Estadística bayesiana

Tarea Examen

Fecha de entrega: 3 de junio

1. (2 puntos) Sean $Y_1, \ldots, Y_n$ una colección de variables aleatorias independientes tales que $Y_i \sim \text{Ga}(\alpha_i, \beta)$, si $V = \sum_i Y_i$.

   (a) ¿Cuál es la distribución del vector aleatorio $W = (Y_1/V, \ldots, Y_n/V)$?

   (b) Asuma que los datos $(x_1, \ldots, x_k)$ siguen una distribución multinomial de parámetros $(\theta_1, \ldots, \theta_k)$ cuya distribución a priori se asume Dirichlet. Sea $\pi = \frac{\theta_1}{\theta_1+\theta_2}$. ¿Cuál es la distribución posterior de π? ¿Qué puedes concluir de este resultado?

   (c) Considere los datos recolectados de dos encuestas (una antes del debate presidencial y otra posterior al debate) realizadas a los alumnos de la Facultad de Ciencias (Tabla 1) [1]. En cada una de estas encuestas fueron entrevistadas 639 personas diferentes. Asumiendo que las encuestas son independientes, modele los datos con dos distribuciones multinomiales diferentes. Para $j = 1, 2$ denote por π_j la proporción de los votantes que prefieren a Claudia por delante de Xóchitl (antes y después del debate). Obtenga un histograma de la distribución posterior de $\pi_2 - \pi_1$. ¿Qué conclusiones puedes obtener de tus resultados?

Tabla 1: Resultados de las encuestas de las preferencias electorales antes y después del debate.

Encuesta	Claudia	Xóchitl	Maynez	Total
antes del debate	307	294	38	639
después del debate	332	288	19	639

2. (1.5 puntos) Asumir que se está interesado en aprender de los hábitos de sueño de los estudiantes del curso de Estadística Bayesiana de la Facultad de Ciencias. En la Tabla 2 se muestras las horas de sueño de 20 alumnos elegidos al azar.

   (a) Ajuste un modelo bayesiano asumiendo que los datos se distribuyen normal con media y varianza desconocidas.

[1] (Disclaimer) Los datos son ficticios, cualquier parecido con la realidad es meramente una coincidencia.

Tabla 2: Horas de sueño de 20 estudiantes elegidos al azar.

9.5	8	7.5	8.5	6	12	6	9	8.5	7.5
8	6.5	9	8	7	9.5	9	7.5	5.5	6.5

(b) Obtener un intervalo de credibilidad al 90% para la media μ y la desviación estándar σ.

(c) ¿Cómo se podría obtener la media y desviación posterior del cuartil $q_{.75}$?

3. (1.5 puntos) Considera la base de datos *transplantesCorazon.txt* en donde se recaba el número de muertes en un lapso de 30 días posteriores a una cirugía de transplante de corazón de 94 hospitales y la exposición esperada de cada hospital, esto es el número esperado de muertes. Se desea estudiar las tasas de mortalidad post-cirugía de transplante de corazón de estos hospitales. Para esto, suponga que las tasas de mortalidad $\lambda_1, \ldots, \lambda_{94}$ vienen de una distribución gamma (α, β) donde se asume que α y β son independientes con distribución

$$g(\alpha, \beta) \quad = \quad \frac{1}{(\alpha+1)^2} \frac{1}{(\beta+1)^2}.$$

(a) A partir de un algoritmo MCMC obtén una muestra de tamaño 1000 de la distribución posterior de $\theta = (\log(\alpha), \log(\beta))$.

(b) Con tus simulaciones construye un intervalo de credibilidad al 95% para las tasas de mortalidad de cada hospital.

4. (3 puntos) Una empresa dedicada a estudiar el rendimiento físico (antes, durante y después del entrenamiento) de caballos de carrera, ha decidido cambiar los sensores de medición de lactato. A raíz del cambio se ha llegado a notar que los dos aparatos dan mediciones diferentes. Calibrar de nueva cuenta el algoritmo de detección de fatiga puede ser bastante costoso por lo que se ha decidido en su lugar encontrar una relación de las mediciones. Utilizando los datos recabados (*Sensores.txt*) ajusta un modelo de regresión lineal bayesiano para darle solución al problema. Tu reporte deberá contener los siguientes elementos.

(a) Procesamiento de los datos (si es requerido) y un análisis de los datos.

(b) La elección de la distribución a priori para los parámetros.

(c) La distribución posterior conjunta, exhibiendo su expresión y su superficies en un gráfico. En el caso de una regresión con dos o más coeficientes, dibujar las curvas de nivel.

(d) Reportar los estimadores bayesianos bajo una pérdida cuadrática y los intervalos de densidad posterior al 95% de probabilidad.

(e) Utilizando los estimadores bayesianos, exhibe en un gráfico tu ajuste junto con sus intervalos de densidad posterior.

(f) Encontrar la distribución predictiva para un nuevo conjunto de covariables. Si para una nueva medición con el dispositivo nuevo se observa un nivel de lactato de 23, ¿qué valor le correspondería en el dispositivo actual?

5. (1 punto) En el modelo de regresión bayesiana una distribución usual para $\beta = (\beta_1, \ldots, \beta_k)$ asume que son condicionalmente independientes dado $\Psi = (\psi_1, \ldots, \psi_k)$ y donde para cada j, $\beta_j \sim \mathcal{N}(0, \Psi_j)$. De esta forma, el valor de Ψ_j se puede pensar como una forma de cuantificar la importancia de la j-ésima variable, con valores grandes Ψ_j representando una mayor importancia.

   (a) Considerando un modelo normal - gamma, esto es,

   $$\beta_j \sim \mathcal{N}(0, \Psi_j), \qquad \Psi_j \sim \text{Ga}(\lambda, \alpha),$$

   ¿Cuál es la varianza a priori de β_j? ¿Qué puedes decir de las colas de la distribución?

   (b) Suponer ahora la siguiente estructura jerárquica

   $$\beta_j \sim \mathcal{N}(0, \Psi_j), \qquad \Psi_j \sim \text{Ga}(\lambda, \alpha_j), \qquad \alpha_j \sim \text{Ga}(\phi, \kappa), \tag{1}$$

   donde κ es un parámetro de escala, λ controla el comportamiento de la distribución cerca del cero y ϕ controla el comportamiento en las colas de la distribución. ¿Cuál es la varianza a priori para β_j para este modelo?

   (c) ¿Puedes encontrar la distribución de las Ψ_j?

   (d) A partir de la distribución de Ψ_j, realiza un estudio de las colas para los siguientes casos[2].

   i. $\lambda = \phi = 0.5$.

   ii. $\lambda = 1$ y $\phi > 1$.

   iii. $\lambda = \kappa = \phi = 1$.

6. (.5 punto) Sea $X \sim \mathcal{N}(\theta, 1)$ y supón que se escoge una distribución inicial con densidad

   $$q(\theta) \;=\; \frac{1}{2}\mathcal{N}(\mu, 1) + \frac{1}{2}\mathcal{N}(-\mu, 1).$$

   (a) Encuentra la distribución posterior.

   (b) Encuentra el estimador bayesiano bajo una función de pérdida cuadrática.

[2]De no poder encontrar la distribución cerrada (¡sí existe!) este análisis se puede intentar resolver de forma empírica, simulando de la estructura jerárquica (1).

(c) Encuentra la distribución predictiva.

7. (1.5 puntos) Considera la base *galaxias.txt*, la cual contiene la velocidad (con respecto a la nuestra) de 82 galaxias observables de seis secciones cónicas bien separadas de nuestro universo. Modela estos datos a partir de un modelo de mezclas finito gaussiano. Para esto:

   (a) Asume una distribución Dirichlet simétrica de parámetro α para los pesos de la mezcla y una distribución normal - inversa gamma para la media y varianza de cada componente de la mezcla.

   (b) Grafica la densidad estimada junto con el histograma de los datos y responde lo siguiente:

      i. ¿Cómo cambia la estimación conforme α tiende a cero?

      ii. ¿Cómo cambia la estimación conforme el número de componentes aumenta?