Tarea 5

Métodos No Paramétricos en Análisis de Supervivencia

Modelos de series de Tiempo y Supervivencia

Profesor: Naranjo Albarrán Lizbeth

Adjuntos: Reyes González Belén

Rivas Godoy Yadira

Integrantes: Cuéllar Chávez Eduardo de Jesús

García Tapia Jesús Eduardo

Miranda Meraz Areli Gissell

Ramírez Maciel José Antonio

Saldaña Morales Ricardo

Grupo: 9249

Fecha: 17/DIC/2021

Utiliza la base de datos de R llamada larynx, del paquete KMsurv, y realice lo siguiente:

Explique sus resultados.

Análisis descriptivo

Ejercicio 1

Realice un análisis descriptivo sobre el tiempo de supervivencia de los sujetos, además de sus características generales y particulares.

```
## Warning: package 'survminer' was built under R version 4.1.2
```

```
## Warning: package 'ggpubr' was built under R version 4.1.2
```

```
## Warning: package 'ggsci' was built under R version 4.1.2
```

Composición de la base

```
attach(larynx)
head(larynx)
```

```
##   stage time age diagyr delta
## 1     1  0.6  77     76     1
## 2     1  1.3  53     71     1
## 3     1  2.4  45     71     1
## 4     1  2.5  57     78     0
## 5     1  3.2  58     74     1
## 6     1  3.2  51     77     0
```

Esta base de datos es acerca del tiempo de supervivencia en pacientes que padecen cáncer de laríngе en distintas etapas.

Tenemos 5 variables y 90 observaciones:

Stage, que representa la etapa en la que se encuentra dicha enfermedad del paciente

Time. que nos dice el tiempo de la observación, en meses

Age, que es edad del paciente cuando se le diagnosticó la enfermedad

Diagyr, que es el año en que se diagnisticó la enfermedad

Delta, que es un indicador tal que es 1 si el paciente murió al tiempo de observación y 0 y se encuentra vivo. En nuestro contexto, la muerte es la falla y el estar vivo la censura

NA y número de observaciones

```
length(larynx$time)
```

```
## [1] 90
```

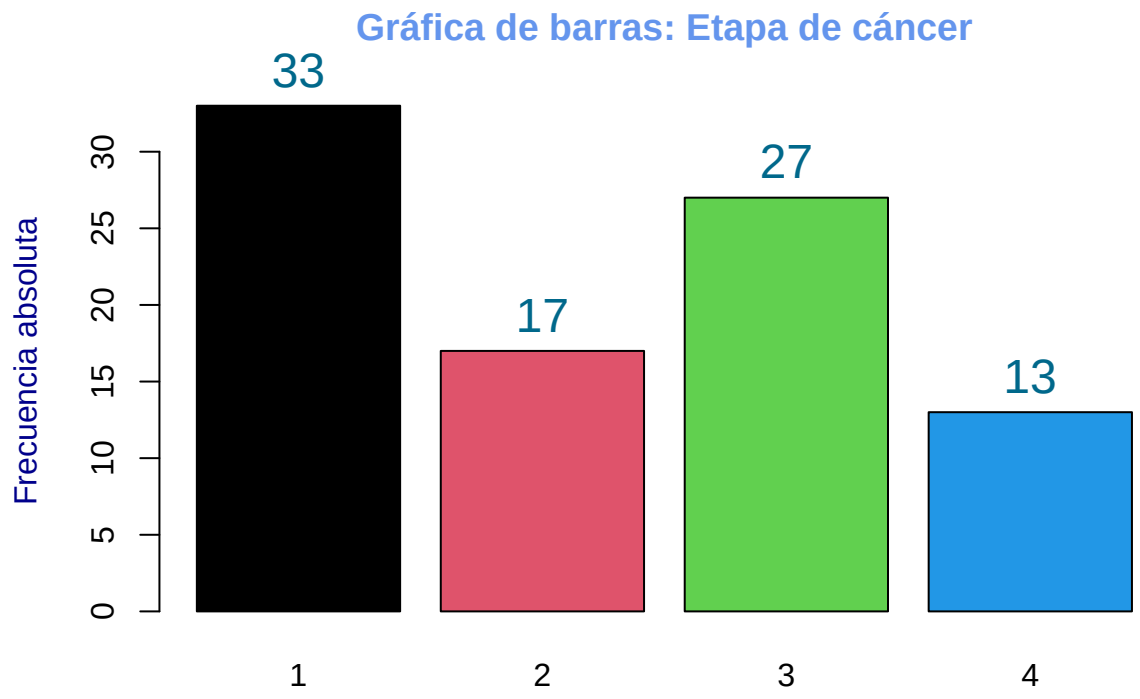
```
sum(is.na.data.frame(larynx))
```

```
## [1] 0
```

Tenemos 5 variables y 90 observaciones, no tenemos ningún NA.

Etapa

```
#Usamos la paleta de lancet porque es una revista de oncología
Edad <- table(stage)
names(Edad) <- unique(larynx$stage)
b <- barplot(Edad,col= palette(pal_lancet("lanonc")(5)),
             main="Gráfica de barras: Etapa de cáncer", col.main="cornflowerblue",
             ylab="Frecuencia absoluta", col.lab="blue4")
text(x=b,y=c(Edad), labels=c(Edad), pos=3, col="deepskyblue4", cex=1.5, xpd=TRUE)
```




```
#Observaciones
Edad
```

```
##  1  2  3  4
## 33 17 27 13
```

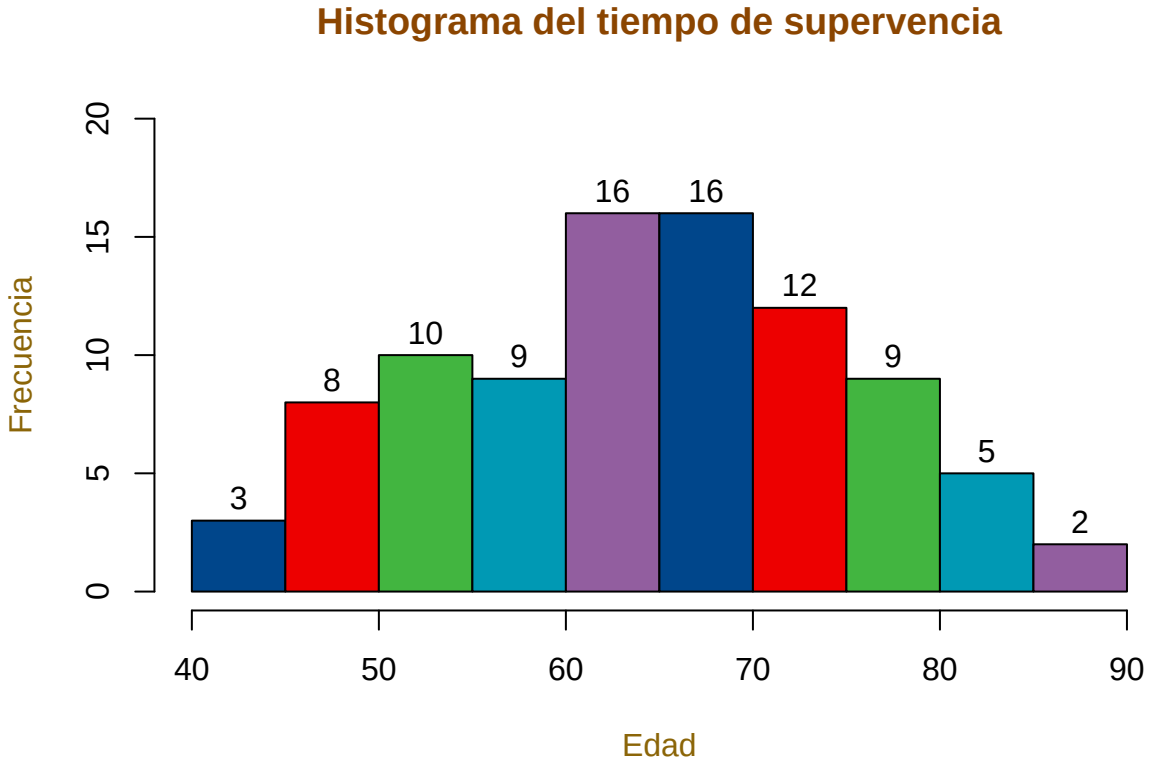
```
#Porcentaje
Edad/.9
```

```
##        1        2        3        4
## 36.66667 18.88889 30.00000 14.44444
```

Podemos ver que la mayoría de las observaciones se concentran en pacientes en etapa 1, que es la inicial, concentrando el 36.66% de las observaciones, después sigue la etapa 3, que tiene 27 observaciones las cuales representan el 30% del total, seguidas de 17 observaciones en etapa 2, representando un 18.8% y al final, como resulta intuitivo, la etapa con menos observaciones es la 4, con 13 que representan el 14.4% del total, esto porque es la etapa más avanzada y grave de la enfermedad, por lo que es entendible que hayan menos pacientes de esta etapa y más de la primera, pues es natural que existan más pacientes en dicho grupo.

Edad

```
edad<-hist(age, probability=F, col=palette(pal_lancet("lanonc")(9)), border="gray0",
     main="Histograma del tiempo de supervencia",col.main="darkorange4",ylab = "Frecuencia", col.lab =
     xlab = "Edad", labels=TRUE,ylim=c(0,20))
```




```
#Puntos de corte
edad$breaks
```

```
##  [1] 40 45 50 55 60 65 70 75 80 85 90
```

```
#Cantidad
edad$counts
```

```
##  [1]  3  8 10  9 16 16 12  9  5  2
```

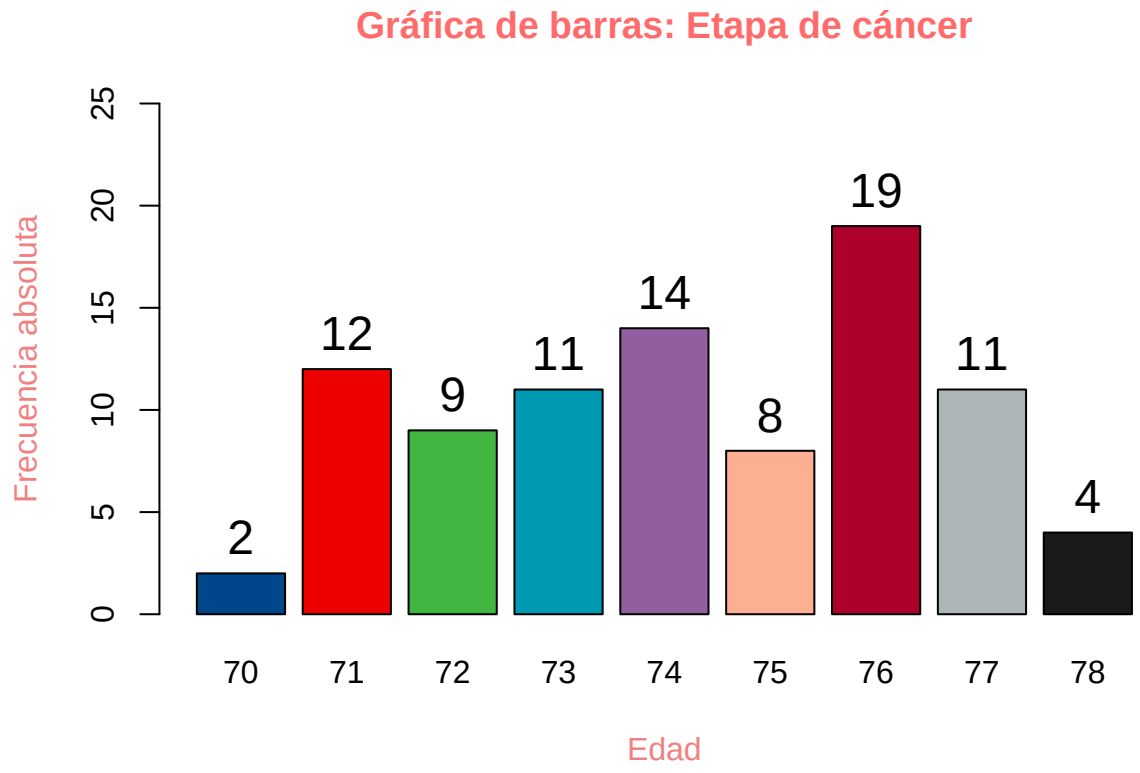
```
#Porcentaje
edad$counts/.9
```

```
##  [1]  3.333333  8.888889 11.111111 10.000000 17.777778 17.777778 13.333333
##  [8] 10.000000  5.555556  2.222222
```

Podemos ver que la mayoría de las observaciones se concentran entre 61 y 70 años, teniendo 32 pacientes en dicho grupo de edad, y notamos que mientras se alejan de este grupo de edad, va decreciendo el número de pacientes en ese grupo de edad, parece tener una forma parecida a una distribución simétrica, o acampanada, pero no parece ser una normal, a primera vista, aunque le da un parecido. el grupo más joven y el más viejo tienen la menor concentración de pacientes.

Año de diagnóstico

```
year <- table(diagyr)
b <- barplot(year,col= palette(pal_lancet("lanonc")(5)),
             main="Gráfica de barras: Etapa de cáncer", col.main="indianred1",
             ylab="Frecuencia absoluta", xlab = "Edad", col.lab="lightcoral",ylim=c(0,25))
text(x=b,y=c(year), labels=c(year), pos=3, col="gray0", cex=1.5, xpd=TRUE)
```




```
#Observaciones
year
```

```
## diagyr
## 70 71 72 73 74 75 76 77 78
##  2 12  9 11 14  8 19 11  4
```

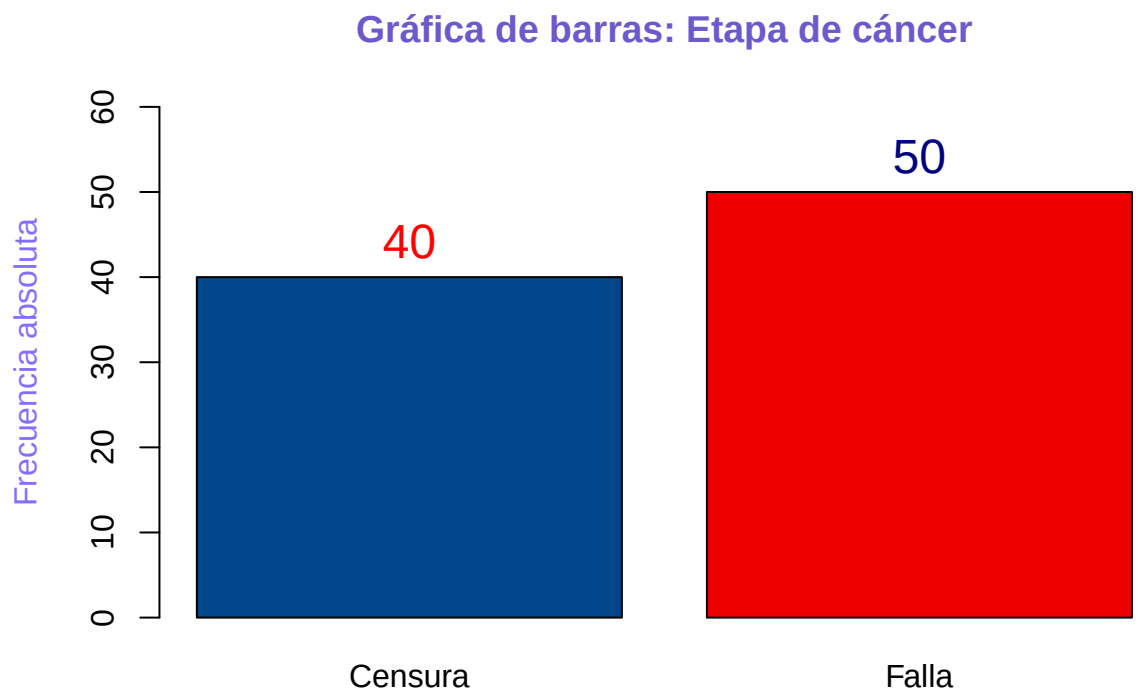
```
#Porcentaje
year/.9
```

```
## diagyr
##        70        71        72        73        74        75        76        77
##  2.222222 13.333333 10.000000 12.222222 15.555556  8.888889 21.111111 12.222222
##        78
##  4.444444
```

Tenemos que el año en el que más personas recibieron su diagnóstico fue en 1976, con 19 personas que representan el 21.11% de los pacientes. Le sigue 1974 con 14 pacientes Hay casi la mosma cantidad de pacientes cuyo diagnóstico fue en 1971,1973 y 1977, siendo de 12 en el primero y 11 en los otros 2. Después de esto, 1972 y 1975 tienen casi los mismos representantes, siendo 9 y 8 respectivamente Posteriormente tenemos

a 1978, que es el año más reciente, con 4 representantes. Esto puede ser natural ya que, podría costar un poco de trabajo contactar a personas para el estudio que hayan recibido su diagnóstico muy recientemente, si es que el estudio se realizó el mismo año. Es de esperar que el año con menor pacientes registrados sea 1970, pues es el año más lejano, con solo 2 pacientes. ##Fallas y censuras

```
fallas <- table(delta)
names(fallas)<-c('Censura','Falla')
b <- barplot(fallas,col= palette(pal_lancet("lanonc")(5)),
             main="Gráfica de barras: Etapa de cáncer", col.main="slateblue",
             ylab="Frecuencia absoluta", col.lab="slateblue1",ylim=c(0,60))
text(x=b,y=c(fallas), labels=c(fallas), pos=3, col= c("red", "navy"), cex=1.5, xpd=TRUE)
```

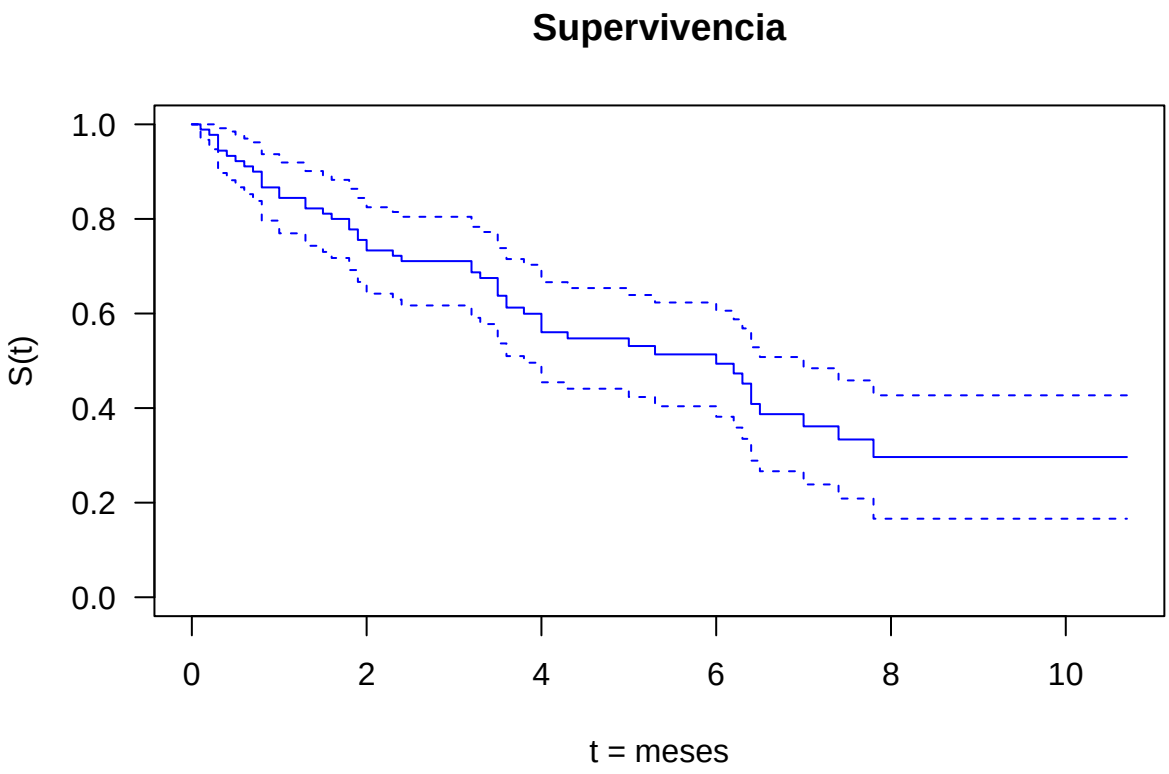


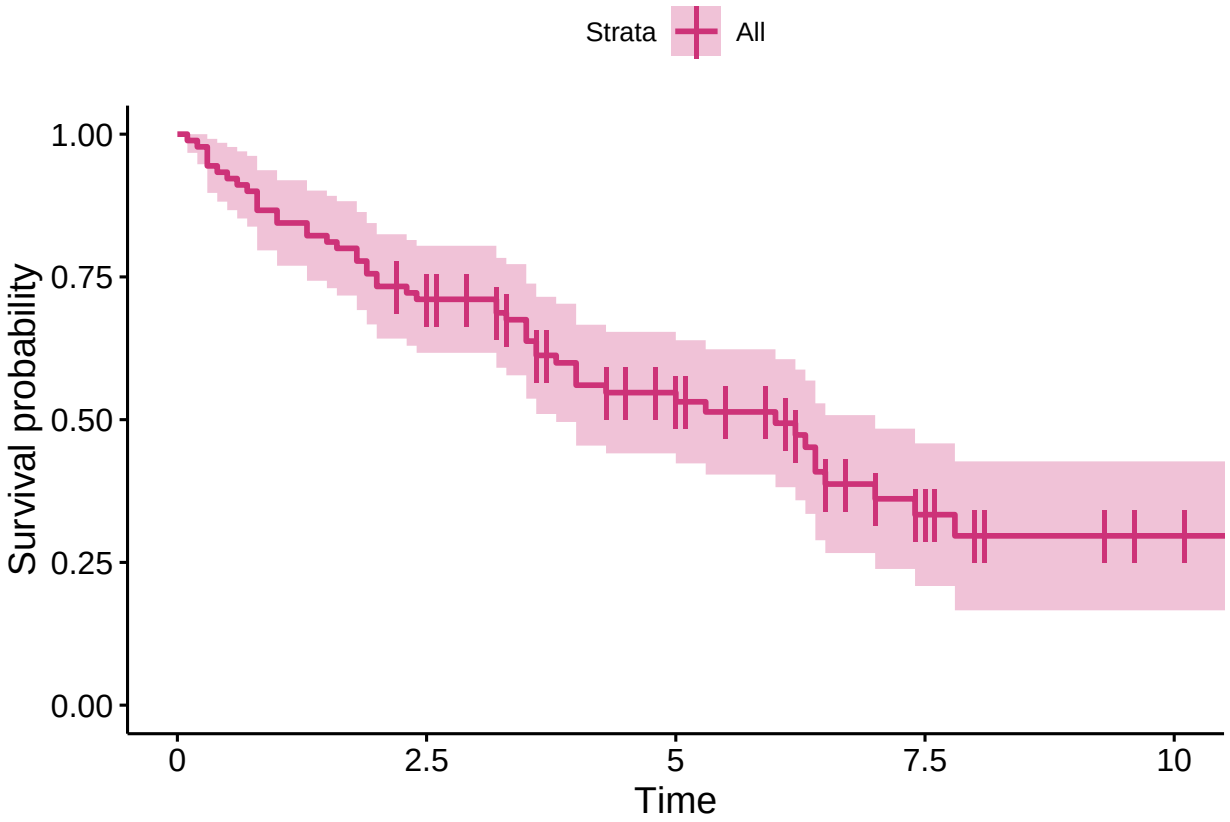

```
fallas/.9
```

```
##  Censura    Falla
## 44.44444 55.55556
```

Tenemos 50 fallas, que representan el 55.55% de nuestros datos, y 40 censuras, que representan el 44.44% de nuestros datos. Tenemos ligeramente más fallas que censuras en nuestro estudio. ## Función de supervivencia

```
Surv_general <- Surv(larynx$time, larynx$delta,type = "right")
general_fit <- survfit(Surv_general ~ 1, type = "kaplan-meier", conf.int = 0.95,
                       conf.type = "plain",data=larynx)
plot(general_fit, main="Supervivencia", xlab =  "t = meses", ylab = "S(t)",
           col="blue1", las=1)
ggsurvplot(general_fit, data=larynx,palette = "violetred3",censor.size=7,
           censor.shape= 124)
```


Supervivencia
1.0
0.8
0.6
0.4
0.2
0.0
S(t)
0
2
4
6
8
10
t = meses



Podemos observar que al inicio del estudio, en los primeros dos meses, se tiene una gran cantidad de fallas, seguido por el periodo de 3 a 4 meses y de 6 a 6 y medio meses, aproximadamente; siendo los periodos donde más se concentran las fallas.

Otro punto a destacar es que no llegamos a 0 en la función de supervivencia, esto porque los últimos datos son censuras, como podemos ver en el dataset (en este contexto es que se encuentran vivos).

Podemos notar un periodo de estabilidad, donde la función de supervivencia se mantiene constante, en el periodo de aproximadamente 4 meses y medio a poco antes de 6, si bien tenemos fallas, estas son pocas, por lo que se mantiene casi constante en el periodo mencionado, aunque existen muchas fallas (bajo este contexto, ello significa que el sujeto de estudio se encontraba vivo al tiempo de observación).

Al inicio, los intervalos de confianza están bastante cercanos al estimador puntual.
Los intervalos de confianza se hacen muy grandes hacia el final de la función de supervivencia, pero sesgados hacia la derecha, es decir, el intervalo de confianza superior tiene una distancia mayor al estimador puntual, respecto del intervalo inferior.

Dicho lo anterior, se puede intuir que los primeros dos meses son los 'más difíciles', puesto que hay muchas fallas en este periodo (son frecuentes y es donde más se acumulan), pero observamos que conforme se acerca a los 8 meses se entra a un periodo "seguro", donde las fallas son casi nulas, y al tiempo de observación vemos sujetos vivos (fallas) únicamente; teniendo un valor de entre 0.22 y 0.40 aproximadamente, la función de supervivencia con los intervalos de confianza, y de aproximadamente 0.26 con el estimador puntual, por lo que tienen todavía una "Buena" probabilidad de sobrevivir al menos hasta este periodo.

Kaplan-Meier

Ejercicio 2

Con el estimador de Kaplan-Meier para la función de supervivencia S(t), calcule y grafique: S(t) poblacional. S(t) por estadio de la enfermedad. S(t) por grupos de edad. Identifique las variables que afectan el tiempo de supervivencia. Incluya los intervalos del 95% confianza.

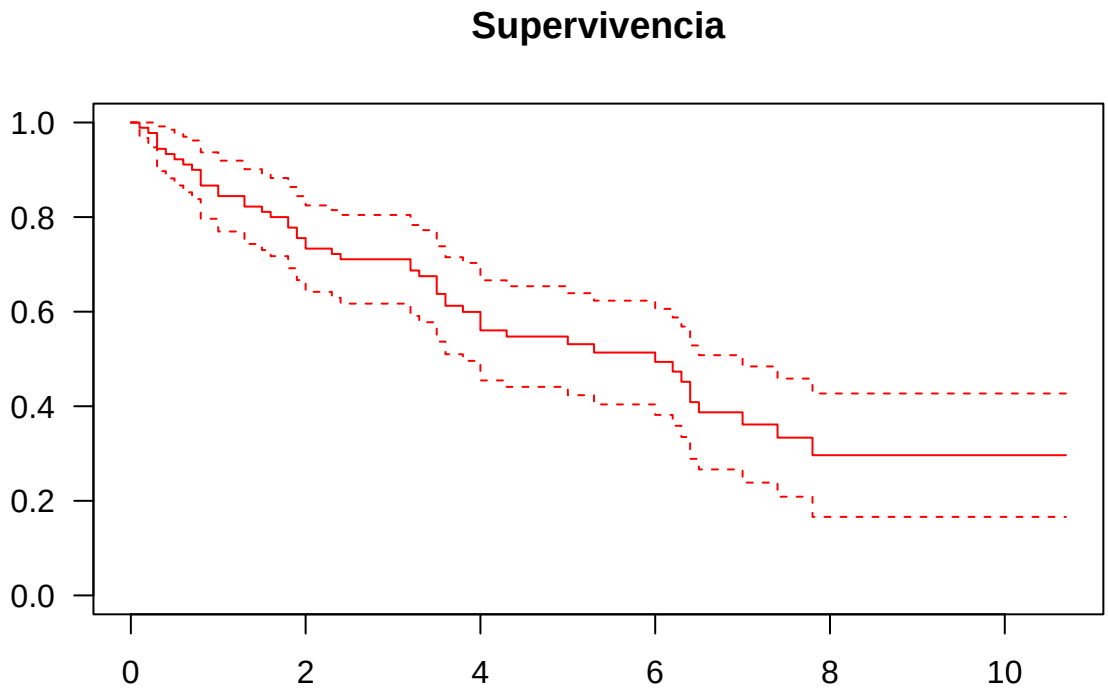
S(t) poblacional

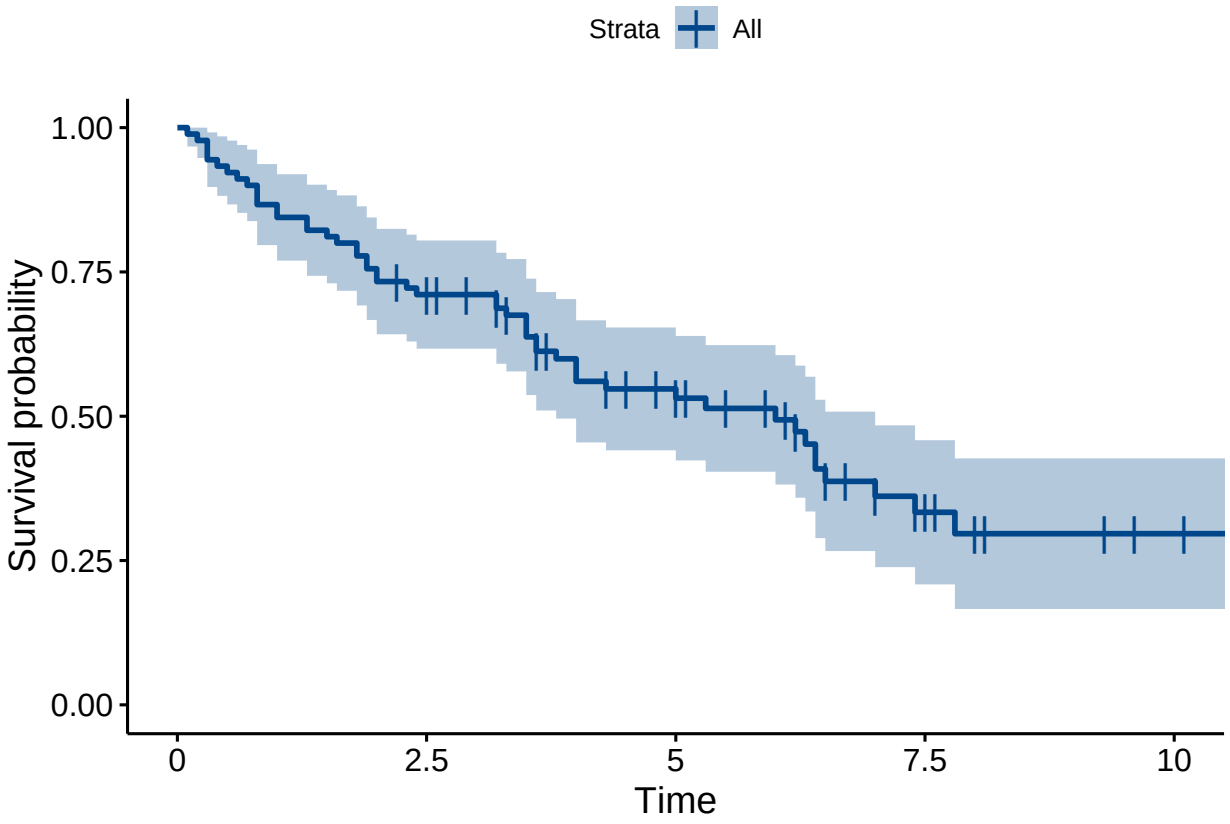
Este ya lo calculamos en el inciso anterior:

```
summary(general_fit)
```

```
## Call: survfit(formula = Surv_general ~ 1, data = larynx, type = "kaplan-meier",
##     conf.int = 0.95, conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.1     90       1    0.989  0.0110        0.967        1.000
##   0.2     89       1    0.978  0.0155        0.947        1.000
##   0.3     88       3    0.944  0.0241        0.897        0.992
##   0.4     85       1    0.933  0.0263        0.882        0.985
##   0.5     84       1    0.922  0.0282        0.867        0.978
##   0.6     83       1    0.911  0.0300        0.852        0.970
##   0.7     82       1    0.900  0.0316        0.838        0.962
##   0.8     81       3    0.867  0.0358        0.796        0.937
##   1.0     78       2    0.844  0.0382        0.770        0.919
##   1.3     76       2    0.822  0.0403        0.743        0.901
##   1.5     74       1    0.811  0.0413        0.730        0.892
##   1.6     73       1    0.800  0.0422        0.717        0.883
##   1.8     72       2    0.778  0.0438        0.692        0.864
##   1.9     70       2    0.756  0.0453        0.667        0.844
##   2.0     68       2    0.733  0.0466        0.642        0.825
##   2.3     65       1    0.722  0.0472        0.629        0.815
##   2.4     64       1    0.711  0.0478        0.617        0.805
##   3.2     60       2    0.687  0.0491        0.591        0.783
##   3.3     57       1    0.675  0.0497        0.578        0.772
##   3.5     54       3    0.638  0.0514        0.537        0.738
##   3.6     51       2    0.613  0.0524        0.510        0.715
##   3.8     47       1    0.599  0.0528        0.496        0.703
##   4.0     46       3    0.560  0.0540        0.455        0.666
##   4.3     43       1    0.547  0.0543        0.441        0.654
##   5.0     34       1    0.531  0.0550        0.423        0.639
##   5.3     30       1    0.514  0.0560        0.404        0.623
##   6.0     26       1    0.494  0.0572        0.382        0.606
##   6.2     24       1    0.473  0.0584        0.359        0.588
##   6.3     22       1    0.452  0.0596        0.335        0.568
##   6.4     21       2    0.409  0.0612        0.289        0.529
##   6.5     19       1    0.387  0.0616        0.266        0.508
##   7.0     15       1    0.361  0.0627        0.239        0.484
##   7.4     13       1    0.334  0.0637        0.209        0.458
##   7.8      9       1    0.297  0.0666        0.166        0.427
```

```
plot(general_fit, main="Supervivencia", col="red", las=1)
ggsurvplot(general_fit, data=larynx,palette = 'lancet',censor.size=4.9,
           censor.shape=124)
```


Supervivencia
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10



S(t) por estadio de la enfermedad

S(t) por grupos de edad

Pruebas Log-Rank con $\alpha = 0.05$

Ejercicio 3

Usando pruebas No paramétricas Log-Rank:

Por estadio

Compare las funciones de supervivencia por estadio, es decir, realice el contraste de hipótesis:

$$H_0 : S_j(t) = S_k(t) \forall t > 0, \forall j, k. \textbf{ vs } .H_a : S_j(t) \neq S_k(t); \textbf{p.a } t > 0, p.aj \neq k.$$

Por grupo de edad

Compare las funciones de supervivencia por grupos de edad