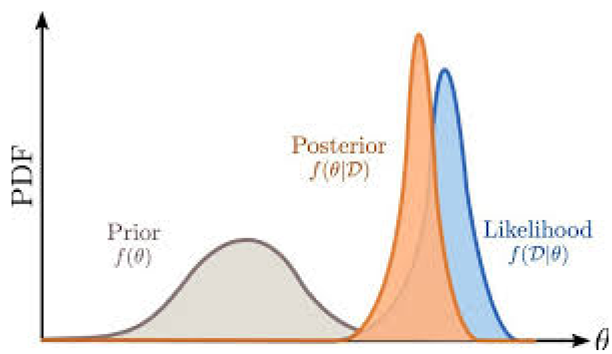
Reglas del TP:

- Este trabajo debe hacerse de forma individual o de a dos.
- Deben enviarse el script de R y un informe que contenga las respuestas a todas las preguntas y los gráficos pedidos, en lo posible en RMarkdown. No hace falta explicar en el informe que es lo que hace cada una de las funciones del script.
- El script debe estar prolijo. Esto en particular implica que las variables tienen que tener un nombre descriptivo (es decir, no llamar `a`, `b`, `c` a las variables).

Markov Chain Monte Carlo: el método de Metrópolis-Hastings



Vamos a plantear dos problemas de estimación Bayesiana:

1. Se quiere saber cuál es la proporción p de fumadores en la Capital Federal. Se asume una distribuión a priori $Beta(3, 5)$ para p. Se encuestan a 300 personas, y resulta que exactamente 50 de ellas fuman. Con esta información, se quiere saber el estimador Bayes para p.
2. Supongamos que para un determinado genotipo, la probabilidad del alelo a es p y la del alelo A es $1 - p$. Si se asume que la población se reproduce aleatoriamente, esto implica que la probabilidad de los genotipos aa, aA y AA son p^2, $2p(1 - p)$ y $(1 - p)^2$ respectivamente. Se asume una distribución a priori $U(0, 1)$ para p. Se observan 13 personas con el genotipo aa, 210 con el genotipo aA y 240 con el genotipo AA. Con esta información, se quiere saber el estimador Bayes para p.

El problema 1. se puede resolver analíticamente, pero el 2 no, porque no se llega a ninguna distribución conocida al calcular la distribución a posteriori. A continuación vamos a describir el Método de Metrópolis-Hastings (simplificado), que permite samplear la distribución a posteriori que se quiere encontrar.

Supongamos que $g(p)$ es la propbabilidiad puntual o densidad de la distribución a priori y $f(\mathbf{x}|p)$ es la función de verosimilitud (donde $\mathbf{x}$ representa la muestra entera). Llamemos P a la distribución a posteriori que queremos aproximar. Vamos a definir una secuencia $p_1, p_2, \ldots$ de la siguiente manera:

- $p_1 = p_{init}$ con un p_{init} a elección
- Para $n \geq 1$, se define
  1. El candidato $w_n = p_n + k_n$ donde $k_n \sim N(0, \sigma_0^2)$ (independientes) con un valor σ_0^2 a elección.
  2. La probabilidad de aceptación $r_n = \min\left(1, \dfrac{g(w_n)f(\mathbf{x}|w_n)}{g(p_n)f(\mathbf{x}|p_n)}\right)$.
  3. Se tira una moneda con probabilidad r_n de éxito. Si se obtiene un éxito, entonces $p_{n+1} = w_n$ (es decir, se acepta al candidato). En caso contrario, $p_{n+1} = y_n$.
- Frenar el proceso luego de N iteraciones (con un N a elección).

El método asegura que la secuencia $p_1, p_2, \ldots$ en un momento se aproxima a ser estacionaria y cuando eso sucede, se estaciona con la distribución a posteriori P que nos interesa. De esta manera, podemos descartar los primeros M valores (cuando consideremos que la cadena se estabilizó) y pensar al resto $(p_{M+1}, \ldots, p_N)$ como una muestra de la distribución a posteriori P. Por ejemplo, podemos calcular $E(P)$ como el promedio de los valores $p_{M+1}, \ldots, p_N$.

Se pide lo siguiente:

1. (a) Encontrar analíticamente la distribución a posteriori para el problema 1.
   (b) Aproximar la distribución a posteriori implementando el método de Metrópolis-Hastings para este problema (elegir $p_{init} = 0.5$ y $\sigma^2 = 0.01$, $N = 50000$ y el M que parezca adecuado, mirando cuando se estabiliza la secuencia.)
   (c) Comprobar que las distribuciones a posteriori de los dos items anteriores son muy similares (hacemos esto para convencernos de que este método funciona bien).
   (d) (Opcional) Explorar probando diferentes valores de p_{init}, σ^2, N y M.
2. (a) Aproximar la distribución a posteriori implementando el método de Metrópolis-Hastings para el problema 2 (elegir $p_{init} = 0.5$).
   (b) Aproximar el estimador Bayes en este problema.
   (c) (Opcional) Probar otras distribuciones a priori y observar las diferencias.