Guía 2

Familias conjugadas

1. Se observa la realización de n variables aleatorias Poisson independientes de parámetro λ. Escribir y graficar la función de Likelihood $L(\lambda)$ para cada caso:

   a. $(y_1, y_2, y_3) = (3, 7, 19)$.
   b. $(y_1, y_2, y_3, y_4) = (12, 12, 12, 0)$.
   c. $y_1 = 12$.
   d. $(y_1, y_2, y_3, y_4, y_5) = (16, 10, 17, 11, 11)$.

2. Suponiendo un prior $\lambda \sim \text{Gamma}(24, 2)$, especificar y graficar la distribución posterior para λ correspondiente a cada escenario del ejercicio anterior.

3. Repetir el ejercicio anterior suponiendo un prior $\lambda \sim \text{Gamma}(2, 2)$

4. Supongamos que uno tiene una variable aleatoria Y que puede modelar con una distribución geométrica. Es decir que $P(Y = y \mid \theta) = \theta\,(1-\theta)^{y-1}$ para $y \in \{1, 2, 3, \dots\}$. Se utiliza un prior $Beta(a, b)$ para θ.

   a. ¿Qué situación se representa con una variable aleatoria geométrica?
   b. Derivar la distribución posterior para θ suponiendo que se observó $Y = y$. Identificar la distribución encontrada y sus parámetros.
   c. El modelo Beta es un prior conjugado de la Geométrica?

5. Supongamos que λ es el promedio del número total de goles por partido del mundial de fútbol femenino. En este ejercicio vamos a analizar λ usando el modelo Gamma-Poisson donde Y_i es el número de goles que se hicieron en un partido i del Mundial. Es decir que $P(Y_i \mid \lambda) = \text{Pois}(\lambda)$. Como prior, tomar: $P(\lambda) = \text{Gamma}(1, 0.25)$.

   a. Graficar el conocimiento previo sobre λ.
   b. ¿Por qué el modelo Poisson es razonable para Y_i?
   c. Escribir la distribución posterior para λ suponiendo que en los primeros tres partidos se meten $(y_1, y_2, y_3) = (3, 7, 4)$ goles.
   d. OPCIONAL: Repetir usando datos reales, por ejemplo el dataset `wwc_2019_matches` del paquete `fivethirtyeight`.

Simulaciones, muestreo de la posterior y algoritmo de Metropolis-Hastings

1. Simular 10000 datos de una distribución Beta(3,7). Graficar un histograma de los datos y la distribución teórica superpuesta. Dar tres medidas de resumen de la distribución usando los datos simulados.

2. Repetir el ejercicio 1 para la distribución Gamma(4,2) y para la Normal(4,1).

3. Supongamos que tomaron datos y y con un modelo bayesiano llegaron a que la distribución posterior para la probabilidad de que respondan un mail dentro de las 24 horas, θ, es Beta(2,5). Les llega un nuevo

mail. ¿Cuál es la probabilidad de que lo respondan en menos de 24 horas? Es decir, se pide $P(\tilde{y} = 1 \mid y)$, la *posterior predictive*. Resolverlo de dos formas:

   a. Calculando analíticamente $P(\tilde{y} = 1 \mid y)$.
   b. Simulando datos. Para esto:
      i. Generar 10000 valores simulados del parámetro θ usando la distribución posterior.
      ii. Para cada valor del parámetro θ_i, simular un y_i predicho.
      iii. ¿Cuál es la frecuencia de $y = 1$ entre los predichos? ¿Coincide con lo calculado en a?.

4. Supongamos que quieren saber si los mails que reciben los lunes los responden más rápido que los que reciben los sábados. Con datos que recolectan de su propia experiencia llegan a que la distribución posterior para θ_L es Beta(3,7) (correspondiente a mails que llegan los lunes) y θ_S es Beta(4,8) para los que llegan los sábados. Para responder la pregunta simulando, sigan estos pasos:

   a. Simular 10000 valores de θ_L y θ_S.
   b. Calcular, para cada para de valores $\theta_L(i)$ y $\theta_S(i)$, la diferencia $\delta_i = \theta_L(i) - \theta_S(i)$.
   c. Usar las muestras aleatorias δ_i para responder la pregunta. Por ejemplo, preguntarse por la probabilidad de que esa diferencia sea positiva.

5. Implementar el algoritmo de Metropolis-Hastings para conseguir muestras de una distribución Normal($\mu = 0.4, \sigma = 0.6$). Construir una función que tome como argumento el largo de la cadena (número de muestras), el valor inicial y algún parámetro necesario de la función de propuesta.

6. (*BayesRules 7.15*) Identificar una pregunta que se pueda responder con un modelo Beta-Binomial para la probabilidad θ de que ocurra algo. Por ejemplo: proporción de colectivos que vienen llenos cuando vienen a Exactas, etc.

   a. Proponer un prior para θ.
   b. Juntar datos (de verdad o inventados).
   c. Simular 2000 valores de θ obtenidos con el algoritmo de Metropolis-Hastings.
   d. Graficar la cadena resultante (secuencia de muestras). ¿Están satisfechos con el resultado? ¿A qué hay que estar atentos para aceptar las muestras?

7. (*BayesRules 7.16*) Identificar una pregunta que se pueda responder con un modelo Normal para μ, un valor medio de interés. Por ejemplo: la temperatura máxima promedio en Otoño en Buenos Aires, el tiempo medio de uso de celular diario, etc.

   a. Proponer un prior para μ.
   b. Juntar datos (de verdad o inventados).
   c. Simular 2000 valores de μ obtenidos con el algoritmo de Metropolis-Hastings. Usar σ fijo.
   d. Graficar la cadena resultante (secuencia de muestras). ¿Están satisfechos con el resultado? ¿A qué hay que estar atentos para aceptar las muestras?

Modelo lineal

1. Para el siguiente modelo, simular observaciones y_{obs} con el prior

$$\begin{aligned} y &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \text{Normal}(0, 10) \\ \sigma &= \text{Exponencial}(1) \end{aligned}$$

2. Escribir el modelo del ejercicio 1 en `brms`.

3. Traducir el siguiente modelo (en sintaxis de `brms`) matemáticamente>

```
    family = gaussian,
    y ~ 1 + x,
    prior = c(prior(normal(0, 10), class = Intercept),
              prior(normal(0, 10), class = b),
              prior(exp(1),  class = sigma))
```

4. Usando el dataset de alturas `Howell1` que viene con el paquete `rethinking`, dar una predicción para la altura de individuos que pesan 46.95, 43.72, 64.78, 32.59 y 54.63 Kg. Dar también un intervalo de credibilidad del 95% para estas predicciones.

5. Del mismo dataset, seleccionar sólo los individuos que tienen menos de 18 años. Ajustar con `brms` un modelo de regresión lineal para la altura teniendo como explicativa al peso de los individuos. Graficar los datos y la recta obtenida con los valores medios de la posterior y un intervalo de credibilidad para esos valores. También graficar un intervalo de credibilidad para la altura predicha por el modelo. Todo en el mismo gráfico. ¿Te parece adecuado el modelo? ¿Qué aspectos podrías cambiar para mejorar el modelo?

6. Usando el dataset `penguins` del paquete `palmerpenguins`, estudiar la relación entre el largo de las aletas de los pinguinos (`flipper_length`) y su peso (`body_mass_g`).

    a. Proponer un modelo en el que el peso tiene una distribución normal con parámetros μ y σ, donde μ se basa en el largo de las aletas.
    b. Correr el modelo en STAN usando `brms` o `rstanarm`, eligiendo priors para los parámetros o usando los priors default.
    c. Diagnosticar las cadenas de muestras de las posteriores para cada parámetro usando el número efectivo de muestras (n_{eff}) y $\hat{R}$.
    d. Encontrar la distribución posterior para el peso esperado de un pinguino que tiene una aleta de largo 200 mm. Graficarla y dar medidas resumen.
    e. Hacer 500 predicciones del peso de un pinguino con una aleta de largo 200 mm. Graficar la distribución de estos pesos predichos y comparar con el resultado del ítem anterior. ¿A qué se debe la diferencia?
    f. Graficar 100 rectas correspondientes al peso esperado de pinguinos con una aleta de largo entre 150 mm y 250 mm. Hacer lo mismo para el peso predicho por el modelo.