Survival Analysis

Análisis de Supervivencia y Fiabilidad

Máster Universitario en Data Science

Profesor: Ángel Berián, Ricard Sierra & Xavier Vilasís

Índice de la Sesión

1. ¿Qué es el análisis de supervivencia?

Estudio del **tiempo hasta la ocurrencia de un evento**.

Diferencia crítica frente a la regresión/clasificación estándar:

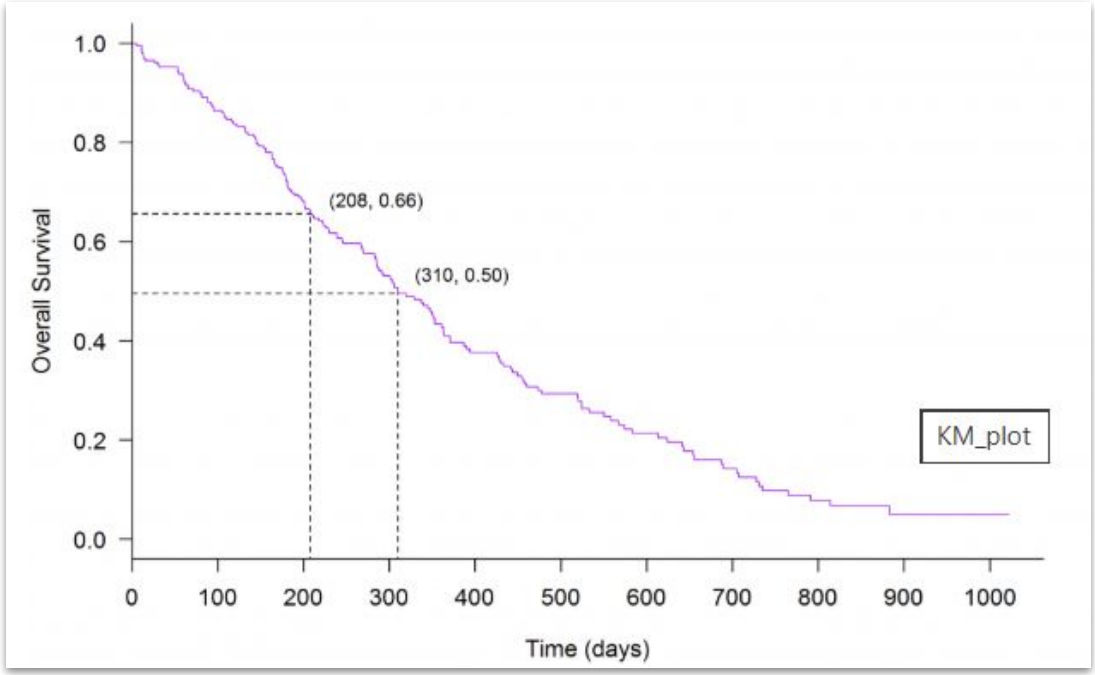
- Manejo explícito del tiempo.
- Tratamiento de la **censura** (datos incompletos).

Ámbitos de Aplicación

Medicina: Tiempo hasta muerte o recaída.

Ingeniería: Fiabilidad y tiempo hasta fallo.

Negocio: Churn de clientes.



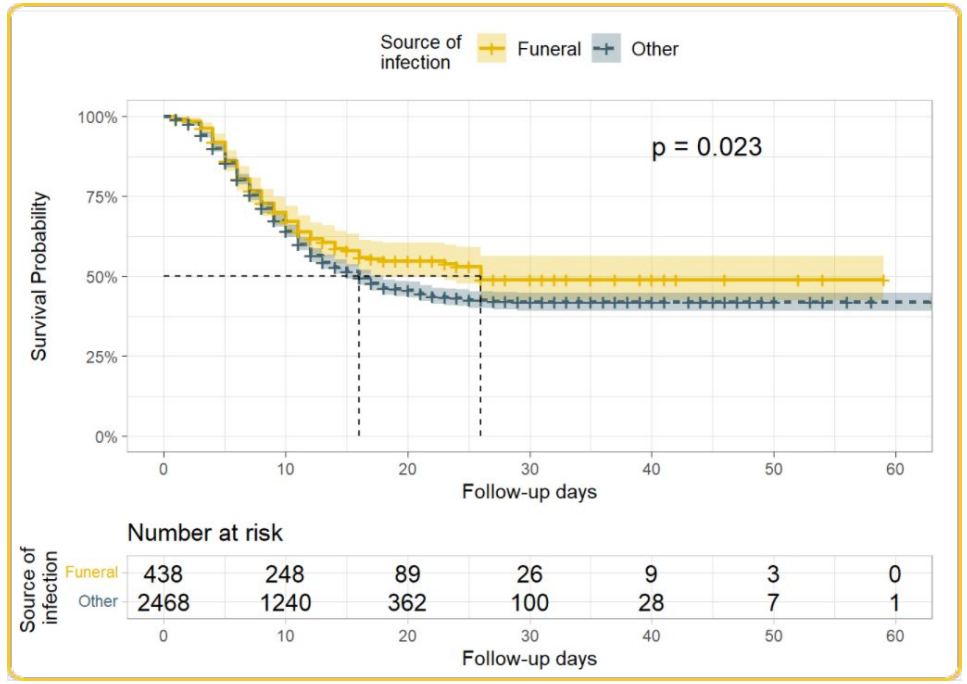
Fiabilidad y Supervivencia

La función fundamental que define la probabilidad de que el evento **no** haya ocurrido aún.

$$S(t) = P(T > t)$$

Interpretación: Probabilidad de "sobrevivir" más allá de un tiempo dado *t*.

Evento (T): Fallo, muerte, abandono, cancelación.



Equivalencia Conceptual

Ingeniería

Análisis de **Fiabilidad**.

Estudio del "Tiempo hasta el fallo" (Time-to-failure).

Medicina / Negocio

Análisis de **Supervivencia**.

Estudio del tiempo hasta el evento (muerte, churn).

Base Matemática

Misma formulación probabilística, distinto contexto semántico.

Por qué es clave en Data Science

Datos Incompletos

Permite trabajar con información **censurada** (sabemos que el cliente sobrevivió hasta hoy, pero no cuándo se irá).

Decisiones Dinámicas: Basadas en riesgo cambiante a lo largo del tiempo, no estático.

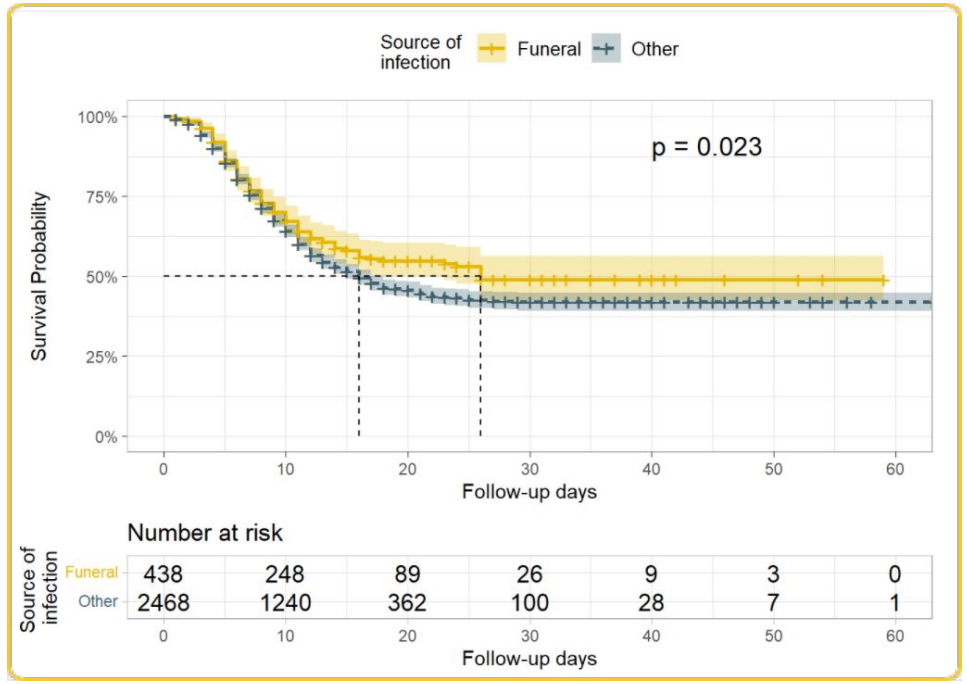
Modelado Interpretable: Entender cómo el tiempo afecta la probabilidad, crucial para mantenimiento predictivo y LTV (Lifetime Value).

2. Función de Supervivencia S(t)

$$S(t) = P(T > t) = 1 - F(t)$$

Propiedades:

- Función no creciente.
- S(0) = 1 (Al inicio, todos sobreviven).
- S(∞) = 0 (Eventualmente, el evento ocurre).



Función de Riesgo (Hazard-ratio)

Definición h(t)

tasa instantánea de ocurrencia del evento en t, condicionada a T ≥ t

- **No es una probabilidad:** Es una tasa (velocidad); puede ser > 1
- **Relación fundamental:**

$$h(t) = -\frac{d}{dt}\log S(t), \qquad H(t) = \int_0^t h(u)\,du$$

- Relación directa con la dinámica del evento (¿el riesgo aumenta o disminuye con el tiempo?).

Riesgo Acumulado H(t)

Definición H(t)

riesgo total acumulado hasta el tiempo t.

- **Relación con el hazard:**

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

- **Conexión clave con la supervivencia:**

$$S(t) = \exp(-H(t))$$

- **Interpretación**: suma de exposiciones al riesgo a lo largo del tiempo.
- facilita el análisis teórico y la estimación de S(t).

Esperanza de Supervivencia *E[T]*

- **Definición:** tiempo medio hasta la ocurrencia del evento.
- **Interpretación**: área bajo la curva de supervivencia.
- **En fiabilidad**: Mean Time Between Failures (MTBF).
- **Expresión general:**

$$E[T] = \int_0^\infty S(t)\,dt$$

Limitación

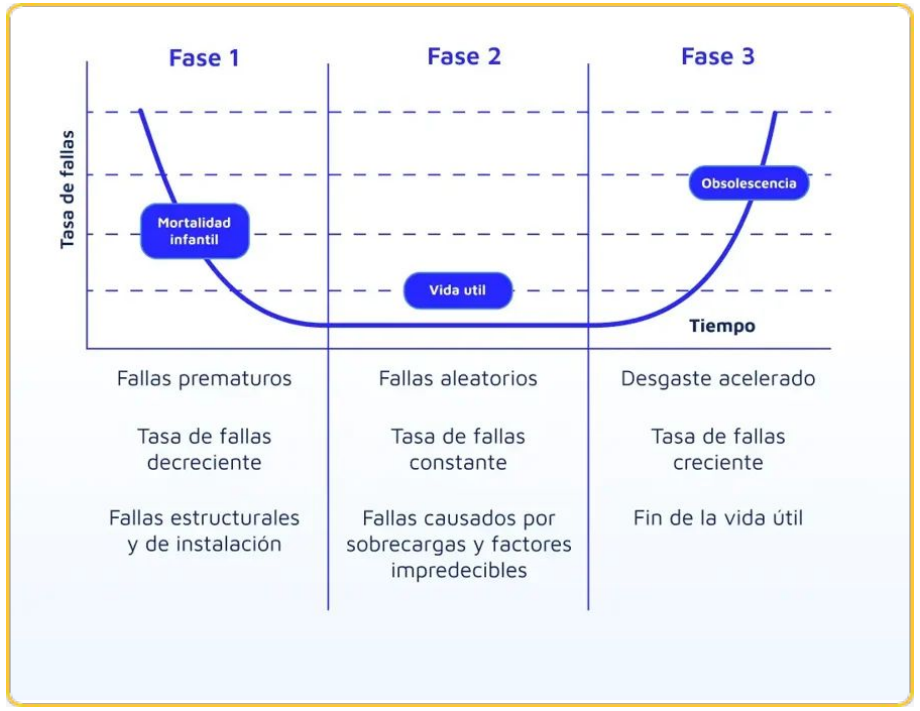
Es muy sensible y difícil de estimar con precisión cuando hay **censura pesada** (si no observamos el final de la vida de muchos sujetos, la media es desconocida).

3. Modelado del Riesgo: Curva de Bañera

Modelo fundamental en ingeniería de fiabilidad que visualiza la tasa de fallos de un activo a lo largo de su ciclo de vida, mostrando tres fases:

- **Fase 1: Mortalidad Temprana.** Fallos tempranos, riesgo decreciente (defectos de fabricación).
- **Fase 2: Vida Útil.** Riesgo constante (fallos aleatorios).
- **Fase 3: Desgaste.** Riesgo creciente (envejecimiento).

En función de la fase, se pueden deducir las causas de los errores.



Modelos Paramétricos

Concepto

Asumimos que el tiempo T sigue una distribución de probabilidad conocida (forma funcional definida).

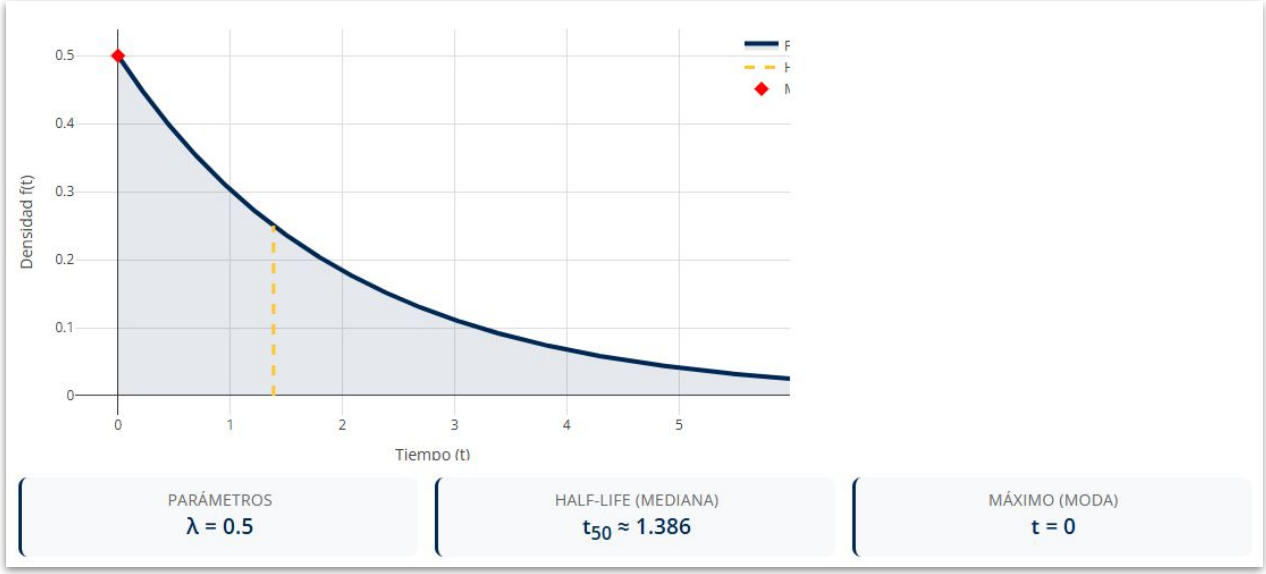
Ventajas

Si el modelo es correcto, las estimaciones son más suaves, precisas y eficientes que los métodos no paramétricos. Permiten extrapolar más allá de los datos observados.

Riesgo Constante: Exponencial

h(t) = λ (constante)

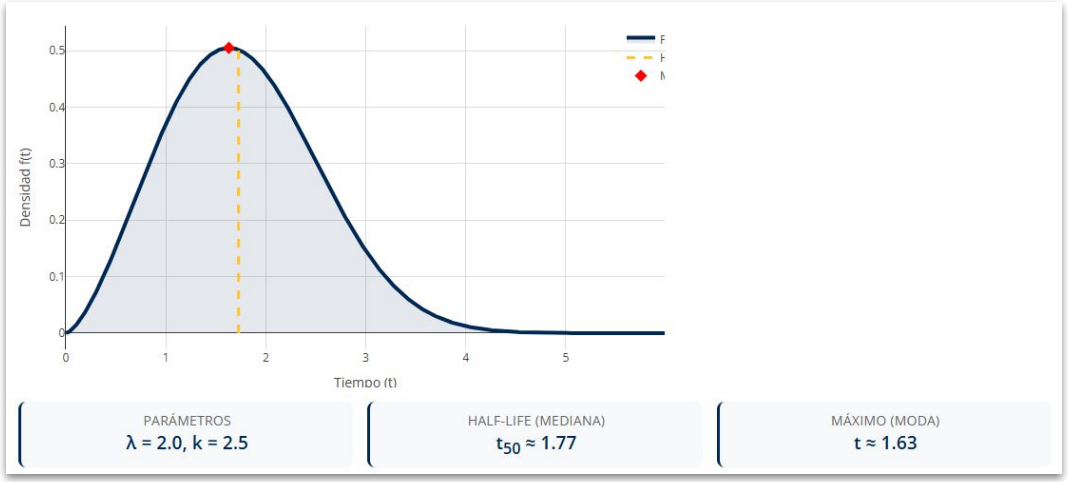
- **Memoria Nula:** La probabilidad de fallar mañana es la misma si el componente es nuevo o si tiene 10 años.
- Caso base más simple (Baseline).
- Útil para la fase central de la curva de bañera.



Modelo Weibull

- **Definición:** modelo paramétrico flexible para tiempos de fallo.
- **Función de riesgo:**

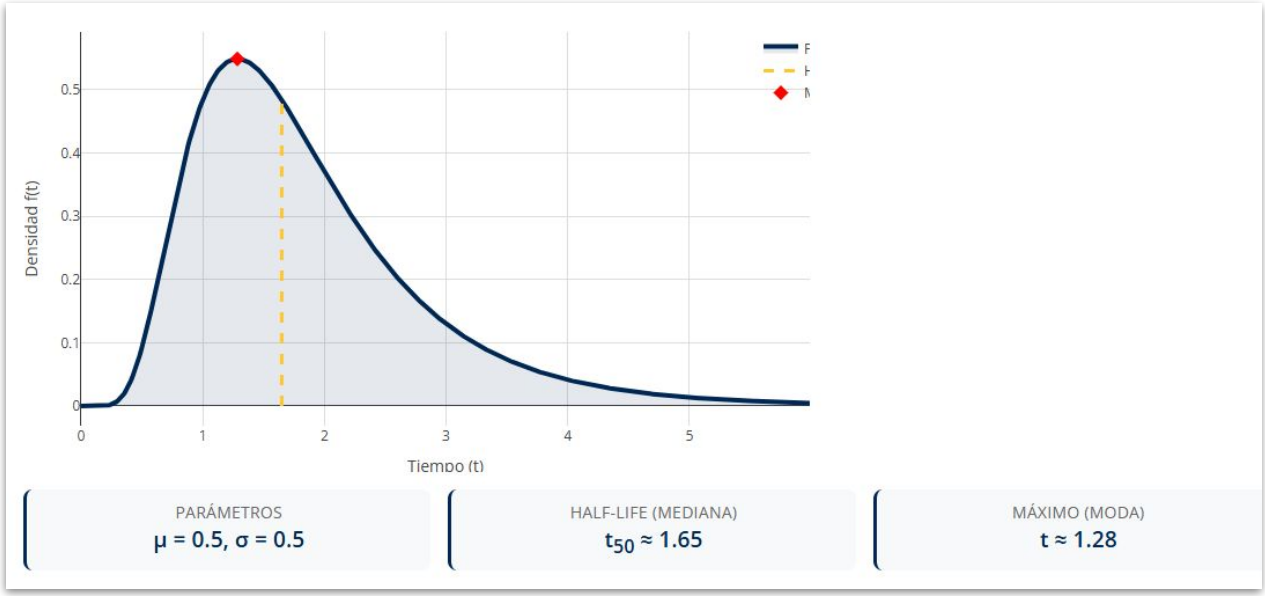
  donde k = shape, λ = scale.
- **Parámetro Shape (k)**
  - k<1: riesgo decreciente (mortalidad infantil).
  - k=1: riesgo constante (caso exponencial).
  - k>1: riesgo creciente (desgaste).
- **Parámetro Scale (λ)**
  - Escala temporal del proceso.
  - A mayor λ, mayor vida característica del sistema
- **Half-life (mediana)**
- **Máximo de la densidad (modo)**
- **Uso típico:** fiabilidad, mantenimiento predictivo y ciclos de vida.



Modelo Lognormal

Riesgo no monótono

- El riesgo aumenta al principio, alcanza un pico y luego disminuye.
- Común en eventos dominados por procesos latentes biológicos (ej. tiempo de incubación).
- Casos de pico intermedio de riesgo donde los supervivientes a largo plazo tienen menor riesgo que al inicio.



4. Modelos con Covariables

¿Por qué no basta con S(t)?

El riesgo no depende solo del paso del tiempo.

Factores externos (edad, tratamiento, temperatura, presión) modifican la probabilidad del evento.

Necesitamos modelos de **Regresión**.

Modelo de Riesgos Proporcionales de Cox

- **Definición**: modelo semi-paramétrico que relaciona covariables con el riesgo.
- **Función de riesgo:**

$$h(t \mid x) = h_0(t) \exp(\beta^\top x)$$

donde $h_0(t)$ es el riesgo base.

- **Interpretación:**
  - Las covariables actúan multiplicando el riesgo base.
  - No se especifica la forma de $h_0(t)$

- **Supuesto clave:**
  - Riesgos proporcionales: el HR es constante en el tiempo.
  - Las curvas de supervivencia no deben cruzarse.
- **Ventaja principal:**
  - Aísla el efecto de covariables sin asumir distribución del tiempo.

Hazard Ratio (HR)

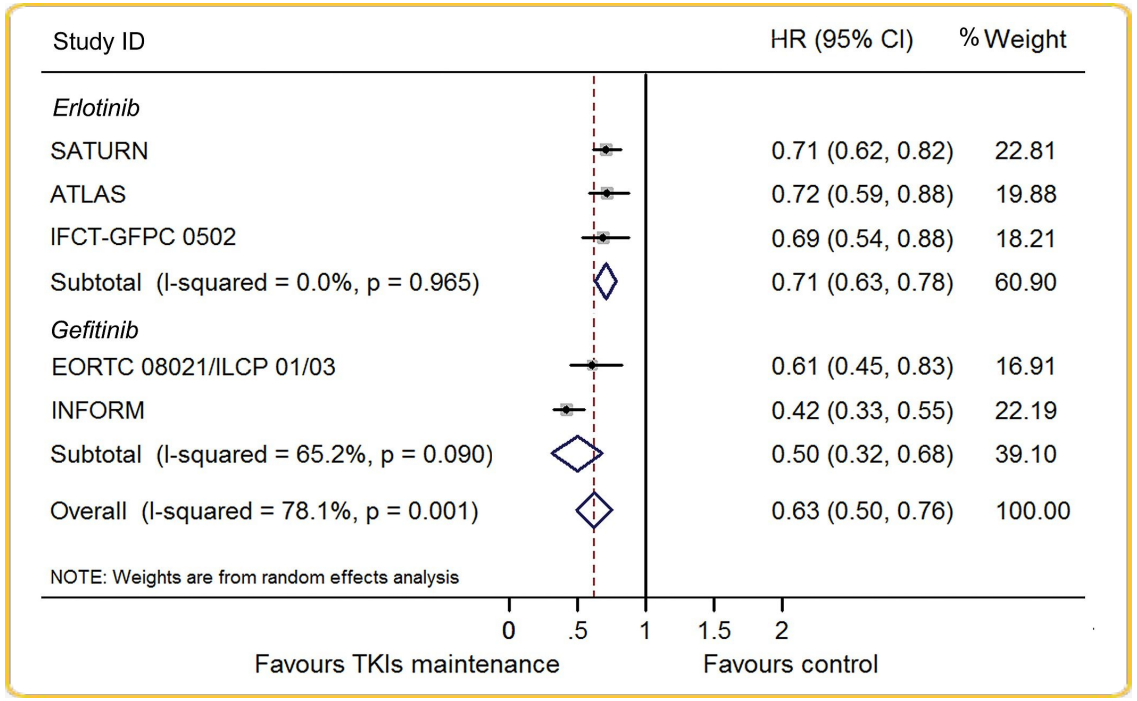
Interpretación directa del efecto de una covariable:

HR = 1: Sin efecto.

HR > 1: Aumenta el riesgo (Factor de riesgo).

HR < 1: Reduce el riesgo (Factor protector).

Uso central en inferencia aplicada (ej. "Fumar aumenta el riesgo 2x").



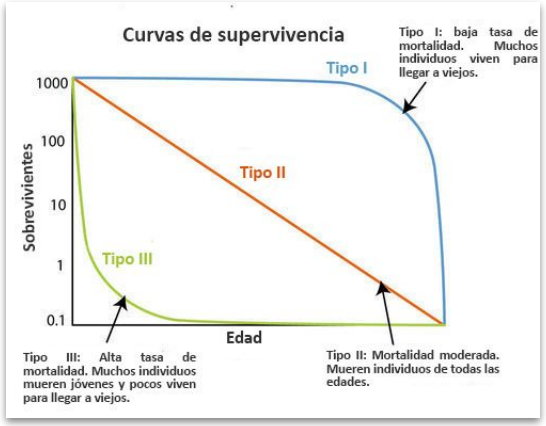
Supuesto de Riesgos Proporcionales

Hipótesis Clave

El Hazard Ratio (HR) es **constante** en el tiempo. Si un tratamiento reduce el riesgo a la mitad hoy, también debe reducirlo a la mitad dentro de un año.

Consecuencias prácticas: Si no se cumple, el modelo de Cox es inválido.

Señal de violación: Curvas de supervivencia que se cruzan.



5. Estimación: Censura & Truncamiento

Censura a la derecha

Ocurre cuando un individuo es observado desde un tiempo inicial t_0 hasta un tiempo t_C sin que el evento de interés haya ocurrido.

- Sabemos que el evento ocurre después de t_C, pero no cuándo.

Censura a la izquierda

Se da cuando sabemos que el evento ocurrió antes de un cierto tiempo, pero desconocemos el momento exacto.

- El evento ocurrió antes del tiempo de censura.

Truncamiento a la derecha

Ocurre cuando la población del estudio solo incluye individuos que ya han experimentado el evento.

- No se observan tiempos largos de supervivencia.

Truncamiento a la izquierda

Se produce cuando no se incluyen en el estudio individuos que ya habían superado un cierto hito antes de comenzar la observación.

- Solo entran individuos “que sobreviven lo suficiente” para ser observados.

Estimación Empírica

Objetivo

Construir funciones (S, h) a partir de datos observados.

Desafío

No podemos usar simplemente "Total Eventos / Total Sujetos" debido al tiempo variable de exposición.

Likelihood

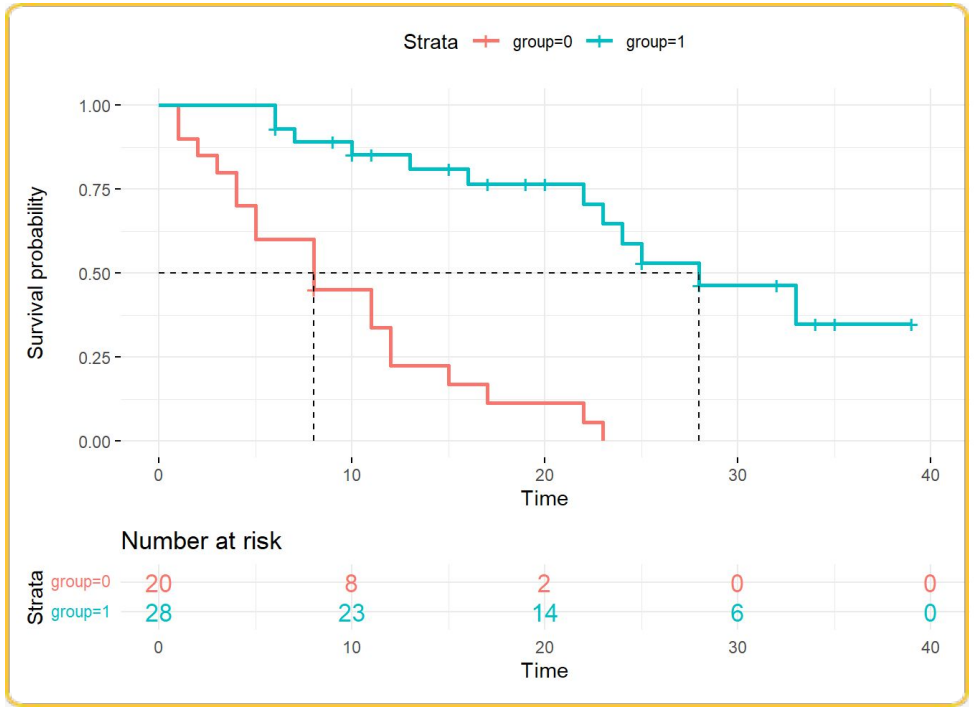
Papel crucial de la función de verosimilitud ajustada por censura.

Estimador Kaplan–Meier

- Estimador **No Paramétrico** de S(t).
- **Construcción:**

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

  donde d_i = eventos en t_i, n_i = individuos en riesgo
- **Manejo de censura:**
  - Incorpora censura a la derecha de forma natural.
  - La censura no reduce S(t), solo n_i
- **Interpretación gráfica:**
  - Curva escalonada decreciente.
  - Caídas solo en tiempos de evento.
- **Uso típico:**
  - Estimación descriptiva y comparación visual de grupos.



6. Comparación e Inferencia

Comparación Visual

Curvas Kaplan–Meier superpuestas por grupo (Tratamiento vs Control).

Test Log-Rank (Mantel-Cox): Prueba de hipótesis no paramétrica.

Hipótesis Nula: No hay diferencia en las curvas de supervivencia entre los grupos.

Se basa en comparar lo observado vs lo esperado en cada instante de tiempo.

Inferencia y Validez

Homogeneidad
Importancia de que los grupos sean comparables.

Confusión
Factores ocultos (Confounders) que afectan tanto al tratamiento como a la supervivencia.

Riesgo

Conclusiones causales erróneas si no se ajusta por covariables (usando Cox, por ejemplo).

Aplicaciones Prácticas

Medicina: Eficacia de tratamientos oncológicos.

Ingeniería: Mantenimiento predictivo en flotas industriales.

Negocio: Modelado de **Churn** y retención de suscriptores.

Churn

En el ámbito empresarial, el churn se define como el momento en que un cliente **cancela un servicio o suscripción** que estaba utilizando (por ejemplo, Spotify o Netflix).

La **predicción de churn** consiste en identificar qué clientes tienen mayor probabilidad de **abandonar la empresa**, a partir de su comportamiento y uso del servicio.

Captar nuevos clientes suele ser **más costoso y difícil** que retener a los actuales. Por ello, la predicción de churn permite **detectar clientes de alto riesgo** y orientar acciones de retención de forma más eficiente.

7. Ejemplo Práctico: Churn

Enunciado

Una empresa de **suscripción digital (SaaS)** quiere analizar el **tiempo hasta la cancelación** de sus clientes desde que se registran en la plataforma.

Objetivo

- Estimar la **probabilidad de que un cliente continúe activo** después de cierto tiempo.
- Comparar la supervivencia entre dos tipos de clientes:
  - Clientes con **plan básico**
  - Clientes con **plan premium**

Variable de interés

T: tiempo (en meses) desde el alta hasta la cancelación del servicio.

Censura

No todos los clientes han cancelado al finalizar el período de observación:

- Algunos siguen activos → **datos censurados a la derecha.**

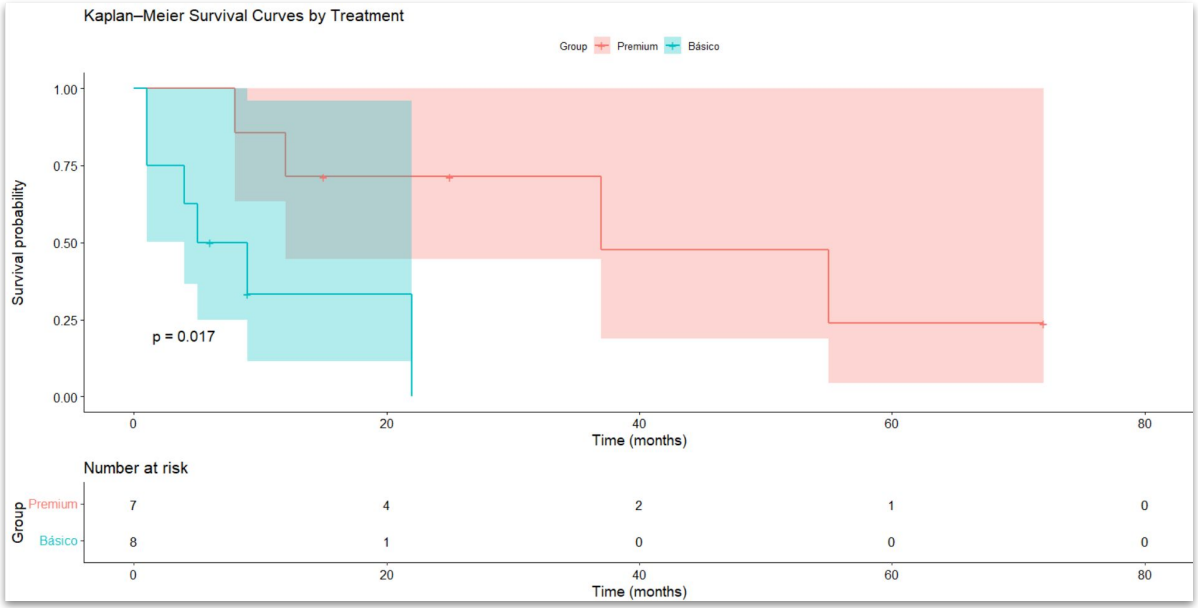
Datos observados

Para cada cliente se registra:

- **Tiempo**: meses desde el alta hasta cancelación o fin del estudio
- **Evento**
  - 1 = cancelación
  - 0 = censurado (cliente sigue activo)
- **Plan**: básico / premium

Resolución

Cliente	Tiempo (meses)	Evento	Plan
1	3	cancelación (1)	Básico
2	5	censura (0)	Premium
3	2	cancelación (1)	Básico
4	8	censura (0)	Premium
5	6	cancelación (1)	Básico
…	…	…	…



Log-rank test

```
Call:
survdiff(formula = Surv(time_months, event) ~ treatment, data = df)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
treatment=Premium 7        4     7.08      1.34      5.68
treatment=Básico  8        6     2.92      3.25      5.68

 Chisq= 5.7  on 1 degrees of freedom, p= 0.02
```

Conclusiones

Resultados

- La curva de **clientes premium** se mantiene por encima de la de **clientes básicos**.
- Para cualquier tiempo t, se observa:
  Sprem(t) > Sbas(t)

Interpretación

- Los clientes premium tienen **menor riesgo de cancelación**
- El plan de suscripción actúa como **factor asociado a la supervivencia**

El Análisis de Supervivencia no modela “si ocurre un evento”, sino cuándo ocurre, incorporando de forma natural la censura, algo crítico en problemas reales de negocio y ciencia de datos.