TP Inferencia Estadística

Jesica Charaf e Ignacio Spiousas

25 de junio de 2023

El problema

Una fuente radiactiva emite partículas alfa de acuerdo con un proceso de Poisson de intensidad λ por segundo. Se tiene una fuerte sospecha de que el parámetro desconocido supera 0.5, y con el objetivo de confirmar dicha sospecha se medirán los tiempos entre emisiones consecutivas, obteniendo de esa manera una muestra de variables aleatorias independientes con distribución $\mathcal{E}(\lambda)$.

Se proponen dos alternativas, con respecto a la construcción del test de hipótesis:

1. Armar un test de nivel exacto α a partir del estadístico $T_1 = \sum_{i=1}^{n} X_i$.
2. Armar un test de nivel asintótico α a partir del estadístico $T_2 = \sqrt{n}\frac{\bar{X}-2}{2}$

El objetivo es comparar ambos métodos mediante una simulación en **R**, para luego elegir uno de ellos y realizar el experimento. Para cada valor de $n \in 10, 30, 100, 1000$ se quiere estudiar el nivel empírico, y compararlo para las dos alternativas propuestas. También se desea aproximar y graficar la función potencia, y comparar.

Consideramos el siguiente procedimiento para la simulación:

- Generar una muestra de tamaño n de una distribución $\mathcal{E}(\frac{1}{2})$
- Calcular el valor del estadístico T para los dos test propuestos. Rechazar H_0 si el p-valor es menor al nivel 0.05.
- Repetir los ítems anteriores $N_{rep} = 10000$ veces, contando el total de veces en que se rechazó H_0. Se puede estimar el nivel de este test como la proporción empírica de rechazos, es decir, dividiendo la cantidad total de rechazos por N_{rep}, para cada valor de n, y armar una tabla con los resultados obtenidos.

Se pide:

a. Plantear claramente las hipótesis.

b. Justificar la elección de los dos estadísticos que propone el enunciado, y determinar sus distribuciones (exacta o asintótica) bajo H_0.

c. Realizar la simulación pedida para los diferentes valores de n, calcular el nivel empírico y expresarlo en una tabla.

d. Para cada uno de los valores propuestos de n y eligiendo una grilla de valores para λ, aproximar y graficar la función potencia.

e. Concluir en base a los resultados observados.

La resolución

a. Plantear claramente las hipótesis.

Consideramos que lo razonable es que la hipótesis nula sea que el parámetro **NO** supera el valor de 0.5 y la hipótesis alternativa que **SÍ** lo supera. Como dice el enunciado, se tiene la "sospecha" de que es mayor, por ende, el *status quo* es que el parámetro es menor. Simbólicamente:

$$H_0 : \lambda \leq 0.5 \qquad y \qquad H_1 : \lambda > 0.5$$

b. Justificar la elección de los dos estadísticos que propone el enunciado, y determinar sus distribuciones (exacta o asintótica) bajo H_0.

Primero vamos a desarrollar el **test de nivel exacto** α.

Partimos de que las mediciones son i.i.d y tienen la siguiente distribución:

$$X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{E}(\lambda)$$

O sea que la función de densidad de cada X es:

$$f_X(x) = \lambda e^{-\lambda x} \mathcal{I}_{x \geq 0}$$

Justamente por ser independientes, la densidad condicional se puede escribir como la productoria de las densidades individuales

$$f(\underset{\sim}{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \mathcal{I}_{x_i \geq 0}$$

Y por ser identicamente distribuidas podemos operar y llegar a:

$$f(\underset{\sim}{x}) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \mathcal{I}_{x_i \geq 0}$$

De este resultado podemos ver que se trata de una distribución de familia exponencial donde:

$$\begin{aligned} A(\lambda) &= \lambda^n \\ C(\lambda) &= -\lambda \\ r(\underset{\sim}{x}) &= \sum_{i=1}^{n} x_i \\ h(\underset{\sim}{x}) &= \prod_{i=1}^{n} \mathcal{I}_{x_i \geq 0} \end{aligned}$$

Entonces el estadístico (T_1 desde ahora) es igual $r(\underset{\sim}{X})$, de manera que el mismo queda definido como:

$$\begin{aligned} T_1 &= r(\underset{\sim}{X}) \\ T_1 &= \sum_{i=1}^{n} X_i \end{aligned}$$

Recuperando el estadístico propuesto en el enunciado del problema y teniendo en cuenta que $C(\lambda)$ es una función decreciente con λ, estamos en condiciones de escribir el test uniformemente más potente (UMP) de la siguiente forma:

$$\delta(\underset{\sim}{X}) = \begin{cases} 1 & si \quad -\sum_{i=1}^{n} X_i > K_\alpha \\ 0 & si \quad no \end{cases} \qquad .$$

con K_α tal que:

$$P_{\lambda=0.5}[\delta(\underset{\sim}{X}) = 1] = \alpha$$

Entonces, lo siguiente que tenemos que hacer es calcular K_α, y para esto vemos a "pararnos" en un mundo en el que H_0 es verdadera. Bajo el mundo H_0, vamos a considerar al parámetro como $\lambda = 0.5$, por lo tanto, la distribución de cada una de las variables aleatorias será $X_i \overset{i.i.d}{\sim} \mathcal{E}(\frac{1}{2})$. Con esto en mente, podemos determinar la distribución de la suma de variables aleatorias (que es el estadístico T_1):

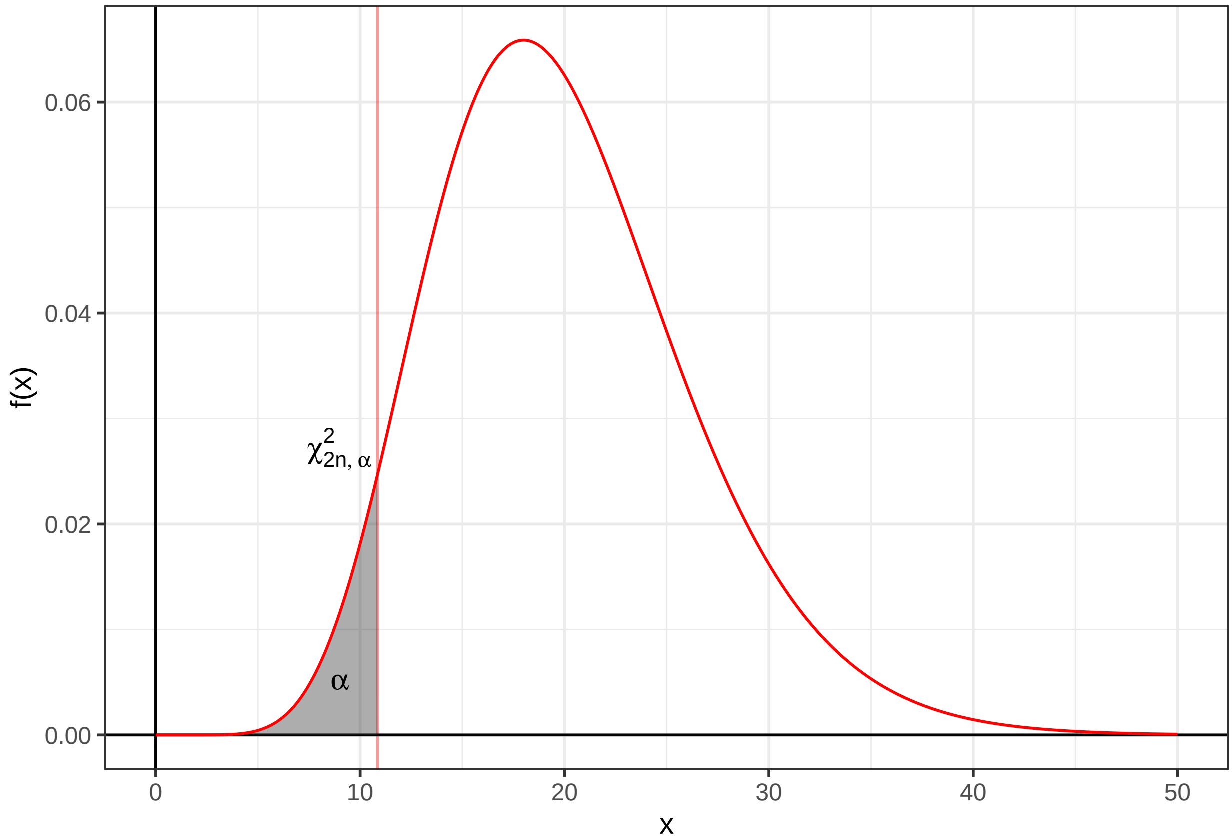
$$\begin{aligned} \sum_{i=1}^{n} X_i &\sim \Gamma(n, \frac{1}{2}) \\ &\sim \chi^2_{2n} \end{aligned}$$

Es decir, la suma de variables aleatorias $X_i \overset{i.i.d}{\sim} \mathcal{E}(\frac{1}{2})$ tiene distribución χ^2_{2n}. Entonces ahora vayamos a despejar K_α de la condición $P_{\lambda=0.5}[\delta(\underset{\sim}{X}) = 1] = \alpha$

$$\begin{aligned} \alpha &= P_{\lambda=0.5}[\delta(\underset{\sim}{X}) = 1] \\ &= P_{\lambda=0.5}(-\sum_{i=1}^{n} X_i > K_\alpha) \\ &= P_{\lambda=0.5}(\sum_{i=1}^{n} X_i < -K_\alpha) \\ &= P_{\lambda=0.5}(\sum_{i=1}^{n} X_i < \chi^2_{2n,\alpha}) \end{aligned}$$

Es decir, $-K_\alpha$ es igual al cuantil α de la distribución Chi cuadrado de grados de libertad $2n$ ($\chi^2_{2n,\alpha}$).

Por ejemplo, en la figura a continuación podemos ver, para $n = 10$ y $\alpha = 0.05$, el cuantil de la distribución (línea roja vertical, $\chi^2_{2n,\alpha}$) junto con el área gris que representa a la probabilidad acumulada de $\alpha = 0.05$.



Finalmente, podemos al test de nivel exacto como:

$$\delta(\underset{\sim}{x}) = \begin{cases} 1 & si \quad \sum_{i=1}^{n} x_i > \chi^2_{2n,\alpha} \\ 0 & si \quad no \end{cases} \quad .$$

Ahora sigamos con el **test de nivel asintótico** α.

Partimos de nuevo de que las mediciones son i.i.d y tienen la siguiente distribución:

$$X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{E}(\lambda)$$

Sabemos que $E(X_i) = \frac{1}{\lambda}$ y que $V(X_i) = \frac{1}{\lambda^2}$, por lo tanto $E(\bar{X}) = \frac{1}{\lambda}$ y que $V(\bar{X}) = \frac{1}{n}\frac{1}{\lambda^2}$. Entonces, usando el teorema central del límite podemos decir que:

$$\frac{\bar{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{n\lambda^2}}} \overset{\mathcal{D}}{\sim} \mathcal{N}(0,1)$$
$$\sqrt{n}\frac{\bar{X} - \frac{1}{\lambda}}{1/\lambda} \overset{\mathcal{D}}{\sim} \mathcal{N}(0,1)$$

Ahora volvemos a “pararnos” en H_0 y a considerar a $\lambda = 0.5$ (el supremo de Θ_1) y nos queda:

$$\sqrt{n}\frac{\bar{X}-\frac{1}{0.5}}{1/0.5} \overset{\mathcal{D}}{\sim} \mathcal{N}(0,1)$$

$$\sqrt{n}\frac{\bar{X}-2}{2} \overset{\mathcal{D}}{\sim} \mathcal{N}(0,1)$$

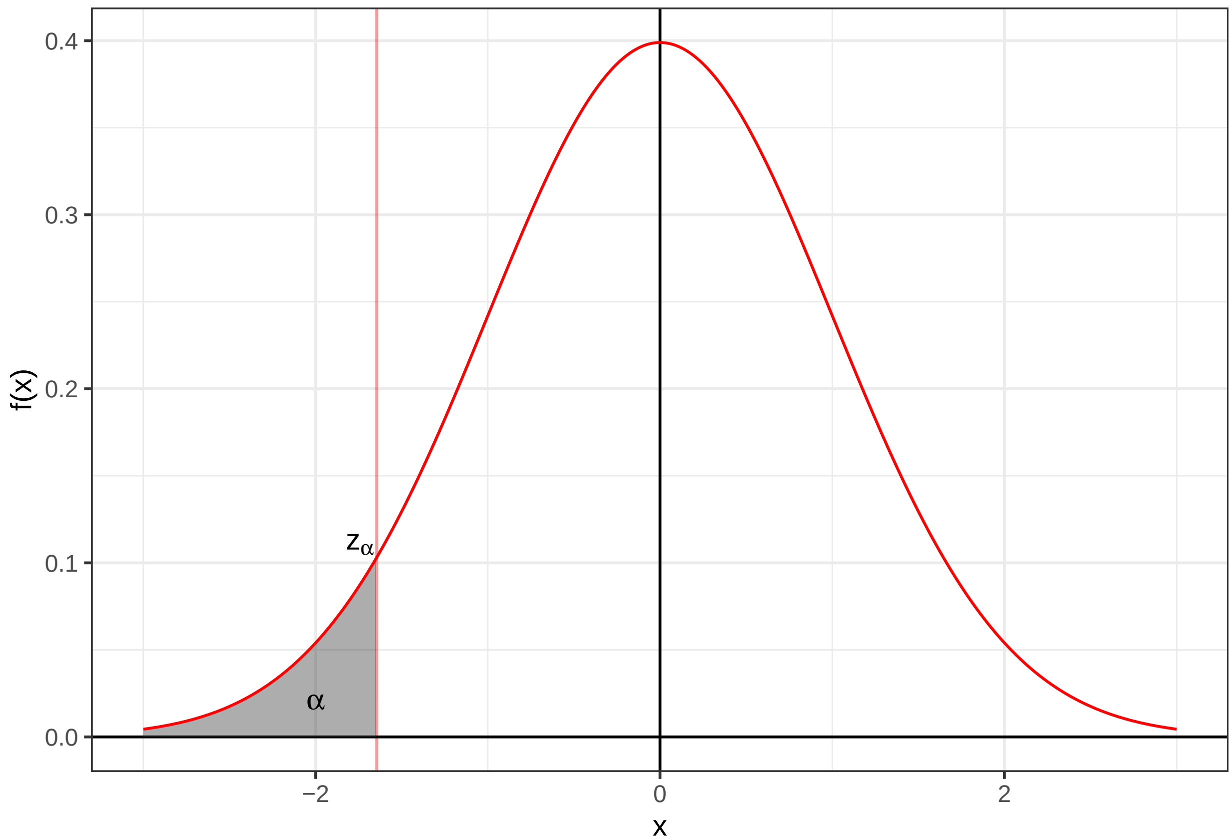
Que corresponde con el estadístico T_2 planteado en el enunciado. Luego, hallamos un K'_α tal que:

$$P_{\lambda=0.5}\left(\sqrt{n}\frac{\bar{X}-2}{2} \Big\langle K'_\alpha\right) \overset{n\to\infty}{\to} \alpha$$

De forma que el test asintótico de nivel α queda definido como:

$$\delta(\underset{\sim}{x}) = \begin{cases} 1 & si \quad \sqrt{n}\frac{\bar{X}-2}{2} < Z_\alpha \\ 0 & si \quad no \end{cases}$$

En este caso el cuantil que tenemos que buscar es el de una distribución $\mathcal{N}(0,1)$. Por ejemplo, para $\alpha = .05$:



c. Realizar la simulación pedida para los diferentes valores de n, calcular el nivel empírico y expresarlo en una tabla.

Ahora lo que vamos a hacer es simular el "experimento" y medir la cantidad de veces que se rechazó H_0 para ver como se relaciona eso con el n para cada uno de los tests propuestos (el UMP y el asintótico).

```r
# La función simulación simula el experimento y se fija el resultado del test
simulacion_UMP <- function(Nrep, n, lambda){
  rechazos <- rep(NA,Nrep)
  for (i in 1:Nrep) {
    muestra <- rexp(n, lambda)
    estadistico <- sum(muestra) # El estadístico del test UMP
    rechazos[i] <- pchisq(estadistico, 2*n)<0.05 # El test compara con Chisq
  }

  alpha_emp <- mean(rechazos)
  return(alpha_emp)
}

simulacion_asin <- function(Nrep, n, lambda){
  rechazos <- rep(NA,Nrep)
  for (i in 1:Nrep) {
    muestra <- rexp(n, lambda)
    estadistico <- sqrt(n)*(mean(muestra)-2)/2 # El estadístico del test asintótico
    rechazos[i] <- pnorm(estadistico, 0, 1)<0.05 # El test compara con Norm
  }

  alpha_emp <- mean(rechazos)
  return(alpha_emp)
}

set.seed(123)
ns <- c(10,30,100,1000)

tabla <- tibble(n = ns) %>%
  rowwise() %>%
  mutate(alpha_emp_UMP = simulacion_UMP(Nrep = 10000, n = n, lambda = 0.5),
         alpha_emp_asin = simulacion_asin(Nrep = 10000, n = n, lambda = 0.5))

tabla %>%
  knitr::kable(col.names = c("n", "alfa empírico UMP", "alfa empírico asintótico"))
```

n	alfa empírico UMP	alfa empírico asintótico
10	0.0504	0.0253
30	0.0505	0.0364
100	0.0482	0.0425
1000	0.0531	0.0494

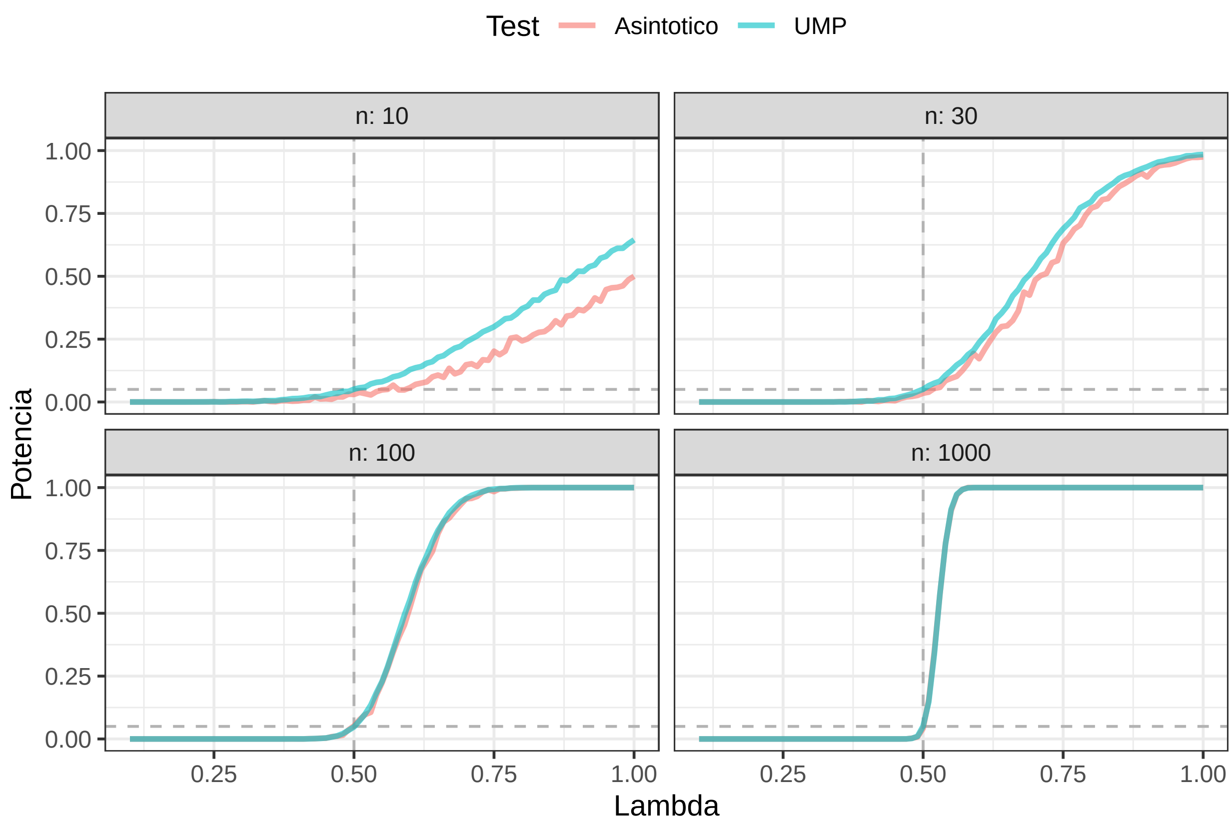
d. Para cada uno de los valores propuestos de n y eligiendo una grilla de valores para λ, aproximar y graficar la función potencia.

```r
# La función simulación simula el experimento y se fija el resultado del test
set.seed(123)
ns <- c(10,30,100,1000)
lambdas <- seq(0.1,1,.01)

potencias <- expand_grid(n = ns, Lambda = lambdas) %>%
  rowwise() %>%
```

```r
  mutate(UMP = simulacion_UMP(Nrep = 10000, n = n, lambda = Lambda),
         Asintotico = simulacion_asin(Nrep = 1000, n = n, lambda = Lambda)) %>%
  pivot_longer(cols = !c(n, Lambda), names_to = "Test", values_to = "Potencia")

potencias %>% ggplot(aes(x = Lambda,
                         y = Potencia,
                         color = Test)) +
  geom_vline(xintercept = 0.5, color = "gray70", linetype = "dashed") +
  geom_hline(yintercept = 0.05, color = "gray70", linetype = "dashed") +
  geom_line(linewidth = 1, alpha = .6) +
  facet_wrap(.~n, labeller = label_both) +
  theme_bw() +
  theme(legend.position = "top")
```



e. Concluir en base a los resultados observados.