TRABAJO DE FIN DE GRADO – "Grado en Estadística" – (2018/2023)

FACULTAD DE CIENCIAS

CONSTRUCCIÓN MATEMÁTICA DE ESTIMADORES DE LA FUNCIÓN DE SUPERVIVENCIA

AUTOR/A: Irene del Valle Ramón

TUTOR/A: D.ª María Jesús Rivas López

AÑO DE PRESENTACIÓN: Curso 2022/2023

Salamanca, julio de 2023

CONSTRUCCIÓN MATEMÁTICA DE ESTIMADORES DE LA FUNCIÓN DE SUPERVIVENCIA

AUTOR/A: Irene del Valle Ramón

TUTOR/A: D.ª María Jesús Rivas López

AÑO DE PRESENTACIÓN: Curso 2022/2023

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

Facvltad D Ciencias
VNiVERSiDAD D SALAMANCA

Certificado de los tutores TFG Grado en Estadística

Dña. María Jesús Rivas López, profesora del Departamento de Estadística de la Universidad de Salamanca,

HACE CONSTAR:

Que el trabajo titulado "*CONSTRUCCIÓN MATEMÁTICA DE ESTIMADORES DE LA FUNCIÓN DE SUPERVIVENCIA*", que se presenta, ha sido realizado por Dña. Irene del Valle Ramón, con DNI 44442704-B y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, a fecha de firma electrónica.

RIVAS LOPEZ MARIA JESUS - 07956354X

Firmado digitalmente por RIVAS LOPEZ MARIA JESUS - 07956354X
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-07956354X, givenName=MARIA JESUS, sn=RIVAS LOPEZ, cn=RIVAS LOPEZ MARIA JESUS - 07956354X
Fecha: 2023.07.03 18:56:23 +02'00'

Fdo.: María Jesús Rivas López

-ÍNDICE

-ÍNDICE DE GRÁFICOS

-ÍNDICE DE TABLAS

-ÍNDICE DE FIGURAS

1 – INTRODUCCIÓN

La **Estadística** está muy presente en diversos aspectos de la vida diaria, ya que para obtener información se hace uso de diferentes análisis estadísticos. De ahí, que Cabriá (1994) defina la Estadística como el **conjunto de métodos, procedimientos y fórmulas** que permiten obtener información para poder analizarla y lograr conclusiones relevantes. Se trata de la Ciencia de los datos, cuyo principal fin es mejorar la compresión de una serie de hechos a partir de la información recopilada.

Por tanto, este documento se centra en el **Análisis de supervivencia**, ya que se trata de una disciplina estadística antigua con raíces en la demografía y la ciencia del siglo XVII.

En la actualidad, en muchos campos de estudio es importante conocer el momento en el que ocurre un evento que se pueda considerar de interés, es decir, el tiempo que tarda en suceder ese evento. Por ejemplo, desde un punto de vista médico, puede ser justamente el momento del fallecimiento, pero esta situación no se limita únicamente a la vida o la muerte, ya que los eventos o situaciones estudiadas se clasifican en: beneficiosas (tolerancia a un medicamento, alta hospitalaria, etc); perjudiciales (muerte, recaída, etc); comparativas (cambio de tratamiento); etc.

Y el tiempo que transcurre hasta que ocurre un evento se denomina de diferentes formas dependiendo del área estudiada, **tiempo de fallo** para el área industrial o **tiempo de supervivencia** o tiempo de muerte para el área de Ciencias de la salud.

Este análisis necesita conocer las variables, “tiempo hasta el evento de interés” y otra variable que “indique si ha ocurrido o no ese evento de interés”, por lo que esta situación, tiene que estar muy bien definida por el investigador.

El **Análisis de supervivencia** es por tanto una forma de estudiar el tiempo de fallo utilizando información sobre hechos que han sucedido con anterioridad en aspectos similares (Chiang, 2001); y ,para poder llevar a cabo este proceso, se necesitan ciertas habilidades estadísticas.

Se trabaja con una variable aleatoria positiva, es decir, que no sigue una distribución normal. De ahí, que sea más conveniente trabajar con distribuciones que tengan en cuenta esta característica de positividad. Además, con el objetivo de incorporar la información que proporcionan a los individuos con tiempos censurados, hay que utilizar métodos desarrollados para este fin.

Pero al igual que ocurre en otras áreas de la Estadística, los métodos utilizados para poder llevar a cabo el análisis de supervivencia son tanto paramétricos como no paramétricos, es decir, no suponen de manera sistemática una distribución sobre los datos observados. De ahí, que se lleve a cabo la construcción matemática de estimadores no paramétricos de la función de supervivencia a través del **Método de Kaplan-Meier (K-M)**, que utiliza los tiempos de observación, tanto censurados como no censurados, introducido por Edward L. Kaplan y Paul Meier en 1958. Dichos autores, formalizaron la influencia de entradas y abandonos en distintos momentos del estudio, a través del ajuste del conjunto de individuos en riesgo, e individuos bajo observación en un determinado tiempo. También, se demuestra matemáticamente que el estimador Kaplan-Meier es el **estimador máximo verosímil** que es el principal objetivo de este trabajo.

De esta manera, se puede estimar la **curva de supervivencia** (probabilidad de que un evento ocurra después de un tiempo dado), ya que representa la probabilidad de supervivencia acumulada frente al tiempo. Se trata de una forma correcta de resumir los datos de manera visual y que además permite estimar la mediana del tiempo de supervivencia.

En la práctica, también se utiliza el **estimador de Nelson Aalen (NA)**, introducido por Nelson (1972) y Aalen (1978), porque es más eficaz con muestras de pequeño tamaño. Y puede demostrarse que los estimadores K-M y NA son asintóticamente equivalentes, es decir, se logra el mismo resultado al sustituir en el límite una función por otra.

Los **paquetes estadísticos** que ayudan a elaborar estas curvas de supervivencia necesitan diferentes datos: el tiempo hasta un evento; el número de individuos que siguen expuestos al suceso tras finalizar el estudio; el número de eventos; y la cantidad de individuos. Y a partir de ellos, devuelven: la función de supervivencia; la función de riesgo; e incluso una representación gráfica de las curvas.

En este proceso es importante considerar contrastes de hipótesis que comparan unas curvas con otras.

Por tanto, para poder llevar a cabo el manejo de toda información asociada a un estudio de supervivencia y realizar su análisis, es necesario utilizar un **paquete estadístico** fiable, en este caso, se utiliza **RStudio**.

Finalmente, y con el objetivo de avanzar en el conocimiento de este análisis y conocer su potencial, se presenta un **documento** organizado en **ocho capítulos**, donde se detalla el estudio de supervivencia de un individuo, dependiendo de sus diferentes funciones y estimadores. De esta manera, se **comienza** con una *"Introducción"* donde se aclaran los conceptos más importantes sobre los que gira el estudio, y se explican las directrices de actuación para llevarlo a cabo, así como las referencias teóricas sobre las que se apoya. Posteriormente, en un **segundo capítulo**, *"Análisis de supervivencia"*, se enmarcan los conceptos y las bases principales para realizar este análisis de supervivencia, en referencia a las censuras y las funciones principales. Se continúa en el **capítulo tres**, *"Modelos paramétricos y no paramétricos"*, donde se detallan los métodos estadísticos que más se utilizan en los modelos no paramétricos (La función de supervivencia empírica, la tabla actuarial de supervivencia, el estimador de Kaplan-Meier y Nelson Aalen), así como los modelos paramétricos (El modelo Exponencial, Weibull y Log-normal). Seguidamente, en el **capítulo cuatro**, *"Estimadores de la función de supervivencia"*, se explica detalladamente la forma en que se puede estimar la función de supervivencia, mediante el estimador Kaplan-Meier, estimador principal para este análisis, y el estimador de Nelson Aalen. También se demuestra que el estimador K-M es el estimador máximo verosímil. En **quinto lugar**, *"Comparación de la supervivencia",* se comparan las funciones de supervivencia relacionadas con el análisis de supervivencia, donde se detallan las dos pruebas estadísticas principales para comparar la supervivencia de dos o más grupos. En el **sexto capítulo**, *"Aplicación del análisis de supervivencia",* se muestra el análisis práctico del estimador Kaplan-Meier que se compara con el estimador de Nelson Aalen para ver cuál de los dos estimadores es más efectivo, aunque se sabe que ambos resultados son equivalentes. En el **capítulo siete**, *"Conclusiones"*, se indican los diferentes aspectos finales que se obtienen tras la realización de este trabajo. Por último y como **octavo capítulo**, *"Referencias bibliográficas"*, se indican las diferentes y múltiples fuentes bibliográficas y webgráficas empleadas para la realización, apoyo y comprobación de este trabajo.

2 - ANÁLISIS DE SUPERVIVENCIA

El **Análisis de supervivencia** es una de las técnicas que más han evolucionado con el tiempo y sobre todo durante estos últimos años dentro del campo de la Estadística. Su objetivo es estudiar el tiempo hasta que ocurre un evento, y a ese tiempo se le denomina **tiempo de supervivencia o tiempo de fallo**. El evento tiene que estar muy bien definido, para poder determinar exactamente la fecha de ocurrencia. Este evento, en el ámbito sanitario puede estar asociado a la muerte de un paciente, a su alta hospitalaria, a la remisión de la enfermedad, etc.

El **tiempo de supervivencia** o de fallo es el tiempo que pasa desde que se inicia un estudio hasta la ocurrencia de un evento. Y para poder determinar el tiempo de fallo, se deben cumplir **tres exigencias**, un tiempo de origen bien definido, una escala temporal y un evento bien determinado. Cabe destacar que para los tiempos de fallo deben ser No negativos.

2.1 – CENSURA

En el Análisis de supervivencia se pueden encontrar una serie de **censuras**, es decir, tiempos en los que el evento de interés no es observado en algunos de los individuos; bien porque los estudios que se terminaron antes de que ocurriera el evento, o porque el individuo decide abandonar el estudio, o por otras causas que no están relacionadas con la investigación.

Dentro del Análisis de supervivencia se pueden encontrar censuras de diferentes tipos, censura por la derecha, censura por la izquierda y censura por intervalos.

- Censura por la derecha: Este tipo de censura es el más común, ya que el tiempo hasta el evento, T, se produce cuando es mayor que el tiempo de censura observado, C, es decir, el evento no se observa hasta que finalice el estudio. Esta censura tiene una característica principal, ya que el tiempo de censura observado tiene que ser menor al tiempo de supervivencia o fallo, lo que se da principalmente en los estudios biomédicos. Por tanto, estos datos se representan por dos variables aleatorias (X, δ):

$$X = \min(T, C), \quad \delta = \begin{cases} 1 \; si \; T \leq C \rightarrow ocurre \; el \; evento \\ 0 \; si \; T > C \rightarrow observación \; de \; la \; censura \end{cases}$$

  La censura por la derecha se divide en los siguientes tipos:

  - ***Censura tipo I***: En la mayoría de los estudios los investigadores predeterminan el tiempo máximo de duración de la investigación. Por ello, los pacientes que no experimenten el evento tienen un tiempo de análisis desconocido y por tanto censurado.

    Es importante destacar que los tiempos de censura pueden variar entre pacientes según se vaya desarrollando el estudio. Por ejemplo, este tipo de censuras es muy común en los ensayos clínicos, ya que se realiza con humanos o animales y se comienza con un número fijo de pacientes a los que se les administra el tratamiento. Pero debido al tiempo de duración, los investigadores suelen terminar el estudio en un tiempo determinado en el que no todos los pacientes experimentan el evento de interés.

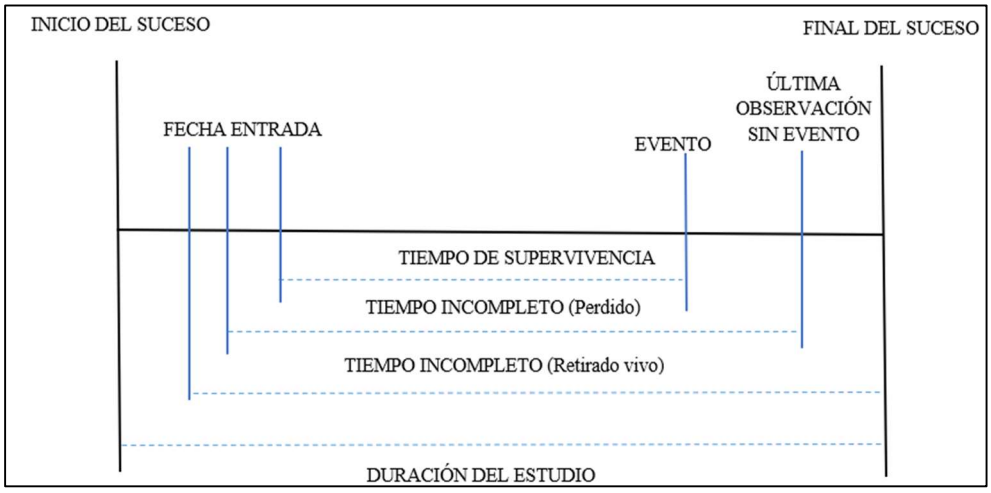
- ***Censura tipo II***: En este tipo de censura, a diferencia del anterior, los investigadores deciden prolongar ese tiempo de observación en los pacientes, hasta que ocurran k fallos de n posibles (siendo n el número de pacientes, $(k \leq n)$). Los pacientes que no experimentan el evento antes de experimentar los primeros k fallos, son los que representan las observaciones censuradas. Y en esta situación, la censura está controlada por el investigador. Por ejemplo, para comprobar la duración que puede llegar a tener un equipo electrónico, es necesario hacer funcionar todos los aparatos al mismo tiempo. La prueba finaliza cuando los primeros k de los n aparatos fallan.
- ***Censura aleatoria***: Se trata de la tipología de censura más diferente, ya que ocurre sin ningún control por parte de los investigadores. Las censuras en este apartado se presentan por abandono del estudio del paciente, lo que significa la pérdida exacta de la fecha en la que ocurre el evento o por la muerte del paciente que se pueda dar por alguna causa que no esté relacionada con el evento final de interés. Por ejemplo, en un ensayo clínico, los pacientes entran en estudio en distintos momentos y cada uno puede llegar a tomar tratamientos distintos, por eso, la censura puede tomar distintas causas, ya sea la muerte accidental o por otra causa distinta al evento.

- Censura por la izquierda: Dentro de este tipo, las censuras ocurren antes de que los pacientes entren al estudio, es decir, los investigadores saben que los pacientes ya han presentado el evento, pero no en el momento exacto de su aparición.
  Por ejemplo, un paciente que haya padecido un infarto, el investigador no sabe cuándo fue el momento exacto en el que se originó, pero con las diferentes pruebas a las que somete al paciente lo acaban concluyendo.

- Censura por intervalo: En este tipo, el investigador no conoce el tiempo exacto del evento de interés, pero si sabe que se produce en un intervalo determinado.
  Por ejemplo, un paciente que sufra de migraña, el investigador no sabe el momento exacto de su desaparición, pero lo que conoce, es que dura entre unas horas o días.

Y de esta manera, en el siguiente esquema, se puede comprobar el **desarrollo** de un estudio de **análisis de supervivencia**.

Figura 1: *Esquema del estudio de un Análisis de supervivencia (Fernández, 1995)*



2.2 – FUNCIONES RELACIONADAS PARA EL ANÁLISIS DE SUPERVIVENCIA

Para poder establecer el **modelo de supervivencia** es necesario saber cómo describir de manera efectiva la **variable aleatoria**, tiempo hasta un evento.

Y para ello, en este apartado, se detalla exhaustivamente todas las **funciones** que están relacionadas con el análisis de supervivencia, donde la función de densidad y supervivencia son las más comunes.

2.2.1 – FUNCIÓN DE SUPERVIVENCIA

La **función de supervivencia** a tiempo t, se define como la probabilidad de que un individuo no experimente el evento de interés antes de un tiempo determinado t. Aunque una definición más matemática es la siguiente.

Definición 1. Sea T una variable aleatoria positiva (no negativa) con función de distribución $F(t)$ y función de densidad $f(t)$. La función de supervivencia $S(t)$ es:

$$S(t) = 1 - F(t) = P(T > t) = \int_t^{\infty} f\,(u)\,du. \rightarrow Caso\ en\ el\ que\ T\ sea\ Continua.$$

$$S(t) = 1 - F(t) = P(T > t) = \sum_{tj>t} f(tj). \rightarrow Caso\ en\ el\ que\ T\ sea\ Discreta$$

(con soporte en $t_1, \dots, t_n$)

Se verifica que, $S(t)$ es una función no creciente, tal que:

$$S(0) = 1 \quad y \quad S(t) = 0 \quad cuando\ t \rightarrow \infty.$$

Lo que se deduce de que la probabilidad de sobrevivir al inicio del estudio es uno y la probabilidad de sobrevivir a un tiempo infinito es nula.

2.2.2 – FUNCIÓN DE RIESGO (HAZARD FUNCTION)

La **función de riesgo Hazard Function**, $h(t)$, es la función más adecuada para describir la dinámica de un proceso de supervivencia.

$$h(t) = P(T = t | T \geq t)$$

Para el **caso continuo**, esta función determina la probabilidad de que a un paciente le ocurra un evento de interés en un incremento de tiempo Δt, dado a que ha sobrevivido hasta el tiempo t. Esta función se define como:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t}$$

Aplicando la definición de **probabilidad condicionada** a la igualdad de la función de riesgo se tiene que:

$$h(t) = \lim_{\Delta t \to 0} \frac{P((t < T \leq t + \Delta t) \cap (T \geq t))/P(T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{P(T \geq t)\Delta t}$$

En este caso:

$$P(t < T \leq t + \Delta t) = \int_t^{t+\Delta t} f(u)\, du = F(t + \Delta t) - F(t)$$

Al sustituir lo anterior en la aplicación de la probabilidad condicionada $h(t)$ y aplicando la definición de derivada, $F'(x) = \lim_{h \to 0} \frac{F(x+h)-F(x)}{h}$, se obtiene:

$$h(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{P(T \geq t)} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}$$

Para el **caso discreto**, donde la variable T está soportada en ciertos tiempos $t_1, t_2, \ldots, t_n$ la función de riesgo es la determina la probabilidad condicional de que el evento ocurra a tiempo t_j, condicionada a que el sujeto permanece vivo antes de dicho tiempo,

$$h(t_j) = \frac{f(t_j)}{S(t_{j-1})}$$

2.2.3 – FUNCIÓN DE RIESGO ACUMULADO

La **función de riesgo acumulado** $H(t)$, se define como:

- Caso continuo: será la integral

$$H(t) = \int_0^t h(u)\, du$$

- Caso discreto: Si T es discreta con valores $t_1 < t_2 < \cdots < t_n$,:

$$H(t) = \sum_{t_j \leq t} h(t_j)$$

Estas tres funciones tienen **relaciones entre sí** que, dependiendo de si están en caso **continuo** o **discreto**, se desarrollan a continuación.

- Caso continuo:

  La relación de la función de supervivencia, $S(t)$, con la función de densidad viene dada por:

$$S(t) = \int_t^{\infty} f(u)\, du \rightarrow f(t) = -\frac{dS(t)}{dt}$$

La **relación** que $S(t)$ tiene con la **función de riesgo**, $h(t)$, es:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t | T > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t \cdot \mathbb{P}(T > t)}$$

$$= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = -\lim_{\Delta t \to 0} \frac{1 - F(t + \Delta t) - \big(1 - F(t)\big)}{\Delta t} \frac{1}{S(t)}$$

$$= -\lim_{\Delta t \to 0} \frac{S(t + \Delta t) - S(t)}{\Delta t} \frac{1}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{dlnS(t)}{dt}$$

Partiendo de lo anterior, la función de riesgo, $h(t)$, en función de la función de supervivencia es:

$$h(t) = -\frac{dlogS(t)}{dt} \rightarrow S(t) = e^{-\int_0^t h(u)\, du}$$

Entonces, se obtienen las relaciones entre la función de riesgo y la **función de densidad**, $f(t)$:

$$h(t) = \frac{f(t)}{S(t)} \rightarrow f(t) = h(t) \cdot S(t) \rightarrow f(t) = h(t) \cdot e^{-\int_0^t h(u)\, du}$$

Finalmente, para la **función de riesgo acumulado**, $H(t)$, se tiene que:

$$H(t) = \int_0^t h(u)\, du = \int_0^t \frac{f(u)}{S(u)} = \int_0^t \frac{-S'(u)}{S(u)}\, du = -\int_0^t \frac{d}{du} (\ln\big(S(u)\big) = -\ln S(t)$$

Por tanto, la función de supervivencia puede expresarse como:

$$S(t) = \exp[-H(t)] = e^{-H(t)}$$

Y, por definición, la relación existente entre la función de riesgo y la función de riesgo acumulado es:

$$h(t) = -\frac{dH(t)}{dt}$$

- <u>Caso discreto</u>: La relación de la **función de supervivencia**, $S(t)$,con el resto de las funciones, viene dada por:

$$S(t) = \mathbb{P}(T \geq t)) = \sum_{t_j \geq t} f\big(t_j\big) \rightarrow f\big(t_j\big) = S(t_{j-1}) - S(t_j)$$

Veamos ahora la relación que tiene la función de supervivencia, $S(t)$, con la **función de riesgo**, $h(t)$:

$$h\big(t_j\big) = \frac{f(t_j)}{S(t_{j-1})} = \frac{S\big(t_{j-1}\big) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}$$

Se verifica, por tanto que:

$$S(t_j) = [1 - h(t_j)]S(t_{j-1})$$

Desarrollando los términos:

$$S(t_1) = [1 - h(t_1)]S(t_{1-1}) = [1 - h(t_1)]S(t_0) = 1 - h(t_1)$$

$$S(t_2) = [1 - h(t_2)]S(t_{2-1}) = [1 - h(t_2)]S(t_1) = [1 - h(t_2)][1 - h(t_1)]$$

$$S(t_3) = [1 - h(t_3)]S(t_{3-1}) = [1 - h(t_3)]S(t_2) = [1 - h(t_3)][1 - h(t_2)][1 - h(t_1)]$$

$$\vdots$$

Y se obtiene que para todo tiempo t:

$$S(t) = \prod_{t_j \le t} [1 - h(t_j)]$$

Por otra parte, para la **función de probabilidad**, $f(t)$, se tiene:

$$h(t_j) = \frac{f(t_j)}{S(t_{j-1})} \rightarrow f(t_j) = h(t_j) \cdot S(t_{j-1})$$

$$\rightarrow f(t_j) = h(t_j) \prod_{k=1}^{j-1} [1 - h(t_k)] = \frac{h(t_j)}{1 - h(t_j)} \prod_{k=1}^{j} [1 - h(t_k)]$$

En tercer lugar, se obtiene una relación entre la función de supervivencia y la **función de riesgo acumulado**, $H(t)$:

$$H(t) = \sum_{t_j \le t} h(t_j) = \sum_{t_j \le t} \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})}$$

3 – MODELOS PARAMÉTRICOS Y NO PARAMÉTRICOS

Para estudiar la supervivencia, se necesitan **modelos específicos**, ya que la variable no sigue una distribución normal y tienen censura, por lo que en el desarrollo del estudio se separa de los modelos de regresión clásicos.

Por ello, el análisis de datos de supervivencia se realiza utilizando **modelos paramétricos**, como la Distribución Exponencial, Distribución Weibull y la Distribución Log-normal que se ven de una forma más globalizada en el siguiente apartado (3.1), y también utilizando **modelos no paramétricos**, como el método de Kaplan-Meier, entre otros.

Para realizar el análisis de supervivencia, los métodos estadísticos que más se utilizan son los **modelos no paramétricos**.

De esta manera, las curvas de supervivencia se obtienen usando **dos métodos**, el análisis actuarial, que divide el tiempo en intervalos y calcula la supervivencia en cada intervalo. Y el método de Kaplan-Meier, que tiene en cuenta los posibles abandonos del estudio. La diferencia de ambos métodos es que el método de Kaplan-Meier proporciona probabilidades, mientras que el análisis actuarial facilita aproximaciones.

Dada la importancia que tiene la función de supervivencia dentro de los modelos paramétricos, se proponen las siguientes estimaciones.

- Función de supervivencia empírica: Si tenemos n individuos en estudio, se define como:

$$S_n(t_j) = \frac{\#\{T \geq t_j\}}{n}, siendo\ \#\ el\ cardinal.$$

  Se considera como un buen estimador, en el caso de no tener censura, para la función de distribución, puesto que:

$$\sup_{t \in \mathbb{R}} |S_n(t) - S(t)| \to 0, cuando\ n \to \infty$$

  La función de supervivencia empírica es una función escalonada decreciente con saltos cuya longitud es igual a $\frac{1}{n}$, solamente después de cada tiempo de supervivencia.

  En el caso de tener d tiempos de evento iguales a cierto tiempo, la función de supervivencia empírica decrece a ese tiempo con salto igual a $\frac{d}{n}$.

  Si t_1 es la observación más pequeña y t_n la observación más grande, se verifica que:

$$S_n(t) = 1\ \ para\ t < t_1$$

$$S_n(t) = 0\ \ para\ \ t > t_n$$

  Y aunque, se considera como un buen estimador, se desperdicia mucha información ya que no tiene en cuenta si el tiempo, t_j, es una censura o un tiempo de fallo.

- Tabla actuarial de vida o "tabla actuarial de supervivencia": Se trata de una extensión de la tabla de frecuencias relativas.

  Para poder obtener la **función de supervivencia** asociada a esta tabla se siguen los siguientes **apartados**:

  1. El tiempo de observación se divide en k intervalos:

  $$0 = a_0 < a_1 < a_2 < \cdots < a_k = \infty$$

  2. Se busca que la probabilidad de sobrevivir o fallar en el periodo $[a_{j-1}, a_j)$, y a partir de la **regla del producto**:

  $$S(a_j) = \mathbb{P}[T \geq a_0] \cdot \mathbb{P}[T \geq a_1 | T \geq a_0] \cdot \mathbb{P}[T \geq a_2 | T \geq a_1] \cdot \ldots$$

  $$\ldots \cdot \mathbb{P}\left[T \geq a_j \middle| T \geq a_{j-1}\right]$$

  Por ejemplo, en el caso de hacerlo **para a_2**, a partir de la definición anterior quedaría:

  $$S(a_2) = \mathbb{P}(T \geq a_0 \cap T \geq a_1 \cap T \geq a_2) =$$

  $$= \mathbb{P}(T \geq a_0) \cdot \mathbb{P}(T \geq a_1 | T \geq a_0) \cdot \mathbb{P}(T \geq a_2 | T \geq a_1)$$

  3. Se busca que la probabilidad de sobrevivir o fallar en el periodo $[a_{j-1}, a_j)$, son sucesos complementarios, por tanto:

  $$\mathbb{P}\left[T \geq a_j \middle| T \geq a_{j-1}\right] = 1 - \mathbb{P}\left[T < a_j \middle| T \geq a_{j-1}\right]$$

  Probabilidad de sobrevivir a $[a_{j-1}, a_j)$ — Probabilidad de fallo en $[a_{j-1}, a_j)$

  4. Para calcular ambas probabilidades se debe tener en cuenta si existe o no censuras:

     - Sin censura

Todos los individuos en riesgo podrían experimentar el evento en el intervalo $[a_{j-1}, a_j)$.

$d_j \rightarrow$ es el número de individuos que mueren en el intervalo.

$n_j \rightarrow$ son los individuos en riesgo (es decir, susceptibles de experimentar el fallo) al inicio del intervalo.

Entonces, se tiene la **probabilidad de fallo en el intervalo**,

$$\mathbb{P}\left[T < a_j \middle| T \geq a_{j-1}\right] = \frac{d_j}{n_j}$$

Por el contrario, se obtiene la **probabilidad de sobrevivir en ese intervalo**, que se calcula de la misma manera que en el caso anterior, tomando el complementario de la probabilidad de fallo:

$$\mathbb{P}\left[T < a_j \middle| T \geq a_{j-1}\right] = 1 - \frac{d_j}{n_j}$$

Por tanto, la **función de supervivencia** se define como la probabilidad de sobrevivir a un conjunto de intervalos consecutivos:

$$S(a_j) = \prod_{i=1}^{j} \left(1 - \frac{d_j}{n_j}\right)$$

- Con censura

Se supone que la censura se distribuye de manera uniforme en el intervalo, por lo que se supone que los individuos se censuran a la mitad del intervalo, $[a_{j-1}, a_j]$,

$d_j, n_j \rightarrow$ como en el caso anterior

$c_j \rightarrow$ son los individuos censurados en ese intervalo.

De esta manera, la **probabilidad de fallo en el intervalo,** $[a_{j-1}, a_j)$, en el caso de tener censuras es:

$$\mathbb{P}\left[T < a_j \middle| T \geq a_{j-1}\right] = \frac{d_j}{n_j - \frac{c_j}{2}}$$

Donde se considera que el número real de individuos en riesgo es

$$n_j' = n_j - \frac{c_j}{2}$$

Por el contrario, se tiene la **probabilidad de sobrevivir en ese intervalo**, que es unos menos la probabilidad de fallo, es decir:

$$\mathbb{P}\left[T < a_j \middle| T \geq a_{j-1}\right] = 1 - \frac{d_j}{n_j - \frac{c_j}{2}}$$

Por tanto, la **función de supervivencia** se define como:

$$S(a_j) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i - \frac{c_i}{2}}\right)$$

Si consideramos un valor cualquiera de tiempo t, la función de supervivencia será:

$$S(t) = S(a_{j-1}), \qquad si\ t \in [a_{j-1}, a_j)$$

La **tabla de vida** se utiliza para calcular la probabilidad de supervivencia de un individuo hasta un momento determinado por el investigador. Para crear la tabla de vida, se debe saber el número y el tamaño de los intervalos estudiados, ya que siguen presentes las mismas dificultades que aparecen para crear un histograma.

A modo de resumen, se presentan todos los datos de manera clara y ordenada en la siguiente tabla:

Tabla 1: *Resumen de la tabla actuarial de vida.*

TABLA DE VIDA				
Inter.	d_j	c_j	n_j	n'_j
$[a_{j-1}, a_{j-2}]$	Mueren	Censurados	En riesgo	$n_j' - \frac{c_j}{2}$
TABLA DE VIDA				
$\widehat{S}(a_{j-1})$		$\hat{q}_j$	$\hat{p}_j$	
$\prod_{i=1}^{j-1}\left(1 - \frac{d_i}{n_j - \frac{c_i}{2}}\right)$		$\frac{d_j}{n'_j}$	$1 - \frac{d_j}{n'_j}$	

- El estimado Kaplan-Meier: Es el estimado por el método de máxima verosimilitud, lo cual se aprecia de una forma más clara y desarrollada en el apartado 4.1.

- El estimador de Nelson Aalen: Se usa principalmente para analizar datos censurados, donde se desconoce el tiempo exacto de ocurrencia de un evento para ciertas observaciones. En el apartado 4.2 aparece desarrollado.

3.1 – MODELOS PARAMÉTRICOS MÁS UTILIZADOS EN SUPERVIVENCIA

Se consideran los **tres modelos paramétricos** más utilizados: Modelo Exponencial, Modelo Weibull y Modelo Log-normal. Vamos a hallar todas las funciones relacionadas con estas tres distribuciones de probabilidad.

3.1.1 – MODELO EXPONENCIAL

Este es el modelo más importante para poder analizar los datos del tiempo de fallo o supervivencia. Una variable exponencial se utiliza para modelar el tiempo que pasa entre dos sucesos aleatorios no muy frecuentes, es decir, de Poisson, con tasa λ

constante, por lo que el riesgo de que suceda ese evento no se modifica con el paso del tiempo.

Y así, se define la **función de riesgo** como: $h(t) = \lambda$

A partir de esta función, $h(t) = \lambda$, se obtiene el resto de las funciones para el análisis de supervivencia:

- Función de riesgo acumulado:

$$H(t) = \int_0^t h(t)\, dt = \int_0^t \lambda\, dt = \lambda t$$

- Función de supervivencia: Es fácil de demostrar al conocer el valor de $H(t)$:

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(t)\, dt} = \mathrm{e}^{-\lambda t}$$

  De forma gráfica, el $\log(t)$, describe una línea recta con pendiente positiva que pasa por el origen $(0,0)$.

- Función de densidad de una variable exponencial: Su fórmula es sencilla de ver, ya que es la multiplicación de $h(t)$ por $S(t)$, valores obtenidos en los puntos anteriores:

$$f(t) = h(t)S(t) = \lambda e^{-\lambda t}$$

- Función de distribución: Se obtiene a partir de la función de supervivencia, $S(t)$:

$$F(t) = 1 - S(t) = 1 - e^{-\lambda}$$

Por tanto, la esperanza y varianza de una variable de esta distribución se calculan como:

$$E[T] = \int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}$$

$$Var[T] = \mathbb{E}[T^2] - \mathbb{E}[T]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

3.1.2 – MODELO WEIBULL

Este modelo, es una extensión del modelo anterior (Modelo Exponencial), con mismo parámetro de escala, λ, pero con la diferencia, que se añade un nuevo parámetro de forma, γ. Por lo que este modelo tiene por dos parámetros positivos, $\lambda\, y\, \gamma$.

El Modelo Weibull, es el más utilizado para calcular los tiempos de supervivencia en el caso de que el riesgo no sea una función constante, sino que es creciente o decreciente.

De esta manera, se define la **función de riesgo** como: $h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$

- Para $\gamma > 0, \lambda > 0\, y\, t > 0$, si se observa que esta función es creciente o decreciente, se puede obtener este modelo. Es decir:

$$Si\ \gamma > 1 \rightarrow h(t) \text{ es una función creciente.}$$

$$Si\ \gamma < 1 \rightarrow h(t) \text{ es una función decreciente.}$$

Cuando $\gamma = 1$, **coincide** con el **modelo exponencial** y, por tanto, la función de riesgo será constante.

A partir de esta función de riesgo, se obtiene el resto de las funciones útiles para el análisis de supervivencia:

- Función de riesgo acumulado:

$$H(t) = \int_0^t h(u)\, du = \int_0^t \lambda\gamma(\lambda u)^{\gamma-1}\, du =$$

$$= \lambda\gamma \int_0^t (\lambda u)^{\gamma-1} du = \lambda\gamma \int_0^t \lambda^{\gamma-1} u\, du = \lambda\gamma\lambda^{\gamma-1} \int_0^t u^{\gamma-1}\, du = \lambda^{\gamma}\gamma \left[\frac{u^{\gamma}}{\gamma}\right]_0^t = (\lambda t)^{\gamma}$$

- Función de supervivencia: Es fácil encontrar esta función al conocer el valor obtenido en el punto anterior, $H(t)$:

$$S(t) = e^{-H(t)} = e^{-\int_0^t \lambda\gamma(\lambda t)} = e^{-(\lambda t)^{\gamma}}$$

De forma gráfica, el $\log(t)$ describe una curva que pasa por el origen.

- Función de distribución: Se obtiene gracias a saber el valor de la función de supervivencia, $S(t)$:

$$F(t) = 1 - S(t) = 1 - e^{-H(t)} = 1 - e^{-(\lambda t)^{\gamma}}$$

- Función de densidad de una variable Weibull: Se obtiene al realizar la derivada de la función de distribución, $F'(t)$:

$$f(t) = F'(t) = -e^{-(\lambda t)^{\gamma}} \cdot (-\lambda\gamma(\lambda t)^{\gamma-1}) = \lambda\gamma(\lambda t)^{\gamma-1} \cdot e^{-(\lambda t)^{\gamma}}$$

Si $T \sim Weibull\ (\gamma, \lambda)$. Su esperanza y varianza son:

$$\mathbb{E}[T] = \frac{1}{\lambda}\Gamma\left(\frac{1}{\gamma} + 1\right)$$

$$Var[T] = \frac{1}{\lambda^2}\Gamma\left(\frac{2}{\gamma} + 1\right) - \left(\frac{1}{\lambda}\Gamma\left(\frac{1}{\gamma} + 1\right)\right)^2 = \frac{1}{\lambda^2}\left[\Gamma\left(\frac{2}{\gamma} + 1\right) - \left(\Gamma\left(\frac{1}{\gamma} + 1\right)\right)^2\right],$$

Donde $\Gamma(t)$ es la función **gamma de Euler** para $t > 0$,

$$\Gamma(t) = \int_0^{\infty} y^{t-1} e^{-y} dy$$

3.1.3 – MODELO LOG-NORMAL

Este modelo se relaciona con la distribución normal, ya que el tiempo de supervivencia T sigue una distribución Log-normal cuando:

$$Ln(T) \rightarrow se\ distribuye\ como\ N(\mu, \sigma^2)$$

El Modelo Log-normal parte de la **función de densidad**, definida como:

$$f(t) = \frac{1}{\sqrt{2\pi\sigma t}} \mathrm{e}^{\left(-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right)}$$

Donde, $\mu\ y\ \sigma$ son parámetros de escala.

A partir de esta función de densidad se obtienen el resto de **funciones** para el análisis de supervivencia:

- Función de supervivencia: estaría definida a partir de la función de densidad, $f(t)$, como:

$$S(t) = \int_t^{\infty} f(t)dt = 1 - \int_0^t f(t)dt = 1 - \int_0^t \frac{1}{\sqrt{2\pi\sigma t}} \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2} dt$$

  Utilizando la notación de la función de distribución de una normal estándar, $\Phi()$, se obtiene la función de supervivencia:

$$S(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right)$$

  Gráficamente, se puede observar de forma aproximada una línea recta con pendiente negativa que pasa por el punto de origen.

- Función de riesgo: Se obtiene al conocer el valor de la función de densidad, $f(t)$ y la función de supervivencia, $S(t)$:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\sqrt{2\pi\sigma t}} \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2}}{1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right)}$$

  Si se representa, toma el valor cero cuando, $t = 0$, crece hasta su máximo valor y decrece de forma que tiende a cero si $t \rightarrow \infty$.

  Si $T \sim Log - normal\ (\mu, \sigma^2)$ su esperanza y varianza son:

$$\mathbb{E}[T] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$Var[T] = \exp(2\mu + 2\sigma^2) - \exp\left(2\mu + \sigma^2\right)$$

4 – ESTIMADORES DE LA FUNCIÓN DE SUPERVIVENCIA

El **estimador principal** de la función de supervivencia es el **estimador Kaplan-Meier**, el cual se compara con el estimador de Nelson Aalen, para comprobar su eficiencia, dependiendo del tamaño de la muestra.

4.1- ESTIMADOR KAPLAN-MEIER (K-M)

Este método fue desarrollado por Kaplan-Meier (1958) y denominado "product-limit estimator" (estimador producto límite); calcula la probabilidad de sobrevivir a un tiempo t_i determinado.

Es un método **no paramétrico** utilizado en el análisis de supervivencia para estimar la función de supervivencia de una población en la que los individuos están sujetos a un evento que puede ocurrir en cualquier momento. Se utiliza principalmente en **estudios médicos** como ensayos clínicos y estudios epidemiológicos.

Existen numerosas formas de estimar el método de Kaplan-Meier, ya que no trabaja con periodos de tiempo, sino que los mismos tiempos de observación van cooperando en la estimación de la función de supervivencia.

Para calcular este estimador, se sigue el siguiente procedimiento:

Tabla 2: *Pasos para realizar el estimador Kaplan-Meier (K-M).*

1	Se ordenan los datos de tiempo de evento $\boldsymbol{t_i}$ en orden creciente.
2	Se identifica el número de individuos en riesgo en cada momento t_i.
3	Se calcula la proporción de individuos que sobreviven en cada momento t_i.
4	Se calcula la probabilidad acumulada de supervivencia en cada t_i multiplicando la proporción de individuos que sobreviven en ese momento t_i por la probabilidad acumulada de supervivencia en el momento anterior t_{i-1}.
5	Finalmente, se repite el segundo y cuarto paso para cada t_i hasta que finalice el periodo de seguimiento.

Este estimador de Kaplan-Meier la supervivencia se estima de dos maneras distintas, dependiendo de que los casos presenten o no censuras:

- Estimador de supervivencia para casos No censurados:

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{ni}$$

Donde:

$n_i \rightarrow$ Es el número de individuos que están en riesgo en un determinado tiempo t_i, o el número de individuos activos y no censurados justo antes del tiempo t_i.

$d_i \rightarrow$ Es el número de individuos que experimentan el evento en un instante preciso de tiempo t_i.

El producto representa la probabilidad acumulada de probabilidades condicionadas, es decir, la probabilidad de que ocurra un evento como consecuencia de que haya ocurrido otro evento relacionado.

De este modo, si la muestra de n individuos no presenta casos censurados, en el estimador se tiene:

$$\hat{S}(t_0) = 1 \ y \ \hat{S}(t_n) = 0$$

La **varianza del estimador Kaplan-Meier para casos no censurado**, se halla, utilizando el método delta y la fórmula de Greenwood:

1. Método delta

   Para $t_k \leq t \leq t_{k-1}$, se obtiene:

$$\log\left(\hat{S}(t)\right) = \sum_{i=1}^{k} \log\left(1 - \frac{d_i}{n_i}\right) = \sum_{i=1}^{k} \log(1 - q_i) = \sum_{i=1}^{k} \log(p_i)$$

   De esta manera:

$$Var\left[\log\left(\hat{S}(t)\right)\right] = Var\left(\sum_{i=1}^{k} \log(p_i)\right) = \sum_{i=1}^{k} Var(\log(p_i))$$

   Donde, los p_i, son independientes, por lo que:

$$Var(\log(p_i)) \approx \left(\frac{1}{p_i}\right)^2 \frac{p_i(1-p_i)}{n_i} = \frac{1-p_i}{n_i p_i}$$

   Entonces, de la misma manera se obtiene para $\hat{S}(t)$:

$$Var\left(\log\left(\hat{S}(t)\right)\right) = \sum_{i=1}^{k} \frac{1-p_i}{n_i p_i}$$

Pero, lo que en realidad se necesita, es la varianza de $\hat{S}(t)$, sin el logaritmo.

Para ello, se vuelve a aplicar el **método delta** para $\log\left(\hat{S}(t)\right)$:

$$Var\left[\log\left(\hat{S}(t)\right)\right] = \left(\frac{1}{\hat{S}(t)}\right)^2 \cdot Var\left(\hat{S}(t)\right)$$

Despejando la varianza de $\hat{S}(t)$ en la ecuación anterior, se tiene:

$$Var\left(\hat{S}(t)\right) = \hat{S}(t)^2 \cdot Var\left[\log\left(\hat{S}(t)\right)\right] = \hat{S}(t)^2 \sum_{i=1}^{k} \frac{1-p_i}{n_i p_i}$$

2. Fórmula de Greenwood

Esta fórmula se obtiene, al devolverle el valor original a p_i:

$$Var\left(\hat{S}(t)\right) = \hat{S}(t)^2 \cdot \sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)}$$

Para cualquier tiempo t, se tiene:

$$Var\left(\hat{S}(t)\right) = \hat{S}(t)^2 \cdot \sum_{i|t_i<t}^{k} \frac{d_i}{n_i(n_i - d_i)}$$

Finalmente, al haber aplicado el método delta y la fórmula de Greenwood, se obtiene la **varianza del estimado K-M para casos No censurados**:

$$Var\left[\hat{S}(t)\right] = \left(\prod_{t_i<t} \frac{n_i - d_i}{ni}\right)^2 \sum_{t_i<t} \frac{d_i}{n_i(n_i - d_i)}$$

- Estimador de supervivencia para casos Si censurados: En el caso de tener datos que presentan censura, la información del tiempo de supervivencia es incompleto ya que algunos individuos han abandonado o se ha perdido información durante el estudio. Se utiliza la misma fórmula que para el caso anterior (casos No censurados):

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{ni}$$

Con la diferencia, de que, si t es un valor censurado, la supervivencia se mantiene constante, por tanto, los tiempos en los que se halla el producto anterior son sólo los tiempos de evento. Y en la curva de supervivencia, se señala justo el momento en que el individuo sale del estudio.

En el caso de tener **muestras grandes**, el estimador de Kaplan-Meier sigue una distribución normal, por el teorema central del límite (modificación de Kaplan-

Meier, (1958): normal con poca censura. Por lo que el intervalo de confianza al $\left(100(1-\propto)\right)\%\ de\ \hat{S}_{KM}(t)$, viene dado por:

$$\hat{S}_{KM}(t)\ \pm\ z_{1-\frac{\propto}{2}}\ SE\left(\hat{S}_{KM}(t)\right)$$

Donde:

$z_{1-\frac{\propto}{2}}$ → Indica el percentil de una distribución normal estándar de dos colas al nivel de significación $1-\frac{\propto}{2}$, es decir:

$$P\left(Z < z_{1-\frac{\propto}{2}}\right) = 1-\frac{\propto}{2}\text{, con } Z\sim N(0,1)$$

$SE\left(\hat{S}_{KM}(t)\right)$ → Es el error estándar de la estimación de Kaplan-Meier.

Por ejemplo: Si en el estudio se utilizan intervalos de confianza del 95%, se tendrá un $\propto= 0{,}05$ y la confianza del estimador Kaplan-Meier es:

$$\hat{S}_{KM}(t)\ \pm\ z_{0{,}975}\ SE\left(\hat{S}_{KM}(t)\right) = \hat{S}_{KM}(t)\ \pm\ 1{,}96\ \cdot\sqrt{V(\hat{S}_{KM}(t))}$$

Si se tienen datos que presenten un porcentaje alto de observaciones censuradas, se debe emplear el **estimador de Kaplan-Meier ponderado, $\boldsymbol{KMp}$**, propuesto por (Bahrawar 2005), quien modificó el método original, ponderando así las observaciones con censura, con un factor o con tasa de no censura. KMp, considera la censura como una parte fundamental del análisis.

$$W_j = \frac{n_j - c_j}{n_j} \rightarrow donde\ 0 \leq W_j \leq 1$$

Donde:

W_j → Es la tasa de no censura para el tiempo t_j.

c_j → Es el número de censuras para el tiempo t_j.

$$\text{En el caso de: } \begin{cases} W_j = 1 \rightarrow No\ hay\ censura\ en\ t_j \\ W_j < 1 \rightarrow\ Hay\ una\ censura\ en\ t_j \end{cases}$$

De esta forma, este **estimador de Kaplan-Meier ponderado, $\boldsymbol{KMp}$**, se concreta como:

$$\hat{S}(t) = \prod_{j:t(j)\leq t} W_j\left(\frac{n_j - d_j}{n_j}\right)$$

Una vez realizado estos cálculos, tanto para los casos censurados o no censurados, el estudio produce una **curva de supervivencia** que muestra la probabilidad acumulada de supervivencia a lo largo del tiempo.

Esta curva, se emplea para comparar las tasas de supervivencia entre diferentes conjuntos de individuos y también para identificar los factores que están asociados con una mayor o menor supervivencia.

Se representan en gráficas escalonadas que comienzan con una supervivencia de uno que se mantiene constante hasta producirse el primer evento y que da un salto decreciente cada vez que da un nuevo evento, es decir, este estimador de supervivencia, no se modifica en el caso de encontrar alguna observación con censura. Peor los tiempos de censura provocan una bajada mayor de la curva de supervivencia afectando así al descenso que experimentan en el siguiente escalón.

Una vez realizada la curva de supervivencia, se encuentran diferentes ventajas y desventajas para este análisis:

Tabla 3: *Ventajas y desventajas sobre la curva de supervivencia K-M.*

VENTAJAS	DESVENTAJAS
El tiempo de observación del individuo puede ser corto.	La parte derecha de la curva, que es en la que se tiene menos individuos, es menos fiable.
Se consideran las observaciones con o sin censuras.	Si las curvas se cruzan al principio significan que los individuos con peor pronóstico fallan antes.
Se tiene el tiempo exacto de cada individuo estudiado. Ya que se apunta el momento en que el individuo entra al estudio.	Las curvan dan la impresión de que el evento ocurre al principio (dependiendo de la base de datos estudiada), esto es porque a algunos individuos no les da tiempo comenzar con el estudio.
Se tienen resultados buenos si se tienen muestras pequeñas. (En el caso de tener un evento).	Se espera que las curvas se crucen si el tratamiento no es efectivo.
Se puede comparar las curvas de supervivencia entre diferentes grupos.	Solo controla una covariable, es decir, solo controla un factor que influye en la supervivencia, lo que afecta a la interpretación del resultado.

A modo de síntesis, el **método de Kaplan-Meier** es un análisis útil para estimar la supervivencia de un conjunto de individuos que están expuestos a un evento determinado.

Y proporciona una forma de visualizar la probabilidad de supervivencia a lo largo del tiempo y de comparar las curvas de supervivencia entre diferentes grupos.

4.1.1 – DEMOSTRACIÓN DE SU MÁXIMA VEROSIMILITUD

Es importante destacar que el **estimador Kaplan-Meier es el estimador máximo verosímil** de la función de supervivencia. La técnica de máxima verosimilitud es utilizada para estimar los parámetros de un modelo probabilístico a partir de un conjunto de datos observados.

Se verifica que $\hat{S}(t)$ es un estimador consistente de $S(t)$, es decir, que a medida que aumenta el tamaño de la muestra, $\hat{S}(t)$ se acerca al verdadero valor de $S(t)$. Por tanto, el error en la estimación se aproxima a cero a medida que el tamaño de la muestra aumenta.

Para demostrar que este estimador K-M es máximo verosímil, primero se debe definir la función de verosimilitud y después maximizarla.

Sea una muestra $t_0, t_1, t_2, \dots, t_n,$ de tiempos observados. La información que aportan a la verosimilitud los diferentes tiempos será: la densidad a tiempo t_i , $f(t_i) = \lambda_i$ para cada tiempo de fallo; es decir, la probabilidad de que ocurra un evento en ese determinado momento, y la supervivencia a tiempo t_i , $S(t_i)$ para cada tiempo de censura; es decir, la probabilidad de que el evento ocurra más allá de este tiempo censurado.

Puede verse que los parámetros de los que depende la función de verosimilitud son precisamente las probabilidades λ_i de que se produzca un evento a tiempo t_i, y veremos que sólo depende de dichas probabilidades en los tiempos de fallo y no de censura. Por tanto, la función de verosimilitud será de la forma:

$$L(\lambda_0, \lambda_1, \dots, \lambda_k)$$

La función de verosimilitud ve cómo de probable es la muestra de tiempos que nos ha salido en el estudio y vamos a ver qué vector de parámetros $(\lambda_0, \lambda_1, \dots, \lambda_k)$ la maximiza

Para **hallar el estimador máximo verosímil de la supervivencia,** se siguen los siguientes pasos:

1. Dividimos en intervalos y contamos las censuras en los mismos:

Suponemos que tenemos $t_0, t_1, t_2, \dots, t_k,$ tiempos de fallo entre los tiempos observados. Los intervalos de tiempo que vamos a considerar para este estimador de Kaplan-Meier contienen un único tiempo de fallo, aunque pueden darse varios fallos (empates) a dicho tiempo.

$$I_j = [t_j, t_{j+1}), donde\ j = 0, \dots, k-1 \qquad I_k = [t_k, \infty)$$

Además, se tiene:

$$d_j\ empates\ en\ t_j \rightarrow número\ de\ fallos\ ocurridos\ en\ t_j,\ j = 0,1, \dots, k.$$

$c_j \rightarrow$ Es el número de censuras entre dos tiempos de fallo consecutivos, es decir en el intervalo $I_j$donde $j = 0,1, \dots, k-1$.

$c_k \rightarrow$ Son las censuras a tiempo de t_k.

Es decir, un intervalo, I_j, presenta c_j censuras para $t_{j1}, t_{j2}, \dots, t_{jc_j}$.

2. Se **construye** la función de **verosimilitud**:

- Si se tiene información de un individuo que muere en t_j:

No ha muerto en $\boldsymbol{t_1}$	$(\mathbf{1}-\boldsymbol{\lambda_1})$
No ha muerto $\boldsymbol{t_2}$	$(1-\lambda_2)$
⋮	⋮
No ha muerto en $\boldsymbol{t_{j-1}}$	$(1-\lambda_{j-1})$
Si muere en $\boldsymbol{t_j}$	λ_j

Por tanto, se obtiene **la verosimilitud para ese tiempo de fallo** como:

$$L_j = \prod_{i=1}^{j-1}(1-\lambda_i)\,\lambda_j = \lambda_j \prod_{i=1}^{j-1}(1-\lambda_i)$$

Si $\boldsymbol{d_j}$ **individuos recogen esta misma información en** $\boldsymbol{t_j}$, entonces:

$$(L_j)^{d_j} = \lambda_j^{d_j} \prod_{i=1}^{j-1}(1-\lambda_i)^{d_j}$$

- Si se tiene información del individuo que sobrevive a $I_j = [t_j, t_{j+1})$:

$$L_{ji} = \prod_{i=1}^{j}(1-\lambda_i) = (1-\lambda_j)\prod_{i=1}^{j-1}(1-\lambda_i)$$

Si se tiene $\boldsymbol{c_j}$ **individuos con censuras en** $\boldsymbol{I_j}$, la información es:

$$(L_{ji})^{c_j} = (1-\lambda_j)^{c_j} \prod_{i=1}^{j-1}(1-\lambda_i)^{c_j}$$

Si se une toda la información de $t_0, \dots, t_1, \dots, t_2, \dots$, se tiene:

$$L(\lambda_0, \lambda_1, \dots, \lambda_k) = (L_j)^{d_j} \cdot (L_{ji})^{c_j} =$$

$$= \prod_{j=0}^{k}\left[\lambda_j^{d_j}(1-\lambda_j)^{c_j}\left(\prod_{i=0}^{j-1}(1-\lambda_i)^{d_j+c_j}\right)\right]$$

Para ver cómo queda la función de verosimilitud de una forma más sencilla, **primero** se desarrolla el producto:

$$L(\lambda_0, \lambda_1, \dots, \lambda_k) = \lambda_0^{d_0}(1-\lambda_0)^{c_0}\,\lambda_1^{d_1}(1-\lambda_1)^{c_1}(1-\lambda_0)^{d_1+c_1}\cdot$$

$$\cdot\,\lambda_2^{d_2}(1-\lambda_2)^{c_2}(1-\lambda_0)^{d_2+c_2}(1-\lambda_1)^{d_2+c_2}\cdot$$

$$\cdot\,\lambda_3^{d_3}(1-\lambda_3)^{c_3}(1-\lambda_0)^{d_3+c_3}(1-\lambda_1)^{d_3+c_3}(1-\lambda_2)^{d_3+c_3}\cdot$$

$$\vdots$$

$$\cdot\,\lambda_k^{d_k}(1-\lambda_k)^{c_k}(1-\lambda_0)^{d_k+c_k}(1-\lambda_1)^{d_k+c_k}\cdot\ldots\cdot(1-\lambda_{k-1})^{d_k+c_k}$$

Segundo, se agrupan los factores y se obtiene:

$$L(\lambda_0, \lambda_1, \dots, \lambda_k) = \lambda_0^{d_0}\lambda_1^{d_1}\lambda_2^{d_2}\cdots\lambda_k^{d_k}(1-\lambda_0)^{c_0+d_1+c_1+d_2+c_2+d_3+c_3+\cdots+d_k+c_k}\cdot$$

$$\cdot(1-\lambda_1)^{c_1+d_2+c_2+d_3+c_3+\cdots+d_k+c_k}\cdot\ldots\cdot(1-\lambda_{k-1})^{c_{k-1}+c_k+d_k}(1-\lambda_k)^{c_k}$$

Puede verse que el número de personas en riesgo al comienzo del intervalo I_j coincide con:

$$n_j = \left(c_j + d_{j+1} + c_{j+1} + d_{j+2} + c_{j+2} + \cdots + c_k\right)$$

Por tanto:

$$L(\lambda_0, \lambda_1, \dots, \lambda_k) = \lambda_0^{d_0}\lambda_1^{d_1}\lambda_2^{d_2}\cdots\,\lambda_k^{d_k}\,(1-\lambda_0)^{n_0-d_0}(1-\lambda_1)^{n_1-d_1}\cdots(1-\lambda_k)^{n_k-d_k}$$

Finalmente, volviendo a agrupar los índices, la **función de verosimilitud** es:

$$L(\lambda_0, \lambda_1, \dots, \lambda_k) = \prod_{i=0}^{k}\lambda_i^{d_i}(1-\lambda_i)^{n_i-d_i}$$

3. <u>Se maximiza la función de verosimilitud:</u>

Para **maximizar la función de verosimilitud**, se maximiza su logaritmo; que es la llamada función de log-verosimilitud

$$l(\lambda_0, \lambda_1, \dots, \lambda_k) = logL(\lambda_0, \lambda_1, \dots, \lambda_k),$$

Los pasos para maximizar la función de verosimilitud son los siguientes:

- Se calcula la función de log-verosimilitud:

$$l(\lambda_0, \lambda_1, \dots, \lambda_k) = \log L(\lambda_0, \lambda_1, \dots, \lambda_k) = \sum_{i=0}^{k}[d_i\log(\lambda_i) + (ni - d_i)\log(1-\lambda_i)]$$

- Se calculan sus derivadas respecto de cada una de las λ_i y se igualan a cero:

$$\frac{\partial l(\lambda_0, \lambda_1, \dots, \lambda_k)}{\partial \lambda_i} = \frac{d_i}{\lambda_i} - \frac{n_i - d_i}{1 - \lambda_i} = 0 \rightarrow \frac{d_i}{\lambda_i} = \frac{n_i - d_i}{1 - \lambda_i} \rightarrow$$

$$\rightarrow d_i - \cancel{d_i \lambda_i} = \lambda_i n_i - \cancel{\lambda_i d_i} \rightarrow$$

$$\hat{\lambda}_i = \frac{d_i}{n_i}$$

Después se calculan las derivadas segundas y se construye su Matriz Hessiana, $H(\lambda_0, \lambda_1, \dots, \lambda_k)$, que es una matriz cuadrada $(k+1)\, x(k+1)$, cuyos elementos son dichas derivadas, y se mira si la matriz $H(\hat{\lambda}_0, \hat{\lambda}_1, \dots, \hat{\lambda}_k)$ es definida positiva o negativa, para determinar si $(\hat{\lambda}_0, \hat{\lambda}_1, \dots, \hat{\lambda}_k)$ es un máximo o un mínimo.

$$H(\lambda_0, \lambda_1, \dots, \lambda_k) = \begin{pmatrix} \frac{\partial^2 l}{\partial \lambda_0^2} & \frac{\partial^2 l}{\partial \lambda_0 \partial \lambda_1} & \cdots & \cdots & \frac{\partial^2 l}{\partial \lambda_0 \partial \lambda_k} \\ \frac{\partial^2}{\partial \lambda_0 \partial^2 \lambda_1} & \frac{\partial^2 l}{\partial \lambda_1^2} & \cdots & \cdots & \frac{\partial^2 l}{\partial \lambda_1 \partial \lambda_k} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \lambda_k \partial^2 \lambda_0} & \frac{\partial^2 l}{\partial \lambda_k \partial \lambda_1} & \cdots & \cdots & \frac{\partial^2 l}{\partial \lambda_k^2} \end{pmatrix}$$

Como

$$\frac{\partial^2 l(\lambda_0, \lambda_1, \dots, \lambda_k)}{\partial \lambda_i^2} = -\frac{d_i}{\lambda_i^2} - \frac{n_i - d_i}{(0 - \lambda_i)^2} \quad y \quad \frac{\partial^2 l(\lambda_0, \lambda_0, \dots, \lambda_k)}{\partial \lambda_i \partial \lambda_j} = 0 \;\; i \neq j$$

Se tiene que:

$$H = \begin{pmatrix} -\frac{d_0}{\lambda_0^2} - \frac{n_0 - d_0}{(0 - \lambda_0)^2} & 0 & \cdots & \cdots & 0 \\ 0 & -\frac{d_1}{\lambda_1^2} - \frac{n_1 - d_1}{(0 - \lambda_1)^2} & \cdots & \cdots & 0 \\ & & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \vdots & \vdots & & & \\ 0 & 0 & \cdots & \cdots & -\frac{d_k}{\lambda_k^2} - \frac{n_k - d_k}{(0 - \lambda_k)^2} \end{pmatrix}$$

Si al sustituir $\hat{\lambda}_i = \frac{d_i}{n_i}$, $i = 0, 1, \dots, k$, la H es definida negativa, entonces $(\hat{\lambda}_0, \hat{\lambda}_1, \dots, \hat{\lambda}_k)$ es un máximo de la función de verosimilitud, y por tanto los estimadores máximos verosímiles de las probabilidades de fallo serán las $\hat{\lambda}_i = \frac{d_i}{n_i}$.

Para un indicador i cualquiera:

$$\frac{\partial^2 l(\lambda_0, \lambda_1, \dots, \lambda_k)}{\partial \lambda_i^2} = \frac{-d_i}{\frac{d_i^2}{n_i^2}} - \frac{n_i - d_i}{\left(\frac{n_i - d_i}{n_i}\right)^2} = \frac{-n_i^2}{d_i} - \frac{n_i^2}{n_i - d_i} = \frac{-n_i^3}{d_i(n_i - d_i)}$$

Por lo que la matriz hessiana valorada en $(\hat{\lambda}_0, \hat{\lambda}_1, \dots, \hat{\lambda}_k)$ queda:

$$H(\hat{\lambda}_0, \hat{\lambda}_1, \dots, \hat{\lambda}_k) = \begin{pmatrix} \frac{-n_0^3}{d_0(n_0-d_0)} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{-n_1^3}{d_1(n_1-d_1)} & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \frac{-n_k^3}{d_k(n_k-d_k)} \end{pmatrix}$$

Como para todo i se verifica que $n_i - d_i > 0$, se tiene que cada uno de los elementos de la diagonal de la matriz es negativo, y, por tanto, esa matriz es definida negativa, por tanto, $(\hat{\lambda}_0, \hat{\lambda}_1, \dots, \hat{\lambda}_k)$ es un máximo.

4. Cálculo del estimador máximo verosímil de la función de supervivencia:

Veamos qué forma tiene la función de supervivencia para un tiempo cualquiera t.

$$\hat{S}(t) = \prod_{t_i<t}(1-\hat{\lambda}_i) = \prod_{t_i<t}\left(1-\frac{d_i}{n_i}\right) = \prod_{t_i<t}\frac{n_i-d_i}{n_i}$$

Podemos comprobar que **el estimador máximo verosímil de $S(t)$** que se obtiene coincide con el **estimador Kaplan-Meier** visto en el apartado anterior (4.0).

Como en su construcción sólo se ha tenido en cuenta los tiempos de fallo, y las censuras para hacer disminuir el tamaño del grupo de riesgo, se verifican estos tres requisitos de la función de supervivencia $S(t)$.

Tabla 4: *Requisitos de la función de supervivencia.*

Tiene que ser continua para los tiempos que presenten censuras.
Tiene que ser discontinua para los tiempos de fallo, es decir, cada vez que se produzca un fallo, "salta", es por ello su discontinuidad.
Tiene que ser constante entre dos tiempos de fallo que estén seguidos.

4.2 – ESTIMADOR DE NELSON AALEN (NA)

El estimador de Nelson Aalen es un estimador **no paramétrico** de la función de riesgo acumulado en el análisis de supervivencia. Se usa principalmente para analizar datos censurados, donde se desconoce el tiempo exacto de ocurrencia de un evento para ciertas observaciones.

Este estimador (NA) fue propuesto en 1969 y quien le dio forma fue Altschuler (1970), tras utilizarlo con técnicas de conteo. Muchas veces se compara con el estimador de Kaplan-Meier para el análisis de supervivencia.

De esta forma, el estimador de **Nelson Aalen** se define con ayuda de la función de riesgo acumulada (2.2.3).

En el caso de tener una **variable aleatoria discreta**, se crea para poder estimarla, la función empírica de riesgo acumulado, $\widehat{H}(t)$, dependiendo del tipo de censura:

- Censura por la derecha: Nelson estima la función de riesgo acumulado, $H(t)$, y la define como:

$$\widehat{H}(t) = \sum_{i=0}^{n} \frac{0_{\{t_i \leq t, \delta_i = 0\}}}{\sum_{j=0}^{n} 0_{\{t_i \leq t_j\}}} = \sum_{j \leq t} \frac{d_j}{n_j}$$

  Donde:

  $d_j \rightarrow$ Es el número de fallos ocurridos en un instante t_j.

  $n_j \rightarrow$ Es el número de individuos que se encuentran en riesgo en t_j.

  Así, el estimador de **Nelson Aalen para la función de supervivencia** se expresa:

$$\hat{S}(t) = e^{-\widehat{H}(t)} = e^{\sum_{j \leq t} \frac{d_j}{n_j}}$$

  Es importante destacar, que, en el año 1984, Flemyn y Harrington recomiendan este estimador como una forma alternativa al estimador no paramétrico de máxima verosimilitud.

- Censura por la izquierda: Una extensión de esta estimación de la función de riesgo acumulado, propuesta en 1998 por Pan y Chapel. Se denomina **estimador de Nelson Aalen extendido, $\widehat{\boldsymbol{H}}_{\boldsymbol{e}}(\boldsymbol{t})$**, y es buena para solucionar los problemas de subestimación que se dan en el estimador no paramétrico de máxima verosimilitud en el caso de haber truncamiento. En este caso, para estimar el riesgo acumulado, se define igual que en el punto anterior (censura por la derecha), con la diferencia:

$$n_j \; pasa \; a \; ser \rightarrow r_j$$

  Donde r_j es el número de individuos en riesgo antes de t_j.

  Así, el **estimador de Nelson Aalen extendido** se expresa como:

$$\widehat{H}_e(t) = \sum_{i=0}^{n} \frac{0_{\{t_i \leq t, \delta_i = 0\}}}{\sum_{j=0}^{n} 0_{\{x_j \leq t_i \leq t_j\}}} = \sum_{t_i \leq t} \frac{d_i}{r_i}$$

  y el estimador de **Nelson Aalen para la función de supervivencia** es:

$$\hat{S}(t) = e^{-\widehat{H}(t)} = e^{-\sum_{t_i \leq t} \frac{d_i}{r_i}}$$

  En el caso de tener una **variable aleatoria continua**, la utilización de este estimador generó muchas dudas por parte de Nelson (1972), Breslow y Crowley (1974), Efron (1977) y Altschuler (1979). Aunque se llegó a la conclusión de que la $\boldsymbol{H}(\boldsymbol{t})$ estimada por Nelson y la $\boldsymbol{S}(\boldsymbol{t})$ estimada por Kaplan-Meier, **son equivalentes**, salvo para los valores altos de t, ya que las estimaciones a dicho tiempo son menos estables.

5 – COMPARACIÓN DE FUNCIONES DE SUPERVIVENCIA

La **comparación de funciones relacionadas con el análisis de supervivencia** se utiliza para encontrar diferencias entre el tiempo de supervivencia hasta el evento de dos o más grupos. Para ello, se necesitan **pruebas estadísticas** apropiadas con el fin de determinar si todos los grupos presentan la misma supervivencia y, en caso de no ser así, cuáles son diferentes.

La **representación de las curvas** de supervivencia avisa si se encuentran diferencias entre ellas, pero con las pruebas estadísticas adecuadas puede determinarse qué diferencias entre las curvas son significativas. Además, sirven para indicar si un factor considerado en el estudio influye de forma importante en el tiempo de supervivencia que se esté estudiando.

Existen varias pruebas estadísticas para poder comparar la supervivencia de dos o más grupos. Se detallan a continuación las más utilizadas.

5.0 – PRUEBAS LOG-RANK

La **prueba Log-Rank**, se utiliza para comparar dos funciones de supervivencia.

Su objetivo es determinar si existen diferencias significativas entre las curvas de supervivencia de dos grupos en estudio.

Para realizar la **prueba Log-Rank**, se tienen que seguir los siguientes pasos:

1. Se definen los dos grupos de interés, como **grupo I** y **grupo II**.

2. Se ordenan los tiempos de fallo de los dos grupos de interés:

$$t_0 < t_2 < \cdots < t_k$$

   Donde:

   $d_{0j}\ y\ d_{2j}$ → Son los individuos del grupo I (d_{0j}) y grupo II (d_{2j}), que fallan en un determinado tiempo t_j, para $j = 0, 2, \ldots, k$. De esta manera, se puede obtener el **total de los individuos que fallan** en un determinado tiempo, es decir:

$$d_j = d_{0j} + d_{2j}$$

   $n_{0j}\ y\ n_{2j}$ → Son individuos que están en riesgo del grupo I (n_{0j}) y grupo II (n_{2j}), antes del tiempo t_j. De este modo, se puede obtener el **total de los individuos que están en riesgo**, es decir:

$$n_j = n_{0j} + n_{2j}$$

2. Se contrasta la hipótesis de igualdad de supervivencia entre los dos grupos:

$$\begin{cases} H_0 : S_0(t) = S_2(t) \\ H_0 : S_0(t) \neq S_2(t) \end{cases}$$

3. Se coloca la información en una tabla de contingencia para cada t_j, donde los grupos se ponen en las filas, en las columnas se encuentra el número de veces que el evento ocurre o no a tiempo t_j y el número de individuos en riesgo a t_j.

Tabla 5: *Tabla de contingencia para el contraste de hipótesis.*

TABLA DE CONTINGENCIA			
GRUPO	OCURRE	NO OCURRE	EN RIESGO
I	d_{0j}	$n_{0j} - d_{0j}$	n_{0j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
TOTAL	d_j	$n_j - d_j$	n_j

4. Una vez hecha la tabla de contingencia, se compara el número de fallos observados dentro de cada grupo y el número de fallos esperados bajo la hipótesis nula.

Vamos a trabajar bajo la hipótesis nula; es decir, considerando que la **hipótesis nula es cierta**, y que la supervivencia para ambos grupos es semejante, por lo que no hay diferencias entre ellos.

Los marginales totales: $d_j, n_j - d_j$ se consideran fijos y, de esta manera:

D_{0j} y $D_{2j} \rightarrow$ Son variables aleatorias con distribución hipergeométrica:

$$D_{0j} \sim hipergeométrica\ (n_j, d_j, n_{0j})$$

$$D_{2j} \sim hipergeométrica\ (n_j, d_j, n_{2j})$$

Como los grupos son independientes, estas variables hipergeométricas también lo son y por tanto se obtiene:

$$P(D_{0j} = d_{0j}, D_{2j} = d_{2j}) =^{indep} = P(D_{0j} = d_{0j}) \cdot P(D_{2j} = d_{2j}) =$$

$$= \frac{\binom{d_j}{d_{0j}} \cdot \binom{n_j - d_j}{n_{0j} - d_{0j}}}{\binom{n_j}{n_{0j}}} \cdot \frac{\binom{d_j}{d_{2j}} \cdot \binom{n_j - d_j}{n_{2j} - d_{2j}}}{\binom{n_j}{n_{2j}}}$$

Bajo la hipótesis nula, las funciones de supervivencia y riesgo son iguales para ambos grupos a cada tiempo t_j, lo que provoca que **el número de fallos esperados y varianza** en t_j se exprese para el **grupo I** como:

$$e_{0j} = \mathbb{E}(D_{0j}) = n_{0j} \cdot \frac{d_j}{n_j} = n_{0j} \cdot h_j$$

$$V_{0j} = \mathbb{V}(D_{0j}) = \frac{n_j - d_j}{n_j - 0} \cdot d_j \cdot \frac{n_{0j}}{n_j} \cdot \left(0 - \frac{n_{0j}}{n_j}\right) = \frac{n_j - d_j}{n_j - 0} \cdot d_j \cdot \frac{n_{0j}}{n_j} \cdot \left(\frac{n_j - n_{0j}}{n_j}\right) =$$

$$= \frac{(n_j - d_j) \cdot \overbrace{(n_j - n_{0j})}^{n_{2j}} \cdot d_j \cdot n_{0j}}{n_j^2 \cdot (n_j - 0)} = \frac{n_{0j} \cdot n_{2j} \cdot d_j \cdot (n_j - d_j)}{n_j^2 \cdot (n_j - 0)}$$

Donde:

e_{0j} → Es el número medio de eventos en el grupo I a tiempo t_j.

Esta varianza, V_{0j}, tiene en cuenta tanto los tamaños de los grupos como el número de eventos y el número total de individuos que se encuentran en riesgo para t_j.

Así el **estadístico Log-Rank** es:

$$U = \sum_{j=0}^{k} (d_{0j} - e_{0j})$$

Bajo la hipótesis nula tenemos que E[U] = 0 y Var[U] es la suma de las V_{0j} . Por tanto:

$$L = \frac{U - E[U]}{\sqrt{Var[U]}} = \frac{\sum_{j=0}^{k} (d_{0j} - e_{0j})}{\sqrt{\sum_j^k V_{0j}}} \to N(0,1)$$

Elevando al cuadrado, se obtiene el **estadístico de la prueba Log-Rank (LR)**, para nuestra colección de tiempos t_j, donde $j = 0, 2, \dots, k$:

$$LR = \frac{\left(\sum_{j=0}^{k} (d_{0j} - e_{0j})\right)^2}{\sum_j^k V_{0j}} \to \chi_1^2$$

Es decir, se comparan los fallos observados (que son los eventos en cada situación) con los fallos esperados si las distribuciones fueran iguales.

Así, si el estadístico observado está en la región de aceptación del contraste aceptamos la hipótesis nula de igualdad de funciones de supervivencia, y en caso contrario debemos considerar que las funciones de supervivencia en ambos grupos son diferentes.

Esto quiere decir que, por ejemplo, si separamos los grupos por la variable sexo y resulta que las funciones de supervivencia son significativamente distintas, el factor sexo está afectando a la supervivencia.

Aunque este test nos indica la posible influencia de factores o covariables en la supervivencia lo que no nos indica es cuánto influyen en la misma. Para ello deberíamos hacer uso de regresión de Cox.

5.2 – PRUEBA DE WILCOXON

La **prueba de Wilcoxon o prueba de rangos** se utiliza para comparar dos muestras que están relacionadas.

En análisis de supervivencia, esta prueba se usa para comparar la supervivencia de dos grupos, siendo esto muy similar a la prueba vista anteriormente (prueba Log-Rank) con la diferencia de que la **prueba de Wilcoxon** pondera las diferencias de la Log-Rank a través del número de individuos en riesgo para cada t_j.

Para realizar la prueba de Wilcoxon, se tiene que seguir los siguientes pasos:

Tabla 6: *Pasos para realizar la prueba de Wilcoxon.*

1	Se definen los grupos de interés: Grupo I y Grupo II
2	Se calculan los tiempos de supervivencia de cada individuo tanto para el grupo I como el grupo II.
3	Si un individuo no experimenta el evento final, se considera una observación censurada.
4	Una vez organizados los tres pasos anteriores, se utiliza la prueba de Wilcoxon para comparar las distribuciones de supervivencia de los dos grupos, de la que se tiene la Log-Rank ponderada: $$U_W = \sum_{j=0}^{k} n_j(d_{0j} - e_{0j}) = \sum_{j=0}^{k} n_j \left(d_{0j} - n_j \frac{d_j}{n_j}\right)$$ Bajo la hipótesis nula de igualdad de funciones de supervivencia se tiene que la esperanza $\mathbb{E}(U_W) = 0$, varianza $V_w = \mathbb{V}(U_w) = \sum_{j=0}^{k} n_j^2 V_{0j}$ (siendo las V_{0j} las de la prueba Log-Rank). Así se tendrá que: $$L_W = \frac{U_W - E[U_W]}{\sqrt{Var[U_W]}} = \frac{\sum_{j=0}^{k} n_j(d_{0j} - e_{0j})}{\sqrt{\sum_j^k n_j^2 \, V_{0j}}} \to N(0,1)$$ Y elevando al cuadrado se obtiene el **estadístico de la prueba Wilcoxon**: $$W_w = \frac{L_W^2}{V_w} = \frac{\left(\sum_{j=0}^{k} n_j \left(d_{0j} - e_{0j}\right)\right)^2}{\sum_j^k n_j^2 V_{0j}} \to \chi_1^2$$ Esta prueba, principalmente se usa para valorar la importancia de los individuos que se encuentran al inicio del estudio, ya que hay un mayor número de ellos que están en riesgo al compararse con el final del estudio.

El estadístico Wilcoxon es menos sensible que el anterior (Log-Rank) y se usa principalmente para detectar diferencias entre las curvas de supervivencia cuando estas se cortan.

Este estadístico da un pero más elevado a las diferencias observadas en los primeros momentos del estudio, pues considera que son más informativas pues al inicio es donde hay número de unidades en riesgo.

Hay diferentes maneras de ponderar el estadístico Log-rank de forma que puede definirse una **familia de pruebas ponderadas Log-Rank** cuyo estadístico de contraste es:

$$\sum_{j=0}^{k} w_j(d_{0j} - e_{0j})$$

donde:

$w = (w_0, w_1, \dots, w_k)$ → Es un vector que pondera las diferencias entre los fallos observados con los esperados a lo largo del tiempo de observación.

Si se consideran varios vectores, se obtienen diferentes pruebas, que se ven reflejados en la siguiente tabla:

Tabla 7: *Pruebas según la ponderación w considerada.*

Si $\boldsymbol{w_j = 1}$ → se obtiene la prueba Log-Rank.
Si $\boldsymbol{w_j = n_j}$ → Se obtiene la prueba de Breslow o Wilcoxon generalizado.
Si $\boldsymbol{w_j = \sqrt{n_j}}$ → Se obtiene la prueba de Tarone Ware.
Si $\boldsymbol{w = \prod_1^j \frac{n_i - d_i + 1}{n_i + 1}}$ → Se obtiene la prueba de Prentice.

6 – APLICACIÓN PRÁCTICA DEL ANÁLISIS DE SUPERVIVENCIA

En este punto se refleja principalmente la aplicación práctica del análisis de supervivencia mediante el método de Kaplan-Meier junto con el de Nelson Aalen, comparando cuál de los dos métodos es más exacto a la hora de realizar la curva de supervivencia.

La base de datos utilizada es *'psych'*,1997, que está dentro del paquete *'KMsurv'*, del programa estadístico Rstudio, utilizado para realizar este análisis de supervivencia. Esta base de datos está formada por los tiempos hasta el alta de un tratamiento psicológico para pacientes mayores de edad que han acudido a terapia psicológica. Se tienen datos de 26 pacientes, tanto hombres como mujeres con una edad entre los 19 y 58 años. La variable respuesta viene dada por la duración del tratamiento psicológico en semanas, es decir, cuanto tiempo tarda en que el paciente finalice la terapia, o el abandono de la misma antes de tiempo.

En la **Tabla 8**, se muestran todas las variables, como las explicativas (Sexo y Edad) o de respuesta (Tiempo y Estado).

Tabla 8: *Variables de la base de datos 'psych'*

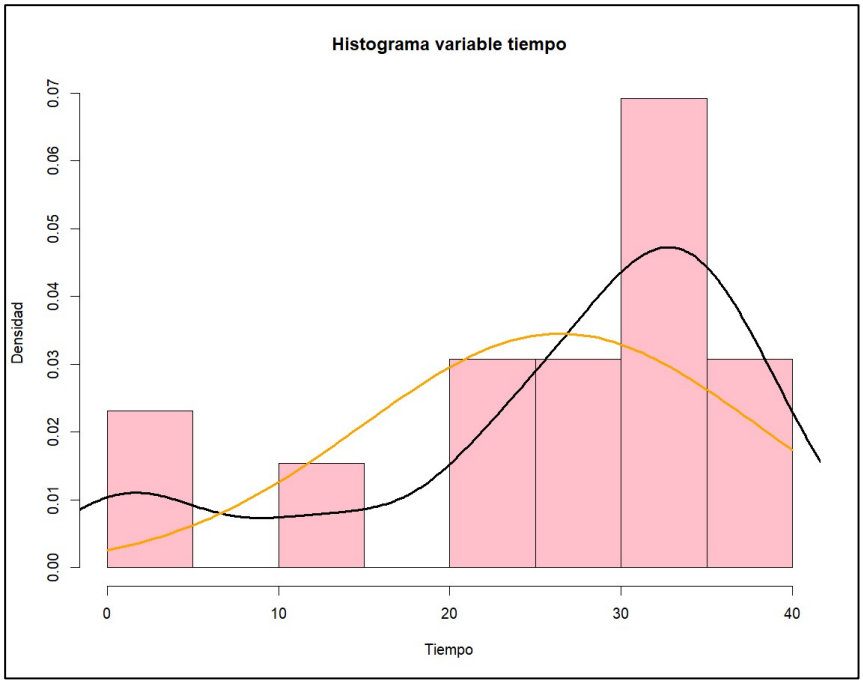
Nombre original	Nombre en R	Definición	Valores
Sex	Sexo	Sexo del paciente en estudio	1 = Masculino 2 = Femenino
Age	Edad	Edad en años del paciente en estudio	Valor numérico (desde los 19 hasta los 58 años).
Time	Tiempo	Tiempo hasta el alta (o abandono) en semanas	Valor numérico
Death	Estado	Indica cómo termina el paciente el estudio.	0 = abandono 1 = alta

La **salida que nos proporciona R** de estos datos da una idea general más clara:

```
Sexo        Edad            Tiempo          Estado
M:11   Min.   :19.00   Min.   : 1.00   Min.   :0.0000
F:15   1st Qu.:28.25   1st Qu.:22.50   1st Qu.:0.0000
       Median :32.50   Median :30.50   Median :1.0000
       Mean   :35.15   Mean   :26.42   Mean   :0.5385
       3rd Qu.:42.50   3rd Qu.:34.75   3rd Qu.:1.0000
       Max.   :58.00   Max.   :40.00   Max.   :1.0000
```

Tras conocer como es la base de datos se comienza con el análisis, para ello, se realiza un **histograma de la variable tiempo** para comprobar si los datos siguen una distribución normal o no.

Gráfico 1: *Histograma de la variable tiempo ('psych')*



Al observar este gráfico se puede observar que el tiempo del tratamiento no se distribuye de forma normal, ya que la línea naranja representa la distribución normal y la línea negra es la función de densidad de la variable tiempo.

Para confirmar que estos datos no siguen una distribución normal, se utiliza el test de Shapiro-Wilk, obteniendo de R un **p-valor = 0.001166**.

Como este p-valor < 0.05, se rechaza la hipótesis nula, es decir, se acepta la hipótesis de que los datos no se distribuyen de manera normal, cosa que se veía reflejada en el gráfico 1.

Una vez que se ha visto que la distribución de los datos no es normal, y que por tanto la distribución de los datos no permitiría hacer un estudio de regresión habitual, se comienza con el **análisis no paramétrico.**

En primer lugar, se realiza la estimación de **Kaplan-Meier**, y para ello se muestran los tiempos en los que se dan los eventos y las censuras:

```
 [1]  1   1   2  22  30+ 28  32  11  14  36+ 31+ 33+
[13] 33+ 37+ 35+ 25  31+ 22  26  24  35+ 34+ 30+ 35
[25] 40  39+
```

Donde los tiempos censurados son los que se representan con el signo '+'.

Con esta estimación de K-M, se obtienen las probabilidades de supervivencia para cada tiempo de evento.

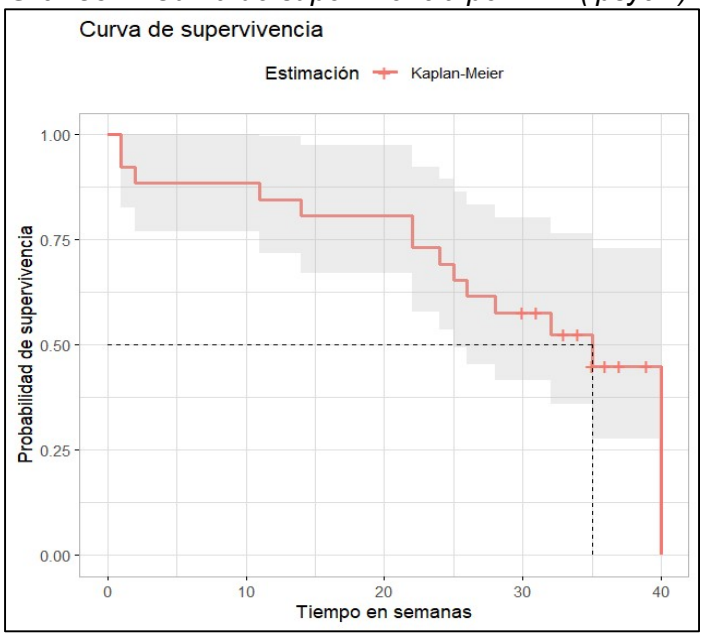
En ella se muestran los 12 tiempos para saber cómo es la estimación de la función de supervivencia, que seguidamente se refleja en el gráfico 2.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	26	2	0.923	0.0523	0.826	1.000
2	24	1	0.885	0.0627	0.770	1.000
11	23	1	0.846	0.0708	0.718	0.997
14	22	1	0.808	0.0773	0.670	0.974
22	21	2	0.731	0.0870	0.579	0.923
24	19	1	0.692	0.0905	0.536	0.895
25	18	1	0.654	0.0933	0.494	0.865
26	17	1	0.615	0.0954	0.454	0.834
28	16	1	0.577	0.0969	0.415	0.802
32	11	1	0.524	0.1013	0.359	0.766
35	7	1	0.450	0.1111	0.277	0.730
40	1	1	0.000	NaN	NA	NA

Mirando **la columna *survival*** (que refleja la función de supervivencia), se observa cómo se obtiene una mediana que está situada en la semana 35 del estudio, con 7 pacientes que siguen en riesgo y 1 evento ocurrido al finalizar ese intervalo de tiempo.

Con esta estimación de la función de supervivencia se representa la curva de supervivencia:

Gráfico 2: *Curva de supervivencia por K-M ('psych').*

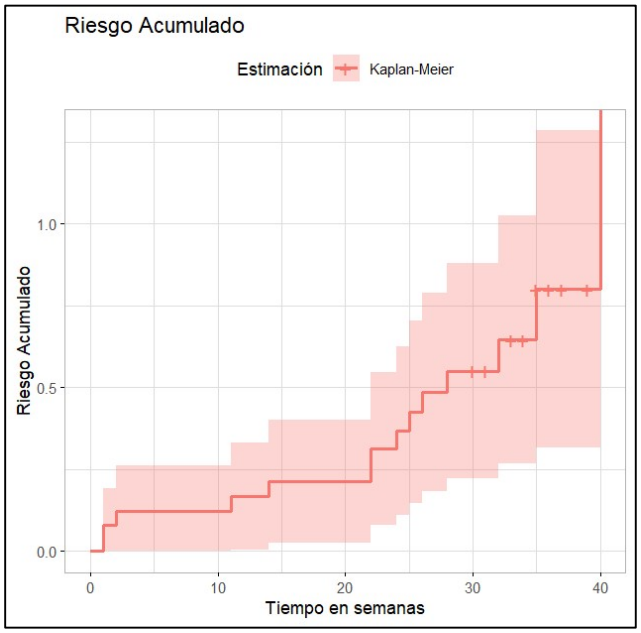


En esta curva de supervivencia, se observa que la probabilidad de que los pacientes sigan en terapia disminuye a medida que el tiempo va pasando, ya que los pacientes están superando los tratamientos psicológicos. Las censuras se encuentran reflejadas con una cruz en los tiempos correspondientes.

La curva comienza con un decrecimiento que no es muy destacable, hasta la semana 22 en que comienza un descenso más pronunciado, provocado por la velocidad con la que se producen los eventos (altas del tratamiento). Por último, esta curva finaliza en la semana 40 sin ningún paciente en tratamiento, lo que implica tener una supervivencia nula y por tanto en la semana 40 todos los pacientes han sido dados de alta o han abandonado el estudio.

El **riesgo acumulado** se ve reflejado en la siguiente gráfica:

Gráfico 3: *Curva de riesgo acumulado por K-M ('psych').*



En esta representación gráfica se observa un aumento a lo largo del tiempo, destacando que crece a partir de la semana 22, ya que en las semanas siguientes se produce un mayor número de eventos entre los pacientes.

De la misma forma, se ha realiza la estimación de Kaplan-Meier con la **influencia de covariables**, con la que se obtiene resultados para esas categorías.

Para ello, se calcula la estimación de la función de supervivencia para los dos valores de la **covariable *'Sexo'***. Los resultados son los siguientes:

```
Call: survfit(formula = Surv(Tiempo, Estado) ~ Sexo, data = psych)

        n events median 0.95LCL 0.95UCL
Sexo=M 11      4     35      35      NA
Sexo=F 15     10     26      14      NA
```

La mediana de los tiempos de alta en ambos grupos es muy diferente en ambos sexos, pues para el sexo masculino se encuentra en la semana 35 y para el sexo femenino, se da en la semana 26. Esto podría inducirnos a pensar que el sexo influye en el tiempo de alta del tratamiento, siendo el tratamiento más largo en los hombres.

Las tablas con la estimación de la supervivencia separadas por sexo se muestran a continuación:

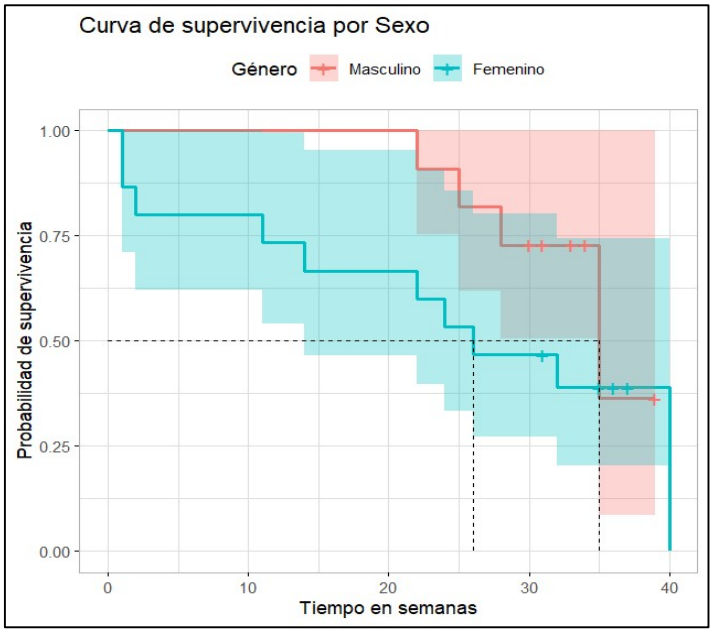
```
Call: survfit(formula = Surv(Tiempo, Estado) ~ Sexo, data = psych)

                Sexo=M
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   22     11       1    0.909  0.0867       0.7541            1
   25     10       1    0.818  0.1163       0.6192            1
   28      9       1    0.727  0.1343       0.5064            1
   35      2       1    0.364  0.2658       0.0868            1
```

Sexo=F						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	15	2	0.867	0.0878	0.711	1.000
2	13	1	0.800	0.1033	0.621	1.000
11	12	1	0.733	0.1142	0.540	0.995
14	11	1	0.667	0.1217	0.466	0.953
22	10	1	0.600	0.1265	0.397	0.907
24	9	1	0.533	0.1288	0.332	0.856
26	8	1	0.467	0.1288	0.272	0.802
32	6	1	0.389	0.1287	0.203	0.744
40	1	1	0.000	NaN	NA	NA

Puede verse que el total de varones que comienzan el tratamiento es de 11, de los cuales solo 4 son dados de alta, los demás son abandonos; por tanto, el porcentaje de censura en los hombres es un 63.6%.

Por otra parte, hay 15 mujeres que comienzan el tratamiento y 10 de ellas experimentan el evento y por tanto reciben el alta médica, así el porcentaje de abandono en mujeres es solamente del 33.33%, la mitad del porcentaje de abandono de tratamiento que el experimentado por los hombres.

Gráfico 4: *Curva de supervivencia para el grupo Sexo por K-M ('psych').*



La curva de supervivencia para el sexo masculino, en color anaranjado, se mantiene constante hasta la semana 22, que es la semana en la que se produce el primer informe de alta de un paciente varón. A partir de esta empieza a decrecer hasta la semana 35. Como no hay altas a partir de la semana 35, la curva se mantiene constante a partir de este momento, aunque no llega al 0 puesto que no todos los hombres finalizan el estudio con un alta a esa semana ya que existe un dato censurado en la semana 39.

La curva de supervivencia para el sexo femenino muestra una tendencia descendente más sostenida que en el caso del sexo masculino, con eventos (altas) repartidas más homogéneamente en el periodo de estudio, y sin ninguna supervivencia tras 40 semanas; es decir todas las mujeres han sido dadas de alta en esas 40 semanas o bien han abandonado el estudio.

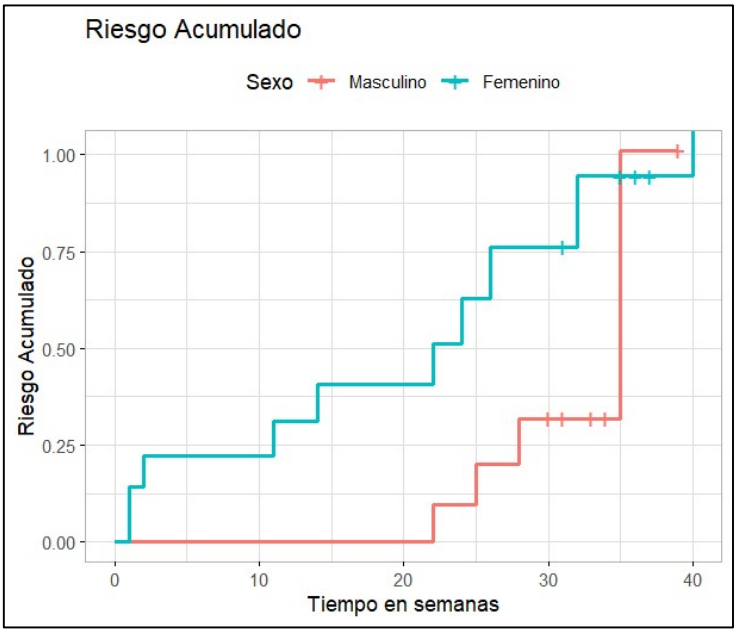
Para ver si la diferencia que se observa en las curvas de supervivencia separadas la covariable 'Sexo' es significativa, se **realiza el test de Log-Rank**, del que se obtiene:

```
Call:
survival::survdiff(formula = Surv(Tiempo, Estado) ~ Sexo, data = psych,
    rho = 0)

        N Observed Expected (O-E)^2/E (O-E)^2/V
Sexo=M 11        4     6.24     0.807      1.61
Sexo=F 15       10     7.76     0.650      1.61

 Chisq= 1.6  on 1 degrees of freedom, p= 0.2
```

Como el p-valor = 0.2 > 0.05, se acepta la hipótesis nula de igualdad entre los grupos, es decir, aunque aparentemente parece que la covariable *'Sexo'* influye en la supervivencia, no hay diferencia significativa entre la supervivencia en ambos grupos los grupos. Por tanto, la diferencia que veíamos en el gráfico puede estar distorsionada por la gran cantidad de abandonos que presenta el grupo de los hombres. Este es un buen ejemplo para mostrar que las representaciones gráficas pueden ser orientativas de diferencias entre grupos sin que esas diferencias deban ser consideradas.

Veamos el **riesgo acumulado** separado por la covariable sexo:

Gráfico 5: *Curva de riesgo acumulado de la covariable Sexo por K-M ('psych').*

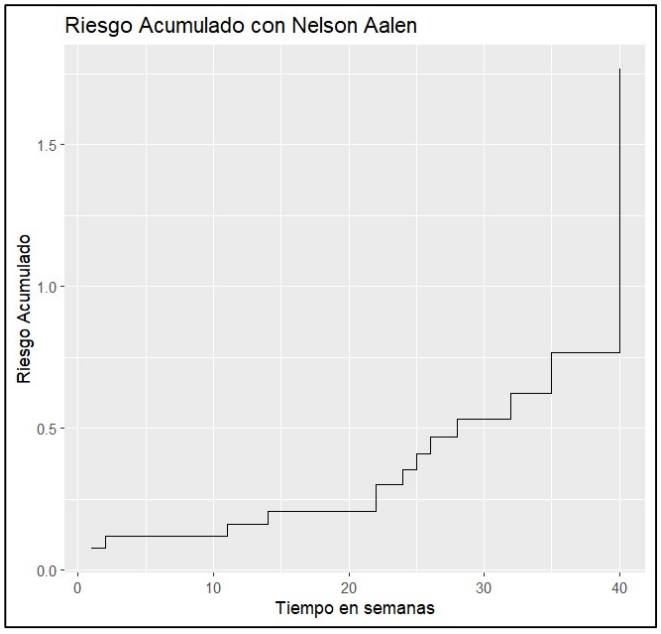


En esta gráfica se observa un aumento a lo largo del tiempo para ambos sexos, pues la función de riesgo acumulado ha de ser creciente.

Cabe destacar que dicho riesgo para el sexo masculino se mantiene constante hasta la semana 22, en el que se produce el primer evento, y en la semana 35 tiene un aumento muy pronunciado, lo que podría parecer un mayor número de eventos entre los pacientes masculinos, pero que en el fondo se debe a la gran disminución del grupo de pacientes en riesgo, pues aunque las semanas 28 y 35 son consecutivas en la tabla de estimación, entre ellas se producen 6 censuras, lo cual supone casi un 50% de los hombres que comenzaron el estudio.

En segundo lugar, se realiza la **estimación de Nelson Aalen**, de la función de riesgo acumulado, que se hace de forma similar a la realizada mediante el método de Kaplan-Meier. Usando así el comando *'qplot'* en R se obtiene el siguiente gráfico:

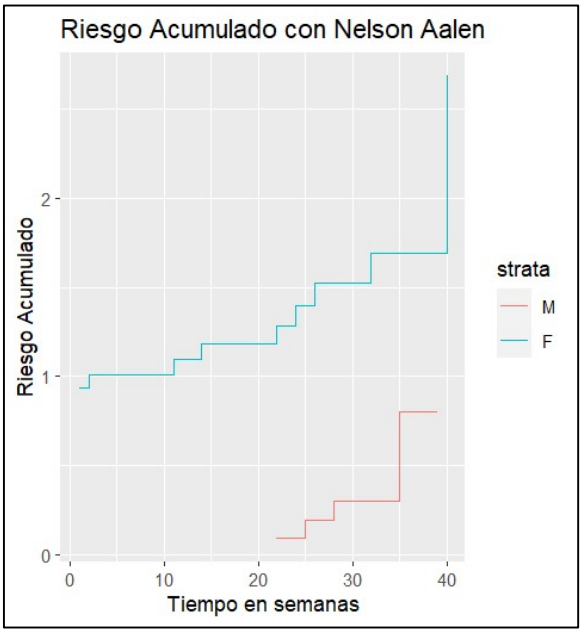
Gráfico 6: *Curva de riesgo acumulado por Nelson Aalen ('psych').*



En esta gráfica, se observa la misma forma de la función de riesgo acumulado que se obtuvo con el estimador Kaplan-Meier (gráfica 3); aunque en este caso el programa no nos muestra bandas de confianza. Una diferencia importante entre ambos gráficos es que este último es menos informativo en el sentido de que no están reflejados los abandonos de ninguna forma.

Separando los riesgos *acumulados por Nelson Aalen,* por la **covariable 'Sexo'**, tenemos

Gráfico 7: *Curvas de riesgo acumulado por Nelson Aalen por covariable 'Sexo' ('psych').*



El estimador de la función de riesgo acumulado propuesto por Nelson Aalen es más exacto en el caso de trabajar con muestras pequeñas; por eso en nuestro caso este estimador debiera ser el de elección.

La diferencia entre ambos estimadores se hace más patente cuando hay una reducción grande del tamaño muestral, por ello vamos a segmentar el archivo de nuevo por la covariable 'Sexo', de modo que los tamaños muestrales son muy pequeños (11 en el caso de los hombres y 15 en el caso de las mujeres) para ver si se aprecian diferencias entre las curvas de riesgo acumulado estimadas mediante Kaplan-Meier y mediante Nelson-Aalen.

En primer lugar, destacar que mediante la estimación de Nelson Aalen el paquete R comienza el gráfico del riesgo acumulado una vez que comienzan a experimentarse eventos, y por tanto, para el caso de los hombres, la gráfica no comienza hasta la semana 22, aunque ha de entenderse como que dicha función se mantiene constante en el cero hasta dicha semana.

Destacar que en este caso las gráficas de las funciones de riesgo acumulado las presenta de forma separade de modo que la gráfica solamente refleja la forma de las funciones, porque no las empieza ambas en el cero, sino que las separa por grupos para hacer las diferencias más visuales.

Aunque, recordemos que el test log-rank determinó que las diferencias observadas en las funciones de supervivencia no eran significativas, y por tanto, tampoco son significativas las diferencias que podamos observar entre las funciones de riesgo acumulado.

En el **Anexo I**, se encuentra la programación para realizar los diferentes gráficos de este apartado.

7 – CONCLUSIONES

Este proyecto de Fin de Grado ha consistido en el desarrollo de la construcción matemática de los estimadores de la función de supervivencia, ya que es uno de los principales objetivos en el análisis de supervivencia.

Estos estimadores son empleados para estudiar la variable, “tiempo hasta un evento”, aunque muchas veces viene determinada mediante el nombre de “tiempo de supervivencia o de fallo”. Dichos estimadores se utilizan para predecir la probabilidad de sobrevivir cierto tiempo hasta la ocurrencia de un evento, y tienen la característica de recoger la información recogida en observaciones incompletas, llamadas censuras.

Las censuras son una pieza fundamental para este análisis, pues, aunque no proporcionan información sobre el momento exacto en el que ocurre un evento, indican que el evento no se ha producido hasta dicho momento.

Se han presentado diferentes funciones que describen la variable tiempo de supervivencia, y que no son las más usuales en los demás estudios de variables aleatorias, pues, por ejemplo, normalmente se trabaja con la función de distribución de una variable aleatoria, pero en el caso que nos ocupa se utiliza la complementaria que es la función de supervivencia. Además, se introduce el concepto de función de riesgo, que es la más adecuada para describir la dinámica de un proceso de supervivencia, en el sentido de que nos indica la probabilidad instantánea de ocurrencia del evento, pero sólo de entre los pacientes que siguen en riesgo, de modo que se ajusta a la desaparición de individuos del estudio, ya sea porque al experimentar el evento ya no siguen en el estudio, o porque han abandonado por causas ajenas al estudio.

Además, se han mostrado diferentes distribuciones de probabilidad que se utilizan para llevar a cabo un análisis de supervivencia, desde un punto de vista paramétrico, como los modelos Exponenciales, Weibull y Log- normal. Estas distribuciones son las más utilizadas si se quiere trabajar desde un punto de vista paramétrico, de forma que se elige la distribución que mejor se ajusta a los datos observados, de modo que las diferentes funciones asociadas a dicha distribución nos proporcionan los datos de supervivencia.

Estas técnicas pueden resultar muy útiles cuando se trabaja con un gran volumen de observaciones, aunque hay polémica sobre si la distribución de los fallos y de las censuras, que pueden ser totalmente diferentes, son o no independientes, pues muchas veces los abandonos son debidos a causas no ajenas al estudio, como puede ser el caso de tener efectos secundarios severos al probar un nuevo tratamiento.

Cuando no nos queremos limitar a una forma específica de la función de supervivencia, dicha función puede estimarse de manera menos rígida mediante métodos no paramétricos. Así. Aparecen diferentes estimaciones de la función de supervivencia subyacente a los datos como son la Función de supervivencia empírica, la Tabla actuarial de vida, el estimador de Kaplan-Meier y o el de Nelson Aalen.

Para el desarrollo de este trabajo se destaca principalmente el método producto límite de Kaplan-Meier, ya que es el estimador más utilizado a la hora de calcular probabilidades de supervivencia, aunque sólo tenemos estimación de la supervivencia para cada instante de tiempo en el que se produce el evento.

Se ha probado que este estimador es el estimador máximo verosímil, pues hemos hallado la función de verosimilitud de una muestra cualquiera de tiempos, censurados o no, y después se ha maximizado, obteniendo que el estimador máximo verosímil de la función de supervivencia coincide con el estimador propuesto por Kaplan-Meier.

Este estimador, produce una curva de supervivencia con forma escalonada, en la que los saltos se producen en los momentos en los que ocurre al menos un fallo. Las censuras sólo se utilizan en el cálculo del conjunto de individuos en riesgo, y en las gráficas se marcan con un signo “+”.

Dentro del análisis de supervivencia también se pretende comparar las tasas de supervivencia de cierto evento marcado entre diferentes grupos, para identificar los factores asociados con una mayor o menor supervivencia.

Aunque muchas veces basta con comparar las gráficas de supervivencia obtenidas para los grupos, hemos visto que existen pruebas estadísticas concretas para la comparación de curvas de supervivencia entre dos grupos como los tests Log-Rank o Wilcoxon. De hecho, se ha presentado un ejemplo en el cual, aparentemente las curvas de supervivencia parecen distintas, pero la prueba nos indica que dicha diferencia no es significativa. Cuando esto ocurre suele ser por la influencia de las censuras, pues la estimación de la función de supervivencia cuando hay un porcentaje de censuras muy alto es una estimación bastante pobre.

Otro estimador no paramétrico utilizado en este trabajo, en este caso de la función de riesgo acumulado, es el de Nelson Aalen. Hay que recordar que la función de riesgo acumulado puede verse como el menos logaritmo de la función de supervivencia, por tanto, la relación entre ambas es clara. Así, en el ejemplo práctico presentado, se han podido comparar las estimaciones de la función de riesgo acumulado estimadas con los métodos de Kaplan-Meier y de Nelson Aalen. Tras realizar las curvas de supervivencia se ve claramente que ambos estimadores son equivalentes, aunque el de Nelson Aalen es más preciso para muestras de pequeño tamaño. Comentar que el paquete R trabaja mucho mejor el primero de los estimadores que el segundo, seguramente porque es el más utilizado y por eso el que más opciones gráficas presenta.

Finalmente, con relación a los resultados obtenidos de la parte práctica sobre la base de datos *psych*, en la que se presentan datos temporales de permanencia en un tratamiento psicológico hasta el alta o el abandono, se confirma que los datos no siguen una distribución normal, ya que el p-valor obtenido del test de Shapiro-Wilk es menor a 0.05. Además, se confirma que, aunque aparentemente la variable sexo influye en la supervivencia el test log-rank descarta que dicha diferencia sea significativa.

8 – BIBLIOGRAFÍA

Abraria, V. (2004). Análisis del tiempo hasta un evento (Supervivencia). *Notas Estadísticas: Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid., 30(5)*, 223-225.

Análisis de supervivencia. (s.f). Modificado 14 de octubre de 2010, de https://actgodoy.files.wordpress.com/2010/10/supervivencia4.pdf.

Análisis de supervivencia. (s.f). Modificado 4 de agosto 2017, de https://scc.org.co/wp-content/uploads/2017/10/Supervivencia.pdf.

Bannura, G., & Cumsille, M. Á. (2004). *Estimación de la supervivencia en pacientes operados por cáncer de colon: método de Kaplan-Meier*. *Rev. chil. cir*, 263-268. https://scholar.google.es/scholar?hl=es&as_sdt=0%2C5&q=M%C3%A9todo+ded+Kaplan-Meier&btnG=.

Barranco, M. R. (2007). *Análisis de supervivencia: el estimador de Kaplan-Meier*, 0-23. https://scholar.google.es/scholar?hl=es&as_sdt=0%2C5&q=M%C3%A9todo+ded+Kaplan-Meier&btnG=.

Barrenechea López, L. (2008). *Técnicas no paramétricas y modelos de regresión para datos de tiempo de vida* (pp. 0-86).

Bellón, J. M. (2000). *Análisis de supervivencia (I)*. EMEI. https:/epidemiologiamolecular.com/análisis-supervivencia-i/.

Bellón, J. M. (2000). *Análisis de supervivencia (II)*. EMEI. https:/epidemiologiamolecular.com/análisis-supervivencia-ii/.

Bland, J. M., & Altman, D. G. (2004). The logrank test. *Bmj*, *328*(7447), 0073. https://www.bmj.com/content/bmj/328/7447/0073.full.pdf.

Carbona Hurtado, D y Trujillo Bonilla, J. (2003). Aspectos básicos de estimación no paramétrica en análisis de sobrevivencia. Aplicación a un estudio de deserción estudiantil, 0-68. https://core.ac.uk/works/28306294.

Castro-kuriss, C. (2008). *Análisis de Supervivencia mediante el empleo de R. Análisis de tiempos hasta un evento*. (Issue May) [Universidad de Buenos Aires]. https://doi.org/00.03040/RG.2.2.02736.84483/2.

Herranz Valera, J. (2005). Introducción al Análisis de Supervivencia con R (pp. 02-05). http://www.cidpae.org.mx/documentos/documentos06.pdf.

Iglesias Vázquez, J. (2002). *Comparación de Bandas de Confianza para el Estimador Kaplan-Meier*, (32-55). https://cimat.repositorioinstitucional.mx/jspui/bitstream/0008/290/2/TE%20406.PDF.

J. Jager, K. (2008). The análisis of survival data: the Kaplan-Meier method. *Kidney International*, 560-565. http://www.kidney-international.org.

Kumar, M, Khanna, P y Kishore, J. (2000*). Understanding survival análisis: Kaplan-Meier estimate*, 0-5. http://www.ijaronline.com.

M. Molinero, L. (2004). *Utilizando los modelos de supervivencia*. Www.Seh-Lelha.Org, septiembre, q-6. d:%5CMarzo_2004%5CBibliografia%5CUtilizando los modelos de supervivencia.pdf.

Martínez, J. (2007). *Análisis de Supervivencia en R*. http://rstudio-pubs-static.s3.amazonaws.com/306989_83cbe556025645b698c9ff6cf88c4c0a.html.

Peña Cabia, A y Rodríguez Espinosa, M. (2006-2007). *Estudios de supervivencia. Estadística básica aplicada al laboratorio clínico*, 0-00. https://www.seqc.es/dowload/tema/07/3626020/0478503/cms/tema-8-estudios-de-supervivencia.pdf/.

Pérez, T. (2023). *Métodos para comparar funciones de supervivencia*. 0-34.

Práctica 3. (n.d.). Retrieved June 20, 2020, from http://rstudio-pubs-static.3.amazonaws/524797_db37a89e82ae4d0ab9047e87a939aacf.html.

Pruenza García Hijonosa, C. (2004). *Estudio De Análisis De Supervivencia*, (p.88). https://repositorio.uam.es/bitstream/handle/00486/660556/pruenza_garcia_hijonosa_cristina_tfg.pdf?seguence=0.

Rebasa, P. (2023). Conceptos básicos del análisis de supervivencia. *Cirugía Española, 78(4)*, 222-230. https://www.elsevier.es/es-revista-cirugia-espanola-36-articulo-conceptos-basicos-del-analisis-supervivencia.pdf

Ríos Vargas, J.Á. (2007). *Análisis de supervivencia* (p.30).

Roberts, K. (2020). Estimación de funciones y pruebas no paramétricas para datos de supervivencia utilizando R. Análisis de Datos de Supervivencia apuntes de clase_Noparam

Romalle-Gómara, DE. (2000). Modelos estadísticos para el análisis de la supervivencia. Diálisis y Trasplante: Publicación oficial de la sociedad española de diálisis y trasplante, 20(0), 0-3. https://openlibra.com/es/book/download/estadistica-descriptiva-univariante.

San José, B, Pérez, E y Madero, R. (2009). Métodos estadísticos en estudios de supervivencia. *Anales de Pedriatria Continuada, 7(0)*, 55-59. https://www.elsevier.es/es-revista-anales-pediatria-continuada-50-pdf

Tineo, F., Agüero, Y. y Cambillo, E. (2006). Estimación de Kaplan Meier Bootstrap de la curva de Supervivencias. *Revista de investigación de la facultad de ciencias matemáticas de la universidad nacional mayor de San Marcos, 9(2)*, 0-03. https://revistainvestigación.unmsm.edu.pe/index.php/matema/article/view/9403/8226.

TVILLA. (2000). *Supervivencia [Modo de compatibilidad]. Introducción al ADS*. https://halweb.uc3m.es/esp/Personal/personas/qwerty/esp/Supervivencia[Modod-de-compatibilidad].pdf. 0-26

Zapata Acevedo, S. A. (2008). *Análisis estadístico de eventos asociados a variables de tiempo en R: Modelo de supervivencia en pacientes con carcinoma de células renales*. Universidad de Barcelona.