Intervalos de confianza frecuentista y bayesianos para H0

Tomas Mastantuono, LU:522/23, DNI:45421427

Objetivo

En este trabajo se busca construir intervalos de confianza frecuentistas y bayesianos para el parámetro H_0 utilizando datos de cronómetros cósmicos. Los datos que me dan para analizar son de la forma $\left(z, H\left(z\right), \sigma_{H(z)}\right)$, donde z es el corrimiento al rojo, $H\left(z\right)$ es la tasa de expansión del universo a ese corrimiento y $\sigma_{H(z)}$ es el error asociado a $H\left(z\right)$.

El modelo cosmológico utilizado es el ΛCDM, el cual me relaciona la tasa de expansión del universo con el parámetro H_0 y la densidad de la materia Ω_m a través de la ecuación (1).

$$H\left(z\right) = H_0\sqrt{\Omega_m\left(1+z\right)^3 + \left(1-\Omega_m\right)} \tag{1}$$

1 Datos observacionales $H\left(z\right)$ vs z

Como primera actividad se pide un gráfico de los datos observacionales $H\left(z\right)$ en función de z, el cual se muestra en la Figura 1, junto con un ajuste utilizando la expresión dada por la ecuación (1).

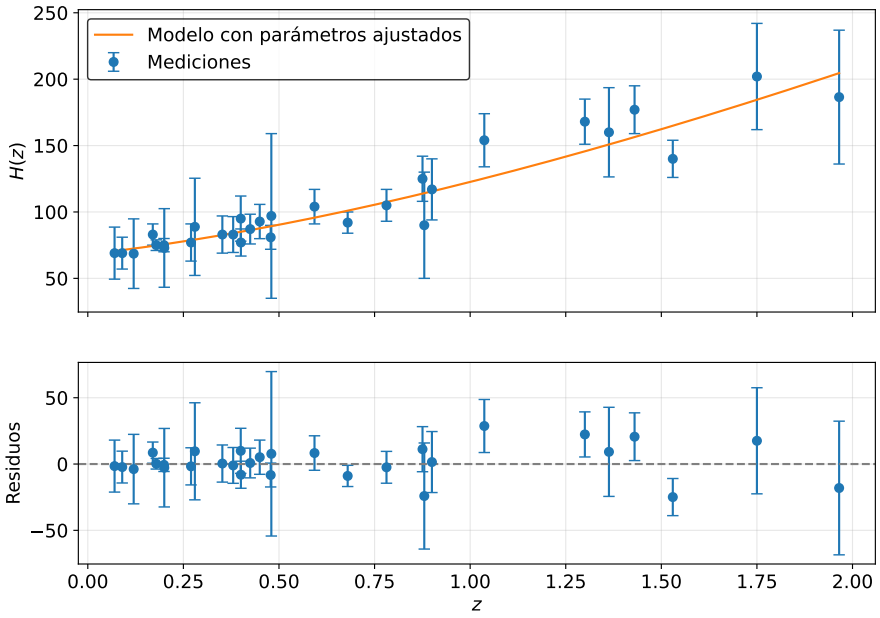

Figure 1: El primer gráfico corresponde a la tasa de expansión del universo $H\left(z\right)$ en función al corrimiento al rojo z junto con su ajuste correspondiente utilizando la ecuación (1). El segundo gráfico corresponde a los residuos obtenido del ajuste y los datos, los cuales parecen no seguir ninguna tendencia específica.

Para realizar el ajuste que se muestra en la Figura utilicé $curvefit$, la cual es una función de una librería de $Scipy$. Del ajuste logré obtener parámetros óptimos para H_0 y para Ω_m, los cuales están dados por $H_0 = (68 \pm 2)$ y $\Omega_m = (0.31 \pm 0.04)$.

2 Intervalo de confianza frecuentista

A partir de este punto se pide explícitamente que se fije el valor de Ω_m a 0.3 y que todo el análisis siguiente sea solo para H_0. Para construir un intervalo de confianza frecuentista, tomé $M = 20$ valores equispaciados dentro del rango $[60, 80]$, esta decisión fue tomada pensando en el momento de tener que visualizar el cinturón de confianza frecuentista, evitándome tener de forma inesperada valores acumulados de H_0 los cuales me agreguen sesgo a la interpretación final.

Para cada uno de estos valores tomados de H_0, realicé $N = 2000$ conjuntos de datos sintéticos construidos por la ecuación (2), utilizando los mismos valores z_i y $\sigma_{H(z_i)}$ de las mediciones. Tener en cuenta que $\epsilon_i \sim N\left(0, \sigma_{H(z_i)}\right)$, por lo que su único objetivo es agregar ruido gaussiano a los datos sintéticos.

$$H_i^{sint} = H_0 \sqrt{\Omega_m \left(1 + z_i\right)^3 + \left(1 - \Omega_m\right)} + \epsilon_i \tag{2}$$

Por lo que hasta el momento tengo M valores diferentes de H_0, de los cuales para cada uno de estos realicé N sets de datos sintéticos. Todo esto tiene como objetivo lograr que para cada valor de H_0 me pueda construir un cinturón de confianza del 68% y del 95% en función de un estimador $\hat{H}_0$. Por la forma en la que fueron construidos los datos sintéticos, los cuales según la ecuación (2) se puede notar un término del cual no depende de ninguna variable aleatoria y un término sumado que posee errores gaussianos, utilizo cuadrados mínimos para estimar el H_0 de cada una de los N sets y repito para los M valores, es decir, uso $curvefit$ y ajusto con la ecuación (1) los N sets de datos sintéticos, obteniendo N valores de $\hat{H}_0$ para cada uno de los M. De esta forma logro obtener numéricamente la distribución de $\hat{H}_0$ para cada uno de los M del rango mencionado, por ejemplo se muestra en la Figura 2 el caso para el primer H_0 dentro del rango $[60, 80]$, donde el primer gráfico muestra contenido un intervalo central del 68% de confianza, mientras que el segundo un 95% de confianza.

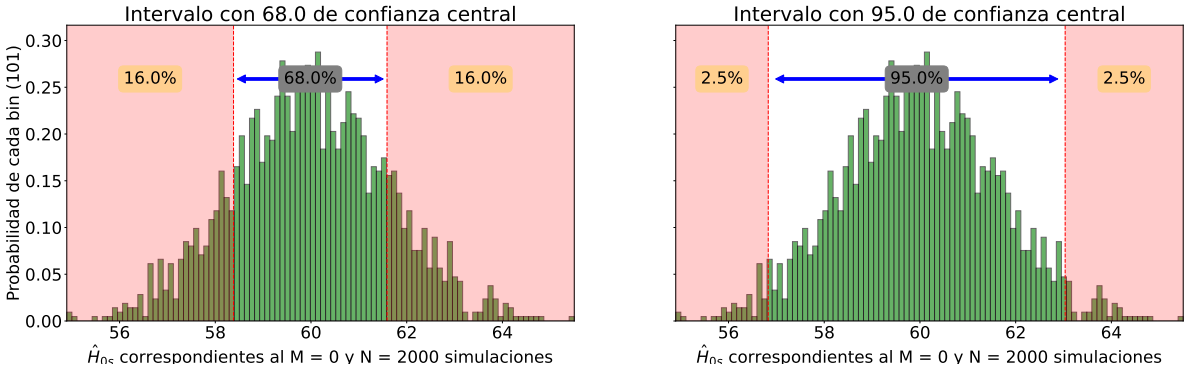

Figure 2: Ambos gráficos corresponde a la misma simulación de datos sintéticos correspondientes de H_0. El gráfico de la izquierda muestra el intervalo que encierra el 68% de probabilidad central. En cambio, el gráfico de la derecha encierra el 95% de probabilidad centrada.

Este proceso fue realizado para los $M = 20$ valores de H_0, donde para cada uno de estas distribuciones obtuve los estimadores $\hat{H}_{Min}$ y $\hat{H}_{Max}$ que marcan los límites del intervalo del 68% y 95% de confianza central.

Utilizando los M valores de H_0, y los estimadores límite de cada distribución puedo construir los cinturones de confianza frecuentista para el 68% y el 95% de confianza, tal como se muestra en la Figura 3.

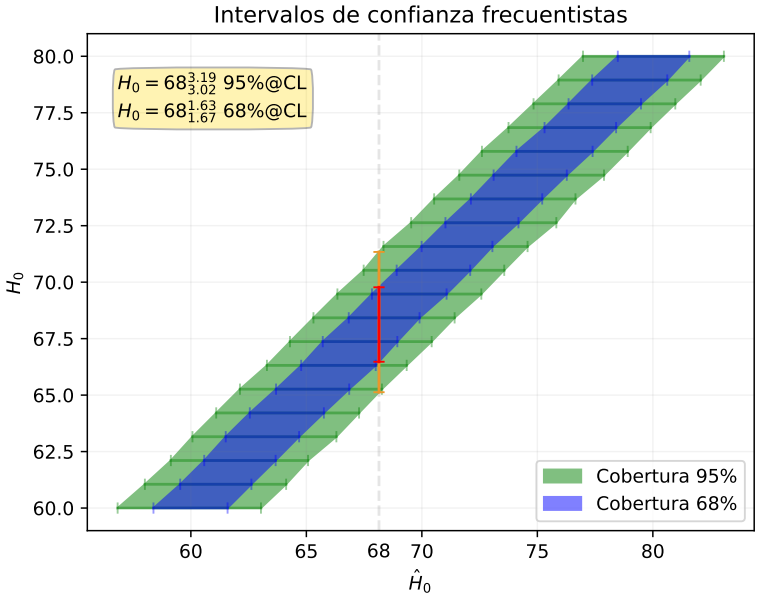

Figure 3: Se muestra H_0 en función de $\hat{H}_0$. El cinturón de color verde corresponde al intervalo el cual encierra una cobertura del 95%@CL. Por otro lado, el cinturón de color azul corresponde al intervalo de confianza que encierra un 68%@CL.

Como me interesaba dar un intervalo de confianza frecuentista, tenía que fijarme donde interseca mi valor obtenido del ajuste con el cinturón correspondiente al 68% o al 95% de confianza. Para esto usé una función de *numpy* llamada *interp* donde al accederle el valor óptimo obtenido del ajuste de la Figura 1 realizaba una interpolación con los valores propios del cinturón, logrando darme un punto aproximado que pertenezca al conjunto del intervalo. Tal como se muestra en la figura, obtuve $H_0 = 68^{+3.19}_{-3.02}$ 95%@CL, y $H_0 = 68^{+1.63}_{-1.67}$ 68%@CL.

Desde una interpretación frecuentista, el cinturón de α% de confianza me dice que solo α de cada 100 veces voy a obtener a un intervalo que contenga al valor real de H_0, por lo que no puedo concluir mas desde esta interpretación.

3 Intervalo de credibilidad bayesiano

La consigna me pide escribir la expresión de la *posterior* para H_0, aún manteniendo fija Ω_m, y suponiendo errores gaussianos e independientes. Por lo que si utilizo el Teorema de Bayes, la expresión de la *posterior* está dada por la ecuación (3).

$$f\left(H_0|\vec{x}\right) = \frac{\prod_i^n f\left(x_i|H_0\right)\pi\left(H_0\right)}{\prod_i^n \int f\left(x_i|H_0\right)\pi\left(H_0\right)d\left(H_0\right)} \tag{3}$$

Donde en esta ecuación $f\left(H_0|\vec{x}\right)$ referencia a la función de distribución de H_0 dado que medí $\vec{x}$, mientras que la función $\pi\left(H_0\right)$ es el *prior* que contiene la información hasta el momento de esa variable. Además a la integral que se encuentra en el denominador se la suele llamar Z, o también evidencia. Hay que tener en cuenta que el enunciado me pide que asuma $f\left(x_i|H_0\right) \sim G\left(H\left(z_i\right), \sigma = \sigma_{H(z_i)}\right)$.

Por lo que el primer problema numérico al cual me enfrento es hallar la función que se encuentra en el numerador de la ecuación (3). El primer paso que realicé fue definirme una función la cual solo necesite los datos medidos y un único valor para H_0, para cada dato de las mediciones realizadas, y el H_0 colocado, me creo una distribución gaussiana centrada en la ecuación (1) evaluado en cada punto z_i y con dispersión $\sigma_{H(z_i)}$. De todo esto obtengo la *pdf* en el punto H_i que obtuve de la medición. Este proceso lo repito para todos los datos de las mediciones, y al ser todas independientes multiplico las funciones de distribución.

Para poder dar la función de distribución del parámetro H_0 necesitaba calcular Z, para realizar esa integral utilicé una función de integración de *Scipy* llamada *quad*, la cual al darle simplemente la función a integrar y el intervalo de integración devuelve el resultado final. Como es un integral numérica, pero los intervalos teóricos son en todos los posibles valores para el parámetro, tomé un atajo y para cada prior pedido por el ejercicio visualicé previamente hasta que punto hay un aporte significativo, por lo que utilicé como intervalo de integración solo esas regiones. Finalmente, utilizando la ecuación (3) logré construir un *posterior* utilizando en un caso el *prior* $U\left[60, 80\right]$, y en otro un *prior* $G\left(70, 5\right)$, tal como se muestra en la Figura 4.

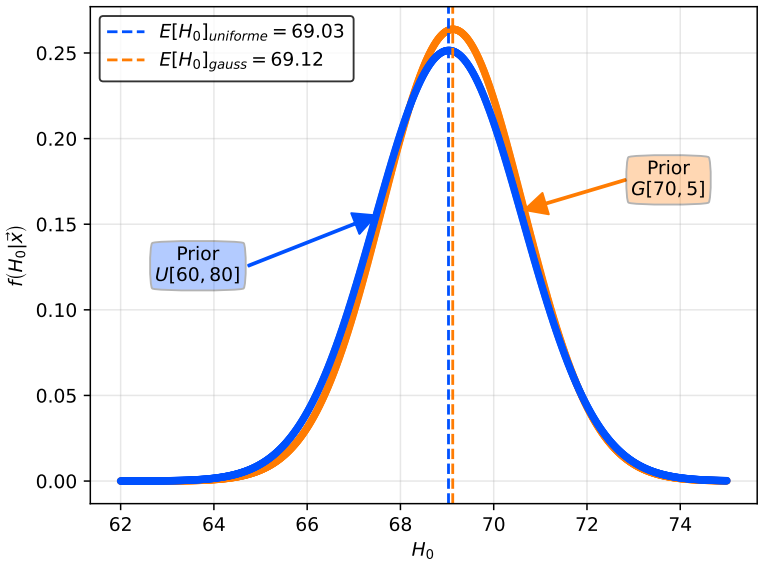

Figure 4: En azul se muestra la función de distribución o *posterior* para H_0 dadas las mediciones $\vec{x}$, la cual surge de suponer un *prior* con distribución $U\,[60, 80]$. Mientras que en naranja se puede ver la función de distribución para H_0 o *posterior* dadas las mediciones $\vec{x}$, la cual surge de suponer un *prior* con distribución $G\,[70, 5]$. Con línea punteada se muestra la *Esperanza* de cada distribución respectivamente.

De la *posterior* proveniente de un *prior* con distribución $U\,[60, 80]$, se obtiene un intervalo de $H_0 = (69.03 \pm 1.60)\,68\%$@CL y $H_0 = (69.03 \pm 3.10)\,95\%$@CL. Mientras que la *posterior* proveniente de un *prior* con distribución $G\,[70, 5]$ se puede obtener un intevalo de $H_0 = (69.12 \pm 1.60)\,68\%$@CL y $H_0 = (69.12 \pm 3.10)\,95\%$@CL.

Como se púede ver, los resultados no son muy diferentes y poseen valores que se intersecan, por lo que si bien a simple vista pareciera que las distribuciones son notariamente diferentes, realmente los resultados comparten un gran intervalo entre ellos. El *prior* uniforme en principio solo agrega información de que el valor de H_0 está contenido en el intervalo $[60, 80]$, mientras que el *prior* gaussiano parece agregar mas información y que el verdadero valor de H_0 está contenido en una campana centrada en 70 y con dispersión de $\sigma = 5$.

Conclusiones

En conclusión obtuve en ambas interpretaciones intervalos los cuales se solapan en ciertos valores. Pude ver que la interpretación frecuentista solo me habla de que en ciertos casos voy a obtener un intervalo que contenga al valor real del parámetro, y me da un intervalo para el estimador medido. Mientras que la interpretación bayesiana me agrega de forma explícita conocimiento previo sobre el parámetro que quiero conocer.