Estudio de caso # 2

Universidad Externado de Colombia

Departamento de Matemáticas

Estadística 2

Juan Sosa, Ph. D.

August 21, 2018

Instrucciones generales

- **Puede hacer el examen solo o puede 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 5 de septiembre de 2018** a las **11:59 p.m.** a la cuenta de correo: `juan.sosa@uexternado.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 (`http://unilearning.uow.edu.au/report/1fi.html`) y proporcionarles un **tamaño adecuado** que no distorsione la información que estas contienen.
- El archivo del reporte final debe ser un **archivo** `pdf` con el siguiente formato: Letra Calibri, tamaño 12, interlineado sencillo con espacio entre párrafos y texto justificado. Márgenes: Normal. Tamaño: Carta. Orientación: Vertical.

- Especificar el software donde se llevó a cabo el computo e **incluir el código** correspondiente como un anexo al final del reporte con el siguiente formato: Letra Courier New, tamaño 10, interlineado sencillo.

- El objetivo principal de este trabajo es la claridad lógica y la interpretación de LOS resultados. **El informe no necesita ser extenso.** Recuerde ser minimalista escribiendo el reporte. Se deben incluir solo aquelLOS gráficos y tablas (¡y valores en la tabla!) que son relevantes para la discusión.

- Hacer el informe ya sea en inglés o español. No ambos!

- **Cualquier evidencia de plagio o copia se castigará severamente** tal y como el reglamento de la Universidad Externado de Colombia lo estipula.

  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 otras clases, personas perdieron la clase debido a una colaboración ilegal; **no deje que le suceda a Usted!**

Intervención de un ente hospitalario

El siguiente corresponde a uno de LOS primeros casos de estudio que David Draper (uno de mis profesores en University of California, SC) compartió conmigo como estudiante de doctorado:

As a small part of a study I [David Draper] worked on at the RAND Corporation in the early 2000s, we obtained data on a random sample of $n = 117$ *women who came to a hospital in Santa Monica, CA, to give birth to premature babies. One outcome of interest was the length of stay (LOS) in the hospital that a woman in this sample experienced, recorded as an integer; it was possible for this variable to be recorded as 0 if the LOS was under 12 hours. The data values are provided in* `LOS.txt`*. The unknown* λ *of principal interest in this problem is the mean LOS for all women giving birth to premature babies at this Santa Monica hospital in 2000.*

Este estudio fue relevante en ese momento porque a partir de los resultados se tomaron decisiones de índole administrativo y financiero que permitieron prestar un mejor servicio a la comunidad de Santa Monica, CA.

a. ¿Cuál es la variable objeto de estudio? ¿Cuáles son las unidades de medición de la variable? ¿Esta variable es discreta o continua? ¿Cuál es la población objetivo?

b. Calcular las medidas de localización (mínimo, máximo, cuartiles, promedio) del LOS (presentar los resultados en una tabla).

c. Calcular las medidas de dispersión (rango, rango intercuartílico, varianza, desviación estándar, coeficiente de variación) del LOS (presentar los resultados en una tabla).

d. Hacer una tabla de frecuencias (relativas) del LOS y el diagrama de barras correspondiente.

e. Comentar brevemente los resultados obtenidos en los numerales b., c. y d.

f. Un posible modelo (estocástico) para el LOS es la **distribución Poisson** con parámetro λ. Este es un modelo popular en la comunidad científica para modelar variables de conteo porque es sencillo y produce resultados

adecuados en muchas situaciones prácticas. Hoy en día existen otros modelos más flexibles, pero más complicados de implementar.

Demostrar que si X es una variable aleatoria (discreta) tal que $X \sim \mathsf{Poisson}(\lambda)$, entonces tanto la media poblacional μ y la varianza poblacional σ^2 son iguales a λ, es decir, $\mu = \mathbb{E}[X] = \lambda$ y $\sigma^2 = \mathbb{V}\mathrm{ar}[X] = \lambda$.

Nota 1: recuerde que $X \sim \mathsf{Poisson}(\lambda)$ se lee "la variable aleatoria X sigue una distribución Poisson con parámetro λ".

Nota 2: si $X \sim \mathsf{Poisson}(\lambda)$, entonces la función de masa de probabilidad de X está dada por

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad \text{para } x \in \{0, 1, 2, \ldots\},$$

donde λ es el parámetro de la distribución y es tal que $\lambda > 0$. Observe que en este caso λ es el parámetro de la población que será objeto de estudio (inferencia).

g. Dado que el promedio poblacional μ es igual a λ, entonces un estimador (puntual) razonable de λ es el promedio muestral, es decir,

$$\hat{\lambda} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Por lo tanto, se estima que en ese hospital de Santa Monica, CA, la estancia promedio de una mujer dando a luz un bebe prematuro es de tres días y medio aproximadamente ($\bar{x} = 3.4359$).

1. Simular $M = 100,000$ muestras aleatorias de tamaño $n = 117$ de la población (distribución Poisson) usando la estimación puntual de λ, es decir, $\lambda = \bar{x} = 3.4359$. Para cada una de las muestras simuladas calcular el promedio muestral $\bar{x}$ correspondiente.

   Nota 1: use `set.seed(1234)` como *semilla* en `R` para ejecutar la simulación.

   Nota 2: para simular muestras aleatorias de una distribución Poisson se utiliza la función `rpois` de `R`.

2. En este caso, de acuerdo con el **Teorema del Límite Central** (TLC), la distribución probabilística (aproximada) de la media muestral $\bar{X}$ es Normal con media

$$\mu_{\bar{X}} = \mathbb{E}\left[\bar{X}\right] = \mu = \lambda = 3.4359$$

y varianza

$$\sigma^2_{\bar{X}} = \mathbb{V}\mathrm{ar}\left[\bar{X}\right] = \frac{\sigma^2}{n} = \frac{\lambda}{n} = \frac{3.4359}{117} = 0.02937\,.$$

En la misma figura, hacer un histograma de la distribución empírica de la media muestral $\bar{X}$ (en color negro) usando los valores simulados en el numeral 1. y sobre este histograma graficar la distribución probabilística de la media muestral $\bar{X}$ de acuerdo con el TLC (en color rojo). ¿Por qué es posible utilizar el TLC en este caso?

3. De acuerdo con el TLC, el intervalo de confianza para λ está dado por

$$\mathrm{IC}_{1-\alpha/2,\mathrm{TLC}}(\lambda) = \hat{\lambda} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}$$

donde $z_{1-\alpha/2}$ es el percentil $100(1-\alpha/2)$ de la distribución Normal estándar. Usando los datos del LOS, calcular e interpretar el intervalo de confianza correspondiente usando una confiabilidad del 95%.

h. Alternativamente, otra forma de hacer inferencia sobre el parámetro λ es seguir a R. Fisher (1890–1962) y usar el **método de máxima verosimilitud**. Para ello observe que que los LOS se asumen como una realización de una muestra aleatoria de una población con distribución Poisson con parámetro λ, es decir, $X_1, \ldots, X_n \overset{\mathsf{IID}}{\sim} \mathsf{Poisson}(\lambda)$.

1. Demostrar que la función de verosimilud es

$$L(\lambda) = \frac{e^{-n\lambda}\,\lambda^s}{\prod_{i=1}^n x_i!}\,,$$

donde $s = \sum_{i=1}^n x_i = 402$ es la suma de todas los LOS.

2. Demostrar que la función de log-verosimilitud es

$$\ell(\lambda) = -n\lambda + s\log(\lambda) - \sum_{i=1}^n \log(x_i!)\,,$$

donde $\log(x)$ es el logaritmo natural de x.

3. Demostrar que el estimador de máxima verosimilitud de λ es

$$\hat{\lambda}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}\,,$$

y por lo tanto la estimación correspondiente es $\bar{x} = 3.4359$.

4. Graficar la log-verosimilitud junto con las primeras dos derivadas (en gráficos aparte) para un rango adecuado de valores de λ. En el gráfico de la log-verosimilitud note que efectivamente en $\lambda = 3.4359$ hay un máximo.

   Nota: no olvide utilizar el criterio de la segunda derivada para demostrar que el punto crítico correspondiente es un máximo.

5. Demostrar que la información (observada) de Fisher es

$$\hat{I} = \frac{n}{\lambda}$$

   Reportar el valor específico de $\hat{I}$ usando $\lambda = \hat{\lambda}_{\text{MLE}}$.

6. De acuerdo con el método de máxima verosimilitud, el intervalo de confianza para λ está dado por

$$\text{IC}_{1-\alpha/2,\text{MLE}}(\lambda) = \hat{\lambda}_{\text{MLE}} \pm z_{1-\alpha/2}\sqrt{\hat{I}}$$

   donde $z_{1-\alpha/2}$ es el percentil $100(1-\alpha/2)$ de la distribución Normal estándar. Usando los datos del LOS, calcular e interpretar el intervalo de confianza correspondiente usando una confiabilidad del 95%.

   Nota: en este caso tanto el TLC como el método de máxima verosimilitud producen el mimo margen de error dado que el parámetro objeto de estudio coincide con el valor esperado del modelo propuesto (Poisson).

i. Finalmente, otra alternativa para hacer inferencia sobre el parámetro λ consiste en seguir a B. Efron (1938 –) usando técnicas de *remuestreo* como el **Bootstrap**.

   (a) Tomar $M = 100,000$ muestras aleatorias **con reemplazo** de tamaño $n = 117$ de la muestra, esto es, tomar una muestra aleatoria de la muestra original. Para cada una de las muestras simuladas calcular el promedio muestral $\bar{x}$ correspondiente.

   Nota 1: use `set.seed(1234)` como *semilla* en `R` para ejecutar la simulación.

   Nota 2: para tomar muestras aleatorias de la muestra original se utiliza la función `sample` de `R`.

   (b) En la misma figura, hacer un histograma de la distribución empírica de los promedios muestrales $\bar{x}^{(i)}$ resultado del proceso de remuestreo y sobre este histograma graficar una aproximación de la función de densidad.

(c) De acuerdo con la metodología Bootstrap, la estimación (puntual) de λ es

$$\hat{\lambda}_{\mathrm{B}} = \frac{1}{M}\sum_{i=1}^{M} \bar{x}^{(i)}$$

donde $\bar{x}^{(i)}$ es el valor la media muestral de la i-ésima muestra de la muestra original. Adicionalmente, el intervalo de confianza para λ de acuerdo con esta técnica está dado por

$$\mathrm{IC}_{1-\alpha/2,\mathrm{B}}(\lambda) = (\bar{x}_{\alpha/2}\,;\, \bar{x}_{1-\alpha/2})\,,$$

donde $\bar{x}_{\alpha/2}$ y $\bar{x}_{1-\alpha/2}$ son los percentiles $100(\alpha/2)$ y $100(1-\alpha/2)$ de la distribución empírica de los valores $\bar{x}^{(i)}$. Usando los datos del LOS, calcular e interpretar el intervalo de confianza correspondiente usando una confiabilidad del 95%.

j. Resumir los métodos anteriores por medio de la siguiente tabla:

Método	$\hat{\lambda}$	$s_{\hat{\lambda}}$	IC inf.	IC sup.
Teorema del Límite Central	3.436			
Máxima Verosimilitud	3.436			
Bootstrap				

donde $\hat{\lambda}$ es la estimación puntual de λ, $s_{\hat{\lambda}}$ es la desviación estándar de la estimación de $\hat{\lambda}$, y IC inf. y IC sup. son los límites inferior y superior del intervalo de confianza, respectivamente. Note que las filas correspondientes al Teorema del Límite Central y Máxima Verosimilitud coinciden!

k. La junta directiva del hospital planeaba hacer una intervención siempre y cuando el promedio (poblacional) del LOS fuera *significativamente* mayor que un día y medio (1.5). Con base en los resultados, ¿existe evidencia en la muestra para asegurar que el promedio poblacional del LOS es aproximadamente de un día y medio? ¿La junta directiva ha debido ejecutar el plan de mejoramiento? ¿Por qué?

l. Completar la siguiente tabla:

	$\widehat{\mathbb{P}}\mathrm{r}\left[X_i = x_i\right]$	
x_i	Empírica	Modelo
0	0.0342	0.0322
1	0.1282	0.1106
2		
3		
4		
5		
6		
7		
8		

Las probabilidades estimadas $\widehat{\mathbb{P}}\mathrm{r}\left[X_i = x_i\right]$ Empíricas son las frecuencias relativas (ya Usted las reportó en el numeral d.). Las probabilidades de la tabla usando el Modelo corresponden a las probabilidades estimadas usando el modelo de Poisson con el valor de $\lambda = \hat{\lambda}_{\mathrm{MLE}} = 3.4359$.

Con base en la tabla anterior y los valores de la media y la varianza muestral (recuerde que bajo la distribución Poisson se tiene que tanto el valor esperado como la varianza son iguales a λ), ¿el modelo de Poisson parece ser un buen modelo para esta población?

Nota: para calcular las probabilidades usando el modelo de Poisson se utiliza la función `dpois` de `R`.