Confiabilidad y Datos Censurados: la otra dimensión de la calidad

Felipe Neira Rojas & Angel Llanos Herrera

2024-10-02

Section 1

Introducción

Introducción

- Existen varias dimensiones para medir la calidad; Desempeño, **confiabilidad** , durabilidad, estética, características incluidas, entre otras. Garvin (1988)
- En el estudio de la confiabilidad y evaluacion de supervivencia o durabilidad a menudo aparecen datos censurados. Se tiene solo información parcial.

Section 2

Confiabilidad

Confiabilidad

- La confiabilidad es crucial para medir la calidad, ya que, esta relacionada con la capacidad del producto para cumplir las expectativas del cliente.
- Responde a la pregunta de ¿Con qué frecuencia falla el producto?
- Un producto de alta confiabilidad es aquel que tiene pocas o ninguna falla dentro de su ciclo de vida esperado, garantizando su funcionalidad.

Formas de medir la confiabilidad

Función de confiabilidad:

$$R_T(t) = 1 - F_T(t), (t > 0)$$

Donde,
$F_T(t) = P(T \leq t) = \int_0^t f_T(u)du$.
$R_T(0) = 1$, debido a que el producto al tiempo 0 debería funcionar con 100% de probabilidad.
$\lim_{t\to\infty} R_T(t) = 0$.

Función de riesgo o tasa de fallas:

$$h_T(t) = \frac{f_T(t)}{R_T(t)}, (t > 0)$$

Donde,
$f_T(t)$, función de densidad de probabilidad de la variable aleatoria T.
$R_T(t)$, función de confiabilidad en el tiempo t de la variable aleatoria T.

Tiempo medio de vida $E(T)$:

$$E(T) = \int_0^\infty tf(t)dt$$

Section 3

Datos censurados

Datos censurados

- Los datos censurados son aquellos en los que no es posible conocer con precisión el valor exacto de una variable de interés, pero se tiene información parcial sobre ella.
- Estas observaciones parciales si pueden aportar información, por lo que es importante incorporarlas de manera adecuada.

Tipos de datos censurados

Censura tipo I:

Se establece un tiempo de seguimiento fijo (el evento es variable)

Censura tipo II:

Se establece un numero de eventos fijo (el tiempo es variable)

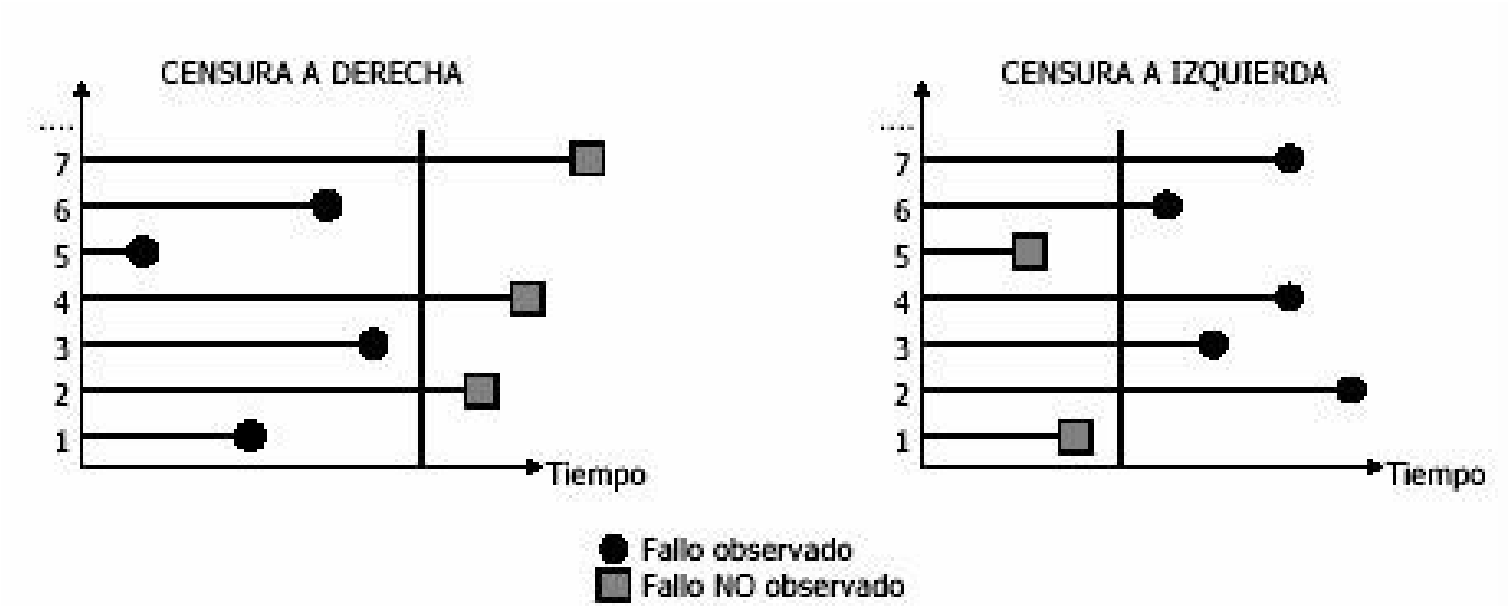
Censura Tipo I

Censura a la derecha

- Si al finalizar el tiempo de seguimiento no se ha observado el evento de interes, se considera que esta observación esta censurada (= tiempo final)

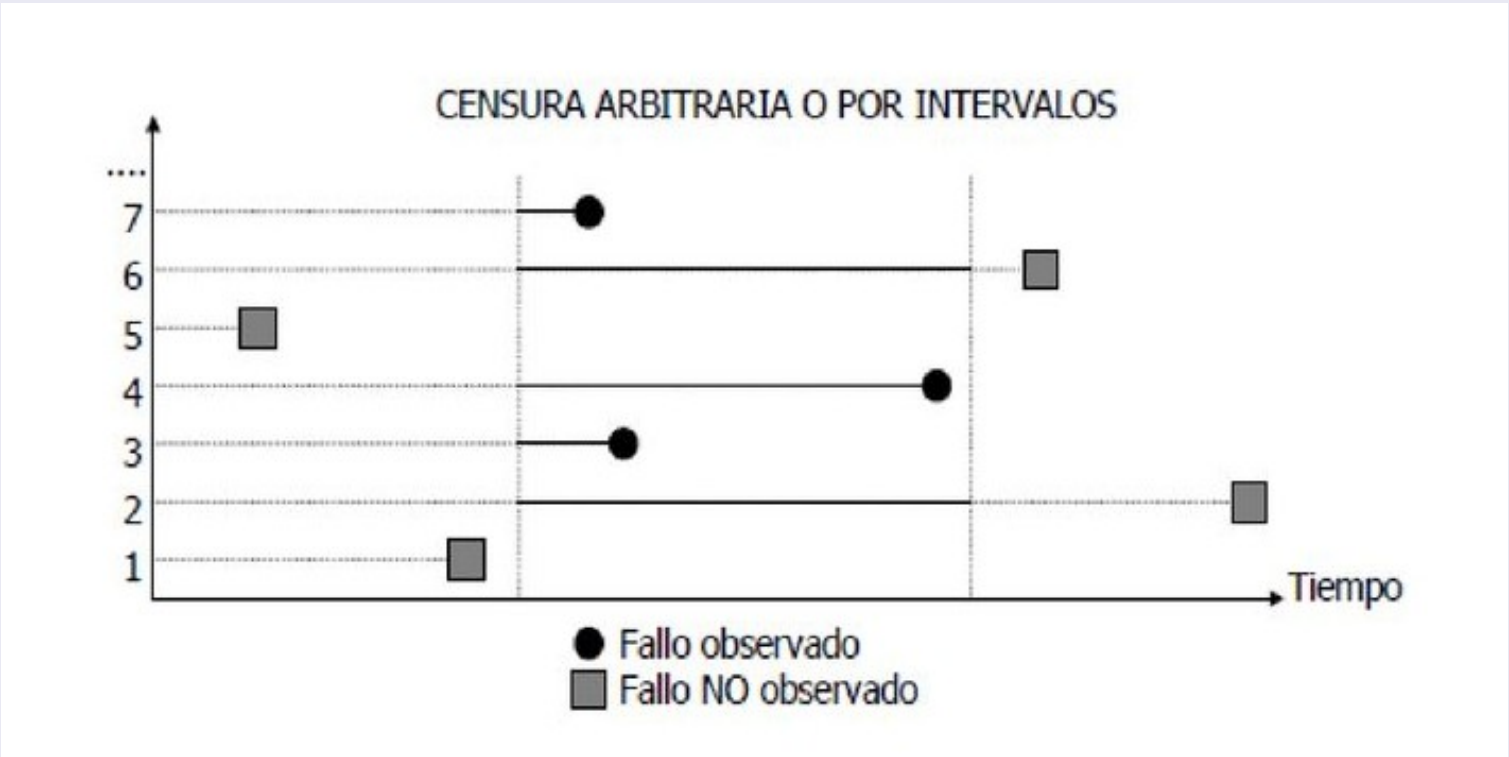
Censura a la izquierda

- Si el evento de interés ya ha ocurrido antes del inicio del seguimiento, se considera que esta observación está censurada a la izquierda.


CENSURA A DERECHA
....
7
6
5
4
3
2
1
Tiempo
CENSURA A IZQUIERDA
....
7
6
5
4
3
2
1
Tiempo
Fallo observado
Fallo NO observado

Censura por intervalos

- Si el evento de interes se observa dentro de un intervalo. Se sabe que el evento ocurrió, pero no se conoce el tiempo exacto. Por lo tanto, se censura dentro de ese intervalo.



Análisis Estadístico para Datos Censurados

- Ignorar la censura en el proceso de estudio de confiabilidad puede llevar a conclusiones equivocadas. Es por esto que es necesario estimar parámetros y funciones que consideren los datos censurados para obtener conclusiones precisas.
- El objetivo del análisis de supervivencia es incorporar esta información parcial que proporcionan los individuos censurados mediante métodos desarrollados para este fin (San José et al., 2009).
- Los modelos estadísticos ayudan a estimar la función de supervivencia (probabilidad que un producto falle en cierto tiempo), o también, estimar tiempos de fallos fututos.

Modelos para Análisis de Supervivencia

- Existen modelos paramétricos y no paramétricos: Estimador de Máxima verosimilitud y Estimador de Kaplan-Meier.
- Los supuestos para estos modelo son: independencia entre los tiempos de vida (hasta el evento) y tiempos de censura.

Estimador de Máxima Verosimilitud

Método de estimación paramétrico para la función de supervivencia.

$$L = \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_r) \prod_{i \in L} [1 - S(C_l)] \prod_{i \in I} [S(L_i) - S(R_i)]$$

Donde,

D: Conjunto de tiempos de muerte

R: Conjunto de observaciones con censura a la derecha

L: Conjunto de observaciones con censura a la izquierda

I: Conjunto de observaciones con censura por intervalos

Estimador de Kaplan-Meier

Método de estimación no paramétrico para la función de supervivencia.

$$\hat{C}(t_{(i)}) = \prod_{j=1}^{i} \left(1 - \frac{f(j)}{n(j)}\right)$$

En dónde,
$\hat{C}(t_{(i)})$: Estima la función de supervivencia.
$1 - \frac{f(j)}{n(j)}$: Probabilidad de que la unidad sobreviva en el tiempo t, dado que sobrevivió el periodo anterior.
f_j: Unidades que fallaron en el tiempo j.
n_j: Unidades de riesgo, la resta del total de unidades con las unidades que ya fallaron en los periodos estudiados.

¿Cúal técnica utilizar?

- El método a elegir dependerá de la distribución de los datos.
- Utilizar el Estimador de Máxima Verosimilitud en datos con distribución conocida entrega un análisis robusto.
- En caso de no cumplir el supuesto de distribución conocida, el Estimador de Kaplan-Meier se adapta a cualquier distribución.

Section 4

Relacion entre Confiabilidad y Datos Censurados

Relacion entre Confiabilidad y Datos Censurados

- Subestimación o sobreestimación de la confiabilidad al ignorar datos censurados.
- El análisis de supervivencia es una herramienta para evaluar la confiabilidad mientras se tienen en cuenta datos censurados.

Section 5

Aplicación

Ejercicio 1

Durabilidad de celulares despues de reparación

Una empresa de telefonía móvil quiere evaluar la eficiencia de tres servicios técnicos diferentes (A, B y C) a los cuales envía sus celulares para reparaciones. Se dispone de datos sobre el tiempo que cada celular funciona correctamente después de ser reparado, y se desea analizar si hay diferencias en la durabilidad de los celulares según el servicio técnico.
El periodo de interés del estudio es desde 2020 a 2023 (36 meses)

Objetivo y Preguntas

Analizar el tiempo que un celular funciona correctamente después de una reparación y evaluar si el servicio técnico influye en la durabilidad.

1. ¿Cuál es la tasa de fallas acumulada para cada servicio técnico?
2. ¿Cuál es la mediana de durabilidad de los celulares después de ser reparados, según el servicio técnico?
3. ¿Existen diferencias significativas en la durabilidad de los celulares según el servicio técnico?

Resolución (Importamos librerías y datos)

Librerías

```
# Importar las librerías necesarias
library(survival)
library(fitdistrplus) # Ajustar distribución
library(tidyr)
library(ggplot2)
library(survminer)
library(flexsurv) # EMV
library(readr)
```

Datos

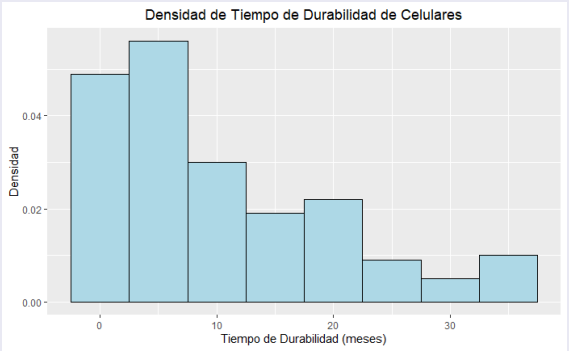
```
head(datos_celulares, 3)
```

```
## # A tibble: 3 x 3
##   Tiempo_Durabilidad Evento Servicio_Tecnico
##                <dbl>  <dbl> <chr>
## 1                  3      1 C
## 2                  3      1 C
## 3                 11      0 C
```

Resolución (Distribución)

Distribución

H_0: Los datos de Tiempo_Durabilidad siguen una distribución Weibull.
H_1: Los datos de Tiempo_Durabilidad no siguen una distribución Weibull.



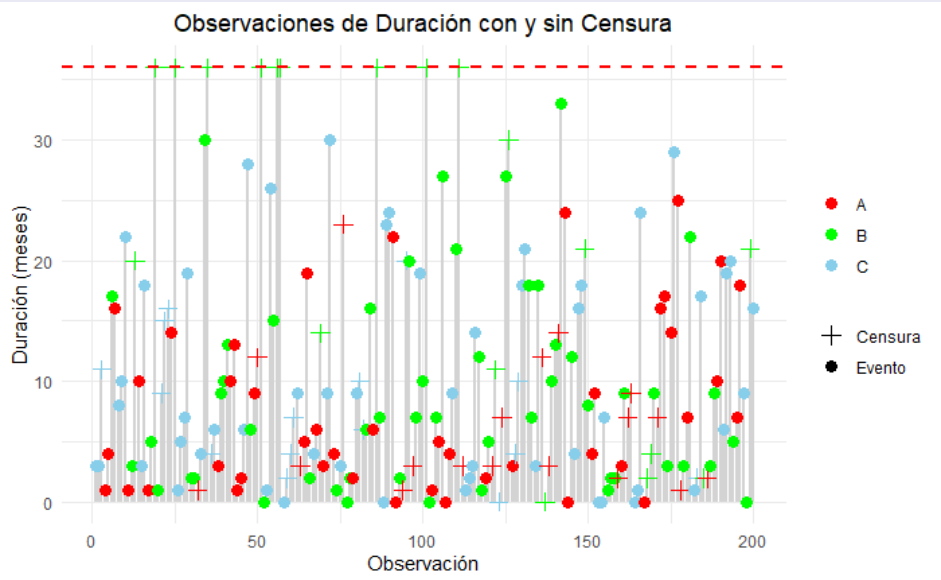

```
datos_celulares$Tiempo_Durabilidad[datos_celulares$Tiempo_Durabilidad==0] <- 0.01
ajuste_weibull <- fitdist(datos_celulares$Tiempo_Durabilidad, "weibull")
forma <- ajuste_weibull$estimate["shape"]
escala <- ajuste_weibull$estimate["scale"]

ks.test(datos_celulares$Tiempo_Durabilidad, "pweibull", shape = forma, scale = escala)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos_celulares$Tiempo_Durabilidad
## D = 0.089475, p-value = 0.08133
## alternative hypothesis: two-sided
```

Resolución (Censura)

Periodo de estudio (36 meses)

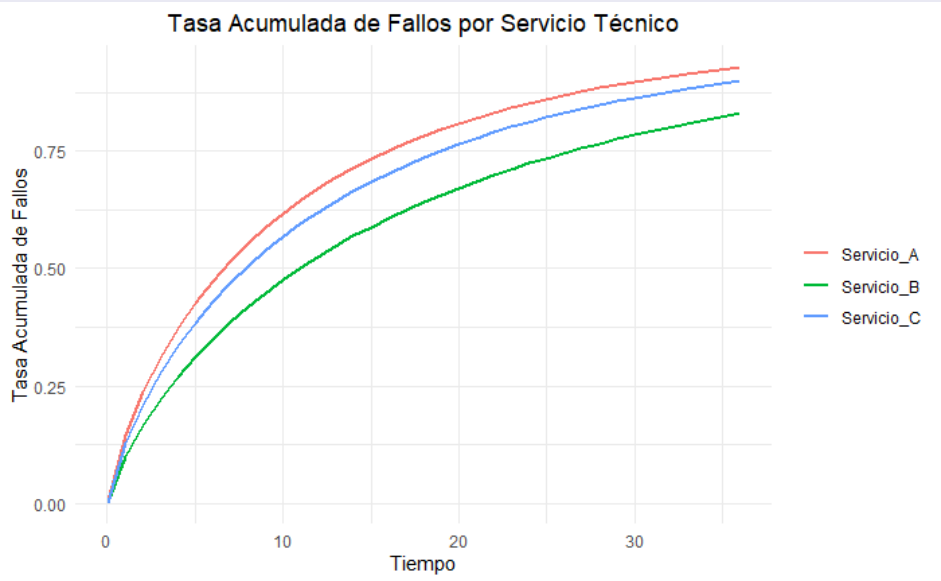
```
# Censurar los datos que exceden los 36 meses (censura a la derecha)
datos_celulares$Tiempo_Durabilidad[datos_celulares$Tiempo_Durabilidad > 36] <- 36
datos_celulares$Evento[datos_celulares$Tiempo_Durabilidad == 36] <- 0 # Censura
```



Resolución (Pregunta 1)

Tasa de fallas acumuladas

```
# Ajuste de la funcion de supervivencia con ajuste Weibull (EMV) Estratificado
flex <- flexsurvreg(Surv(datos_celulares$Tiempo_Durabilidad,
                         datos_celulares$Evento) ~ datos_celulares$Servicio_Tecnico,
                    data = datos_celulares, dist = "weibull") # Ajuste weibull
flexgg <- flex %>% summary(type = "survival") %>% data.frame

tasa_fallos_acumuladas <- data.frame(
  tiempo = flexgg$datos_celulares.Servicio_Tecnico.A.time,
  Servicio_A = (1- flexgg$datos_celulares.Servicio_Tecnico.A.est),
  Servicio_B = (1- flexgg$datos_celulares.Servicio_Tecnico.B.est),
  Servicio_C = (1- flexgg$datos_celulares.Servicio_Tecnico.C.est))
```



Resolución (Pregunta 2)

Mediana de durabilidad de los celulares según servicio tecnico

```
#Se hace lo mismo para Servicio Tecnico B y C
ajuste_weibull_A <- fitdist(
  datos_celulares$Tiempo_Durabilidad[datos_celulares$Servicio_Tecnico=="A"],
  "weibull")
forma_A <- ajuste_weibull_A$estimate["shape"]
escala_A <- ajuste_weibull_A$estimate["scale"]
mediana_A <- qweibull(0.5, shape = forma_A, scale = escala_A)


print("------------Medianas-----------")
```

```
## [1] "------------Medianas-----------"
```

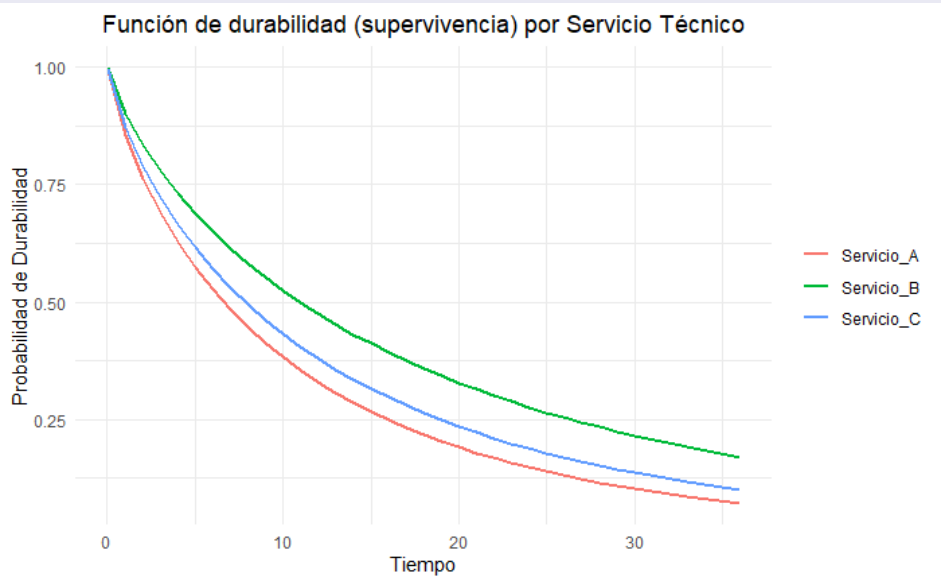
```
paste("A = ", round(mediana_A,2), "B = ", round(mediana_B,2), "C = ", round(mediana_C,2))
```

```
## [1] "A =  4.73 B =  7.67 C =  6.1"
```

Resolución (Pregunta 3)

Curva de durabilidad (Supervivencia)

```
supervivencia <- data.frame(
  tiempo = flexgg$datos_celulares.Servicio_Tecnico.A.time,
  Servicio_A = flexgg$datos_celulares.Servicio_Tecnico.A.est,
  Servicio_B = flexgg$datos_celulares.Servicio_Tecnico.B.est,
  Servicio_C = flexgg$datos_celulares.Servicio_Tecnico.C.est)
```



Resolución (Pregunta 3)

Diferencias significativas: Durabilidad según servicio técnico

Aplicamos una Prueba de Log-Rank, donde las hipotesis son:

- H_0: No existe diferencia en la función de supervivencia entre los servicios tecnicos.
- H_1: Existe al menos una diferencia en la función de supervivencia entre los servicios tecnicos.

```
surv_obj <- Surv(datos_celulares$Tiempo_Durabilidad, datos_celulares$Evento)

# Prueba de log-rank
log_rank_test <- survdiff(surv_obj ~ Servicio_Tecnico, data = datos_celulares)
log_rank_test
```

```
## Call:
## survdiff(formula = surv_obj ~ Servicio_Tecnico, data = datos_celulares)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## Servicio_Tecnico=A 61       43     33.9     2.463     3.551
## Servicio_Tecnico=B 71       52     64.8     2.520     5.121
## Servicio_Tecnico=C 68       53     49.4     0.269     0.438
##
##  Chisq= 6  on 2 degrees of freedom, p= 0.05
```

Resolución (Pregunta 3)

¿Cuáles difieren significativamente?

```
log_rank_result <- survdiff(surv_obj ~ Servicio_Tecnico, data = datos_celulares,
  subset = Servicio_Tecnico %in% c("A", "C"))
log_rank_result
```

```
## Call:
## survdiff(formula = surv_obj ~ Servicio_Tecnico, data = datos_celulares,
##     subset = Servicio_Tecnico %in% c("A", "C"))
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## Servicio_Tecnico=A 61       43     36.8      1.05      1.93
## Servicio_Tecnico=C 68       53     59.2      0.65      1.93
##
##  Chisq= 1.9  on 1 degrees of freedom, p= 0.2
```

```
log_rank_result <- survdiff(surv_obj ~ Servicio_Tecnico, data = datos_celulares,
  subset = Servicio_Tecnico %in% c("B", "A"))
log_rank_result
```

```
## Call:
## survdiff(formula = surv_obj ~ Servicio_Tecnico, data = datos_celulares,
##     subset = Servicio_Tecnico %in% c("B", "A"))
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## Servicio_Tecnico=A 61       43     34.3      2.19      3.87
## Servicio_Tecnico=B 71       52     60.7      1.24      3.87
##
##  Chisq= 3.9  on 1 degrees of freedom, p= 0.05
```

Ejercicio 2

Permanencia de clientes (Senior / No Senior) en una compañia telefónica

Una empresa de telecomunicaciones desea predecir el tiempo hasta que un cliente se dé de baja de la compañía. El tiempo que un cliente ha sido parte de esta se encuentra en la variable tenure, y el indicador de sí el cliente ha abandonado está en la variable Churn (1 si el cliente se dio de baja, 0 si sigue activo). Los clientes que no se han dado de baja son ejemplos de datos censurados, ya que desconocemos cuanto tiempo más permanecerán como clientes.

Objetivo y Preguntas

El objetivo de este ejercicio es utilizar el análisis de Kaplan-Meier, para estimar la función de supervivencia del tiempo hasta el abandono de clientes en una empresa de telecomunicaciones y comparar según rango etario.

1. ¿Cúal es la tasa de abandono acumulado?
2. ¿Cuál es la función de supervivencia estimada utilizando el análisis de Kaplan-Meier para el tiempo hasta el abandono de los clientes?
3. ¿Cuál es el tiempo medio de permanencia de los clientes en el servicio antes de hacer el abandono?

Resolución

Importar librerías

```
library(readr) #Librería para leer datos
library(ggplot2) # Para gráficos
library(survival) #Para función de supervivencia
library(survminer)
```

Importar Datos

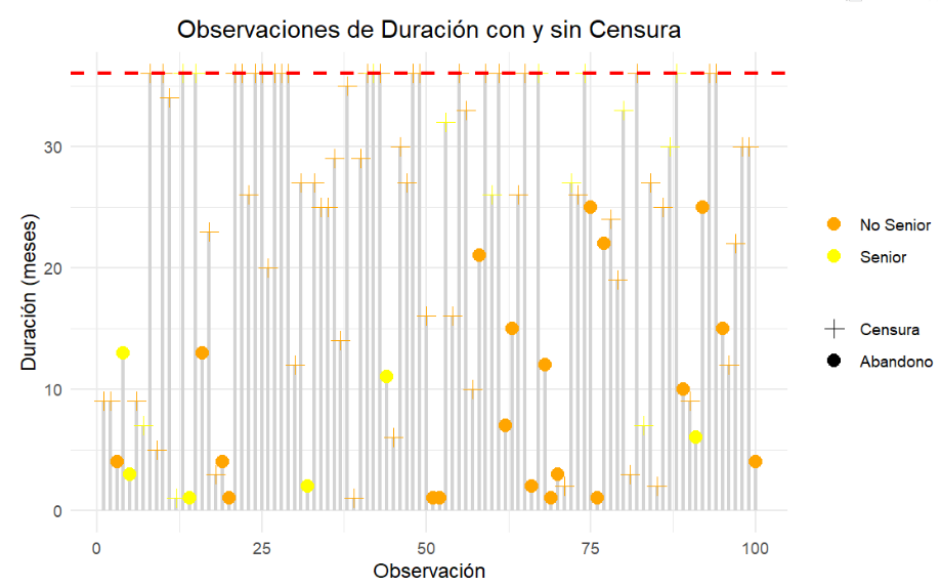
```
# Mostrar los primeros 3 datos del DataFrame
head(Telecomunicaciones, 3)
```

```
## # A tibble: 3 x 21
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
##   <chr>      <chr>          <dbl> <chr>   <chr>       <dbl> <chr>
## 1 0002-ORFBO Female             0 Yes     Yes             9 Yes
## 2 0003-MKNFE Male               0 No      No              9 Yes
## 3 0004-TLHLJ Male               0 No      No              4 Yes
## # i 14 more variables: MultipleLines <chr>, InternetService <chr>,
## #   OnlineSecurity <chr>, OnlineBackup <chr>, DeviceProtection <chr>,
## #   TechSupport <chr>, StreamingTV <chr>, StreamingMovies <chr>,
## #   Contract <chr>, PaperlessBilling <chr>, PaymentMethod <chr>,
## #   MonthlyCharges <dbl>, TotalCharges <dbl>, Churn <chr>
```

Resolución (Censura)

Periodo de estudio (36 meses)

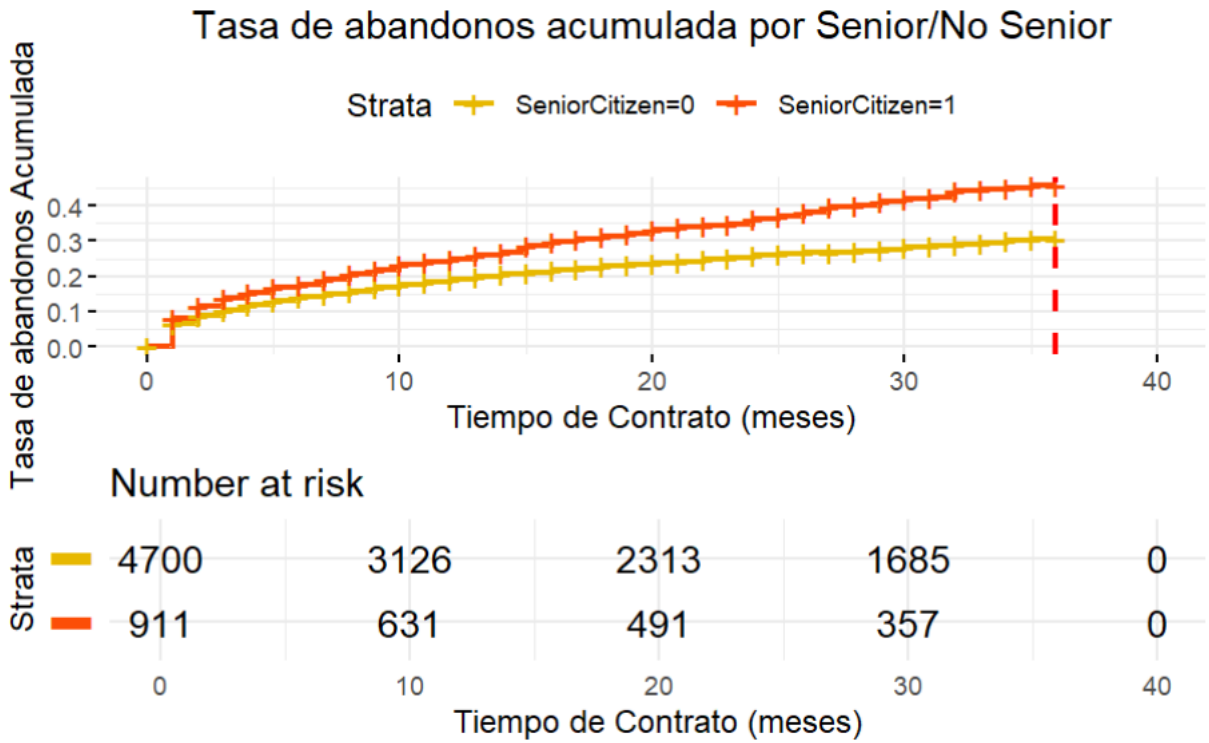
```r
# Cambiamos a variable indicadora
Telecomunicaciones$Churn <- ifelse(Telecomunicaciones$Churn == "No", 0, 1)

# Censurar los datos que exceden los 36 meses
Telecomunicaciones$tenure[Telecomunicaciones$tenure > 36] <- 36
Telecomunicaciones$Churn[Telecomunicaciones$tenure == 36] <- 0
# Censurado si no abandono en 36 meses
```



Resolución (Pregunta 1)

Tasa de fallas acumuladas

```
# Crear un objeto de supervivencia
surv_obj <- Surv(time = Telecomunicaciones$tenure,
                 event = Telecomunicaciones$Churn)

# Ajustar el modelo de supervivencia por categoría de edad
fit <- survfit(surv_obj ~ SeniorCitizen,
               data = Telecomunicaciones)

ggsurv <- ggsurvplot(fit,
           data = Telecomunicaciones,
           fun = "event",
           conf.int = FALSE,
           palette = c("#E7B800", "#FC4E07"),
           ggtheme = theme_minimal(),
           title = "Tasa de abandonos acumulada por Senior/No Senior",
           xlab = "Tiempo de Contrato (meses)",
           ylab = "Tasa de abandonos Acumulada",
           risk.table = TRUE,
           risk.table.height = 0.4,
           # Ajustar la altura de la tabla de riesgo
           tables.y.text = FALSE)
```


Tasa de abandonos acumulada por Senior/No Senior
Strata
SeniorCitizen=0
SeniorCitizen=1
Tasa de abandonos Acumulada
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
Tiempo de Contrato (meses)
Number at risk
Strata
4700
3126
2313
1685
0
911
631
491
357
0
0
10
20
30
40
Tiempo de Contrato (meses)

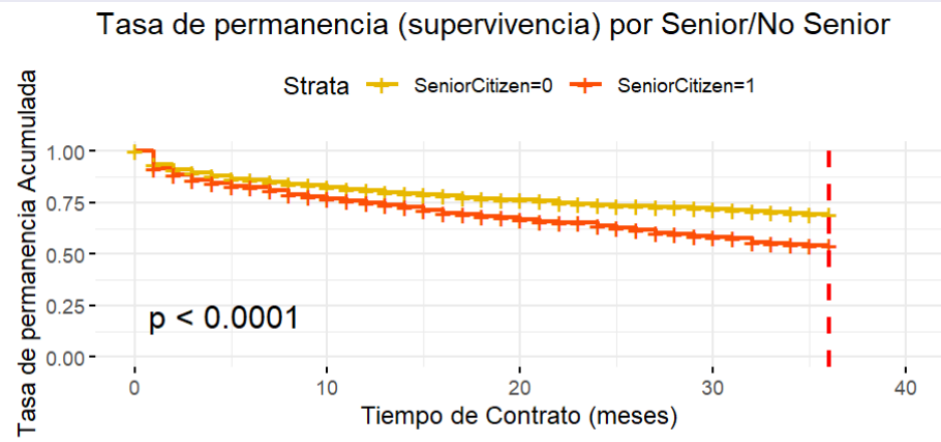
Resolución (Pregunta 2)

Función de Supervivencia

```r
# Objeto de supervivencia
surv_obj <- Surv(time = Telecomunicaciones$tenure,
                 event = Telecomunicaciones$Churn)

# Ajustar
fit <- survfit(surv_obj ~ SeniorCitizen,
               data = Telecomunicaciones)

ggsurv <- ggsurvplot(fit,
                     data = Telecomunicaciones,
                     conf.int = FALSE,
                     ggtheme = theme_minimal(),
                     palette = c("#E7B800", "#FC4E07"),
                     ncensor.plot = TRUE,
                     pval = TRUE,
                     title = "Tasa de permanencia (supervivencia) por Senior/No Senior",
                     xlab = "Tiempo de Contrato (meses)",
                     ylab = "Tasa de permanencia")
```

Resolución (Pregunta 2)

Tasa de Permanencia (Supervivencia)



Resolución (Pregunta 3)

Medias

```
# Calcular la media de permanencia para cada grupo
medias <- summary(fit)$table[, "rmean"]

# Mostrar las medias
print(medias)
```

```
## SeniorCitizen=0 SeniorCitizen=1
##        28.47155        25.44849
```

Ahora, determinamos si hay diferencias significativas en la durabilidades entre servicios tecnicos. Para esto aplicamos una Prueba de Log-Rank, donde las hipotesis son:
H_0: No existe diferencia en la función de supervivencia entre grupo etario.
H_1: Existe al menos una diferencia en la función de supervivencia entre grupo etario.

```
survdiff(formula = surv_obj ~ SeniorCitizen,
         data = Telecomunicaciones)
```

```
## Call:
## survdiff(formula = surv_obj ~ SeniorCitizen, data = Telecomunicaciones)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## SeniorCitizen=0 4700     1141     1249      9.34      57.1
## SeniorCitizen=1  911      360      252     46.28      57.1
##
##  Chisq= 57.1  on 1 degrees of freedom, p= 4e-14
```

Section 6

Conclusiones

Conclusiones

- El análisis de confiabilidad se ve presente en gran parte de los procesos de control estadístico.
- Los datos censurados al estar presentes en procesos de control pueden generar sesgos en el estudio.
- La estimación de Máxima Verosimilitud y el Análisis de Kaplan-Meier ofrecen las herramientas para cualquier tipo de dato.
- Es importante considerar los datos censurados dentro del estudio para la toma de decisiones eficiente.

Section 7

Referencias

Referencias

- Garvin, D.A. (1988): Managing Quality, Nueva York: The Free Press.
- Montgomery, D. C. (2004). Control estadístico de la calidad (3.ª ed.). Limusa Wiley.
- San José, B., Pérez, E., & Madero, R. (2009). Métodos estadísticos en estudios de supervivencia. Anales de Pediatría Continuada, 7(1), 55-59. https://doi.org/10.1016/S1696-2818(09)70453-6

-Moore, D. F. (2016). Applied survival analysis using R. Springer International Publishing.