of operation is very close to 1.Therefore,the manufacturer should say that we will focu on improving the reliability of the product and providing good after-sales services to avoid customer dissatisfaction.~~

(correct) The product use $10 \times 365 \times 2.3/1000 = 8.395$ thousand of cycles in 10 years. So,the fraction of failure is P(T < 8.395) = 1/25=0.04.

(g)

(wrong) ~~Assume that there have different rates,λ_1, λ_2 ,and X : failure time of units.~~

The density of X,f(x) $= p \times f_1(x) + (1-p) \times f_2(x)$, where p : unknown or known weight proportion

$f_i(x)$: the p.d.f. of Exp(λ_i)

$$
\begin{aligned}
&F(x) = P(X \leq x) = p(1 - \exp(-\lambda_1 x)) + (1-p)(1 - \exp(-\lambda_2 x)) \\
&E(X) = \int_0^\infty (1 - F(x))dx = p/\lambda_1 + (1-p)/\lambda_2 \\
&S(x) = 1 - F(x) = pe^{-\lambda_1 x} + (1-p)e^{-\lambda_2 x} \\
&\text{h.f. } \lambda(x) = \frac{p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}}{pe^{-\lambda_1 x} + (1-p)e^{-\lambda_2 x}}
\end{aligned}
$$

~~From the above,we see that the expected value can be obtained from the weights,and the hazard rate function is not constant when use rate varies in the population of units.~~

(correct) Suppose that the failure time in **cycles** is denoted by C with a cdf $F_C(c)$ and the failure time in days is denoted by T with cdf $F_T(t)$.Then,conditional on a fixed use rate of r cycles per day,T=C/r.That is ,the cdf of T is $F_T(t) = F_C(tr)$.
Now,if there is a pooulation of K groups each having it own use rate r_k and the population of units in group k is π_k,then the cdf for the population
$F_T(t) = \sum_{k=1}^{K} \pi_k F_C(tr_k)$.
This is known as a discrete mixture distribution.

2.

(a)

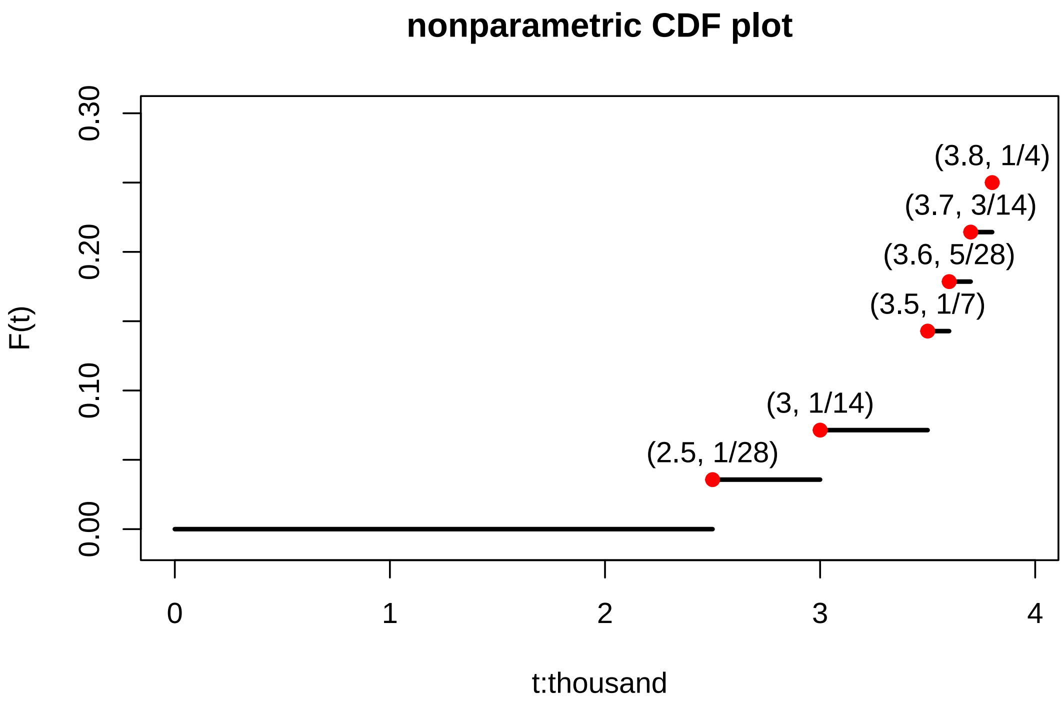
Based on this information,we can infer about the distribution of silicon photodiode detectors lifetime. Thus,it could be used to construct the statistical models to predict the probability of lifetime at different time intervals.This may be more suitable for statisticians or reliability engineers to study it.

(b)

```r
# wrong
data_2 <- read.csv("PhotoDetector.csv",header = T)
x <- c(0,2500,3000,3500,3600,3700,3800,3900)
y <- fractions(c(0,1/28,2/28,3/28,4/28,5/28,6/28,6/28))
plot(x,y,col=0,
     xlim=c(0,3950),ylim=c(-0.01,0.25),
     xlab="t",ylab="F(t)",lwd=2,
     main="nonparametric CDF plot")
for(i in 1:6){
  lines(x[i:(i+1)],y[c(i,i)],lwd=2.5)
}
points(x[-c(1,8)], y[-c(1,8)], pch = 19, col = "red")
text(x=x[-c(1,8)], y=y[-c(1,8)],
     labels = paste("(", x[-c(1,8)], ", ", y[-c(1,8)], ")", sep = ""),
     pos = 3)
```

```r
# correct
data_2 <- read.csv("PhotoDetector.csv",header = T)
x <- c(0,2500,3000,3500,3600,3700,3800,3900)/1000
y <- c(0,1/28,2/28,4/28,5/28,6/28,7/28,7/28)
plot(x,y,col=0,
     xlim=c(0,3.95),ylim=c(-0.01,0.3),
     xlab="t:thousand",ylab="F(t)",lwd=2,
     main="nonparametric CDF plot")
for(i in 1:6){
  lines(x[i:(i+1)],y[c(i,i)],lwd=2.5)
}
points(x[-c(1,8)], y[-c(1,8)], pch = 19, col = "red")
text(x=x[-c(1,8)], y=y[-c(1,8)],
```

```
    labels = paste("(", x[-c(1,8)], ", ", fractions(y[-c(1,8)]), ")", sep = ""),
    pos = 3)
```



(c)

(wrong) ~~By Greenwood's formula,~~

$$\hat{\text{Var}}(\hat{F}(t_i)) = (\hat{S}(t_i))^2 \sum_{j:t_j \leq t_i} \frac{\hat{p}_j}{n_j(1-\hat{p}_j)}, \text{where } \hat{p}_j = \frac{d_j}{n_j}$$

```
#wrong
d <- c(1,1,2,1,1,1) # num. of failed
n <- c(27,26,24,23,22,21) # num. of entered
S <- 1-y[-c(1,8)] #each survival
est_VarF <- 0
for (i in 1:length(d)){
  est_VarF[i] <- S[i]*sum(d[1:i]/(n[1:i]*(n[1:i]-d[1:i])))
}

result <- data.frame("time"=x[-c(1,8)],"Failed"=d,
                     "Entered"=n,
                     "F"=y[-c(1,8)],
                     "est"=est_VarF)
knitr::kable(result,row.names = FALSE,
          col.names = c(colnames(result)[-c(4,5)],
```

```
                        "$\\hat{F}(t_i)$",
                        "$\\text{Var}(\\hat{F}(t_i))$"))
```

(correct) This is inspection data,its variance:

$$\frac{\hat{F}(t_i)(1-\hat{F}(t_i)}{n}$$

With logit trandformation,the 100(1-α)% CI for logit($\hat{F}(t)$) is

$$\log\frac{\hat{F}(t)}{1-\hat{F}(t)} \pm z_{1-\alpha/2} \times (\hat{F}(t)(1-\hat{F}(t)))^{-1} \times \text{s.e.}_{\hat{F}(t)}$$

```
# correct
d <- c(1,1,2,1,1,1) # num. of failed
n <- c(27,26,24,23,22,21) # num. of entered
S <- 1-y[-c(1,8)] #each survival
est_VarF <- y[-c(1,8)]*(1-y[-c(1,8)])/28
l.lower <- y[-c(1,8)]/(y[-c(1,8)]+(1-y[-c(1,8)])*exp(qnorm(p = 0.975) *1/(y[-c(1,8)]*(1-y[-c(1,8)]
l.upper <- y[-c(1,8)]/(y[-c(1,8)]+(1-y[-c(1,8)])/exp(qnorm(p = 0.975) *1/(y[-c(1,8)]*(1-y[-c(1,8)]
result <- data.frame("t(i-1)"=c(x[-c(8)]),
                     "t(i)"=c(x[-c(1,8)],""),
                     "Failed"=c(d,0),
                     "Entered"=c(28,n),
                     "F"=c(y[-c(1)]),
                     "var"=c(est_VarF,est_VarF[6]),
                     "lower(logit)"=c(l.lower,l.lower[6]),
                     "upper(logit)"=c(l.upper,l.upper[6]))
knitr::kable(result,row.names = FALSE,
             col.names = c("$t_{i-1}$",
                        "$t_i$",
                        colnames(result)[-c(1,2,5,6,7,8)],
                        "$\\hat{F}(t_i)$",
                        "$\\text{Var}(\\hat{F}(t_i))$",
                        colnames(result)[c(7,8)]))
```

t_{i-1}	t_i	Failed	Entered	$\hat{F}(t_i)$	$\text{Var}(\hat{F}(t_i))$	lower.logit.	upper.logit.
0.0	2.5	1	28	0.0357143	0.0012300	0.0050077	0.2141806
2.5	3	1	27	0.0714286	0.0023688	0.0179303	0.2447653
3.0	3.5	2	26	0.1428571	0.0043732	0.0546678	0.3244802
3.5	3.6	1	24	0.1785714	0.0052387	0.0763383	0.3637925
3.6	3.7	1	23	0.2142857	0.0060131	0.0995732	0.4021319

t_{i-1}	t_i	Failed	Entered	$\hat{F}(t_i)$	$\text{Var}(\hat{F}(t_i))$	lower.logit.	upper.logit.
3.7	3.8	1	22	0.2500000	0.0066964	0.1241167	0.4394945
3.8		0	21	0.2500000	0.0066964	0.1241167	0.4394945

(d)(e)

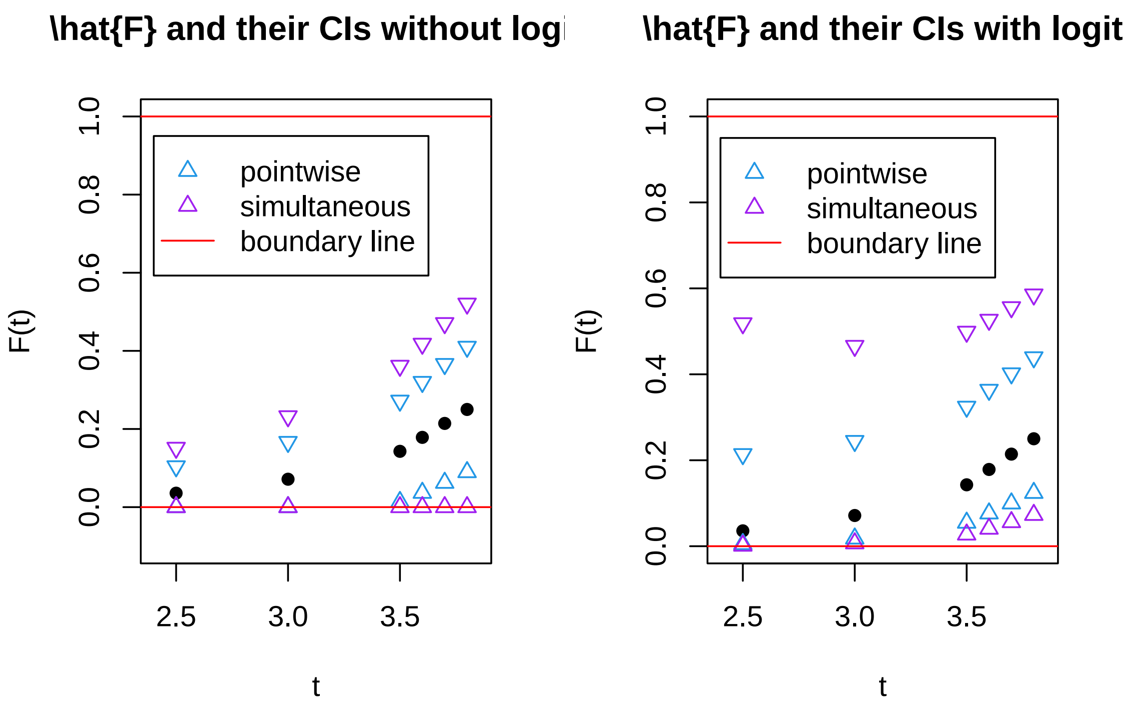
- pointwise 95% CI: $\hat{F}(t_i) \pm z_{(0.975)} \times se_{\hat{F}}(t_i)$
- simultaneous 95% CI:$\hat{F}(t_i) \pm 3.31 \times se_{\hat{F}}(t_i)$,where $3.31 = e_{(0.01,0.99,0.975)}$.

```r
# wrong
plot(result[,1],result[,4],col=0,xlim=c(2400,3850),ylim=c(-0.1,1.1),
     xlab="t",ylab="F(t)")
Vi = result[,5]
e=3.31
for(i in 1:5){
  lines(x=result[i:(i+1),1],y=result[c(i,i),4],lwd=2.5)
  lines(x=result[i:(i+1),1],
        y=result[c(i,i),4]+qnorm(0.975)*sqrt(Vi)[c(i,i)],
        lty=5,col=4,lwd=2.5)
  lines(x=result[i:(i+1),1],
        y=result[c(i,i),4]-qnorm(0.975)*sqrt(Vi)[c(i,i)],
        lty=5,col=4,lwd=2.5)
  lines(x=result[i:(i+1),1],
        result[c(i,i),4]+e*sqrt(Vi)[c(i,i)],
        lty=5,col="purple",lwd=2.5)
  lines(x=result[i:(i+1),1],
        result[c(i,i),4]-e*sqrt(Vi)[c(i,i)],
        lty=5,col="purple",lwd=2.5)
}
abline(h=c(0,1),lwd=1,col="gray")
legend(x=2500,y=1,c("pointwise","simultaneous"),
       col=c(4,"purple"),lwd=2.5)
```

```r
#correct
par(mfrow=c(1,2))
### plot 1
plot(as.numeric((result[,2])[-7]),result[1:6,5],col=1,
     xlim=c(2.4,3.85),ylim=c(-0.1,1),xlab="t",ylab="F(t)",
     pch=16,main="\\hat{F} and their CIs without logit")
```

```
Vi = result[,6]
e=3.31
p.l <- ifelse((result[,5])[-7] - qnorm(p = 0.975)*sqrt(Vi[-7]) <= 0,0,(result[,5])[-7] - qnorm(p =
p.u <- result[,5][-7] + qnorm(p = 0.975)*sqrt(Vi[-7])
s.l <- ifelse((result[,5])[-7] -e*sqrt(Vi[-7]) <= 0,0,(result[,5])[-7] -e*sqrt(Vi[-7]))
s.u <- (result[,5])[-7] +e*sqrt(Vi[-7])
points(x = as.numeric((result[,2])[-7]),y=p.l,
       col=4,type="p",lty=2,pch=24) #pointwise lower
points(x = as.numeric((result[,2])[-7]),y=p.u,
       col=4,type="p",lty=2,pch=25) # pointwise upper

points(x = as.numeric((result[,2])[-7]),y=s.l,
       col="purple",type="p",lty=2,pch=24) # simulanteous lower
points(x = as.numeric((result[,2])[-7]),y=s.u,
       col="purple",type="p",lty=2,pch=25) # simulanteous upper
abline(h=c(0,1),lwd=1,col="red")
legend(x=2.4,y=0.95,c("pointwise","simultaneous","boundary line"),
       col=c(4,"purple","red"),pch=c(24,24,NA),lty=c(NA,NA,1))
### plot2
plot(as.numeric((result[,2])[-7]),result[1:6,5],col=1,
     xlim=c(2.4,3.85),ylim=c(0,1),xlab="t",ylab="F(t)",
     pch=16,main="\\hat{F} and their CIs with logit")
points(x = as.numeric((result[,2])[-7]),y=result[-7,7],
       col=4,type="p",lty=2,pch=24) #pointwise lower
points(x = as.numeric((result[,2])[-7]),y=result[-7,8],
       col=4,type="p",lty=2,pch=25) # pointwise upper
l.si <- y[-c(1,8)]/(y[-c(1,8)]+(1-y[-c(1,8)])*exp(e *1/(y[-c(1,8)]*(1-y[-c(1,8)]))*sqrt(est_VarF))
u.si <- y[-c(1,8)]/(y[-c(1,8)]+(1-y[-c(1,8)])/exp(e *1/(y[-c(1,8)]*(1-y[-c(1,8)]))*sqrt(est_VarF))
points(x = as.numeric((result[,2])[-7]),y=l.si,
       col="purple",type="p",lty=2,pch=24) # simulanteous lower
points(x = as.numeric((result[,2])[-7]),y=u.si,
       col="purple",type="p",lty=2,pch=25) # simulanteous upper
abline(h=c(0,1),lwd=1,col="red")
legend(x=2.4,y=0.95,c("pointwise","simultaneous","boundary line"),
       col=c(4,"purple","red"),pch=c(24,24,NA),lty=c(NA,NA,1))
```



(f)

Pointwise CI and simultaneous confidence bands are two different approaches to contructing CI. Pointwise CIs are constructed based on the concept of treating each failure sample likely as independent asymptotic normal samples. However,simultaneous confidence bands are constructed by considering the Bonferroni correction concept,which is usually more conservative and makes the width longer.

3.

(b)(c)

3. $\mathbb{P} = (p_1, p_2, \cdots, p_m)$

(b) $L(\mathbb{P}) = \left[p_1 \cdot (1-p_1)^{r_1}\right] \times \left[((1-p_1)p_2)^{d_2} \cdot ((1-p_1)(1-p_2))^{r_2}\right] \cdots \cdot$

$$\left(\prod_{i=1}^{m-1}(1-p_i)\, p_m\right)^{d_m} \left(\prod_{i=1}^{m}(1-p_i)\right)^{r_m}$$

$$= \left(\prod_{i=1}^{m} p_i^{d_i}\right) \cdot \left[\prod_{i=1}^{m}(1-p_i)^{\overset{n_i - d_i}{\cancel{n - \sum_{j=1}^{i} d_j - \sum_{j=1}^{i-1} r_j}}}\right]$$

$$= \prod_{i=1}^{m} p_i^{d_i}(1-p_i)^{n_i - d_i} \quad . \#$$

(c)

let

$$\ell(\mathbb{P}) = \ln L(\mathbb{P}) = \sum_{i=1}^{m}\left[d_i \ln p_i + (n_i - d_i)\ln(1-p_i)\right]$$

$$\frac{\partial \ell(\mathbb{P})}{\partial p_i} = \frac{d_i}{p_i} - \frac{n_i - d_i}{1-p_i} = 0 \quad \Rightarrow \quad \hat{p}_i = \frac{d_i}{n_i}$$

$$\left.\frac{\partial^2 \ell(\mathbb{P})}{\partial p_i^2}\right|_{p_i = \hat{p}_i} = \frac{-d_i}{(d_i/n_i)^2} - \frac{n_i - d_i}{(1 - d_i/n_i)^2} = -n_i^2\left(\frac{1}{d_i} + \frac{1}{n_i - d_i}\right) < 0$$

Then, for each j, the MLE of p_j is $\hat{p}_j = \frac{p_j}{n_j}$ #

(d)

(d)

$$-\frac{\partial^2 \log(L(p))}{\partial p_i^2} = \frac{n_i - d_i}{(1-p_i)^2} + \frac{d_i}{p_i^2} = \frac{n_i}{p_i(1-p_i)}$$

$$-\frac{\partial^2 \log(L(p))}{\partial p_i \partial p_j} = 0 \text{ is trivial by part (c).}$$

So, the observation information matrix is diagonal matrix.

Recall $S(t_i) = \prod_{j=1}^{i} (1-p_j)$

$$\frac{\partial \ln S(t_i)}{\partial p_j} = \frac{S'(t_i)}{S(t_i)} = -\frac{1}{1-p_j} \Rightarrow \frac{\partial S(t_i)}{\partial p_j} = -\frac{S(t_i)}{1-p_j}$$

By Taylor expansion,

$$\hat{S}(t_i) \approx S(t_i) + \sum_{j|t_j \le t_i} \frac{\partial S(t_i)}{\partial p_j}\bigg|_{p_j = \hat{p}_j} (\hat{p}_j - p_j)$$

$$\Rightarrow Var(\hat{S}(t_i)) \approx \sum_{j:t_j \le t_i} \left(\frac{S(t_i)}{1-\hat{p}_j}\right)^2 \cdot Var(\hat{p}_j)$$

$$= (S(t_i))^2 \sum_{j:t_j \le t_i} \left(\frac{1}{1-\hat{p}_j}\right)^2 \cdot \frac{\hat{p}_j \cdot (1-\hat{p}_j)}{n_j}$$

$$= (S(t_i))^2 \sum_{j:t_j \le t_i} \frac{\hat{p}_j}{n_j(1-\hat{p}_j)} \quad . \#$$

4.

(a)

Why? From the concept of limit,

$$\frac{-\log(1-p)}{p} \xrightarrow[p\to 0]{L.H.} \frac{1}{1-p} \xrightarrow{p\to 0} 1$$

The condition to assure a good agreement between $\hat{H}(t_i)$ and $\hat{\hat{H}}(t_i)$ is $\hat{p}_j$ for each j is small and close to 0,and thus agree between $\hat{F}(t_i)$ and $\hat{\hat{F}}(t_i)$.

(correct) By Maclaurin series,

$$-\log(1-x) \approx x + \frac{x^2}{2} + \frac{x^3}{3} + ...$$

If x is small enough,above approxiate x.Thus,

$$\hat{H}(t_i) \approx \hat{\hat{H}}(t_i) \text{ and } \hat{F}(t) = 1 - \exp(-\hat{H}(t)) \approx \hat{\hat{F}}(t)$$

(b)

4. (b)

The Taylor expansion for $\log(1-\hat{p}_j)$:

$$\log(1-\hat{p}_j) \approx \log(1-p_j) - \frac{1}{1-p_j}(\hat{p}_j - p_j)$$

$$Var(\log(1-\hat{p}_j)) \approx \left(\frac{1}{1-p_j}\right)^2 Var(\hat{p}_j)$$

$$= \frac{1}{(1-p_j)^2} \cdot \frac{n_j \cdot \hat{p}_j(1-\hat{p}_j)}{n_j^2} = \frac{\hat{p}_j}{n_j(1-\hat{p}_j)}$$

So, $$Var(\hat{H}(t_i)) = \sum_{j=1}^{i} Var(\log(1-\hat{p}_j))$$

$$= \sum_{j=1}^{i} \frac{\hat{p}_j}{n_j(1-\hat{p}_j)}$$

and $Var(\hat{\hat{H}}(t_i)) = \sum_{j=1}^{i} \frac{\hat{p}_j(1-\hat{p}_j)}{n_j}$ is trivial.

For each j,

$$\frac{1-\hat{p}_j}{\frac{1}{1-\hat{p}_j}} = (1-\hat{p}_j)^2 \xrightarrow{\hat{p}_j \to 0} 1 \text{ which}$$

implies for small $\hat{p}_j$, two variance are approximate.

(c)

Do the organization for “Fan.csv”:

```r
# wrong
data_4c <-read.csv("Fan.csv")
di_index <- which(data_4c$Censoring.Indicator == "Fail")
di <- rep(0,length(data_4c[,1]))
di[di_index]=data_4c$Count[di_index]
ri <- rep(0,length(data_4c[,1]))
ri[-di_index] =data_4c$Count[-di_index]
ni <- 0
ni[1] <- sum(data_4c$Count)
for(i in 2:length(data_4c[,1])){
  ni[i] <- ni[1] - sum(data_4c$Count[1:i-1])
}
Si=cumprod(1-di/ni);Fi=1-Si
Hi_KM=-log(Si)
Hi_NA=cumsum(di/ni)
tab <- round(cbind(data_4c[,1],di,ni,di/ni,
                   "H(ti)K.M"=Hi_KM,"H(ti)N.A"=Hi_NA),4)
colnames(tab)[c(1,4)]=c("ti","di/ni")
tab=rbind(c(0,0,70,0,0,0),tab)
knitr::kable(tab)
```

$$1 - F(t) = \exp\left(-H(t)\right) \Rightarrow F(t) = 1 - \exp\left(-H(t)\right)$$

Let's present the table for comparison between K.M F(t) and N.A F(t):

```r
#wrong
compar_F <- data.frame("ti"=tab[,1],
                       "cd"=round(c(0,di/ni),4),
                       "V1"=round(1-exp(-tab[,5]),4),
                       "V2"=round(1-exp(-tab[,6]),4))
colnames(compar_F)[2:4] <- c("di/ni","K.M","N.A")
knitr::kable(compar_F)
```

```r
# correct
data_4c <-read.csv("Fan.csv")
t.index <- unique(data_4c$Hours)
di <- rep(0,length(t.index))
label=0
for (i in 1:35){
  label <- which(data_4c$Hours == t.index[i] & data_4c$Censoring.Indicator=="Fail")
  di[i] <- ifelse(length(label)==0,0,data_4c$Count[label])
```

```r
}
ri <- rep(0,length(t.index))
for (i in 1:35){
  label <- which(data_4c$Hours == t.index[i] & data_4c$Censoring.Indicator=="Censored")
  ri[i] <- ifelse(length(label)==0,0,data_4c$Count[label])
}

ni <- 0
ni[1] <- sum(data_4c$Count)
for(i in 2:length(di)){
  ni[i] <- ni[1] - sum(ri[1:i-1])-sum(di[1:i-1])
}
Si=cumprod(1-di/ni);Fi=1-Si
Hi_KM=-log(Si)
Hi_NA=cumsum(di/ni)
K.M = 1-exp(-Hi_KM)
N.A = 1-exp(-Hi_NA)

result <- data.frame("ti"=t.index,
                     "di"=di,"ni"=ni,
                     "pi"=round(di/ni,4),
                     "V1"=round(K.M,4),
                     "V2"=round(N.A,4))
colnames(result)[5:6] <- c("K.M","N.A")
knitr::kable(result)
```

ti	di	ni	pi	K.M	N.A
450	1	70	0.0143	0.0143	0.0142
460	0	69	0.0000	0.0143	0.0142
1150	2	68	0.0294	0.0433	0.0428
1560	0	66	0.0000	0.0433	0.0428
1600	1	65	0.0154	0.0580	0.0574
1660	0	64	0.0000	0.0580	0.0574
1850	0	63	0.0000	0.0580	0.0574
2030	0	58	0.0000	0.0580	0.0574
2070	2	55	0.0364	0.0923	0.0910
2080	1	53	0.0189	0.1094	0.1080
2200	0	52	0.0000	0.1094	0.1080
3000	0	51	0.0000	0.1094	0.1080
3100	1	47	0.0213	0.1283	0.1268

ti	di	ni	pi	K.M	N.A
3200	0	46	0.0000	0.1283	0.1268
3450	1	45	0.0222	0.1477	0.1460
3750	0	44	0.0000	0.1477	0.1460
4150	0	42	0.0000	0.1477	0.1460
4300	0	38	0.0000	0.1477	0.1460
4600	1	34	0.0294	0.1728	0.1707
4850	0	33	0.0000	0.1728	0.1707
5000	0	29	0.0000	0.1728	0.1707
6100	1	26	0.0385	0.2046	0.2020
6300	0	22	0.0000	0.2046	0.2020
6450	0	21	0.0000	0.2046	0.2020
6700	0	19	0.0000	0.2046	0.2020
7450	0	18	0.0000	0.2046	0.2020
7800	0	17	0.0000	0.2046	0.2020
8100	0	15	0.0000	0.2046	0.2020
8200	0	13	0.0000	0.2046	0.2020
8500	0	12	0.0000	0.2046	0.2020
8750	1	9	0.1111	0.2930	0.2859
9400	0	6	0.0000	0.2930	0.2859
9900	0	5	0.0000	0.2930	0.2859
10100	0	4	0.0000	0.2930	0.2859
11500	0	1	0.0000	0.2930	0.2859

From the above,(K.M) CDF and (N.A) CDF are very approximate,but at t=8750, there is a slight difference between the two values because $\hat{p}$ at t=8750 is not approximated to 0 from the discussion of Problem 4,(a).

(d)

4. (d)

Let $f(t) = \log \frac{1}{1-t} - t$, $t \in [0,1)$

$f'(t) = \frac{1}{1-t} - 1 \geq 0$ and $f(0) = 0$

$\Rightarrow -\log(1-\hat{p}_j) > \hat{p}_j$ for all $\hat{p}_j \in (0,1)$.

From the above, $\hat{\hat{H}}(t_i) < \hat{H}(t_i)$

$e^{-H(t)} = 1 - F(t) \Rightarrow F(t) = 1 - e^{-H(t)}$

$\because e^{-x}$ is an nonincreasing function.

why? $\therefore \hat{F}(t_i) > \hat{\hat{F}}(t_i)$ by $\hat{\hat{H}}(t_i) < \hat{H}(t_i)$ #

(e)

~~For this situation,each n_i's are unknown,we can estimation the CDF within each interval. Suppose that for each interval,with probability π_i. The likelihood:~~

$$
\begin{aligned}
L(\pi) &\propto p_1^{l_1}(1-p_1)^{r_1}(p_1+p_2)^{l_2}(1-(p_1+p_2))^{r_2}\ldots \\
&= \prod_{i=1}^{n} \xi_i^{l_1}(1-\xi_i)^{r_i}, \\
&\text{where } \xi_i = \sum_{j=1}^{i} p_j, l_i : \text{numbers of failed}, r_i : \text{numbers of censored}
\end{aligned}
$$

~~The MLE of $\xi_i = \frac{l_i}{l_i+r_i}$,furthermore,we can get all $\hat{p_j}$.Thus,the estimator of $\hat{H}(t)$ and $\hat{\hat{H}}(t)$ can be obtained.~~

(correct) At time $t_i, n_i = n - \sum_{j=1}^{i-1}(d_j + r_j)$.Hence,$\hat{p}_j = d_j/n_j$ can be used when failure and censoring times are grouped into common intervals.

5.

5. $$\hat{\delta}(t) = n \cdot \sum_{j: t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

$$\hat{k}(t) = \hat{\delta}(t) / (1 + \hat{\delta}(t))$$

By Greenwood's formula,

$$\sum_{j: t_j \le t} \frac{d_j}{n_j(n_j - d_j)} = \sum_{j: t_j \le t} \frac{\hat{p}_j}{n_j(1 - \hat{p}_j)} \approx \frac{Var(\hat{S}(t))}{(\hat{S}(t))^2}$$

Then, $$\hat{\delta}(t) = n \cdot \frac{Var(\hat{S}(t))}{(\hat{S}(t))^2} = \frac{n\, Var(\hat{F}(t))}{(1 - F(t))^2}$$

$$= \frac{F(t)(1 - F(t))}{(1 - F(t))^2}, \quad \because n\hat{F}(t) \sim Bin(n, F(t))$$

$$= \frac{F(t)}{(1 - F(t))}$$

Thus,

$$\hat{k}(t) = \frac{F(t)/(1 - F(t))}{1/(1 - F(t))} = F(t). \quad \#$$