Pruebas de exponencialidad

Lucia Coudet - Daniel Czarnievicz

Octubre de 2018

Análisis de supervivencia

Análisis de supervivencia

- Métodos utilizados para el análisis de información donde la variable de resultado es el tiempo hasta la ocurrencia de un evento de interés.
- Requiere de técnicas especiales dado que:

1. La información suele seguir distribuciones asimétricas, por lo que la normalidad no es un supuesto razonable.
2. Algunas unidades muestrales pueden no haber llegado al fin del experimiento, por lo que se estará en presencia de datos censurados.

Función de supervivencia

- Se define la **función de supervivencia** como la probabilidad de que el tiempo de supervivencia T sea mayor a un tiempo t:

$$S(t) = P(T \geq t)$$

- Ante presencia de datos censurados, la misma se estima mediante el estimador de Kaplan-Meier:

$$\hat{S}(t) = \prod_{j/t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

donde d_j es el número de individuos que experimentaron el suceso de interés en el momento $t_{(j)}$, y r_j es el número de individuos en riesgo inmediatamente antes del momento $t_{(j)}$.

Función de riesgo

- El objetivo de la función de riesgo es estudiar qué períodos tienen mayor probabilidad de ocurrencia del suceso de interés.
- Se define la **función de riesgo** como la probabilidad de que un individuo experimente el suceso de interés en un intervalo pequeño de tiempo, s, dado que ha sobrevivido hasta el inicio del intervalo, para cualquier intervalo cuya longitud tiende a cero:

$$h(t) = \lim_{s \to 0} \frac{P(t \leq T \leq t + s | T \geq t)}{s}$$

donde T son los tiempos de supervivencia de los individuos.

- Estimador de la función de riesgo:

$$\hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}$$

Tiempo de vida

- El complemento de la función de supervivencia es la **función de tiempo de vida**: $F(t) = P(T \leq t) = 1 - S(t)$.
- Toda función para la cual $F(a) = 0 \ \forall a < 0$ es una distribución de vida.
- Existen muchas clases de distribuciones de vida, por ejemplo:
  - IFR: increasing failure rate.
  - DFR: decreasing failure rate.
  - NBU: new better than used.
  - DMRL: decreasing mean residual life.

Tiempo de vida

- A su vez, existen muchas familias de distribución que pueden describir distintas clases de funciones de vida, dependiendo de su parametrización. Por ejemplo:

	DFR	CFR	IFR
Exponencial		$\forall\lambda > 0$	
Weibull	$\alpha < 1$	$\alpha = 1$	$\alpha > 1$
Gamma	$\alpha < 1$	$\alpha = 1$	$\alpha > 1$

- Debido a estas relaciones es que el análisis de supervivencia está intimamente relacionado con las pruebas de exponencialidad. Los investigadores buscan estudiar cómo cambia la probabilidad de ocurrencia del evento de interés.

Failure rate function

$$r(x) = \frac{f(x)}{\bar{F}(x)}$$

en donde $\bar{F}(x) = 1 - F(x)$

Interpretaci?n

$r(x)\delta_x$ es la probabilidad de que un ítem (unidad, persona, parte) viva a la edad x fallezca en el intervalo $(x, x + \delta_x)$ con δ_x pequeño. Si:

- $r(x)$ creciente, entonces la tasa de fallecimiento crece conforme crece la edad.
- $r(x)$ decreciente, entonces la tasa de fallecimiento decrece conforme a la edad.
- $r(x)$ constante, entonces la tasa de fallecimiento ni crece ni decrece con la edad (independencia).

Clases de distribuciones de vida

Una distribución de vida F es de la clase:

- **IFR (increasing failure rate)**, si $r(x)$ es estrictamente no decreciente.
- **DFR (decreasing failure rate)**, si $r(x)$ es estrictamente no creciente.

Es decir:

- $r(x)$ es IFR si $r(x) \leq r(y) \, \forall \, x < y$.
- $r(x)$ es DFR si $r(x) \geq r(y) \, \forall \, x < y$.

The increasing failure rate average class (IFRA)

$r(x)$ la tasa de fallo puede tener una tendencia creciente pero no ser necesariamente no decreciente, como es requerido en la clase IFR.

La clase IFRA considera el caso en el que $r(x)$ fluctúa, por ejemplo, debido a variaiones estacionales. En aplicaciones médicas una tasa $r(x)$ inicialmente creciente puede decrecer debido a una intervención médica.

De forma análoga definimos la clase **DFRA** (decreasing failure rate average).

Observe que:

- $IFR \subset IFRA$
- $DFR \subset DFRA$

Distribución exponencial

Observación

Las distribución exponencial pertenece a la clase *IFR* y también a la *DFR* y a su vez es la única distribución que es *IFR* y *DFR* a la vez.

Lo mismo para las clases *IFRA* y *DFRA*.

Test contra alternativas IFR y DFR

Hipótesis nula de exponencialidad

$$H_0 : r(x) = \lambda$$

La hipótesis nula de exponencialidad implica suponer una **tasa constante**, es decir **independiente de la edad**, el cual es el supuesto de la distribución exponencial. A veces se dice que la distribución exponencial *no tiene memoria*, en el sentido que la edad no afecta la probabilidad de muerte.

Por lo tanto, podemos escribir la hipótesis nula como:

$$H_0 : F(x) = (1 - e^{-\lambda x})\mathrm{I}_{[x \geq 0]}$$

ó,

$$H_0 : \bar{F}(x) = e^{-\lambda x}\mathrm{I}_{[x \geq 0]}$$

Prueba y estadístico de prueba

Sea $X_1, \ldots, X_n$ una muestra de X y sean $X_{(1)}, \ldots, X_{(n)}$ los estadísticos de orden, con $X_{(0)} = 0$. Se consideran los espacios normalizados $D_1, \ldots, D_n$:

$$D_i = (n - i + 1)(X_{(i)} - X_{(i-1)})$$

Entonces tenemos que:

$$\begin{aligned} D_1 &= nX_{(1)} \\ D_2 &= (n-1)(X_{(2)} - X_{(1)}) \\ D_3 &= (n-2)(X_{(3)} - X_{(2)}) \\ &\vdots \\ D_n &= 1(X_{(n)} - X_{(n-1)}) \end{aligned}$$

Estadístico de prueba

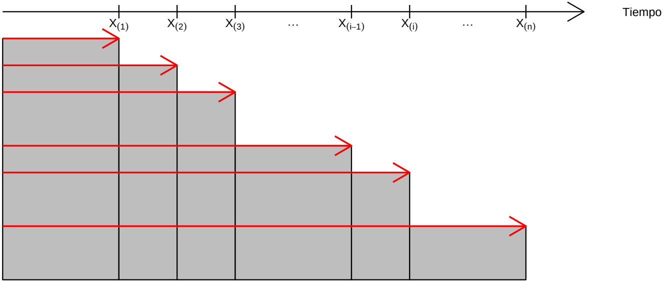
A partir del tiempo total en prueba al momento X_i:

$$S_i = \sum_{u=1}^{i} D_u \quad i = 1, \ldots, n$$

se define el estadístico de prueba de la siguiente manera:

$$\varepsilon = \frac{1}{S_n} \sum_{i=1}^{n-1} S_i$$

Interpretación del tiempo total en prueba al momento $X_{(i)}$



$S_1 = nX_{(1)}$

$+$

$S_2 = (n-1)(X_{(2)}-X_{(1)})$

$+$

$S_3 = (n-2)(X_{(3)}-X_{(2)})$

$+$

$\ldots \qquad \ldots$

$+$

$S_i = (n-i+1)(X_{(i)}-X_{(i-1)})$

$+$

$\ldots \qquad \ldots$

$+$

$S_n = (1)(X_{(n)}-X_{(n-1)})$

Test a una cola con alternativa IFR

- H_0 : F es exponencial
- H_1 : F es IFR (y no exponencial)

Zona de rechazo

$\varepsilon \geq e_\alpha$

Con α tal que $P(\varepsilon \geq e_\alpha) = \alpha$

Test a una cola con alternativa DFR

- H_0 : F es exponencial
- H_1 : F es DFR (y no exponencial)

Zona de rechazo

$\varepsilon \leq \frac{n-1}{2} - e_\alpha$

Test a dos colas con alternativa IFR o DFR

- H_0 : F es exponencial
- H_1 : F es IFR o DFR (y por lo tanto no exponencial)

<u>**Zona de rechazo**</u>:

- $\varepsilon \geq e_{\frac{\alpha}{2}}$
- $\varepsilon \leq \frac{n-1}{2} - e_{\frac{\alpha}{2}}$

Aproximación para muestras grandes

Se había definido al estadístico de prueba ε como:

$$\varepsilon = \frac{1}{S_n} \sum_{i=1}^{n-1} S_i$$

Ahora bien, es posible expresarlo en función de $T_1, \ldots, T_n$ dónde:

$$T_i = \frac{S_i}{S_n} = \frac{\sum_{u=1}^{i} D_u}{\sum_{u=1}^{n} D_u}$$

Entonces:

$$\varepsilon = \sum_{i=1}^{n-1} T_i$$

El cual bajo H_0 corresponde a una suma de VA $U(0,1)$ iid y por lo tanto: $E_0(\varepsilon) = \frac{n-1}{2}$ y $V_0(\varepsilon) = \frac{n-1}{12}$.

Aproximación para muestras grandes

Sea entonces $\varepsilon^* = \frac{\varepsilon - E_0(\varepsilon)}{\sqrt{V_0(\varepsilon)}} = \frac{\varepsilon - \frac{n-1}{2}}{\sqrt{\frac{n-1}{12}}}$.

Aplicando el TLC se cumple que:

$$\varepsilon^* \overset{n\to\infty}{\to} N(0,1)$$

Zonas de rechazo para muestras grandes:

- Alternativa IFR: $\varepsilon^* \geq Z_\alpha$.
- Alternativa DFR: $\varepsilon^* \leq -Z_\alpha$.
- Alternativa IFR o DFR: $|\varepsilon^*| \geq Z_{\frac{\alpha}{2}}$.

Ejemplo: methylmercury poisoning

Methylmercury Poisoning

Se quiere estudiar el efecto de la aplicación de dosis de metilmercurio (un veneno) en el tiempo de vida de los peces, los cuales fueron sometidos a varias dosis del veneno.

Al nivel de una dosis, se obtuvieron los siguientes **tiempos de vida hasta la muerte ordenados** (en días):

$$42, 43, 51, 61, 66, 69, 71, 81, 82, 82$$

Methylmercury poisoning: Test a una cola con alternativa IFR

Se quiere testear la hipótesis nula de exponencialidad (tasa $r(x)$ constante) contra la alternativa de tasa creciente, debido a que se sospecha que cuantos más días pasan más crece la probabilidad de que los peces mueran.

- $H_0 : F$ es exponencial
- $H_1 : F$ es IFR (y por lo tanto no exponencial)

Methylmercury poisoning: Test a una cola con alternativa IFR

```
datos <- sort(c(42, 43, 51, 61, 66, 69, 71, 81, 82, 82))
datos_o <- c(0, datos[-length(datos)])
si <- vector('double', length = length(datos))
for (i in 1:length(datos)){
      if (i==1){
            si[i] <- (length(datos) - i + 1) *
                  (datos[i] - datos_o[i])
      } else {
            si[i] <- (length(datos) - i + 1) *
                  (datos[i] - datos_o[i]) + si[i-1]
      }
}
suma_si <- sum(si[-length(si)])
sn <- si[length(si)]
epsilon <- suma_si / sn
```

Methylmercury poisoning: Test a una cola con alternativa IFR

El estadístico de prueba ε toma el valor 7.7407. Buscamos en la tabla para $n = 10$ y encontramos $P < 0.005$ por lo cual existe evidencia empírica en contra de la hipótesis nula de exponencialidad y a favor de la alternativa IFR.

Observe que $n \geq 9$ por lo cual "podemos" usar la aproximación para muestras grandes de la distribución del estadístico de prueba.

```
en <- (epsilon-(length(datos)-1)/2) /
      sqrt((length(datos)-1)/12)
```

en toma el valor 3.7421 por lo que $P < 0.002$ y por lo tanto la aproximación para muestras grandes confirma la evidencia contra H_0 y a favor de IFR.

Test contra alternativas NBU y NWU

Las clases NBU y NWU

- Las clases **New better than used** (NBU) y **New worst than used** (NWU) buscan modelar fenómemos en los cuales la probabilidad de supervivencia de nuevas unidades es mejor (peor) que aquella de las unidades ya existentes.

- Para ello deben considerarse (al igual que las pruebas anteriores), los objetos de todas las edades que han sobrevidido hasta el momento x.

Relaciones entre clases

- $IFR \subset IFRA \subset NBU \subset NBUE$
- $DFR \subset DFRA \subset NWU \subset NWUE$

Hipótesis de la prueba

Lo anterior se refleja en la hipótesis:

$$H_0 : P(X \geq x + y | X \geq x) = P(X \geq y) \quad \forall x; y \geq 0$$

la cual podemos reescribir utilizando las funciones de vida como:

$$H_0 : \frac{\bar{F}(x+y)}{\bar{F}(x)} = \bar{F}(y) \quad \forall x; y \geq 0$$

Relación con las clases NBU y NWU

- Si $\bar{F}(x+y) \leq \bar{F}(x)\bar{F}(y) \quad \forall x; y \geq 0$, entonces la clase es NBU.
- Si $\bar{F}(x+y) \geq \bar{F}(x)\bar{F}(y) \quad \forall x; y \geq 0$, entonces la clase es NWU.

Estadístico de prueba

- Para construir el estadístico de prueba primero debemos ordenar la muestra.
- Luego computamos el estadístico:

$$T = \sum_{i>j>k} \psi(X_{(i)}; X_{(j)} + X_{(k)})$$

donde:

$$\psi(a; b) = \begin{cases} 1 & \text{si} \quad a > b \\ {}^{1}/_{2} & \text{si} \quad a = b \\ 0 & \text{si} \quad a < b \end{cases}$$

Zonas de rechazo

- Prueba a una cola contra alternativa de NBU: rechazar H0 si $T \leq t_{1;\alpha}$ para un nivel de significación α.
- Prueba a una cola contra alternativa de NWU: rechazar H0 si $T \geq t_{2;\alpha}$ para un nivel de significación α.
- Prueba a dos colas contra alternativa de NBU o NWB: rechazar H_0 si $T \leq t_{1;\alpha_1}$ o si $T \geq t_{1;\alpha_2}$ para un nivel de significación $\alpha_1 + \alpha_2 = \alpha$.

Aproximaciones para muestras grandes

- Se define el estadístico:

$$T^* = \frac{T - E_{H_0}(T)}{\sqrt{V_{H_0}(T)}} \xrightarrow{d} N(0;\,1)$$

- Prueba a una cola contra alternativa de NBU: rechazar H_0 si $T^* \leq -z_\alpha$ para un nivel de significación α.

- Prueba a una cola contra alternativa de NWU: rechazar H_0 si $T^* \geq z_\alpha$ para un nivel de significación α.

- Prueba a dos colas contra alternativa de NBU o NWU: rechazar H_0 si $T^* \leq -z_{\alpha_1}$ o si $T^* \geq z_{\alpha_2}$ para un nivel de significación $\alpha_1 + \alpha_2 = \alpha$.

Ejemplo: methylmercury poisoning (continuación)

Ejemplo

Para este ejemplo no es necesario calcular ψ para todos los trios $i > j > k$ dado que $X_{(10)} < X_{(1)} + X_{(2)}$, por lo que todas valen 0. Por lo tanto, $T = 0$.

Hollander y Proschan (1972) prueban que, para $n \geq 3$, $P_{H0}(T = 0) = \binom{2n-2}{n}^{-1}$

Por lo tanto, $P_{H0}(T = 0) = 0.00002$ para nuestro caso con $n = 10$. Dicho p-valor permite rechazar H0 (exponencialidad), a favor de NBU.

Bibliografía

Bibliografía

- Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
- Hollander, M., Wolfe, D. A., & Chicken, E. (2013). Nonparametric statistical methods (Vol. 751). John Wiley & Sons.
- Hothorn, T., & Everitt, B. S. (2009). A handbook of statistical analyses using R. Chapman and Hall/CRC.
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.