Regresión Avanzada ITAM
26 de abril de 2024

Examen Parcial II

A continuación se dan las reglas de este examen.

- El examen es individual y no puede haber comunicación entre los estudiantes.
- La fecha de entrega es a más tardar el viernes 3 de mayo antes de las 12:00am
- El examen deberá subirse a Canvas en el lugar creado para este examen.
- Se abrirá un foro de discusión para aclarar preguntas del examen. No se atenderán las preguntas que no se hagan por este medio. La intención es que todos cuenten con la misma información. En este foro, el único que puede responder a las dudas es el instructor.

Preguntas

1. La siguiente tabla muestra el tamaño de muestra n_iy el tiempo-medio-al-servicio $\bar{y}_i$ (en segundos) para seis jugadores profesionales de tenis. Supongan que la media de la muestra para el jugador i $\bar{y}_i$ se distribuye normal con media μ_i y desviación estándar $\sigma/\sqrt{n_i}$ donde se supone que $\sigma = 5{,}5$ segundos.

Jugador	n	$\bar{y}$
Murray	731	23.56
Simon	570	18.07
Federer	491	16.21
Ferrer	456	21.70
Isner	403	22.32
Kyrgios	274	14.11

   a. Se está interesado en estimar el tiempo-medio-al-servicio de Murray μ_1 usando solo el tiempo al servicio de Murray. Supóngase que la creencia inicial sobre μ_1 se representa por una densidad normal con media 20 y desviación estándar 10. Encontrar la distribución posterior de μ_1 y construir un intervalo de credibilidad de 90 % para μ_1

b. Suponer ahora que se cree que no hay diferencias entre los tiempos-medios-al servicio y $\mu_1 = \ldots = \mu_6 = \mu$. El tiempo-medio-al-servicio global es $\bar{y} = 19{,}9$ con un tamaño de muestra combinado de $n = 2925$. Suponiendo que μ tiene una distribución inicial $\mathcal{N}(20, 10)$, encontrar la posterior de μ y construir un intervalo de credibilidad de 90 % para μ.

c. ¿Qué enfoque, parte (a) o parte (b), parece más razonable en esta situación?

2. Continuando con el problema anterior, suponer que se quiere estimar el tiempo-medio-al servicio para los seis tenistas siguiendo un modelo jerárquico. Recordar que se supone $\sigma = 5{,}5$ segundos.

$$
\begin{aligned}
\bar{y}_i &\sim \mathcal{N}(\mu_i, \sigma/\sqrt{n_i}), \quad i = 1, \ldots, 6 \\
\mu_i &\sim \mathcal{N}(\mu, \tau), \quad i = 1, \ldots, 6 \\
\mu &\sim \mathcal{N}(20, 1/0{,}0001), \\
1/\tau^2 &\sim \mathcal{G}(0{,}01, 0{,}1)
\end{aligned}
$$

a. Usar JAGS o Stan para simular una muestra de tamaño 1000 de la distribución posterior del modelo jerárquico

b. Construir un intervalo de credibilidad para cada una de las medias.

c. Comparar los intervalos de credibilidad para Murray con los intervalos obtenidos en el ejercicio 1.

3. Resolver el siguiente ejercicio. Pueden usar JAGS o Stan

Table 10.7 displays the number of fire calls and the number of building fires for ten counties in Montgomery County, Pennsylvania from 2015 through 2019. This data is currently described as Emergency - 911 Calls" from `kaggle.com}. Suppose that the number of building fires for the j-th zip code is Poisson with mean $n_j \lambda_j$, where n_j and λ_j are respectively the number of fire calls and rate of building fires for the j-th zip code.

Table 10.7. The number of fire calls and building fires for ten zip codes in Montgomery County, Pennsylvania.

Zip Code	Fire Calls	Building Fires
18054	266	12
18103	1	0
19010	1470	59
19025	246	11
19040	1093	47
19066	435	26
19116	2	0
19406	2092	113
19428	2025	73
19474	4	1

a. Suppose that the building fire rates $\lambda_1, \ldots, \lambda_{10}$ follow a common Gamma(α, β) distribution where the hyperparameters α and β follow weakly informative distributions. Use JAGS to simulate a sample of size 5000 from the joint posterior distribution of all parameters of the model.
b. The individual estimates of the building rates for zip codes 18054 and 19010 are $12/266$ and $59/1470$, respectively. Contrast these estimates with the posterior means of the rates λ_1 and λ_3.
c. The parameter $\mu = \alpha/\beta$ represents the mean building fire rates across zip codes. Construct a density estimate of the posterior distribution of μ.
d. Suppose that the county has 50 fire calls to the zip code 19066. Use the simulated predictive distribution to construct a 90% predictive interval for the number of building fires.