

Resumen del video “Introducción a PyMC”

La estadística bayesiana es una filosofía distinta a la estadística frecuentista ya que esta piensa la probabilidad como la frecuencia con la que pasa un evento, en cambio, la bayesiana piensa la probabilidad como una forma de medir incertidumbre.



En alguna ocasión hemos escuchado hablar de los test de hipótesis, los cuales se clasifican en dos tipos de pruebas de hipótesis: paramétricas (el análisis de varianza, la prueba T y el estadístico Z) y no paramétricas (la prueba de chi cuadrado y la U de Mann-Whitney), cosas que pertenecen a la frecuentista y son poco intuitivas en comparación con la bayesiana que responde las preguntas de una manera más práctica.

Toda la estadística bayesiana parte del Teorema de Bayes que es un principio fundamental de la teoría de la probabilidad que se utiliza para calcular la probabilidad de un evento a partir de información previa. Se basa en la idea de que la probabilidad de un evento es igual a la probabilidad de que ese evento suceda, dado que ya sucedió un evento previo.

La fórmula matemática del teorema de Bayes es:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$




Donde:

-  A representa el evento que se está estudiando.
-  B representa el evento previo.

Lo que representa el núcleo de la estadística bayesiana es:

$$P(\text{Parámetro}|\text{Datos}) \propto P(\text{Datos}|\text{Parámetro}) * P(\text{Parámetro})$$

Donde:

-  $P(\text{Parámetro}|\text{Datos})$ representa lo que queremos calcular, es decir, la distribución de probabilidad de los parámetros después de observar los datos (probabilidad a posteriori)
-  $P(\text{Datos}|\text{Parámetro})$ es la verosimilitud de los datos dado el parámetro, que mide qué tan probable es observar los datos si el parámetro es cierto (likelihood).
-  $P(\text{Parámetro})$ refleja el conocimiento o las creencias previas sobre los valores del parámetro antes de observar los datos (probabilidad a priori).

Los datos que le tenemos que dar a PyMC son likelihood y la probabilidad a priori de nuestro problema para que estime un modelo Bayesiano.

La ventaja de usar PyMC principalmente es que hace toda la matemática compleja por nosotros, ya que $P(\text{Datos}|\text{Parámetro}) * P(\text{Parámetro})$ es una parte muy difícil de resolver matemáticamente y nosotros solo tenemos que saber cual es el modelo que genera los datos y la información que sabemos a priori de lo que queremos estimar, entonces, PyMC nos permite definir modelos bayesianos usando código.

Lo anterior ha llevado a la democratización de la estadística pues ya no necesitas saber calculo, integrales, etcétera para hacer este tipo de modelos, aunque, si se necesita saber bien que es un likelihood y una probabilidad a priori.

En estadística bayesiana, el likelihood (verosimilitud) se refiere a la probabilidad de observar los datos dados los parámetros del modelo. Es decir, es una función que describe qué tan probable es obtener los datos observados bajo diferentes valores de los parámetros del modelo.

Ahora, la probabilidad a priori es una distribución de probabilidad que refleja nuestro conocimiento o creencias previas sobre los parámetros de un modelo antes de observar cualquier dato. Es una forma de formalizar lo que ya sabemos, o creemos saber, acerca de los parámetros del modelo antes de la recopilación de evidencia nueva.

Veamos un ejemplo para entender mejor lo anterior, es mayo del 2020, hace unos meses empezó la pandemia y el gobierno quiere saber cuánta gente se ha contagiado de COVID en Santiago, Chile. Dependiendo de este número el gobierno seguirá o no una estrategia de inmunidad de rebaño.

Para esto se usaron tests que detectan anticuerpos de SARS-CoV-2 que fueron aleatorizados y se hicieron las encuestas.

Partiendo con una distribución binomial donde hay N “experimentos” que resultan en un éxito (1) o un fracaso (0) con probabilidad p . Se tomó una muestra aleatoria de 50 personas en Santiago, donde 40 son negativas, es decir, nuestro test no detecta anticuerpos y 10 son positivas.

A partir de esto construiremos 3 modelos:

- **Modelo 1** asume que el test es perfecto

¿Cuánta gente tuvo COVID-19?

20% (10 de 50)

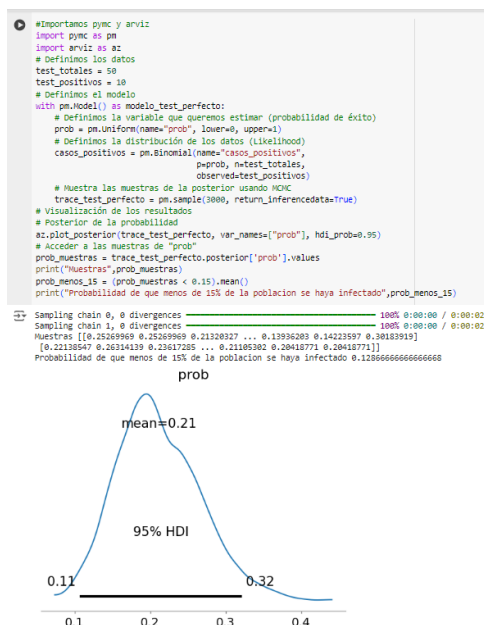
¿Qué tan seguro estamos de este resultado?

Respondamos esto creando este modelo en PyMC.

Con la gráfica podemos ver que las muestras son muy poderosas.

Ministra de Salud: ¿Cuál es la probabilidad de que menos de 15% de la población se haya infectado?

Si observamos en la imagen en este caso nuestra respuesta es 12.87%



- **Modelo 2** incluiremos que el test a veces da falsos positivos

El laboratorio que creó el test hizo 100 pruebas y nos informa que la tasa de falsos negativos es de 10%, entonces, le diremos al modelo que la proporción de tests positivos no es lo mismo que la proporción de gente de COVID.

Ahora la probabilidad de un test positivo es afectada por dos factores:

- La probabilidad de tener COVID (lo llamaremos `prob_cov`)
- La probabilidad de un falso positivo (lo llamaremos `prob_fp`)

Ministra de Salud: ¿Cuál es la probabilidad de que menos de 15% de la población se haya infectado?

Si observamos en la imagen en este caso nuestra respuesta es 68.17%

- **Modelo 3** que incluye incertidumbre sobre la tasa de falsos positivo

Ahora también modelamos la tasa de falsos positivos con una distribución binomial

La estadística frecuentista y la estadística bayesiana son dos enfoques fundamentales para realizar inferencia en estadística. Aunque ambos tienen objetivos similares hacer inferencias sobre parámetros poblacionales a partir de datos muestrales, sus enfoques, fundamentos y métodos son bastante diferentes como pudimos observarlo con los gráficos y códigos.

Ambos enfoques tienen sus aplicaciones y ventajas, y la elección entre ellos depende del contexto del problema y de las preferencias del analista.

