Práctica 1.
Especificación y evaluación de argumentos causales.

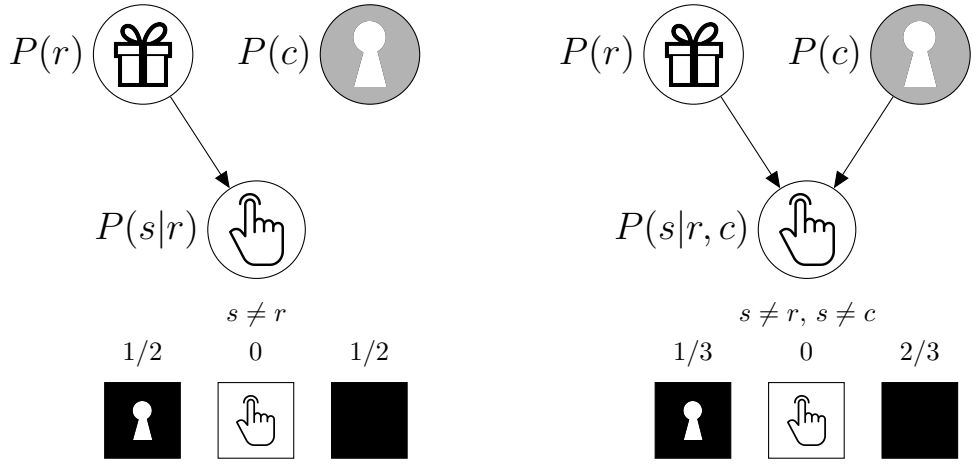
Docente: Gustavo Landfried

Inferencia Bayesiana Causal 1
1er cuatrimestre 2025
UNSAM

Índice

1. Modelo Base vs Modelo Monty Hall

En la siguiente figura se puede observar la especificación gráfica del modelo "Monty Hall" (derecha) y el modelo "Base" (izquierda) vistos en la primera semana. Abajo de ellos se muestra la distribución de creencias *a posteriori* sobre la posición del regalo luego de haber reservado la caja 1 y luego de que nos hayan mostrado que en la caja 2 no estaba el regalo, $P(r|s=2,c=1)$.



(a) Modelo Base (M_0)

(b) Modelo Monty Hall (M_1)

Figura 1: Modelos causales alternativos

La única restricción que supone el modelo Base (izquierda) es q que la pista s no puede señalar la caja en la que se encuentra el regalo $s \neq r$. El modelo Monty Hall (derecha) incluye esta restricción y le agrega una restricción adicional, que la pista s tampoco puede señalar la caja que hemos reservado previamente $s \neq c$.

El objetivo de este ejercicio es actualizar nuestras creencias sobre los modelos causales alternativos luego de observar un conjunto datos, $P(\text{Modelo}|\text{Datos})$.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Datos}|\text{Modelo})P(\text{Modelo})}{P(\text{Datos})}$$

Para ello deberemos calcular:

- La predicción que hace el modelo sobre los datos: $P(\text{Datos}|\text{Modelo})$
- La predicción de los datos realizada con la contribución de todos los modelos: $P(\text{Datos})$
- La creencia previa "honesta" sobre los modelos: $P(\text{Modelo})$

Vayamos paso a paso.

1.1. Definir la distribución de creencia conjunta como producto de las distribuciones condicionales del modelo

Las especificaciones gráficas de los argumentos causales representan dos descomposiciones alternativas de la distribución de creencias conjuntas $P(r,c,s|M)$.

$$\underbrace{P(r,c,s|M_0)}_{\substack{\text{Prior conjunto}\\ \text{hipótesis en } M_0}} = \underbrace{P(r|M_0)P(c|M_0)P(s|r,M_0)}_{\substack{\text{Relaciones causales probabilísticas}\\ \text{propuestas por el modelo } M_0}} \quad , \quad \underbrace{P(r,c,s|M_1)}_{\substack{\text{Prior conjunto}\\ \text{hipótesis en } M_1}} = \underbrace{P(r|M_1)P(c|M_1)P(s|r,c,M_1)}_{\substack{\text{Relaciones causales probabilísticas}\\ \text{propuestas por el modelo } M_1}}$$

Ambos modelos suponen que r y c son variables independientes. El modelo Base M_0, sin embargo, supone que s depende únicamente de r, ($s \neq r$) mientras que el modelo Monty Hall M_1 considera que s depende tanto de r como de s, ($s \neq r$, $s \neq c$).

Las distribuciones condicionales a priori sobre r y c son iguales en ambos modelos.

$$P(r|M) = P(r) = \begin{array}{c|c|c} r=0 & r=1 & r=2 \\ \hline 1/3 & 1/3 & 1/3 \end{array} \qquad P(c|M) = P(c) = \begin{array}{c|c|c} c=0 & c=1 & c=2 \\ \hline 1/3 & 1/3 & 1/3 \end{array}$$

La única diferencia entre los modelos aparece en la distribución condicional sobre la pista. Notar que una distribución condicional, a diferencia de una conjunta, tiene que integrar 1 para cada uno de los valores posibles del condicional . En el modelo Base solo depende del regalo r.

$$P(s|r, M_0) = \begin{array}{c|c|c|c} & s=0 & s=1 & s=2 \\ \hline r=0 & 0 & 1/2 & 1/2 \\ \hline r=1 & 1/2 & 0 & 1/2 \\ \hline r=2 & 1/2 & 1/2 & 0 \end{array}$$

Y en el modelo Monty Hall depende del regalo r y la caja cerrada c. Para simplificar, mostraremos los valores cuando $c = 1$.

$$P(s|r, c=1, M_1) = \begin{array}{c|c|c|c} (c=0) & s=0 & s=1 & s=2 \\ \hline r=0 & 0 & 1/2 & 1/2 \\ \hline r=1 & 0 & 0 & 1 \\ \hline r=2 & 0 & 1 & 0 \end{array}$$

Notar que cada renglón suma 1, pues cada condicional representa una distribución de probabilidad distinta.

1.2. Simular datos con el modelo Monty Hall

Antes de evaluar los modelos necesitamos un conjunto de datos que provengan de la realidad subyacente oculta. Podríamos buscar los datos reales del programa de televisión Monty Hall y revisar si efectivamente el modelo Monty Hall propuesto es mejor que el modelo Base. Para simplificar vamos a suponer que nuestro modelo Monty Hall representa perfectamente la realidad causal subyacente y vamos a generar entonces los datos usando de nuestro propio modelo Monty Hall. Generar un conjunto de datos con $T = 16$ episodios.

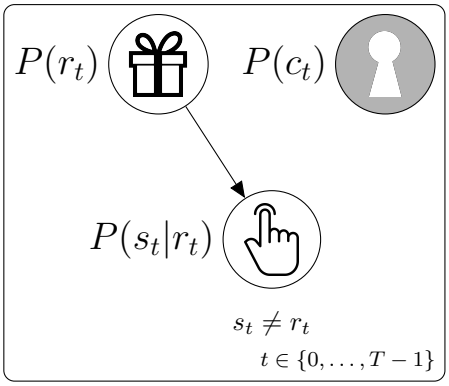
$$\text{Datos} = \{(c_0, s_0, r_0), \ldots, (c_{T-1}, s_{T-1}, r_{T-1})\}$$

1.3. Calcular la predicción a priori que hace cada uno de los modelos sobre la totalidad de la base de datos simulada

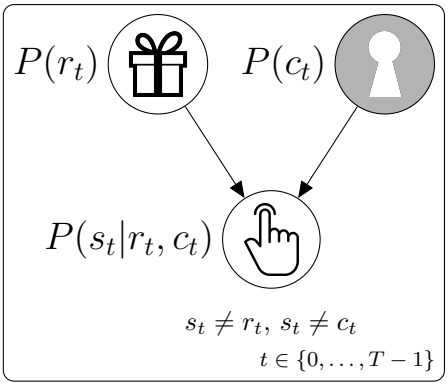
Ahora sí, podemos calcular la predicción del conjunto de datos que hace cada uno de los modelos con la contribución de todas sus hipótesis internas.

$$P(\text{Datos} = \{\underbrace{(c_0, s_0, r_0)}_{\substack{\text{Primer} \\ \text{episodio}}}, \underbrace{(c_1, s_1, r_1)}_{\substack{\text{Segundo} \\ \text{episodio}}}, \ldots \underbrace{(c_{T-1}, s_{T-1}, r_{T-1})}_{\substack{\text{T-ésimo} \\ \text{episodio}}}\}|\text{Modelo})$$

Los modelos causales expresados en la figura 1 proponen relaciones causales probabilísticas entre las variables al interior de un episodio. En principio, estos modelos sólo están definidos para un único episodio. Para extenderlos a T episodios vamos a considerar que contamos con T repeticiones de esa misma estructura causales. Las repeticiones se especifican gráficamente mediante el uso de "placas".



(a) Modelo Base (M_0)



(b) Modelo Monty Hall (M_1)

Figura 2: Extensión de los modelos causales alternativos a T episodios mediante la notación de "placas". El subíndice t representa las repeticiones.

Es decir, entre episodios no hay flechas que vinculen las estructuras causales, por lo que ninguno modelo usa los datos de un episodio para predecir lo datos de otro episodio. Esto permite descomponer la predicción sobre el conjunto de datos sobre todos los episodios como el producto de las predicciones que lo modelos hacen al interior de cada episodio.

$$P(\text{Datos}|\text{Modelo}) = \prod_{t \in \{0,\ldots,T-1\}} P(c_t|\text{Modelo}) P(s_t|c_t, \text{Modelo}) P(r_t|s_t, c_t, \text{Modelo})$$

Calcular $P(\text{Datos}|\text{Modelo})$.

1.4. Calcular la predicción de los datos con la contribución de todos los modelos.

Para actualizar la creencia de los modelos vamos a necesitar la probabilidad de los datos, $P(\text{Datos})$, que no es más que la predicción hecha con la contribución de todos los modelos.

$$P(\text{Datos}) \overset{\text{Regla de la suma}}{=} \sum_{\text{Modelo}} P(\text{Modelo}, \text{Datos}) \overset{\text{Regla de la producto}}{=} \sum_{\text{Modelo}} \underbrace{P(\text{Datos}|\text{Modelo})}_{\substack{\text{Predicción hecha} \\ \text{por el modelo}}} \underbrace{P(\text{Modelo})}_{\substack{\text{Creencia en} \\ \text{el modelo}}}$$

Aprovechar que tenemos la secuencia de predicciones hechas en cada uno de los episodios para calcular cómo se va actualizando la predicción conjunta a medida que se van incorporando nuevos datos.

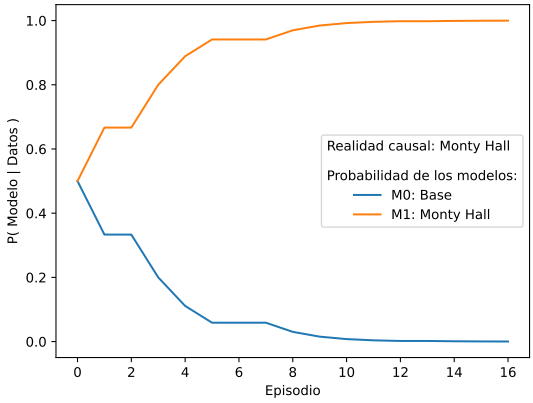
1.5. Calcular el posterior de los modelos

Ahora sí. Tenemos todos los elementos necesarios para calcular el posterior de los modelos.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Modelo}, \text{Datos})}{P(\text{Datos})}$$

1.6. Graficar el valor del posterior a medida que se observan nuevos episodios

Para graficar cómo se va actualizando el posterior deberemos tener guardado el valor del posterior luego de observar cada uno de los episodios.



2. Modelo Alternativo

Se proveé una archivo de datos `NoMontyHall.csv` que contienen 2000 episodios. Los datos fueron generados con la siguiente realidad causal subyacente. La persona que da la pista tiene una probabilidad $p \in [0, 1]$ de acordarse de tener en cuenta la caja que reservamos a la hora de dar la pista. Esta probabilidad es general a todos los episodios. En cada episodio particular, la persona se acuerda o no de tener en cuenta la pista, $a \in 0, 1$. Cuando se acuerda, la persona usa la distribución de probabilidad condicional del modelo Monty Hall para dar la pista. Cuando se olvida usa la distribución de probabilidad condicional del modelo Base para dar la pista.

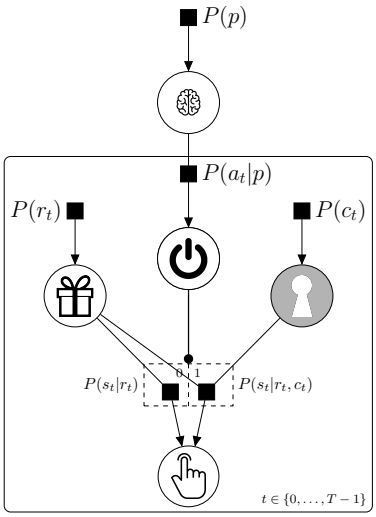


Figura 4: Realidad causal subyacente del conjunto de datos del archivo `NoMontyHall.csv`.

Este tipo de modelo gráfico se conoce como *factor graph*. Los factor graph, a diferencia de las redes bayesianas, incorporan los mecanismos causales o distribuciones de probabilidad condicional a la red causal, formando un grafo bipartito en el cual las variables quedan vinculadas con las distribuciones de probabilidad de las cuales son parámetro.

Al igual que las redes bayesianas, en los factor graph el producto de las distribuciones de probabilidad condicional es la especificación matemática de la distribución de probabilidad conjunta. Para ello es necesario interpretar la notación de compuertas introducida en el artículo *Causality with gates*. Esta notación permite definir mecanismos causales dinámicos, cambian en función del contexto. En este caso, estamos definiendo en detalle la distribución de probabilidad condicional

de la pista.

$$P(s_t|r_t, c_t, a_t) = P(s_t|r_t)^{\mathbb{I}(a_t=0)} P(s_t|r_t, c_t)^{\mathbb{I}(a_t=0)} \tag{1}$$

donde $\mathbb{I}()$ es la función indicadora y $P(s_t|r_t)$ y $P(s_t|r_t, c_t)$ son las distribuciones de probabilidad condicional del modelo Base y del modelo Monty Hall. Luego, la probabilidad conjunta es

$$P(r_1, c_1, s_1, a_1, \ldots, r_T, c_T, s_T, a_T, p|M_2) = P(p) \prod_{t=1}^{T} P(r_t)P(c_t)P(s_t|r_t, M_0)^{1-a_t} P(s_t|r_t, c_t, M_1)_t^a P(a_t|p)$$

En este caso, cuando queremos hacer la predicción al interior de un episodio t vamos a tener dos variables ocultas, a_t y p que deberemos integrar usando la regla de la suma. A diferencia de lo que ocurre en los modelos Base y Monty Hall, en el cual los datos de los diferentes episodios son independientes entre sí, en este modelo hay una variable, la probabilidad p de acordarse, que es común a todos los episodios y los conectan entre sí. Si abrimos las placas con el subíndice t que representa la repetición de los episodios vamos a ver que ellos quedan conectados entre sí por la variable p.

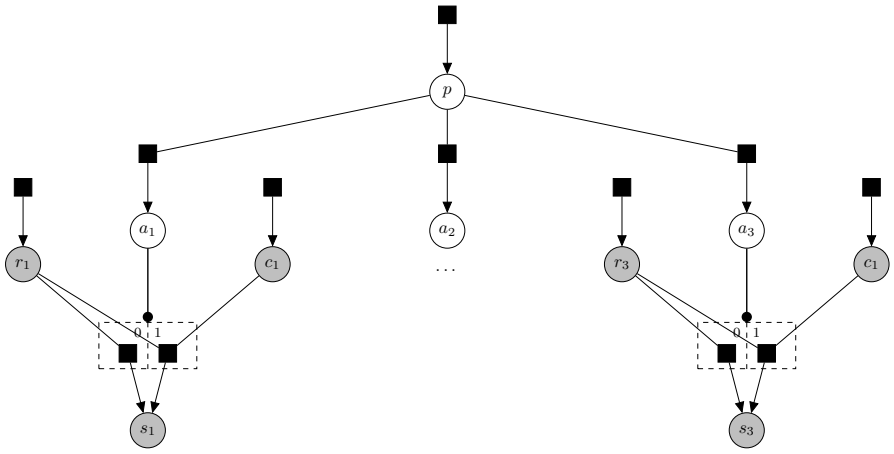


Figura 5: Factor graph del modelo alternativo desplegado. Las variables en blanco son ocultas, y las variables en gris son observadas (disponibles en la base de datos).

El archivo `NoMontyHall.csv` tiene los datos de los episodios: la caja que reservamos c_t, la pista que nos ofrecen s_t, y la posición del regalo r_t. Las variables que hemos agregado para modelar el problema, la probabilidad de acordarse p y las variables de cada episodio que representan si que efectivamente se acuerda o no a, permanecen ocultas.

2.1. Calcular el posterior sobre la probabilidad p de acordarse.

Para inferir la probabilidad de acordarse p es necesario calcular su posterior dado los datos. En particular nos interesa calcular el posterior de la probabilidad de acordarse p dado los datos de todos los episodios.

$$P(p|\text{Datos} = \{(c_0, s_0, r_0), (c_1, s_1, r_1), \ldots\}) = \frac{\overbrace{\prod_t P(c_t, s_t, r_t|p)}^{\text{Verosimilitud}} \overbrace{P(p)}^{\text{Prior}}}{\sum_p P(p) \prod_t P(c_t, s_t, r_t|p)} \tag{2}$$

Cuando estudiemos el flujo de inferencia en estructuras causal vamos a ver la justificación que nos permite descomponer la verosimilitud del conjunto de datos como el producto de cada uno de los episodios.

$$P(\text{Datos} = \{(c_0, s_0, r_0), (c_1, s_1, r_1), \dots\}|p) \stackrel{?}{=} \prod_t P(c_t, s_t, r_t|p) \tag{3}$$

Proponemos esta simplificación válida para que puedan calcular cada uno de los elementos de la verosimilitud de la siguiente forma.

$$\begin{aligned} P(c_t, s_t, r_t|p) &= \sum_a P(c_t, s_t, r_t, a_t|p) \\ &= \sum_a P(r_t)P(c_t)P(a_t|p)P(s_t|r_t, c_t, a_t|p) \end{aligned} \tag{4}$$

Si bien la probabilidad de acordarse puede ser considerada una variable continua, para resolver este problema es suficiente que evalúen un conjunto finito de valores, 11 valores desde 0 a 1 equidistantes.

2.2. Calcular la predicción de un episodio dado los datos de los episodios anteriores

Para calcular la predicción del siguiente episodio dada la información de los eventos anteriores vamos a usar el último posterior de p como priori para el nuevo episodio.

$$P(\text{Episodio}_T = (c_T, s_T, r_T)|\text{Datos}_{1:T-1} = \{(c_0, s_0, r_0), \dots, (c_{T-1}, s_{T-1}, r_{T-1})\}) \tag{5}$$

Para calcular la predicción del episodio T vamos aplicar las siguientes transformaciones.

$$\begin{aligned} P(c_T, s_T, r_T|\text{Datos}_{1:T-1}) &= \sum_p \sum_{a_T} P(c_T, s_T, r_T, a_T, p|\text{Datos}_{1:T-1}) \\ &\stackrel{?}{=} \sum_p \sum_{a_T} P(r_T)P(c_T)P(s_T|r_T, c_T, a_T)P(a_T|p)P(p|\text{Datos}_{1:T-1}) \end{aligned} \tag{6}$$

Cuando estudiemos el flujo de inferencia en estructuras causales vamos a ver la justificación ($\stackrel{?}{=}$) que nos permite descomponer la verosimilitud del conjunto de datos como el producto de cada uno de los episodios.

2.3. Calcular la predicción que hace el modelo alternativo M_A sobre todo el conjunto de datos.

Calcular la verosimilitud del modelo alternativo como el producto de las predicciones de cada uno de los episodios dado los episodios anteriores.

$$P(\text{Datos}_{1:T}|M_A) = P(\text{Episodio}_1|M_A)P(\text{Episodio}_2|\text{Datos}_{1:1}, M_A)P(\text{Episodio}_3|\text{Datos}_{1:2}, M_A)\dots$$

La predicción sobre un conjunto de datos grandes necesariamente resulta ser un número muy cercano a 0. Esto ocurre porque los elementos del producto son probabilidades, números entre 0 y 1, por lo que a medida que vamos agregando episodios este número se va acercando tanto al cero que deja de poder ser representado por una computadora. Para poder expresarlo en una computadora, vamos a calcular el exponente asociado a ese número, que crece exponencialmente más lentamente.

$$\log_{10} P(\text{Datos}_{1:T}|M_A) = \log_{10} P(\text{Episodio}_1|M_A) + \log_{10} P(\text{Episodio}_2|\text{Datos}_{1:1}, M_A) + \dots$$

2.4. Comparar el desempeño del modelo alternativo respecto de los modelos Base y el modelo MontyHall.

Cuando trabajamos con el exponente de las predicciones no vamos a poder calcular el posterior de los modelos directamente. En estos casos, para comparar el desempeño de los modelos lo que hacemos es comparar modelos de a pares.

$$\frac{P(M_i|\text{Datos})}{P(M_j|\text{Datos})} = \frac{P(\text{Datos}|M_i)P(M_i)}{P(\text{Datos}|M_j)P(M_j)} \overset{*}{=} \frac{P(\text{Datos}|M_i)}{P(\text{Datos}|M_j)} \tag{7}$$

Al dividir el valor a posteriori de dos modelos alternativos i y j, se cancela el denominador constante teorema de Bayes $P(\text{Datos})$. Además, en el caso de que tengamos un prior uniforme entre modelos ($\overset{*}{=}$), también se cancelan los priors y la comparación del posterior de los modelos se reduce a la comparación de sus predicciones. Además, si calculamos el exponente, obtendremos la diferencia de desempeño como la diferencia en órdenes de magnitud de las predicciones que hacen los modelos, lo que se conoce como *log Bayes Factor*. Por ejemplo, si comparamos en órdenes de magnitud el desempeño predictivo de los modelos Base M_0 y Monty Hall M_1 sobre los datos generados en el ejercicio anterior obtendremos una diferencia de aproximadamente 4 órdenes de magnitud.

$$\log_{10} \underbrace{\frac{P(\text{Datos}|M_1)}{P(\text{Datos}|M_0)}}_{\text{Bayes factor}} = \underbrace{\log_{10} P(\text{Datos}|M_1) - \log_{10} P(\text{Datos}|M_0)}_{\text{Diferencia predicitva en ordenes de magnitud}} \approx (-17) - (-21) = 4 \tag{8}$$

pues en el ejercicio anterior la predicción que el modelo base hizo del conjunto de datos era $P(\text{Datos}|M_0) \approx 3{,}37 \times 10^{-17}$ y la predicción que el modelo Monty Hall hizo era de $P(\text{Datos}|M_0) \approx 8{,}23 \times 10^{-21}$.

Para interpretar el significado del exponente del Bayes factor (8) es importante recordar que la verosimilitud de los modelos $P(\text{Datos}|M)$ funciona como filtro de la creencia previa. En base 10, una diferencia de un orden de magnitud significa que uno de los modelos preservó 10 veces más creencia que el otro, dos ordenes de magnitud significa que un modelos preservó 100 veces más creencia que el otro, y así sucesivamente. Aunque estos números parezcan extraordinarios, cuatro órdenes de magnitud se considera en el límite de una diferencia no concluyente. Cuando las bases de datos crecen, la diferencia en órdenes de magnitud continúan creciendo, por lo que es normal ver diferencia de 10000, pero en órdenes de magnitud! En esos casos, para ganar intuición es útil calcular la predicción "típica".

2.5. Calcular la predicción típica que hace el modelo de los episodios.

Cuando decimos predicción típica nos referimos a la media geométrica de las predicciones.

$$\text{Predicción típica} = \underbrace{(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M))^{1/N}}_{\text{Media geométrica}} \tag{9}$$

Decimos que es típica porque podemos reemplazar cada una de las predicciones que componen en la secuencia de predicciones por la media geometrica sin alterar el valor final. Es decir,

$$\begin{aligned} P(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M) &= P(d_1|M)P(d_2|d_1, M)\ldots \\ &= \prod_{i \in \{1,\ldots,N\}} \text{Predicción típica} \end{aligned} \tag{10}$$

Así expresada, la media geométrica tendría el mismo problema de representación computacional que señalamos al inicio de este ejercicio. Para calcularla hay que trabajar con el exponente de la predicción.

$$\begin{aligned}\text{Predicción típica} &= 10^{\overbrace{\log_{10}(P(d_1|M)P(d_2|d_1,M)\dots)^{1/N}}^{\text{Exponente de la predicción típica}}} \\ &= 10^{\frac{1}{N}\left(\log_{10} P(d_1|M)+\log_{10} P(d_2|d_1,M)+\dots\right)}\end{aligned} \tag{11}$$

Si calculan la predicción típica del modelo Base en los datos del ejercicio anterior verán que es de 0,382 y la del modelo Monty Hall de 0,454. Dado que observamos en total $N = 48$ datos (3 datos en cada una de los $T = 16$ episodios), podemos usar la predicción típica para recuperar la predicción conjunta.

$$P(\text{Datos}|M_0) = 0{,}382^{48} \qquad\qquad P(\text{Datos}|M_1) = 0{,}454^{48}$$

¿Por qué podemos decir que en promedio (geométrico) el modelo base preserva solo el 38,2 % de la creencia previa luego de cada nueva observación, mientras que el modelo Monty Hall preserva 45,4 %?

2.6. Calcular el posterior en los primeros episodios y graficar

Para poder graficar cómo cambia la creencia de los modelos a medida que vamos observando episodios vamos a calcular el posterior de los tres modelos en los primeros 60 episodios. Debería quedar algo similar a lo siguiente.