Estadística Bayesiana
Examen Parcial # 1

Instrucciones generales

- Este caso de estudio constituye el 60% de la calificación del Examen Parcial 1.
- Debe asociarse con otra persona, entendiendo que la calificación del examen será la misma para ambas personas.
- El reporte final se debe enviar a más tardar el **miércoles 15 de marzo de 2023** a las 11:59 am a la cuenta de correo:

  `jcsosam@unal.edu.co`
- Reportar las cifras utilizando la cantidad adecuada de decimales, dependiendo de lo que se quiera mostrar y las necesidades del problema.
- Numerar figuras y tablas y proporcionarles un tamaño adecuado que no distorsione la información que estas contienen.
- El archivo del reporte final debe ser un archivo `pdf`.
- Usar `LateX` o `Markdown` (en `R` o `Python`) para escribir el informe.
- El código fuente de `R` o `Python` debe reproducir exactamente todos los resultados (incluir semillas donde sea necesario).
- La presentación, la organización, la redacción, y la ortografía serán parte integral de la calificación.

- Si los estudiantes Juan Sosa y Ernesto Perez trabajan juntos, tanto el archivo `pdf` del informe, así como el código fuente, y el asunto del e-mail donde se adjuntan estos archivos, se deben llamar de la siguiente manera:

  `bayes - parcial 1 - juan sosa - ernesto perez`

  Esta condición es indispensable para que su examen sea calificado.

- Usar reglas APA para hacer las referencias correspondientes. No copiar texto de libros o internet sin hacer la cita correspondiente.

- El informe no tiene que ser extenso. Recuerde ser minimalista escribiendo el reporte. Se deben incluir solo aquellos gráficos, tablas, y ecuaciones que sean relevantes para la discusión.

- Cualquier evidencia de plagio o copia se castigará severamente tal y como el reglamento de la Universidad Nacional de Colombia lo estipula. Dejo a mi discreción el uso de software especializado para evaluar si hay copia o plagio de otros informes o internet.

  Si está claro que (por ejemplo) dos grupos han trabajado juntos en una parte de un problema que vale 20 puntos, y cada respuesta habría ganado 16 puntos (si no hubiera surgido de una colaboración ilegal), entonces cada grupo recibirá 8 de los 16 puntos obtenidos colectivamente (para una puntuación total de 8 de 20), y me reservo el derecho de imponer penalidades adicionales a mi discreción.

  Si un grupo resuelve un problema por su cuenta y luego comparte su solución con cualquier otro grupo (porque rutinariamente Usted hace esto, o por lástima, o bondad, o por cualquier motivo que pueda creer tener; ¡no importa!), Usted es tan culpable de colaboración ilegal como la persona que tomó su solución, y ambos recibirán la misma penalidad. Este tipo de cosas es necesario hacerlas ya que muchas personas no hacen trampa, y debo asegurarme de que sus puntajes son obtenidos de manera genuina. En otros semestres, unos estudiantes perdieron la materia debido a una colaboración ilegal; ¡no deje que le suceda a Usted!

Conteo de victimas

Considere la base de datos disponible en

https://www.datos.gov.co/Justicia-y-Derecho/Conteo-de-V-ctimas/sft7-9im5

que contiene el total de víctimas según las entradas de noticias criminales por delito al Sistema Penal Oral Acusatorio en la Ley 906 de 2004 y Ley 1098 de 2006 desde hechos ocurridos en 2010. La base también se puede descargar directamente de la página web del curso en

https://sites.google.com/view/juansosa/bayesian-statistics

bajo el nombre `victimas-justicia-derecho.csv`. A la fecha, la base contiene 3,651,193 registros y 25 campos. Toda la información se encuentra disponible en la página web de referencia.

El objetivo de este caso de estudio es modelar el conteo total de víctimas en Bogotá D. C. en 2022 para establecer si existen diferencias significativas por sexo respecto a delitos sexuales en menores de edad.

Pre-procesamiento de la base de datos

Para ajustar los modelos propuestos, se consideran únicamente los individuos tales que:

- El proceso sí corresponde a un hecho.
- El estado de la noticia criminal es activo.
- El año en que se denunció el hecho es 2022.
- El año en que entró a la Fiscalía la noticia criminal es 2022.
- El año en que presuntamente ocurrió el hecho es 2022.
- El departamento es Bogotá D. C..
- La agrupación de los delitos del código penal es delitos sexuales.
- El país de nacimiento de la víctima es Colombia.
- El país en donde presuntamente ocurrieron los hechos que conoció la Fiscalía es Colombia.

- La agrupación de edad a la que pertenece la víctima es primera infancia, infancia, preadolecente, o adolescente.

La base de datos filtrada de esta manera sin tener en cuenta ningún otro aspecto de los demás campos contiene 394 registros (116 hombres, 270 mujeres, 8 sin información). Finalmente, se remueven los registros sin información de sexo, y acto seguido, también se remueven los *outliers* asociados con el conteo total de víctimas. Así, se obtienen conteos asociados con 115 hombres y 237 mujeres.

Nota: una observación de una variable de interés se denomina *outier* extremo si la observación es bien sea inferior a $q_1 - 3.0RI$ o superior a $q_3 + 3.0RI$, donde q_1 y q_3 son el percentil 25 y 75 de la variable, respectivamente, y $RI = q_3 - q_1$ es el rango intercuartílico.

PARTE 1: Análisis Bayesiano en 2022

Sea $\boldsymbol{y}_k = (y_{k,1}, \ldots, y_{k,n_k})$ el vector de observaciones correspondientes al conteo total de víctimas asociados con la población k, con $k = 1$ (hombres) y $k = 2$ (mujeres). Considere modelos Gamma-Poisson de la forma

$$\begin{aligned} y_{k,i} \mid \theta_k &\overset{\text{iid}}{\sim} \mathsf{Poisson}(\theta_k)\,, \quad i = 1, \ldots, n_k\,, \\ \theta_k &\sim \mathsf{Gamma}(a_k, b_k) \end{aligned}$$

donde a_k y b_k son hiperparámetros, para $k = 1, 2$.

1. Ajustar los modelos Gamma-Poisson de manera independiente con $a_k = b_k = 0.01$, para $k = 1, 2$. Hacer una visualización donde se presenten simultáneamente las distribuciones posteriores y las distribuciones previas correspondientes.

   Nota: usar un solo panel para la visualización.

2. Sea $\eta = (\theta_2 - \theta_1)/\theta_1$. Obtener la distribución posterior de η. Reportar la media, el coeficiente de variación, un intervalo de credibilidad al 95%. Presentar los resultados visual y tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

   Nota: usar métodos de Monte Carlo con una cantidad de muestras adecuada.

3. Llevar a cabo un análisis de sensitividad. Para ello, considere los siguientes estados de información externos al conjunto de datos:

- Distr. Previa 1: $a_k = b_k = 0.01$, para $k = 1, 2$.
- Distr. Previa 2: $a_k = b_k = 0.10$, para $k = 1, 2$.
- Distr. Previa 3: $a_k = b_k = 1.00$, para $k = 1, 2$.
- Distr. Previa 4: $a_k = 1.00$ y $b_k = 1/2$, para $k = 1, 2$.
- Distr. Previa 5: $a_k = 1.00$ y $b_k = 1/3$, para $k = 1, 2$.
- Distr. Previa 6: $a_k = 1.00$ y $b_k = 1/4$, para $k = 1, 2$.

En cada caso calcular la media y el coeficiente de variación a priori, y repetir el númeral anterior. Presentar los resultados visual y tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Nota: usar un solo panel para la visualización.

4. En cada población, evaluar la bondad de ajuste del modelo propuesto utilizando como estadísticos de prueba la media y la desviación estándar. Presentar los resultados visual y tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

   Nota: calcular los valores p predictivos posteriores.

PARTE 2: Análisis frecuentista en 2022

1. Repetir el númeral 2. de la PARTE 1 usando *Bootstrap* paramétrico [1].

   Nota: usar una cantidad de remuestras adecuada.

2. Simular 100,000 muestras aleatorias de poblaciones Poisson bajo los siguientes escenarios:

   - Escenario 1: $n_1 = 10$, $n_2 = 10$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.
   - Escenario 2: $n_1 = 20$, $n_2 = 20$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.
   - Escenario 3: $n_1 = 50$, $n_2 = 50$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.
   - Escenario 4: $n_1 = 100$, $n_2 = 100$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.

   donde $\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i}$ es la media muestral observada de la población k, para $k = 1, 2$. En cada escenario el valor verdadero de η es $\eta = (\bar{y}_2 - \bar{y}_1)/\bar{y}_1$.

   Usando cada muestra, ajustar el modelo de manera tanto Bayesiana (usando la primera configuración previa) como frecuentista (usando *Bootstrap* paramétrico), y en cada caso

[1] http://www.math.chalmers.se/Stat/Grundutb/CTH/tms150/1112/Boot.pdf, pág. 15.

calcular la proporción de veces que el intervalo de credibilidad/confianza al 95% contiene el valor verdadero de η. Reportar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

PARTE 3: Análisis Bayesiano y frecuentista en 2012-2022

Para cada año de 2012 a 2022 (inclusive), ajustar el modelo de manera tanto Bayesiana (usando la primera configuración previa) como frecuentista (usando *Bootstrap* paramétrico), y obtener tanto una estimación puntual como intervalos de credibilidad/confianza al 95% y 99% para η. Presentar los resultados visualmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Nota: usar un solo panel para la visualización.