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 14 de septiembre de 2023** a las 11:59 pm 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!

Introducción

Verizon es la principal compañía telefónica local (ILEC, *incumbent local exchange carrier*) para una gran área del este de Estados Unidos. Como tal, es responsable de brindar servicio de reparación a los clientes de otras empresas telefónicas de la competencia (CLEC, *competing local exchange carrier*) en esta región.

Verizon está sujeto a multas si los tiempos de reparación (el tiempo que lleva solucionar un problema) para los clientes de alguna CLEC son sustancialmente peores que los de los clientes de Verizon. El conjunto de datos `Verizon.csv` (Chihara and Hesterberg, 2019, Sec. 1.3) contiene una muestra de los tiempos de reparación de $n_1 = 1664$ clientes de Verizon (ILEC) y $n_2 = 23$ clientes de la competencia (CLEC). De acuerdo con los datos, los tiempos medios de reparación son $\bar{x}_1 = 8.41$ y $\bar{x}_2 = 16.51$ horas para ILEC y CLEC, respectivamente.

El objetivo de este caso es determinar si la diferencia entre los tiempos promedio de reparación es lo suficientemente grande para declararse como significativa, y por tanto, ser tenida en cuenta como evidencia para llevar a cabo una intervensión y multar a Verizon.

Modelo

La distribución Exponencial es popular para modelar tiempos dado que este modelo permite producir distribuciones con diferentes tasas de decaimiento y variedades de sesgo (para más información acerca de este modelo probabilístico, ver por ejemplo Sosa et al. 2014, Sec. 10.4).

Así, considere modelos Exponenciales independientes de la forma

$$y_{k,i} \mid \lambda_k \overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda_k) \quad \Longleftrightarrow \quad p(y_{k,i} \mid \lambda_k) = \frac{1}{\lambda_k} \exp\left(-\frac{y_{k,i}}{\lambda_k}\right), \quad y_{k,i} > 0\,, \quad \lambda_k > 0\,, \tag{1}$$

para $i = 1, \ldots, n_k$ y $k = 1, 2$ (1: ILEC, 2: CLEC), donde $y_{k,i}$ es el tiempo de reparación (en horas) del individuo i en el grupo k, n_k es el tamaño de la muestra del grupo k, y finalmente, $\mathbf{y}_k = (y_{k,1}, \ldots, y_{k,n_k})$ es el vector columna de observaciones correspondiente.

Preliminares

Los siguientes numerales son insumos importantes para llevar a cabo el desarrollo del caso. Estos numerales no hacen parte de la sección de preguntas y por lo tanto no es necesario entregarlos desarrollados en el informe.

- Muestre que la distribución Exponencial pertenece a la familia exponencial de densidades uniparamétrica.
- Muestre que $s_k = \sum_{i=1}^{n} y_{k,i}$ es un estadístico suficiente para λ_k.
- Muestre que si $X \sim \mathsf{Gamma}(\alpha, \beta)$, entonces $\frac{1}{X} \sim \mathsf{GI}(\alpha, \beta)$.

  Nota: la variable aleatoria X tiene distribución Gamma-Inversa con parámetros $\alpha > 0$ y $\beta > 0$, si la función de densidad de X está dada por:

  $$X \sim \mathsf{GI}(\alpha, \beta) \quad \Longleftrightarrow \quad p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right), \quad x > 0\,.$$

- Considere el modelo Bayesiano Gamma-Inversa-Exponencial dado por la distribución muestral (1) junto con la distribución previa $\lambda_k \sim \mathsf{GI}(a_k, b_k)$, donde a_k y b_k son los hiperparámetros del modelo:

  1. Represente el modelo por medio de un grafo acíclico dirigido (DAG, por sus siglas en inglés).
  2. Halle la distribución posterior $p(\lambda_k \mid \mathbf{y}_k)$.
  3. Halle la distribución marginal $p(\mathbf{y}_k)$.
  4. Muestre que la media posterior $\mathsf{E}(\lambda_k \mid \mathbf{y}_k)$ es un promedio ponderado entre la media previa $\mathsf{E}(\lambda_k)$ y la media muestral $\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i}$.

- Halle el estimador de máxima verosimilitud (MLE, por sus siglas en inglés) de λ_k y la información observada (¡no esperada!) de Fisher correspondiente.

Preguntas

Sea $\eta = \lambda_1 - \lambda_2$. A continuación se hace inferencia estadística sobre η son el fin de responder al objetivo propuesto.

PARTE 1: Análisis Bayesiano

1. Ajuste los modelos Gamma-Inversa-Exponencial con $a_k = 3$ y $b_k = 17$ en cada grupo. A partir de las distribuciones posteriores obtenga la distribución posterior de η. Reporte la media, el coeficiente de variación y un intervalo de credibilidad al 95% para η. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

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

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

   - Distribución previa 1: $a_k = 3$ y $b_k = 17$, para $k = 1, 2$.
   - Distribución previa 2: $a_k = 2$ y $b_k = 8.5$, para $k = 1, 2$.
   - Distribución previa 3: $a_k = 3$ y $b_1 = 16.8$ y $b_2 = 33$, para $k = 1, 2$.
   - Distribución previa 4: $a_k = 2$ y $b_1 = 8.4$ y $b_2 = 16.5$, para $k = 1, 2$.

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

3. En cada población, evalúe la bondad de ajuste del modelo propuesto utilizando la distribución previa 1, utilizando como estadísticos de prueba la media y la desviación estándar. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

   Nota: calcule los valores p predictivos posteriores y en cada grupo realice la visualización de las distribuciones predictivas de los estadísticos de prueba de manera conjunta (dispersograma con histogramas marginales).

PARTE 2: Análisis frecuentista

Repita el númeral 1. de la PARTE 1 usando la Normalidad asintótica del MLE, *Bootstrap* paramétrico y *Bootstrap* no paramétrico. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Nota: asintóticamente se tiene que $\hat{\lambda}_{\text{MLE}} \approx \mathsf{N}(\lambda, \hat{I}^{-1})$, donde $\hat{\lambda}_{\text{MLE}}$ es el MLE de λ y $\hat{I}$ es la información observada de Fisher.

Nota: cuando utilice *Bootstrap*, use una cantidad de remuestras adecuada y el método de los percentiles para calcular los intervalos de confianza.

PARTE 3: Simulación

Simule 100000 muestras aleatorias de poblaciones Exponenciales bajo los siguientes escenarios:

- Escenario 1: $n_1 = 10$, $n_2 = 10$, $\lambda_1 = \bar{y}_1$, y $\lambda_2 = \bar{y}_2$.
- Escenario 2: $n_1 = 20$, $n_2 = 20$, $\lambda_1 = \bar{y}_1$, y $\lambda_2 = \bar{y}_2$.
- Escenario 3: $n_1 = 50$, $n_2 = 50$, $\lambda_1 = \bar{y}_1$, y $\lambda_2 = \bar{y}_2$.
- Escenario 4: $n_1 = 100$, $n_2 = 100$, $\lambda_1 = \bar{y}_1$, y $\lambda_2 = \bar{y}_2$.

donde $\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i}$ es la media muestral observada del grupo k. Observe que el valor verdadero de η en cada caso es $\eta = \lambda_1 - \lambda_2 = \bar{y}_1 - \bar{y}_2$.

Usando cada muestra, ajuste el modelo de manera tanto Bayesiana (usando la distribución previa 1) como frecuentista (usando la Normalidad asintótica, *Bootstrap* paramétrico, *Bootstrap* no paramétrico), y en cada caso calcule la proporción de veces que el intervalo de credibilidad/confianza al 95% contiene el valor verdadero de η. Reporte los resultados tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Referencias

Chihara, L. M. and Hesterberg, T. C. (2019). *Mathematical statistics with resampling and R.* John Wiley & Sons.

Sosa, J. C., Ospina, L. E., and Berdugo, E. P. (2014). *Estadística descriptiva y probabilidades*. U. Externado de Colombia.