Notas Curso de Estadística (Parte I)

Maikol Solís

Actualizado el 19 August, 2020

Índice general

Capítulo 1

Introducción

Capítulo 2

Inferencia estadística

Definición: Hacer afirmaciones probabilísticas respecto a (acerca de) cantidades desconocidas.

2.1. Ejemplo

***Pregunta**: ¿Será posible modelar cuánto dura un componente electrónico en fallar?

Solución: Podemos responder esta pregunta dividiéndola en dos partes:

1. **Modelo probabilístico:** Asuma que los tiempos de vida del componente son exponenciales (en años).
2. **Parámetro:** Sea $\theta > 0$ la tasa de fallo (unidades: 1/Tiempo(años)).

Es decir, tenemos un modelo (exponencial) y estamos decretando que su información estará concentrada en el parámetro θ.

Nota: El parámetro θ contiene la información del modelo, pero ¿Cómo obtenemos esa información

Muestra: Secuencia (sucesión) de variables aleatorias independientes $X_1, X_2, \ldots, X_n, \ldots$. Tomemos una muestra $X_1, X_2, \ldots, X_n, \ldots \overset{i.i.d}{\sim} \text{Exp}(\theta)$.

Objetivos

- Estimar $X_m, X_{m+1}, \ldots$ si se observa $X_1, X_{m-1}, \ldots$ (Predicción).

- Estimar θ usando información.

Datos: Realizaciones de variables aleatorias $X_1, \ldots, X_m$ pertenecientes a la muestra.

Estimación de θ

Dado que $\mathbb{E}(X) = \dfrac{1}{\theta}$ con $X \sim \text{Exp}(\theta)$, por la ley de grandes números se tiene que

$$\underbrace{\frac{1}{n}\sum_{i=1} X_i}_{\bar{X}_n} \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}(X) = \frac{1}{\theta}$$

por propiedad de convergencia en probabilidad.

Un posible candidato para estimar θ es $\dfrac{1}{\bar{X}_n}$, bajo el supuesto por Ley de Grandes Números que θ es una constante (frecuentista).

Realidad: θ no necesariamente es determinístico (factores externos, por la naturaleza del fenómeno).

Asumimos un modelo probabilístico para θ (tasa siempre positiva):

$$\theta \sim \Gamma(\alpha_0, \beta_0)$$

Luego, según estudios previos la tasa esperada es 0.5/año

$$\mathbb{E}(\theta) = \frac{1}{2} = \frac{\alpha_0}{\beta_0}.$$

Un primer indicio de que se podría establecer que $\alpha_0 = 1$ y de $\beta_0 = 2$.

2.2. Modelo estadístico

Vamos a definir como típicamente se define un modelo estadístico.

1. Variables aleatorias observables / hipotéticamente observables:

$$\underbrace{X_t}_{\text{Observable}} = \underbrace{Y_t}_{\text{Hip. observable}} + \underbrace{\epsilon}_{\text{Ruido}}$$

En otras palabras Y_t sería la el dato *"verdadero"* que pasó exactamente en el fenómeno analizado. Esta observación es afectada por muchos factores no observables (por ejemplo: errores de medición, cambio de las condiciones de la economía, etc.). La variable ϵ captura toda esa aleatoriedad que no es parte del fénomeno.

Claramente ni Y_t ni ϵ se pueden medir y la mejor representación del nuestro es fenómeno es a partir de X_t.

2. Distribución conjunta de una muestra de variables observables.

   Es decir cuál es el supuesto general que estoy usando para describir mis observaciones.

3. Parámetros que son hipotéticamente observables (desconocidos).

   ¿Cuál sería la mejor calibración de los componentes del modelo anterior de modo que mi modelo se ajuste a los datos?

4. (Opcional) Distribución conjunta de los parámetros.

   En el caso de Bayes, los parámetro dejan de ser simple valores puntuales y se convierten en distribuciones completas.

- **Inferencia estadística**: procedimiento que genera afirmaciones probabilísticas de un modelo estadístico.

Ejemplo de inferencias:

1. Estimar θ a través de $\dfrac{1}{\bar{X}_n}$.

2. ¿Qué tan probable es que el promedio de las siguientes observaciones es al menos 2?

$$\frac{1}{10}\sum_{i=m+1}^{m+10} X_i > 2$$

3. ¿Qué tan cierto es que $\theta \leq 0{,}4$ después de observar la muestra?

- **Parámetro**: característica (s) que determinan la distribución conjunta de las variables aleatorias de interés.
- **Espacio paramétrico** Ω (espacio de parámetros, puede ser de probabilidad)

Ejemplos:

- $\theta > 0$ (ejemplo anterior); $\Omega = (0, +\infty)$.
- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, (μ, σ^2) parámetros; $\Omega = \mathbb{R} \times [0, +\infty)$.

Ejemplo: Clientes de un banco

¿Qué tan probable es que un cliente no pague su crédito hoy?

- **Datos**: $X_i = \begin{cases} 1 & \text{el cliente } \#i \text{ no pagó} \\ 0 & \text{el cliente } \#i \text{ pagó} \end{cases}.$
- **Muestra**: $X_1, \ldots, X_{10000}$ (realización al día de hoy).
- **Modelos**: $X_1, \ldots, X_{10000} \overset{i.i.d}{\sim} \text{Ber}(p)$ con $p \in [0, 1]$.
- **Parámetro**: p, $\Omega = [0, 1]$.
- **Inferencias:**
  - Estimar p (probabilidad de impago).
  - Suponga que $L(X_i)$ es el saldo en la cuenta del cliente $\#i$.

$$\mathbb{P}\left(\sum_{i=1}^{10000} L(X_i) > u\right) = \text{Probabilidad de ruina}$$

2.3. Estadístico

Definición. Si $X_1, \ldots, X_n$ es una muestra observable. Sea r una función real de n variables:

$$T = r(X_1, \ldots, X_n)$$

es un estadístico.

Nota: T también es aleatorio.

Ejemplos:

- $\hat{p} = \dfrac{1}{10000} \sum_{i=1}^{10000} X_i = \dfrac{\#\text{ no pagan}}{\text{Total}} = r(X_1, \ldots, X_{10000})$
- $L_m = \text{máx}\, L(X_i)$ (saldo del cliente más riesgoso).
- $R_m = \text{máx}\, L(X_i) - \text{mín}\, L(X_i), 1 \leq i \leq 10000$

Capítulo 3

Densidades previas conjugadas y estimadores de Bayes

3.1. Distribución previa (distribución a priori)

Suponga que tenemos un modelo estadístico con parámetro θ. Su θ es aleatorio entonces su densidad (antes de observar cualquier muestra) se llama **densidad previa**: π.

Ejemplo: $X_1, \ldots, X_n \sim \text{Exp}(\theta)$ y θ es aleatorio tal que $\theta \sim \Gamma(\overset{\alpha}{1}, \overset{\beta}{2})$ entonces

$$\pi(\theta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha \theta^{\alpha-1} e^{\beta\theta} = 2e^{-2\theta}, \quad \theta > 0$$

Ejemplo: Sea θ la probabilidad de obtener cara al tirar una moneda.

En este caso antes de modelar exactamente el θ, lo importante es modelar el tipo de moneda. Es decir, supongamos que tenemos dos opciones

- *Moneda justa:* $\theta = \dfrac{1}{2}$ con probabilidad previa 0,8 ($\pi(\frac{1}{2}) = 0{,}8$).
- *Moneda con solo una cara:* $\theta = 1$ con probabilidad previa 0,2 ($\pi(1) = 0{,}2$).

En este ejemplo si tuviéramos 100 monedas con probabilidad previa π entonces 20 tendrían solo una cara y 80 serían monedas normales.

Notas:

- π está definida en Ω (espacio paramétrico).
- π es definida antes de obtener la muestra.

Ejemplo (Componentes eléctricos) Supoga que se quiere conocer el tiempo de vida de cierto componente eléctrico. Sabemos que este tiempo se puede modelar con una distribución exponencial con parámetro θ desconocido. Este parámetro asumimos que tiene una distribución previa Gamma.

Un experto en componentes eléctricos conoce mucho de su área y sabe que el parámetro θ tiene las siguientes características:

$$\mathbb{E}[\theta] = 0{,}0002, \quad \sqrt{\operatorname{Var}(\theta)} = 0{,}0001.$$

Como sabemos que la previa π es Gamma, podemos deducir lo siguiente:

$$\mathbb{E}[\theta] = \frac{\alpha}{\beta}, \operatorname{Var}(\theta) = \frac{\alpha}{\beta^2}$$

$$\implies \begin{cases} \dfrac{\alpha}{\beta} = 2 \times 10^{-4} \\ \sqrt{\dfrac{\alpha}{\beta^2}} = 1 \times 10^{-4} \end{cases} \implies \beta = 20000, \alpha = 4$$

Notación:

- $X = (X_1, \ldots, X_n)$: vector que contiene la muestra aleatoria.
- Densidad conjunta de X: $f_\theta(x)$.
- Densidad de X condicional en θ: $f_n(x|\theta)$.

Supuesto: X viene de una muestra aleatoria si y solo si X es condicionalmente independiente dado θ.

Consecuencia:

$$f_n(X|\theta) = f(X_1|\theta) \cdot f(X_2|\theta) \cdots f(X_n|\theta)$$

Ejemplo

Si $X = (X_1, \ldots, X_n)$ es una muestra tal que $X_i \sim \text{Exp}(\theta)$,

$$
\begin{aligned}
f_n(X|\theta) &= \begin{cases} \prod_{i=1}^{n} \theta e^{-\theta X_i} & \text{si } X_i > 0 \\ 0 & \text{si no} \end{cases} \\
&= \begin{cases} \theta^n e^{-\theta \sum_{i=1}^{n} X_i} & X_i > 0 \\ 0 & \text{si no} \end{cases}
\end{aligned}
$$

3.2. Densidad posterior

Definición. Considere un modelo estadístico con parámetro θ y muestra aleatoria $X_1, \ldots, X_n$. La densidad condicional de θ dado $X_1, \ldots, X_n$ se llama *densidad posterior*: $\pi(\theta|X)$

Teorema. Bajo las condiciones anteriores:

$$\pi(\theta|X) = \frac{f(X_1|\theta) \cdots f(X_n|\theta)\pi(\theta)}{g_n(X)}$$

para $\theta \in \Omega$, donde g_n es una constante de normalización.

Prueba:

$$
\begin{aligned}
\pi(\theta|X) = \frac{\pi(\theta, X)}{\text{marginal de X}} = \frac{\pi(\theta, X)}{\int \pi(\theta, X)\, d\theta} = \frac{P(X|\theta) \cdot \pi(\theta)}{\int \pi(\theta, X)\, d\theta} \\
\frac{f_n(X|\theta) \cdot \pi(\theta)}{g_n(X)} = \frac{f(X_1|\theta) \cdots f(X_n|\theta)\pi(\theta)}{g_n(X)}
\end{aligned}
$$

Del ejemplo anterior,

$$f_n(X|\theta) = \theta^n e^{-\theta y}, y = \sum X_i \text{ (estadístico)}$$

Numerador:

$$f_n(X|\theta)\pi(\theta) = \underbrace{\theta^n e^{-\theta y}}_{f_n(X|\theta)} \cdot \underbrace{\frac{200000^4}{3!} \theta^3 e^{-20000 \cdot \theta}}_{\pi(\theta)} = \frac{20000^4}{3!} \theta^{n+3} e^{(20000+y)\theta}$$

Denominador:

$$g_n(x) = \int_0^{+\infty} \theta^{n+3} e^{-(20000+y)\theta}\, d\theta = \frac{\Gamma(n+4)}{(20000+y)^{n+4}}$$

Entonces la posterior corresponde a

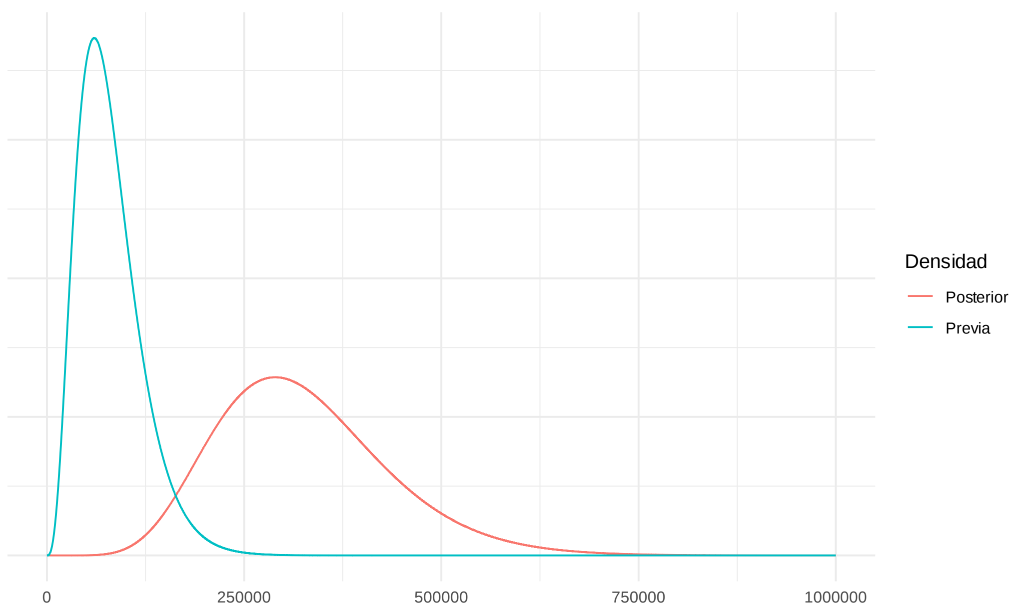
$$\pi(\theta|X) = \frac{\theta^{n+3} e^{-(20000+y)\theta}}{\Gamma(n+4)} (20000+y)^{n+4}$$

que es una $\Gamma(n+4, 20000+y)$.

Con 5 observaciones (horas): 2911, 3403, 3237, 3509, 3118.

$$y = \sum_{i=1}^{5} X_i = 16478, \quad n = 5$$

por lo que $\theta|X \sim \Gamma(9, 36178)$



Es sensible al tamaño de la muestra (una muestra grande implica un efecto de la previa menor).

Hiperparámetros: parámetros de la previa o posterior.

3.3. Proceso de modelación de parámetros.

De ahora en adelante vamos a entender un modelo como el conjunto de los datos $X_1, \ldots, X_n$, la función de densidad f y el parámetro de la densidad θ. Estos dos últimos resumen el comportamiento de los datos.

Ahora para identificar este modelo se hace por partes,

1. La información previa $\pi(\theta)$ es la información extra o basado en la experiencia que tengo del mdoelo.
2. Los datos es la información observada. La función de densidad f filtra y mejora la información de la previa.
3. La densidad posterior es la "mezcla" entre la información y los datos observados. Es una versión más informada de la distribución del parámetro.

3.4. Función de verosimilitud

Bajo el modelo estadístico anterior a $f_n(X|\theta)$ se le llama **verosimilitud** o **función de verosimilitud**.

Observación. En el caso de una función de verosimilitud, el argumento es θ.

Ejemplo.

Sea θ la proporción de aparatos defectuosos, con $\theta \in [0, 1]$

$$X_i = \begin{cases} 0 & \text{falló} \\ 1 & \text{no falló} \end{cases}$$

$\{X_i\}_{i=1}^n$ es una muestra aleatoria y $X_i \sim Ber(\theta)$.

- **Verosimilitud**

$$f_n(X|\theta) = \prod_{i=1}^n f(X_i|\theta) = \begin{cases} \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} & X_i = 0, 1 \ \forall i \\ 0 & \text{si no} \end{cases}$$

- **Previa**:

$$\pi(\theta) = 1_{\{0 \leq \theta \leq 1\}}$$

- **Posterior**:

Por el teorema de Bayes,

$\pi(\theta|X) \propto \theta^y(1-\theta)^{n-y} \cdot 1$

$$= \theta^{\overbrace{y+1}^{\alpha}-1}(1-\theta)^{\overbrace{n-y+1}^{\beta}-1} \implies \theta|X \sim \text{Beta}(y+1, n-y+1)$$

- **Predicción**.

Supuesto: los datos son secuenciales. Calculamos la distribución posterior secuencialmente:

$$\begin{aligned}
\pi(\theta|X_1) &\propto \pi(\theta)f(X_1|\theta)\\
\pi(\theta|X_1, X_2) &\propto \pi(\theta)f(X_1, X_2|\theta)\\
&= \pi(\theta)f(X_1|\theta)f(X_2|\theta) \text{ (por independencia condicional)}\\
&= \pi(\theta|X_1)f(X_2|\theta)\\
&\vdots\\
\pi(\theta|X_1, \ldots, X_n) &\propto f(X_n|\theta)\pi(\theta|X_1, \ldots, X_{n-1})
\end{aligned}$$

Bajo independencia condicional no hay diferencia en la posterior si los datos son secuenciales.

Luego,

$$\begin{aligned}
g_n(X) &= \int_{\Omega} f(X_n|\theta)\pi(\theta|X_1, \ldots, X_{n-1})\, d\theta\\
&= P(X_n|X_1, \ldots, X_{n-1}) \text{ (Predicción para } X_n)
\end{aligned}$$

Continuando con el ejemplo de los artefactos, $P(X_6 > 3000|X_1, X_2, X_3, X_4, X_5)$. Se necesita calcular $f(X_6|X)$. Dado que

$$\pi(\theta|X) = 2{,}6 \times 10^{36}\theta^8 e^{-361778\theta}$$

se tiene

$$f(X_6|X) = 2{,}6 \times 10^{36} \int_0^1 \underbrace{\theta e^{-\theta X_6}}_{\text{Densidad de } X_6} \theta^8 e^{-36178\theta}\, d\theta = \frac{9{,}55 \times 10^{41}}{(X_6 + 36178)^{10}}$$

Entonces,

$$P(X_6 > 3000) = \int_{3000}^{\infty} \frac{9{,}55 \times 10^{41}}{(X_6 + 36178)^{10}}\, dX_6 = 0{,}4882$$

La vida media se calcula como $\frac{1}{2} = P(X_6 > u|X)$.

3.5. Familias conjugadas

Definición. Sea $X_1, \ldots, X_n$ i.i.d. condicional dado θ con densidad $f(X|\theta)$. Sea ψ la familia de posibles densidades previas sobre Ω. Si, sin importar los datos, la posterior pertenece a ψ, entonces decimos que ψ es una familia conjugada de previas.

Ejemplos:

- La familia Beta es familia conjugada para muestras según una Bernoulli.
- La familia Gama es familia conjugada para muestras exponenciales.
- Para el caso Poisson, si $X_1, \ldots, X_n \sim Poi(\lambda)$,entonces la familia Gamma es familia conjugada.

La función de densidad de una Poisson es $P(X_i = k) = e^{-\lambda}\frac{\lambda^k}{k!}$. La verosimilitud corresponde a

$$f_n(X|\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda_i^X}{X_i!} = \frac{e^{-n\lambda\lambda^y}}{\prod_{i=1}^n X_i}.$$

La previa de λ está definida por $\pi(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$. Por lo tanto, la posterior es

$$\pi(\lambda|X) \propto \lambda^{y+\alpha-1}e^{-(\beta+n)\lambda} \implies \lambda|X \sim \Gamma(y+\alpha, \beta+n)$$

- En el caso normal, si $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$,entonces la familia normal es conjugada si σ^2 es conocido.

Si $\theta \sim N(\mu_0, V_0^2) \implies \theta|X \sim N(\mu_1, V_1^2)$ donde,

$$\mu_1 = \frac{\sigma^2\mu_0 + nV_0^2\bar{X}_n}{\sigma^2 + nV_0^2} = \frac{\sigma^2}{\sigma^2 + nV_0^2}\mu_0 + \frac{nV_0^2}{\sigma^2 + nV_0^2}\bar{X}_n$$

Combina de manera ponderada la previa y la de los datos.

Ejemplo

Considere una verosimilitud Poisson(λ) y una previa

$$\pi(\lambda) = \begin{cases} 2e^{-2\lambda} & \lambda > 0 \\ 0 & \lambda \geq 0 \end{cases} \quad \lambda \sim \Gamma(1, 2)$$

Supongamos que es una muestra aleatoria de tamaño n. ¿Cuál es el número de observciones para reducir la varianza, a lo sumo, a 0.01?

Por teorema de Bayes, la posterior $\lambda|x \sim \Gamma(y + 1, n + 2)$. Luego, la varianza de la Gamma es

$$\frac{\alpha}{\beta^2} = \frac{\sum x_i + 1}{(n+2)^2} \leq 0{,}01 \implies \frac{1}{(n+2)^2} \leq \frac{\sum x_i + 1}{(n+2)^2} \leq 0{,}01 \implies 100 \leq (n+2)^2 \implies n \geq 8$$

Teorema. Si $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ con σ^2 conocido y la previa es $\theta \sim N(\mu_0, V_0^2)$, entonces $\theta|X \sim N(\mu_1, V_1^2)$ donde

$$\mu_1 = \frac{\sigma^2\mu_0 + nV_0^2\bar{X}_n}{\sigma^2 + nV_0^2}, \quad V_1^2 = \frac{\sigma^2 V_0^2}{\sigma^2 + nV_0^2}$$

Prueba:

- **Verosimilitud**:

$$f_n(X|\theta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i\theta)^2\right]$$

Luego,

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\theta)^2 &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\theta)^2\\ &= n(\bar{X}+\theta)^2+\sum_{i=1}^{n}(X_i-\bar{X})^2+\underbrace{2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\theta)}_{=0 \text{ pues } \sum Xi=n\bar{X})}\end{aligned}$$

Entonces

$$f_n(X|\theta)\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{X}-\theta)^2\right].$$

- **Previa**:

$$\pi(\theta)\propto \exp\left[-\frac{1}{2V_0^2}(\theta-\mu_0)^2\right].$$

- **Posterior**:

$$\pi(\theta|X)\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{X}-\theta)^2-\frac{1}{2V_0^2}(\theta-\mu_0)^2\right].$$

Con μ_1 y V_1^2 definidos anteriormente, se puede comprobar la siguiente identidad:

$$-\frac{n}{\sigma^2}(\bar{X}-\theta)^2-\frac{1}{V_0^2}(\theta-\mu_0)^2=\frac{1}{V_1^2}(\theta-\mu_1)^2+\underbrace{\frac{n}{\sigma^2+nV_0^2}(\bar{X}_n-\mu_0)^2}_{\text{Constante con respecto a }\theta}$$

Por lo tanto,

$$\pi(\theta|X)\propto \exp\left[-\frac{n}{2V_1^2}(\theta-\mu_1)^2\right]$$

Media posterior:

$$\mu_1=\underbrace{\frac{\sigma^2}{\sigma^2+nV_0^2}}_{W_1}\mu_0+\underbrace{\frac{nV_0^2}{\sigma^2+nV_0^2}}_{W_2}\bar{X}_n$$

Afirmaciones:

1) Si V_0^2 y σ^2 son fijos, entonces $W_1 \xrightarrow[n\to\infty]{} 0$ (la importancia de la media empírica crece conforme aumenta n).
2) Si V_0^2 y n son fijos, entonces $W_2 \xrightarrow[\sigma^2\to\infty]{} 0$ (la importancia de la media empírica decrece conforme la muestra es menos precisa).
3) Si σ^2 y n son fijos, entonces $W_2 \xrightarrow[V_0^2\to\infty]{} 1$ (la importancia de la media empírica crece conforma la previa es menos precisa).

Ejemplo (determinación de n)

Sean $X_1, \dots, X_n \sim N(\theta, 1)$ y $\theta \sim N(\mu_0, 4)$. Sabemos que

$$V_1^2 = \frac{\sigma^2 V_0^2}{\sigma^2 + nV_0^2}.$$

Buscamos que $V_1 \leq 0{,}01$, entonces

$$\frac{4}{4n+1} \leq 0{,}01 \implies n \geq 99{,}75 \text{ (al menos 100 observaciones)}$$

3.6. Densidades previas impropias

Definición. Sea π una función positiva cuyo dominio está en Ω. Suponga que $\int \pi(\theta)\, d\theta = \infty$. Entonces decimos que π es una **densidad impropia**.

Ejemplo: $\theta \sim \text{Unif}(\mathbb{R})$, $\lambda \sim \text{Unif}(0, \infty)$.

Una técnica para seleccionar distribuciones impropia es sustituir los hiperparámetros previos por 0.

Ejemplo:

Se presenta el número de soldados prusianos muertos por una patada de caballo (280 conteros, unidades de combate en 20 años).

Unidades	Ocurrencias
144	0
91	1
32	2

Unidades	Ocurrencias
11	3
2	4

- Muestra de Poisson: $X_1 = 0, X_2 = 1, X_3 = 1, \ldots, X_{280} = 0 \sim \text{Poi}(\lambda)$.
- Previa: $\lambda \sim \Gamma(\alpha, \beta)$.
- Posterior: $\lambda|X \sim \Gamma(y + \alpha, n + \beta) = \Gamma(196 + \alpha, 280 + \beta)$.

Sustituyendo, $\alpha = \beta = 0$

$$\pi(\lambda) = \frac{1}{\Gamma(\alpha)} \beta^\alpha \lambda^{\alpha-1} e^{\beta\lambda}$$

$$\propto \lambda^{\alpha-1} e^{-\lambda\beta}$$

$$= \frac{1}{\lambda}$$

donde $\displaystyle\int_0^\infty \frac{1}{\lambda} d\lambda = \infty$.

Por teorema de Bayes,

$$\theta|X \sim \Gamma(196, 280)$$

3.7. Funciones de pérdida

Definición. Sean $X_1, \ldots, X_n$ datos observables cuyo modelo está indexado por $\theta \in \Omega$. Un estimador de θ es cualquier estadístico $\delta(X_1, \ldots, X_n)$.

Notación:

- Estimador $\rightarrow \delta(X_1, \ldots, X_n)$.
- Estimación o estimado: $\delta(X_1, \ldots, X_n)(\omega) = \delta(\overbrace{x_1, \ldots, x_n}^{datos})$

Definición. Una **función de pérdida** es una función de dos variables:

$$L(\theta, a), \quad \theta \in \Omega$$

con a un número real.

Interpretación: es lo que pierde un analista cuando el parámetro es θ y el estimador es a.

Asuma que θ tiene una previa. La pérdida esperada es

$$\mathbb{E}[L(\theta, a)] = \int_\Omega L(\theta, a)\pi(\theta)\, d\theta$$

la cual es una función de a, que a su vez es función de $X_1, \ldots, X_n$. Asuma que a se selecciona el minimizar esta esperanza. A ese estimador $a = \delta^*(X_1, \ldots, X_n)$ se le llama **estimador bayesiano**, si ponderamos los parámetros con respecto a la posterior.

$$\mathbb{E}[L(\theta, \delta^*)|X] = \int_\Omega L(\theta, a)\pi(\theta)\, d\theta = \min_a \mathbb{E}[L(\theta|a)X].$$

3.7.1. Función de pérdida cuadrática

$$L(\theta, a) = (\theta - a)^2$$

En el caso en que θ es real y $\mathbb{E}[\theta|X]$ es finita, entonces

$$\delta^*(X_1, \ldots, X_n) = \mathbb{E}[\theta|X] \text{ cuando } L(\theta, a) = (\theta - a)^2.$$

Ejemplo: $X_1, \ldots, X_n \sim \text{Ber}(\theta)$, $\theta \sim \text{Beta}(\alpha, \beta) \implies \theta|X \sim \text{Beta}(\alpha + y, \beta + n - y)$.

El estimador de θ es

$$\delta^*(X_1, \ldots, X_n) = \frac{\alpha + y}{\alpha + \beta + n} = \overbrace{\frac{\alpha}{\alpha + \beta}}^{\text{Esperanza previa}} \cdot \frac{\alpha + \beta}{\alpha + \beta + n} + \overbrace{\frac{y}{n}}^{\bar{X}} \cdot \frac{n}{\alpha + \beta + n}.$$

3.7.2. Función de pérdida absoluta

$$L(\theta, a) = |\theta - a|$$

La pérdida esperada es

$$f(a) = \mathbb{E}[L(\theta, a)|X] = \int_{-\infty}^{+\infty} |\theta - a|\pi(\theta|X)\, d\theta = \int_a^{+\infty} (\theta - a)\pi(\theta|X)\, d\theta + \int_{-\infty}^a (a - \theta)\pi(\theta|X)\, d\theta$$

Usando el teorema fundamental del cálculo,

$$F_\pi(a|X) = \int_{-\infty}^{\hat{a}} \pi(\theta|X)\, d\theta = \frac{1}{2} \Leftrightarrow \hat{a} = \underset{a}{\operatorname{argmin}}\, f(a)$$

La **mediana** es el punto de $X_{0,5}$ tal que $F(X_{0,5}) = \dfrac{1}{2}$.

Corolario. Bajo la función de pérdida absoluta, el estimador bayesiano es la mediana posterior.

Ejemplo: Bernoulli.

$$\frac{1}{\text{Beta}(\alpha+y, \beta+n-y)} \int_{-\infty}^{X_{0,5}} \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}\, d\theta = \frac{1}{2}$$

Resuelva para $X_{0,5}$.

3.7.3. Otras funciones de pérdida

- $L(\theta, a) = |\theta - a|^k$, $k \neq 1, 2$, $0 < k < 1$.
- $L(\theta, a) = \lambda(\theta)|\theta - a|^2$ ($\lambda(\theta)$ penaliza la magnitud del parámetro).
- $L(\theta, a) = \begin{cases} 3(\theta-a)^2 & \theta \leq a \text{ (sobreestima)} \\ (\theta-a)^2 & \theta \geq a \text{ (subestima)} \end{cases}$

3.8. Efecto de muestras grandes

Ejemplo: ítemes malos (proporción: θ), $\theta \in [0, 1]$. Función de pérdida cuadrática. El tamaño de muestra son $n = 100$ ítemes, de los cuales $y = 10$ están malos.

$$X_1, \ldots, X_n \sim \text{Ber}(\theta)$$

- Primer previa. $\alpha = \beta = 1$ (Beta). El estimador bayesiano corresponde a

$$\mathbb{E}[\theta|X] = \frac{\alpha + y}{\alpha + \beta + n} = \frac{1 + 10}{2 + 100} = 0{,}108$$

- Segunda previa. $\alpha = 1, \beta = 2 \implies \pi(\theta) = 2e^{-2\theta}, \theta > 0$.

$$\mathbb{E}[\theta|X] = \frac{1+10}{1+2+100} = \frac{11}{103} = 0{,}107$$

La media es $\bar{X}_n = \dfrac{10}{100} = 0{,}1$.

3.9. Consistencia

Definición. Un estimador de θ $\delta(X_1, \ldots, X_n)$ es consistente si

$$\delta(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{\mathbb{P}} \theta.$$

Bajo pérdida cuadrática, $\mathbb{E}[\theta|X] = W_1\mathbb{E}[\theta] + X_2\bar{X}_n = \delta^*$. Sabemos, por ley de grandes números, que $\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta$. Además, $W_1 \xrightarrow[n\to\infty]{} 0$ y $W_2 \xrightarrow[n\to\infty]{} 1$.

En los ejemplos que hemos analizado

$$\delta^* \xrightarrow[n\to\infty]{\mathbb{P}} \theta$$

Teorema. Bajo condiciones generales, los estimadores bayesianos son consistentes.

Estimador. Si $X_1, \ldots, X_n$ es una muestra en un modelo indexado por θ, $\theta \in \Omega$ (k-dimensiones), sea

$$h : \Omega \to H \subset \mathbb{R}^d.$$

Sea $\psi = h(\theta)$. Un **estimador** de ψ es un estadístico $\delta^*(X_1, \ldots, X_n) \in H$. A $\delta^*(X_1, \ldots, X_n)$ estimador de ψ se puede evaluar y construir estimadores nuevos.

Ejemplo. $X_1, \ldots, X_n \sim \text{Exp}(\theta), \theta|X \sim \Gamma(\alpha, \beta) = \Gamma(4, 8{,}6)$. La característica de interés es $\psi = \dfrac{1}{\theta}$, el valor esperado del tiempo de fallo.

Es estimador se calcula de la siguiente manera:

$$
\begin{aligned}
\delta^*(x) = \mathbb{E}[\psi|x] &= \int_0^\infty \frac{1}{\theta}\pi(\theta|x)\; d\theta \\
&= \int_0^\infty \frac{1}{\theta}\frac{8{,}6^4}{\Gamma(4)}\theta^3 e^{-8{,}6\theta}\; d\theta \\
&= \frac{8{,}6^4}{6}\underbrace{\int_0^\infty \theta^2 e^{-8{,}6\theta}\; d\theta}_{\frac{\Gamma(3)}{8{,}6^3}} \\
&= \frac{8{,}6^4}{6}\frac{2}{8{,}6^3} = 2{,}867 \text{ unidades de tiempo.}
\end{aligned}
$$

Por otro lado, vea que $\mathbb{E}(\theta|X) = \dfrac{4}{8{,}6}$. El estimador *plug-in* correspondería a

$$
\frac{1}{\mathbb{E}(\theta|X)} = \frac{8{,}6}{4} = 2{,}15.
$$

3.10. Laboratorio

Lo primero es cargar los paquetes necesarios que usaremos en todo el curso

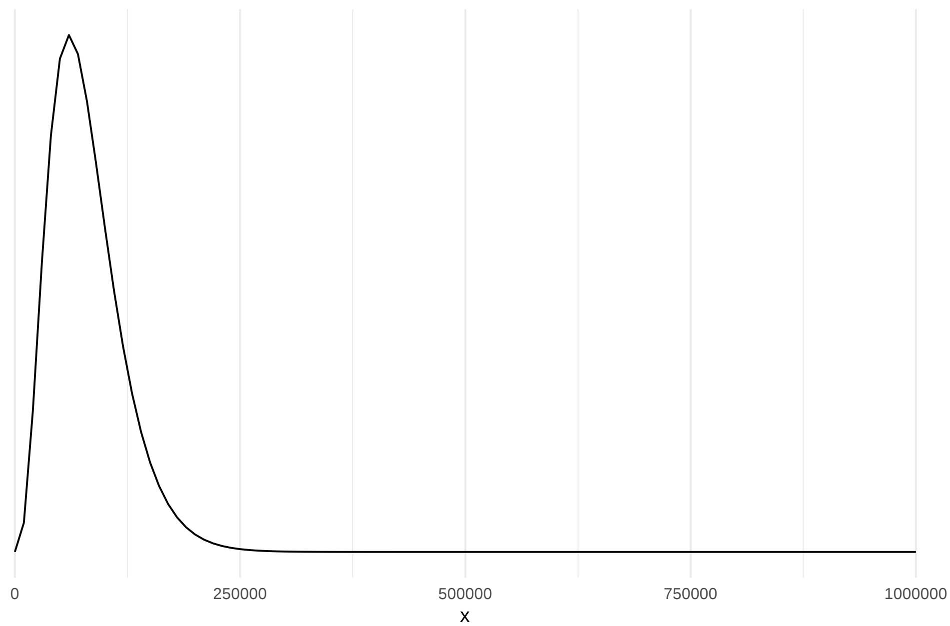
```
library(tidyverse)
```

3.10.1. Distribución previa

En nuestro ejemplo se tenía que $\mathbb{E}[\theta] = 0{,}0002$ y $\mathrm{Var}(\theta) = 0{,}001$. Suponiendo que θ es gamma se puede resolver el sistema de ecuaciones obtenemos que $\beta = 20000$ y $\alpha = 4$.

```
alpha_previa <- 4
beta_previa <- 20000

ggplot(data = data.frame(x = c(0, 1e+06)), aes(x)) +
    stat_function(fun = dgamma, args = list(shape = alpha_previa,
        scale = beta_previa)) + ylab("") + scale_y_continuous(breaks = NULL) +
    theme_minimal()
```



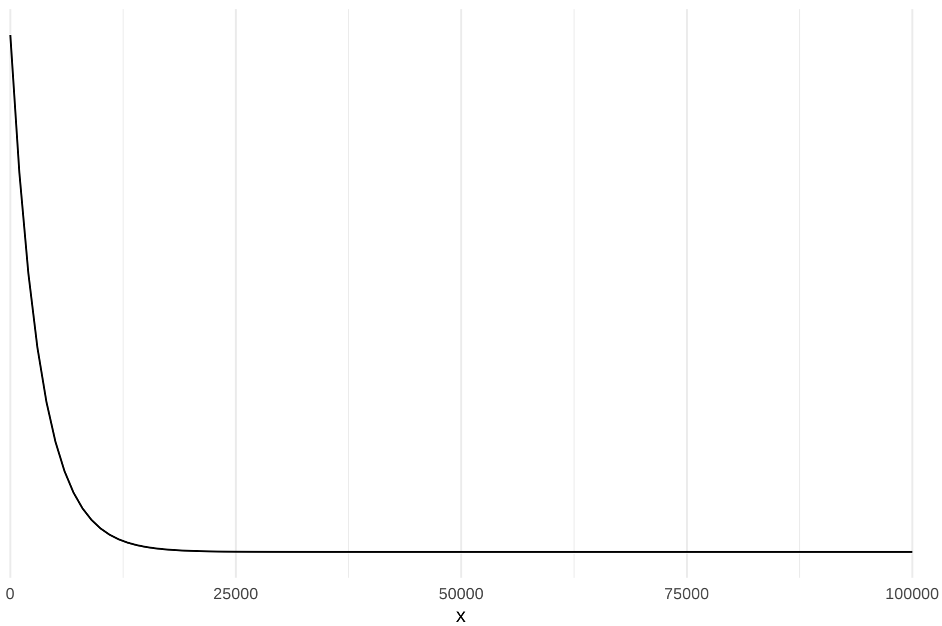
3.10.2. Distribución conjunta

Asumiendo que tenemos algunos datos $X_1, ..., X_n$, asumimos que estos son exponencial recordando que $\mathbb{E}[X] = 1/\theta$, entonces una aproximación de esta densidad es

```
x <- c(2911, 3403, 3237, 3509, 3118)

theta <- 1/mean(x)

ggplot(data = data.frame(x = c(0, 1e+05)), aes(x)) +
    stat_function(fun = dexp, args = list(rate = theta)) +
    ylab("") + scale_y_continuous(breaks = NULL) +
    theme_minimal()
```



3.10.3. Distribución posterior

Según los contenidos del curso, se puede estimar los parámetros de la densidad posterior de la forma

```
(y <- sum(x))
```

```
## [1] 16178
```

```
(n <- length(x))
```

```
## [1] 5
```

```
(alpha_posterior <- n + alpha_previa)
```

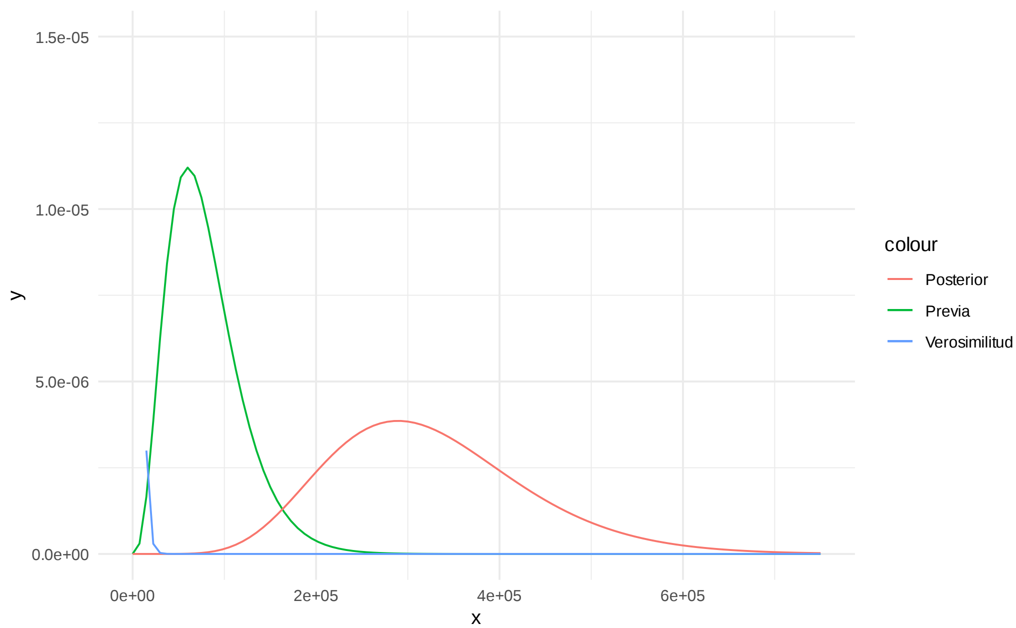
```
## [1] 9
```

```
(beta_posterior <- beta_previa + y)
```

```
## [1] 36178
```

```
ggplot(data = data.frame(x = c(0, 750000)), aes(x)) +
    stat_function(fun = dgamma, args = list(shape = alpha_previa,
```

```
        scale = beta_previa), aes(color = "Previa")) +
    stat_function(fun = dgamma, args = list(shape = alpha_posterior,
        scale = beta_posterior), aes(color = "Posterior")) +
    stat_function(fun = dexp, args = list(rate = theta),
        aes(color = "Verosimilitud")) + ylim(0, 1.5e-05) +
    theme_minimal()
```



3.10.4. Agregando nuevos datos

Si tenemos un 6to dato, y queremos ver cual es su distribución posterior. Lo primero es estimar la densidad posterior de este 6to dato, pero asumiendo que la previa es la densidad que obtuvimos en el caso anterior.

Suponga que $X_6 = 3000$

```
(alpha_previa <- alpha_posterior)
```

```
## [1] 9
```

```
(beta_previa <- beta_posterior)
```

```
## [1] 36178
```

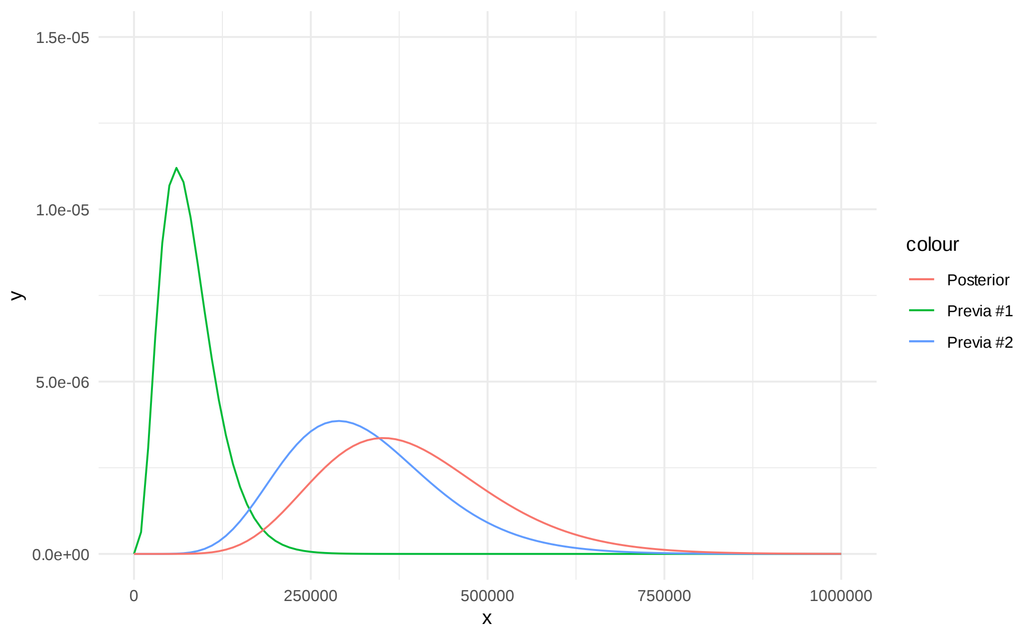
```
(alpha_posterior <- alpha_previa + 1)
```

```
## [1] 10
```

```
(beta_posterior <- beta_previa + 3000)
```

```
## [1] 39178
```

```
ggplot(data = data.frame(x = c(0, 1e+06)), aes(x)) +
    stat_function(fun = dgamma, args = list(shape = 4,
        scale = 20000), aes(color = "Previa #1")) +
    stat_function(fun = dgamma, args = list(shape = alpha_previa,
        scale = beta_previa), aes(color = "Previa #2")) +
    stat_function(fun = dgamma, args = list(shape = alpha_posterior,
        scale = beta_posterior), aes(color = "Posterior")) +
    ylim(0, 1.5e-05) + theme_minimal()
```



3.10.5. Familias conjugadas normales

Si tenemos pocos datos, la información previa es la que “prevalece”.

```
x <- rnorm(n = 3, mean = 10, sd = 1)

(mu <- mean(x))
```

```
## [1] 10.22127
```

```
(sigma <- sd(x))
```

```
## [1] 1.185713
```

```
(n <- length(x))
```

```
## [1] 3
```

```
(mu_previa <- 0)
```

```
## [1] 0
```

```
(sigma_previa <- 1)
```

```
## [1] 1
```

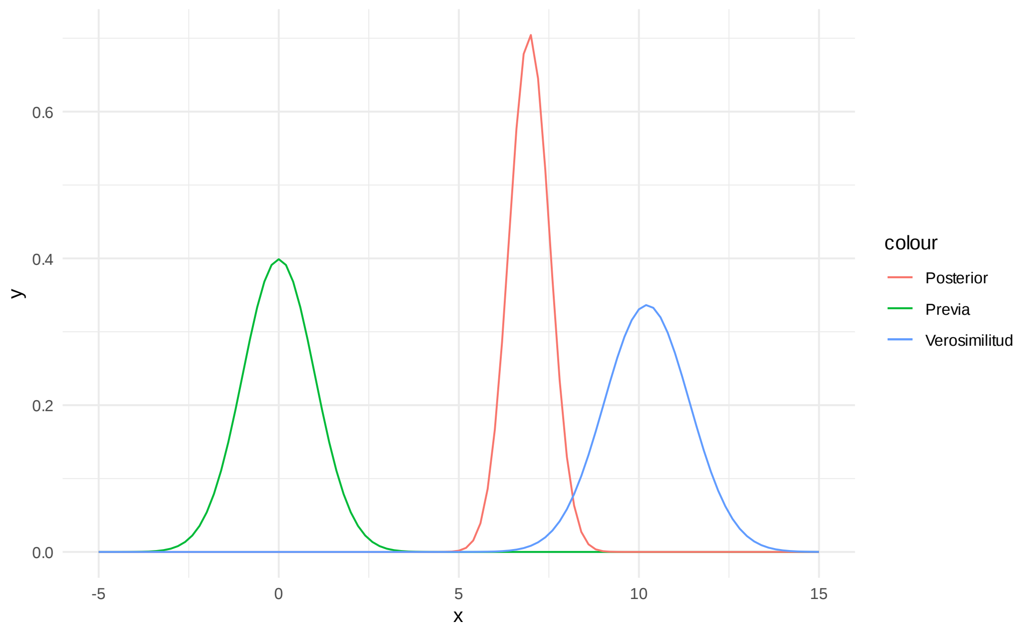
```
(mu_posterior <- ((sigma^2)/(sigma^2 + n * sigma_previa^2)) *
    mu_previa + ((n * sigma_previa^2)/(sigma^2 + n *
    sigma_previa^2)) * mu)
```

```
## [1] 6.959693
```

```
(sigma2_posterior <- (sigma^2 * sigma_previa^2)/(sigma^2 +
    n * sigma_previa^2))
```

```
## [1] 0.3190971
```

```
ggplot(data = data.frame(x = c(-5, 15)), aes(x)) +
    stat_function(fun = dnorm, args = list(mean = mu_previa,
        sd = sigma_previa), aes(color = "Previa")) +
    stat_function(fun = dnorm, args = list(mean = mu_posterior,
        sd = sqrt(sigma2_posterior)), aes(color = "Posterior")) +
    stat_function(fun = dnorm, args = list(mean = mu,
        sd = sigma), aes(color = "Verosimilitud")) +
    theme_minimal()
```



Con más datos, la distribución se ajusta a esto y le quita importancia a la información previa.

```
x <- rnorm(n = 100, mean = 10, sd = 1)

(mu <- mean(x))
```

```
## [1] 9.890422
```

```
(sigma <- sd(x))
```

```
## [1] 1.134588
```

```
(n <- length(x))
```

```
## [1] 100
```

```
(mu_previa <- 0)
```

```
## [1] 0
```

```
(sigma_previa <- 1)
```

```
## [1] 1
```

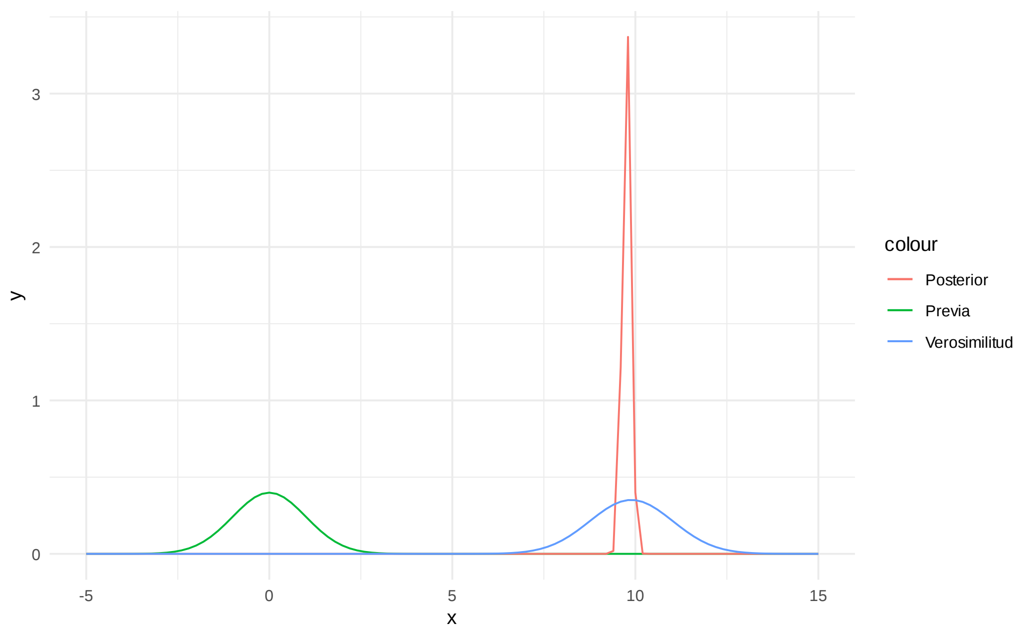
```
(mu_posterior <- ((sigma^2)/(sigma^2 + n * sigma_previa^2)) *
    mu_previa + ((n * sigma_previa^2)/(sigma^2 + n *
    sigma_previa^2)) * mu)
```

```
## [1] 9.764722
```

```
(sigma2_posterior <- (sigma^2 * sigma_previa^2)/(sigma^2 +
    n * sigma_previa^2))
```

```
## [1] 0.01270929
```

```
ggplot(data = data.frame(x = c(-5, 15)), aes(x)) +
    stat_function(fun = dnorm, args = list(mean = mu_previa,
        sd = sigma_previa), aes(color = "Previa")) +
    stat_function(fun = dnorm, args = list(mean = mu_posterior,
        sd = sqrt(sigma2_posterior)), aes(color = "Posterior")) +
    stat_function(fun = dnorm, args = list(mean = mu,
        sd = sigma), aes(color = "Verosimilitud")) +
    theme_minimal()
```



Si los datos por si solo son muy variable, la posterior tiende a parecerse a la distribución previa en lugar que a la verosimilitud.

```
x <- rnorm(n = 10, mean = 10, sd = 5)

(mu <- mean(x))
```

```
## [1] 10.90214
```

```
(sigma <- sd(x))
```

```
## [1] 5.107251
```

```
(n <- length(x))
```

```
## [1] 10
```

```
(mu_previa <- 0)
```

```
## [1] 0
```

```
(sigma_previa <- 1)
```

```
## [1] 1
```

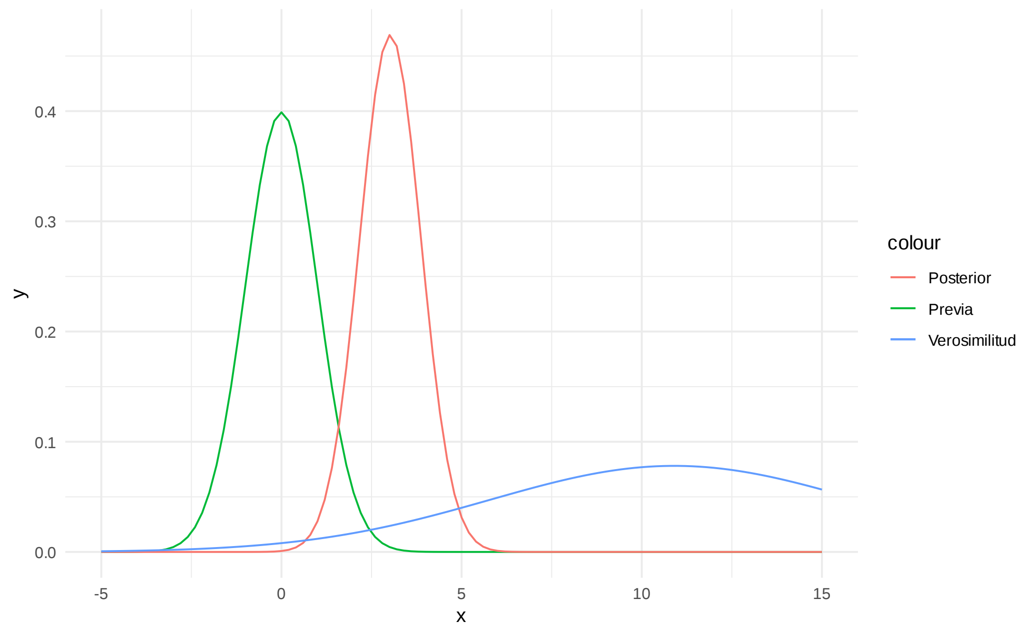
```
(mu_posterior <- ((sigma^2)/(sigma^2 + n * sigma_previa^2)) *
    mu_previa + ((n * sigma_previa^2)/(sigma^2 + n *
    sigma_previa^2)) * mu)
```

```
## [1] 3.021321
```

```
(sigma2_posterior <- (sigma^2 * sigma_previa^2)/(sigma^2 +
    n * sigma_previa^2))
```

```
## [1] 0.722869
```

```
ggplot(data = data.frame(x = c(-5, 15)), aes(x)) +
    stat_function(fun = dnorm, args = list(mean = mu_previa,
        sd = sigma_previa), aes(color = "Previa")) +
    stat_function(fun = dnorm, args = list(mean = mu_posterior,
        sd = sqrt(sigma2_posterior)), aes(color = "Posterior")) +
    stat_function(fun = dnorm, args = list(mean = mu,
        sd = sigma), aes(color = "Verosimilitud")) +
    theme_minimal()
```



3.10.6. Funciones de pérdida

Lo más importante acá es que dependiendo de la función de pérdida podemos construir una estimador para θ. En el caso de los componentes electrónicos recordemos que la posterior nos daba

```
alpha <- 9
beta <- 36178
```

- **Pérdida cuadrática:** Recoremos que la media de una gamma es α/β entonces

```
(theta <- alpha/beta)
```

```
## [1] 0.00024877
```

Y por lo tanto el tiempo promedio del componente electrónico es $1/\theta$=4019.7777778.

- **Pérdidad absoluta:** La distribución Gamma no tiene una forma cerrada para la mediana, por que se puede aproximar así,

```
m <- rgamma(n = 1000, scale = beta, shape = alpha)
(theta <- median(m))
```

```
## [1] 317434.4
```

Y por lo tanto el tiempo promedio del componente electrónico es $1/\theta$=3,1502569 $\times 10^{-6}$.

OJO: En este caso la pérdida cuadrática ajusta mejor ya que la distribución que la pérdida absoluta ya que la distribución NO es simétrica. En el caso simétrico los resultados serían muy similares.

Capítulo 4

Estimación por máxima verosimilitud

¿Será posible estimar sin una densidad previa? Se debería ajustar la noción de muestra a independencia dado el valor de un parámetro.

Recuerde que, para $X_1, \ldots, X_n \stackrel{i.i.d}{\sim} f(X|\theta)$ con θ fijo, la **función de verosimilitud** se define como

$$f_n(X|\theta) = \pi(X_i|\theta) = G(\theta|X).$$

Si $\theta_1, \theta_2 \in \Omega$, θ es el valor real del parámetro. Si la muestra es fija, evaluamos, para θ_1, $f_n(X|\theta_1) = G(\theta_1|X)$ y, de igual forma para θ_2, $f_n(X|\theta_2) = G(\theta_2|X)$. Supongamos que

$$f_n(X|\theta_1) > f_n(X|\theta_2) \implies G(\theta_1|X) > G(\theta_2|X) \text{ (principio de verosimilitud)}$$

Interpretación. Es más verosímil (realista) que el verdadero parámetro sea θ_1 que θ_2 dada la muestra.

Definición. Para cada $x \in \mathcal{X}$ (espacio muestral), sea $\delta(x) \in \delta$ estimador de θ tal que $f_n(x|\theta)$ es máximo. A $\delta(x)$ se le llama **MLE (estimador de máxima verosimilitud)**.

Ejemplo. Si $X_1, \ldots, X_n \sim \text{Exp}(\theta)$, estime θ.

Determinamos la función de verosimilitud,

$$f_n(X|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-X_i/\theta} = \frac{1}{\theta^n} \exp\left(\frac{1}{\theta}\sum_{i=1}^{n} X_i\right) = \theta^{-n} e^{-y/\theta}.$$

Considere la **log-verosimilitud**

$$L(\theta|X) = \ln f_n(X|\theta) = -n \ln\theta - \frac{y}{\theta}$$

Como es una transformación monótona creciente, la función de verosimilitud se maximiza si la log-verosimilitud es máxima. Entonces,

$$\frac{\partial}{\partial\theta} L(\theta|X) = \frac{-n}{\theta} + \frac{y}{\theta^2} = 0 \implies \frac{1}{\theta}\left(-n + \frac{y}{\theta}\right) = 0 \implies \hat{\theta} = \frac{y}{n} = \bar{X}_n.$$

Para verificar que es un máximo:

$$\frac{\partial^2 L}{\partial\theta^2} = \frac{n}{\theta^2} - \frac{2y}{\theta^3}\bigg|_{\theta=\frac{y}{n}} = \frac{1}{\hat{\theta}^2}\left[n - \frac{2y}{\frac{y}{n}}\right] = \frac{-n}{\hat{\theta}^2} < 0.$$

Entonces $\hat{\theta} = \bar{X}_n$ es el MLE de θ.

Ejemplo. En una prueba sobre alguna enfermedad, en un 90 % da la verdadera condición (enfermo) y en un 10 % la prueba se equivoca (que diga que la persona esté enferma cuando está sana). Considere una variable aleatoria Bernoulli(θ),$\theta \in \{0{,}9, 0{,}1\}$ Una muestra sería

$$x = \begin{cases} 1 & \text{si la prueba es positiva} \\ 0 & \text{si no} \end{cases}$$

Si $x = 0$, entonces $f(0|\theta) = \begin{cases} 0{,}9 & \text{si } \theta = 0{,}1 \\ 0{,}1 & \text{si } \theta = 0{,}9 \end{cases}.$

Si $x = 1$, entonces $f(1|\theta) = \begin{cases} 0{,}1 & \text{si } \theta = 0{,}1 \\ 0{,}9 & \text{si } \theta = 0{,}9 \end{cases}.$

El MLE corresponde a

$$\hat{\theta} = \begin{cases} 0{,}1 & \text{si } x = 0 \\ 0{,}9 & \text{si } x = 1 \end{cases}$$

Ejemplo. Para el caso normal, $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 conocida, estime μ.

$$f_n(x|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

La log-verosimilitud es de la forma

$$L(\mu|x) = \frac{-n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Basta con minimizar $Q(\mu) = \sum_{i=1}^{n}(x_i-\mu)^2$.

$$\frac{\partial Q}{\partial \mu} = -2\sum_{i=1}^{n}(x_i-\mu) \implies n\mu = \sum_{i=1}^{n} x_i \implies \hat{\mu} = \bar{x}_n.$$

No hace falta verificar la condición de segundo orden, pues Q es una función cuadrática de μ y tiene un único máximo.

$$\hat{\mu}_{MLE} = \bar{x}_n \quad (*)$$

Ahora, si $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$ desconocido, por $(*)$,

$$L(\sigma^2|X_1, \ldots, X_n) = \frac{n}{2}\ln(2\pi\sigma^2) - -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = 0$$

Entonces

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \text{ (varianza muestral)}$$

Las condiciones de segundo orden quedan como ejercicio.

Nota. Si θ_{MLE} de θ, entonces $h(\theta_{MLE})$ es el MLE de $h(\theta)$.

Sea $h(x,y) = \sqrt{y}$ (es inyectiva). $h(\bar{x}_n, \hat{\sigma}^2) = \sqrt{\hat{\sigma}^2} = \hat{\sigma}$.

El MLE de $\dfrac{\sigma}{\mu} = \dfrac{\hat{\sigma}}{\bar{x}_n}$.

Ejemplo. $X_1, \ldots, X_n \overset{i.i.d}{\sim}$ Unif(0). Estime θ ($\theta > 0$). Suponga que $x_i > 0 \forall i$.

$$f(X|\theta) = \frac{1}{\theta} \cdot 1_{[0,\theta]}(x)$$

La verosimilitud es

$$f_n(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} 1_{\{0 \leq x_i \leq \theta\}} \quad 0 \leq x_i \leq \theta \ \forall i$$

Vea que $f_n(x|\theta)$ es positivo si y solo si $O \leq X_{(n)} \leq \theta$.

El valor de la muestra $\{X_1, \ldots, X_n\}$ en la i-ésima posición cuando los datos se ordenan de menor a mayor se denota $X_{(i)}$ (estadístico de orden). En este caso, $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Entonces $\hat{\theta}_{MLE} = x_{(n)}$.

Capítulo 5

Propiedades del MLE

5.1. Propiedad de invarianza

Teorema. Si $\hat{\theta}$ es el MLE de $\hat{\theta}$ y si g es biyectiva, entonces $g(\theta)$ es el MLE de $g(\theta)$.

Prueba:

Sea Γ el espacio paramétrico $g(\Omega)$. Como g es biyectiva entonces h : inversa de g: $\theta = h(\psi), \psi \in \Gamma$.

Reparametrizando la verosimilitud,

$$f_n(x|\theta) = f_n(x|h(\psi)).$$

El MLE de $\psi : \hat{\psi}$ satisface que $f_n(x|h(\hat{\psi}))$ es máximo.

Como $f_n(x|\theta)$ se maximiza cuando $\theta = \hat{\theta}$, entonces $f_n(x|h(\psi))$ se maximiza cuando $\hat{\theta} = h(\psi)$ para algún ψ.

Se concluye que $\hat{\theta} = h(\hat{\psi}) \implies \hat{\psi} = g(\hat{\theta})$.

Ejemplo: $g(\theta) = \dfrac{1}{\theta}$ es biyectiva si $\theta > 0$. Así,

$$\frac{\hat{1}}{\theta} = \frac{1}{\hat{\theta}} = \frac{1}{\frac{1}{\hat{X}_n}} = \hat{X}_n \quad (\theta \text{ es parámetro de tasa}).$$

¿Qué pasa si h no es biyectiva?

Definicion (Generalización del MLE). Si g es una función de θ y G la imagen de Ω bajo g. Para cada $t \in G$ defina

$$G_t = \{\theta : g(\theta) = t\}$$

Defina $L^*(t) = \max_{\theta \in G_t} \ln f_n(x|\theta)$. El MLE de $g(\theta)(= \hat{t})$ satisface $L^*(\hat{t}) = \max_{t \in G} L^*(t)$.

Teorema. Si $\hat{\theta}$ es el MLE de θ entonces $g(\hat{\theta})$ es el MLE de $g(\theta)$ (g es arbitraria).

Prueba. Basta probar $L^*(\hat{t}) = \ln f_n(x|\hat{\theta})$. Se cumple que $\hat{\theta} \in G_{\hat{t}}$. Como $\hat{\theta}$ maximiza $f_n(x|\theta)$ $\forall \theta$, también lo hace si $\theta \in G_{\hat{t}}$. Entonces $\hat{t} = g(\hat{\theta})$ (no pueden existir 2 máximos en un conjunto con la misma imagen).

Ejemplos.$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

- Si $h(\mu, \sigma^2) = \sigma$ (no es biyectiva) $\implies h(\hat{X}_n, \hat{\sigma}^2) = \sqrt{\hat{\sigma}^2}$ es el MLE de σ.
- $h(\mu, \sigma^2) = \dfrac{\sigma^*}{\mu}$ (coeficiente de variación). $\dfrac{\hat{\sigma}}{\bar{X}_n}$ es el MLE de CV.
- $h(\mu, \sigma^2) = \mu^2 + \sigma^2$. $\mathbb{E}[X^2] - \mu^2 = \sigma^2 \implies \mathbb{E}[X^2] = \mu^2 + \sigma^2$. El MLE de $\mathbb{E}[X^2]$ es $\bar{X}_n^2 + \hat{\sigma}^2$.

5.2. Consistencia

Los estimadores bayesianos son de la forma

$$EB = W_1 \mathbb{E}[\text{Previa}] + W_2 \hat{X}_n.$$

El estimador bayesiano "combina" la esperanza de la previa y el $\hat{\theta}_{MLE}$. El $\hat{\theta}_{MLE}$ "hereda la consistencia del estimador bayesiano".

$$EB = W_1 \mathbb{E}[\text{Previa}] + W_2 \hat{\theta}_{MLE}.$$

Afirmación. Bajo "condiciones usuales",

$$\hat{\theta}_{MLE} \xrightarrow[n \to \infty]{\mathbb{P}} \theta.$$

Capítulo 6

Cálculo numérico

6.1. Método de los momentos

Ejemplo. $X_1, \dots, X_n \sim \Gamma(\alpha, 1)$. Estime α.

$$f_n(x|\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}.$$

Verosimilitud: $f_n(x|\alpha) = \dfrac{1}{\Gamma(\alpha)^n} (\prod x_i) e^{\sum x_i}$.

$$\begin{aligned} \frac{\partial}{\partial\alpha} L(\alpha|x) &= \frac{\partial}{\partial\alpha}\Big[- n \ln\Gamma(\alpha) + (\alpha-1)\ln(\pi x_i) - \sum x_i \Big] \\ &= -n \frac{1}{\Gamma(\alpha)} \frac{d}{d\alpha} \Gamma(\alpha) + \ln(\prod x_i) = 0 \end{aligned}$$

Definición. Asumimos que $X_1, \dots, X_n \sim F$ indexada con un parámetro $\theta \in \mathbb{R}^k$ y que al menos tiene k momentos finitos. Para $j = 1, \dots, k$ sea $\mu_j(\theta) = \mathbb{E}[X_1^j|\theta]$. Suponga que $\mu(\theta) = (\mu_1(\theta), \dots, \mu_2(\theta))$ es biyectiva. Sea M la inversa de μ,

$$M(\mu(\theta)) = \theta = M(\mu_1(\theta), \dots, \mu_2(\theta))$$

y defina los momentos empíricos

$$m_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j, \quad j = 1, \dots, k.$$

El estimador según el método de los momentos es

$$\hat{\theta} = M(m_1, \ldots, m_k).$$

Del ejemplo anterior, $\mu_1(\alpha) = \mathbb{E}[x_1|\alpha] = \alpha$.Dado que $m_1 = \hat{x}_n$, el sistema por resolver es

$$\mu_1(\alpha) = m_1 \Longleftrightarrow \alpha = \bar{x}_n$$

El estimador por método de momentos es $\hat{\alpha} = \bar{X}_n$.

Ejemplo. $X_1, \ldots, X_n \overset{i.i.d}{\sim} \Gamma(\alpha, \beta)$. La varianza de X es

$$\frac{\alpha}{\beta^2} = \mathrm{Var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \frac{\alpha^2}{\beta^2}.$$

Se debe resolver el sistema

$$\begin{cases} \mu_1(\theta) = \dfrac{\alpha}{\beta} = \bar{X}_n = m_1 & (1) \\ \mu_2(\theta) = \dfrac{\alpha(\alpha+1)}{\beta^2} = m_2 & (2) \end{cases}$$

De (1), $\alpha = m_1\beta$. Sustituyendo en (2),

$$m_2 = \frac{m_1\beta(m_1\beta+1)}{\beta^2} = m_1^2 + \frac{m_1}{\beta} = m_2 \implies m_2 - m_1^2 = \frac{m_1}{\beta}.$$

De esta manera,

$$\hat{\beta} = \frac{m_1}{m_2 - m_1^2}, \quad \hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}$$

Teorema. Si $X_1, X_2, \ldots$ i.i.d con distribución indexada por $\theta \in \mathbb{R}^k$. Suponga que los k momoentos teóricos son finitos $\forall\theta$ y suponga que M es continua. Entonces el estimador por el método de momentos es consistente.

¿Cuál es el comportamiento en la distribución de $\hat{\theta}$ cuando la muestra es grande?

Del teorema del límite central,

$$\frac{\bar{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sigma} \xrightarrow{d} N(0,1)$$

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2}\sum \text{Var}(X_1) = \frac{\sigma^2}{n}$$

Implica que se debe multiplicar la media muestral por una constante para hacer la desviación visibile y, con ello, hacer inferencia del parámetro.

Caso general. Si $f(X|\theta)$ es "suficientemente suave" como función de θ, es puede comprobar que la verosimilitud tiende a una normal conforme $n \to \infty$. Es decir,

$$f(X|\theta) \propto \exp\left[\frac{-1}{2\frac{V_n(\theta)}{n}}(\theta - \hat{\theta})\right], \quad n \to \infty \quad (*)$$

donde $\hat{\theta}$ es el MLE de θ.

$$V_n(\theta) \xrightarrow[n\to\infty]{} V_\infty(\theta) < \infty$$

Notas:

1) Si $n \to \infty$ la normal en $(*)$ tiene muchísima precisión y es concentrada alrededor de $\hat{\theta}$.

2) En el caso bayesiano, ninguna previa en θ puede anular el efecto en la verosimilitud cuando $n \to \infty$.

3) Por $(*)$ el MLE se distribute asintóticamente como

$$N\left(\theta, \frac{V_\infty(\theta)}{n}\right),$$

$Var(X_n) \xrightarrow[n\to\infty]{} 0$ y $\mathbb{E}[X_n] = X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ (confirma que el MLE es consistente).

6.2. Método Delta

Si $Y_1, Y_2, \ldots$ es una sucesión de variables aleatorias y sea F^* su c.d.f. continua. Sea $\theta \in \mathbb{R}$ y $\{a_n\}$ sucesión de números positivos tal que $a_n \nearrow \infty$. Suponga que $a_n(Y_n - \theta) \xrightarrow{d} F^*$. Si α es una función tal que $\alpha'(\theta) \neq 0$, entonces

$$\frac{a_n}{\alpha'(\theta)}[\alpha(Y_n) - \alpha(\theta)] \xrightarrow{d} F^*$$

Ejemplo. $X_1, X_2, \dots$ i.i.d. de variables con media μ y varianza σ^2. Sea α una función tal que $\alpha'(\mu) \neq 0$. Por el T.L.C,

$$\frac{\sqrt{n}}{\sigma}(X_n - \mu) \xrightarrow{d} N(0,1)$$

Entonces por el método Delta

$$\frac{\sqrt{n}}{\sigma\alpha'(\mu)}[\alpha(\bar{X}_n) - \alpha(\mu)] \xrightarrow{d} N(0,1)$$

Si $\alpha(\mu) = \dfrac{1}{\mu}$ $(\mu \neq 0) \implies -\dfrac{1}{\mu^2} = \alpha'(\mu)$. Entonces por el método Delta

$$\frac{\sqrt{n}}{\sigma}\mu^2\left[\frac{1}{\bar{X}_n} - \frac{1}{\mu}\right] \xrightarrow{d} N(0,1)$$

Ejemplo (7.6.11)

Si $X_1, X_2 \dots \overset{i.i.d}{\sim} \text{Exp}(\theta)$. Sea $T_n = \sum X_i \implies \hat{\theta} = \dfrac{1}{\bar{X}_n} = \dfrac{n}{T_n}$.

Note que $\dfrac{1}{\hat{\theta}} = \bar{X}_n$ y

$$\frac{\sqrt{n}}{\sigma}\left[\bar{X}_n - \frac{1}{\theta}\right] \xrightarrow[n\to\infty]{d} N(0,1).$$

La varianza de una exponencial es $\sigma^2 = \text{Var}(X_1) = \dfrac{1}{\theta^2}$, entonces

$$\theta\sqrt{n}\left[\bar{X}_n - \frac{1}{\theta}\right] \xrightarrow[n\to\infty]{d} N(0,1).$$

El método Delta nos dice, con $\alpha(\mu) = \dfrac{1}{\mu}$, $\alpha'(\mu) = -\dfrac{1}{\mu^2}$, el comportamiento asintótico de MLE:

$$\frac{\theta\sqrt{n}}{\alpha'(1/\theta)}\left[\bar{\alpha}(X_n) - \alpha\left(\frac{1}{\theta}\right)\right] = \frac{\theta\sqrt{n}}{\frac{1}{1/\theta}}\left[\frac{1}{\bar{X}_n} - \theta\right] \xrightarrow[n\to\infty]{d} N(0,1)$$
$$= \frac{\sqrt{n}}{\theta}\left[\frac{1}{\bar{X}_n} - \theta\right] \xrightarrow[n\to\infty]{d} N(0,1)$$

El MLE $\hat{\theta} = \frac{1}{\bar{X}_n}$ es asintóticamente normal con media θ y varianza $\frac{V_n(\theta)}{n} = \frac{\theta^2}{n}$.

Caso bayesiano. Tome una previa conjugada $\theta \sim \Gamma(\alpha, \beta)$, posterior $\theta \sim \Gamma(\alpha + n, \beta + y)$, $y = \sum X_i$. Supongamos que es entero positivo.

$$\Gamma(\alpha + n, \beta + y) \sim \sum_{i=1}^{\alpha+n} e^{\beta+y}$$

Por el T.L.C., la distribución posterior $\theta|X$ se distribuye como una normal con media $\frac{\alpha + n}{\beta + y}$ y varianza $\frac{\alpha + n}{(\beta + y)^2}$. Tomando una previa poco informativa, (α, β son pequeños), la media es

$$\frac{n}{y} = \frac{1}{\bar{X}_1} = \hat{\theta}_{MLE}$$

y la varianza

$$\frac{1}{y^2/n} = \frac{\theta^2}{n} = \frac{V_n(\hat{\theta})}{n}.$$

Capítulo 7

Estadísticos Suficientes y Criterio de Factorización

Capítulo 8

Estadísticos suficientes

Una función de verosimilitud se va a describir a través de un número. El objetivo es buscar un estadístico $T = r(X_1, \ldots, X_n)$ que resuma de manera óptima la información de $X_1, \ldots, X_n$

Definición. Sea $X_1, \ldots, X_n$ una muestra indexada por θ. Sea T un estadístico, suponga que para cada $\theta \in \Omega$ y para cada t en la imagen de T, $X_1 \cdots X_n | T = t$ depende solamente de t y no de θ. Entonces T es suficiente.

8.1. Teorema de Factorización de Fisher

Teorema. Si $X_1, \ldots, X_n$ es una muestra aleatoria de $f(X|\theta)$, el parámetro θ es desconocido. Un estadístico $T = r(X_1, \ldots, X_n)$ es suficiente si y solo si

$$f_n(x|\theta) = u(x)v(r(x), \theta) \;\forall x \in \mathbb{R},\; \forall \theta \in \mathbb{R}.$$

Prueba (Discreta). $f_n(x|\theta) = \mathbb{P}(X = x|\theta)$

"$\Leftarrow$" Sea $A(t) = \{x \in \mathbb{R} | r(x) = t\}$. Para $\theta \in \mathbb{R}$, $x \in A(t)$,

$$\begin{aligned}
\mathbb{P}(X=x|T=t) &= \frac{\mathbb{P}(X=x\cap T=t)}{\mathbb{P}(T=t)} = \frac{\mathbb{P}(X=x)}{P(T=t)} = \frac{f_n(x|\theta)}{\sum\limits_{y\in A(t)} f_n(y|\theta)} \\
&= \frac{u(x)v(r(x),\theta)}{\sum\limits_{y\in A(t)} u(y)v(r(y),\theta)} = \frac{u(x)}{\sum\limits_{y\in A(t)} u(y)}
\end{aligned}$$

no depende de θ.

Si $x \notin A(t) \implies \mathbb{P}(X=x|T=t) = 0$ no depende de θ.

"$\Rightarrow$" Si T es un estadístico suficiente, $u(x) = \mathbb{P}(X=x|T=t)$ no depende de θ. Sea $v(t,\theta) = \mathbb{P}_\theta(T=t)$. Entonces

$$f_n(x|\theta) = (X=x|\theta) = \frac{\mathbb{P}(X=x|\theta)}{\mathbb{P}(T=t)}\mathbb{P}(T=t) = u(x)v(t,\theta).$$

Consecuencia: $f_n(x|\theta) \propto v(r(x),\theta)$ ($u(x)$ es una constante con respecto a θ). Aplicando el teorema de Bayes,

$$\pi(\theta|x) \propto \pi(\theta)v(r(x),\theta).$$

Corolario. Un estadístico $r(x)$ es suficiente si y solo si no importa cuál previa de θ se use, la posterior depende solamente de $r(x)$ a través de los datos.

Ejemplo. $X_1,\ldots,X_n \sim \text{Poi}(\lambda)$,

$$f_n(x|\theta) = \prod_{i=1}^n \frac{e^{-\lambda}}{x_i!} = \frac{e^{-\lambda n}\lambda^{\sum x_i = r(x)}}{\prod x_i!} = \underbrace{\frac{1}{\prod_{i=1}^n x_i!}}_{u(x)} \underbrace{e^{-\lambda n}\lambda^{r(x)}}_{v(r(x),\lambda)}$$

Si $x_i < 0$ para al menos un i, entonces $f_n(x|\theta) = 0$. Tome $u(x) = 0$. Por el teorema de factorización, $r(x) = \sum x_i$ es un estadístico suficiente para λ.

Ejemplo. $X_1,\ldots,X_n \sim f(x|\theta)$

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & 0<x<1 \\ 0 & \text{otro caso} \end{cases}$$

Verosimilitud: $(0 < x_i < 1 \; \forall i)$

$$f_n(x|\theta) = \theta^n \Big[\underbrace{\prod (x_i)}_{r(x)} \Big]^{\theta-1} = \underbrace{\theta^n (r(x))^{\theta-1}}_{v(r(x),\theta)} \cdot \underbrace{1}_{u(x)}$$

Por el teorema de factorización $r(x) = \prod x_i$ es un estadístico suficiente.,

Ejemplo. $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ (σ^2 conocido).

$$\begin{aligned} f_n(x|\theta) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n X_i^2 + \frac{\mu}{\sigma^2} \sum X_i - \frac{\mu^2 n}{2\sigma^2} \right] \end{aligned}$$

Tome

$$\begin{aligned} u(x) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n X_i^2 \right], \\ v(r(x), \mu) &= \exp\left[\frac{\mu}{\sigma^2} r(x) - \frac{n\mu^2}{2\sigma^2} \right]. \end{aligned}$$

Por teorema de factorización, $r(x) = \sum X_i$ es un estadístico suficiente para μ.

Con σ^2 desconocido, $\theta = (\mu, \sigma^2)$, tome $u(x) = 1$,

$$v(r_1(x), r_2(x), \theta) = (2\pi\sigma^2)^{-n/2} \exp\left[\frac{-r_2(x)}{2\sigma^2} + \frac{\mu r_1(x)}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \right]$$

Entonces

$$(r_1(x), r_2(x)) = \left(\sum x_i, \sum x_i^2 \right)$$

es un estadístico suficiente para (μ, σ^2).

Ejemplo. $X_1, \dots, X_n \overset{i.i.d}{\sim} \text{Unif}(0, \theta)$, $\theta > 0$, $f(x|\theta) = 1_{[0,\theta]}(x) \dfrac{1}{\theta}$.

$$f_n(x|\theta) = \prod_{i=1}^n 1_{[0,\theta]}(x_i) \left(\frac{1}{\theta} \right)$$

Nota: si al menos uno de los $x_i < 0$ o $x_i > \theta$, $u(x) = 0$ ($f(x|\theta) = 0$) (Trivial).

Si $0 < x_i < \theta \ \forall i \implies f_n(x|\theta) = 1_{[0,\theta]}(\text{máx}\{x_i\}) \left(\frac{1}{\theta}\right)^n$.

Si $T = r(x) = X_{(n)} \implies f_n(x|\theta) = u(x)v(r(x),\theta)$, $u(x) = 1$. Por teorema de factorización, $r(x) = x_{(n)}$ es un estadístico suficiente para θ.

Capítulo 9

Estadístico suficiente multivariado.

Si $\theta \in \mathbb{R}^k$, $k \geq 1$ se necesita al menos k estadísticos $(T_1, \dots, T_k)$ para cada $i = 1, \dots, k$, $T_i = r_i(X_1, \dots, X_n)$.

Definición. Suponga que para cada $\theta \in \Omega$ y $(t_1, \dots, t_k) \in \mathbb{R}^k$ valor del estadístico $(T_1, \dots, T_k)$, la distribución condicional de $X_1, \dots, X_n$ dado $(T_1, \dots, T_k) = (t_1, \dots, t_k)$ no depende de θ, entonces $(T_1, \dots, T_k)$ es un **estadístico suficiente** para θ.

Criterio de factorización:

$$f_n(x|\theta) = u(x)v(r_1(x), \dots, r_k(x), \theta) \Leftrightarrow T = (r_1(x), \dots, r_k(x)) \text{ es suficiente}$$

Si $(T_1, \dots, T_k)$ es suficiente para θ y si $(T_1', \dots, T_k') = g(T_1, \dots, T_k)$ donde g es biyectiva, entonces $(T_1', \dots, T_k')$ es suficiente para θ.

$$u(x)v(r(x)|\theta) = u(x)v(g^{-1}(g(r(x))), \theta).$$

Ejemplo. Considere $(T_1', T_2') = g(T_1, T_2) = \left(\frac{1}{n}T_1, \frac{1}{n}T_2 - \frac{1}{n^2}T_1^2\right)$.

De la primera entrada,

$$T_1' = \frac{1}{n}T_1 \implies T_1 = nT_1'.$$

De la segunda,

$$\begin{aligned} T_2' = \frac{1}{n}T_2 - \frac{1}{n^2} &= \frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2 \\ &= \frac{1}{n}\sum X_i^2 - 2X_i\bar{X}_n^2 + \bar{X}_n \\ &= \frac{1}{n}\sum (X_i - \bar{X}_n)^2 = \hat{\sigma}_n^2 \end{aligned}$$

Como g es biyectiva entonces $(\bar{X}_n, \sigma_n^2)$ es un estadístico suficiente para (μ, σ^2).

Ejemplo. $X_1, \ldots, X_n \sim \text{Unif}(a, b)$, $a < b$. Encuentre un estadístico sufienciente.

- Si $x_i \leq a$ o $x_i > b$, tome $u(x) = 0$.
- Si $a < x_i < b \ \forall i$,
- $x_i > a \ \forall i \Leftrightarrow x_{(1)} > a$.
- $x_i < b \ \forall i \Leftrightarrow x_{(n)} < b$.

La verosimilitud es de la forma

$$f_n(x|(a,b)) = \prod_{i=1}^{n} 1_{[a,b]}(x_i) = \underbrace{\frac{1}{(b-a)^n} 1_{\{(Z,W):Z>a,W<b\}}(x_{(1)}, x_{(n)})}_{v(r_1,r_2,(a,b))} \cdot \underbrace{1}_{u(x)}$$

Por teorema de factorización $(X_{(1)}, X_{(n)})$ es un estadístico suficiente para (a, b).

Capítulo 10

Estadísticos minimales

Idea: un estadístico suficiente que garantice una partición de $\mathcal{X}$ (espacio muestral) de la manera más simple posible.

Definición (Estadístico de orden). Sean $X_1, \ldots, X_n \overset{i.i.d}{\sim} f$. Al ordenar los datos

$$(Y_1, \ldots, Y_n) = (X_{(1)}, \ldots, X_{(n)}) \text{ tal que } Y_1 < \ldots < Y_n$$

Nota: $(X_{(1)}, \ldots, X_{(n)})$ es un estadístico suficiente de θ.

Ejemplo. $X_1, \ldots, X_n \sim \text{Cauchy}(\alpha)$.

$$f(x) = \frac{1}{\pi}[1 + (x - \alpha)^2]^{-1}, x \in \mathbb{R}$$

Busque un estimador suficiente para $\alpha \in \mathbb{R}$.

$$f_n(x|\alpha) = \prod(x|\alpha) = \frac{1}{\pi}[1 + (x_i - \alpha)^2]^{-1} = \underbrace{\frac{1}{\pi^n}}_{u(x)} \underbrace{\prod_{i=1}^{n}[1 + (x_i - \alpha)^2]^{-1}}_{v(y,\alpha)}$$

donde $y = (X_{(1)}, \ldots, X_{(n)})$ es suficiente para α.

Ejercicio: estime α usando R o usando método de momentos.

Definición. Un estadístico T es **suficiente minimal** si T es suficiente y es función de cualquier otro estadístico suficiente.

Teorema. Si $T = r(X_1, \ldots, X_n)$ es un estadístico suficiente para θ, entonces el MLE $\hat{\theta}$ de θ depende de $X_1, \ldots, X_n$ solamente a través de T. Además, si $\hat{\theta}$ es suficiente entonces $\hat{\theta}$ es minimal.

Prueba. Por teorema de factorización, $f_n(x|\theta) = u(x)v(r(x), \theta)$ de $T = r(x)$ es suficiente y

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f_n(x|\theta) = \underset{\theta}{\operatorname{argmax}}\, v(r(x), \theta) \quad (\Delta)$$

Como $\hat{\theta} = g(T)$ para cualquier T estadístico suficiente, entonces $\hat{\theta}$ es minimal.

Teorema. Si $T = r(X_1, \ldots, X_n)$ es un estadístico suficiente para θ entonces el estimador bayesiano (bajo una escogencia de L) depende de $X_1, \ldots, X_n$ solamente a través de T (el estimador bayesiano es minimal).

Prueba. Sustituya (Δ) por $\pi(\theta|x) \propto v(r(x), \theta) \cdot \pi(\theta)$. Como cualquier estimador bayesiano depende de $\pi(\theta|x)$, cualquier estimador bayesiano depende de los datos a través de $r(x)$.

Capítulo 11

Mejorando estimadores

¿Existirá otra medida de comparación entre estimadores?

Considere una **función de riesgo**

$$R(\theta, \delta) = \mathbb{E}[(\delta(x) - \theta)^2]$$

Si $\delta(x)$ estima una característica de F:

$$R(\theta, \delta) = \mathbb{E}[(\delta(x) - h(\theta))^2] \quad (\Delta\Delta)$$

donde h es la característica.

Nota: la función de riesgo puede ser calculada con una posterior $\pi(\theta|X)$.

Definición.

- Decimos que δ es **inadmisible** si $\exists\delta_0$ (otro estimador) tal que $R(\theta, \delta)$ $\forall\theta \in \Omega$.
- Decimos que δ_0 **"domina"** a δ en el caso anterior.
- A $(\Delta\Delta)$ se le llama **MSE** o **error cuadrático medio**.

Teorema (Rao-Blackwell). Sea $\delta(X)$ un estimador y T un estadístico suficiente para θ y sea $\delta_0 = \mathbb{E}[\delta(X)|T]$. Entonces

$$R(\theta, \delta_0) \leq R(\theta, \delta) \; \forall\theta \in \Omega$$

Prueba. Por la desigualdad de Jensen,

$$\mathbb{E}_\theta[(\delta(x)-\theta)^2] \geq (E_\theta[(\delta(x)-\theta)])^2.$$

También,

$$\mathbb{E}[(\delta(x)-\theta)^2|T] \geq (E[(\delta(x)|T)]-\theta)^2 = (\delta_0(T)-\theta)^2.$$

Entonces,

$$\mathbb{E}[(\delta(x)-\theta)^2] \leq \mathbb{E}[\mathbb{E}[(\delta(x)-\theta)^2|T]] = \mathbb{E}[(\delta(x)-\theta)^2] = R(\theta,\delta).$$

Nota. Si cambiamos a $R(\theta,\delta) = \mathbb{E}[|\delta(x)-\theta|]$ (error medio absoluto), el resultado anterior es cierto.

Ejemplo. Sean $X_1,\ldots,X_n \overset{i.i.d}{\sim}$ Poisson(θ) donde θ es la tasa de "visitas" de clientes por hora.

A partir de la verosimilitud,

$$f_n(X|\theta) = \frac{e^{-\theta n}\theta^{\sum X_i}}{\prod X_i!}$$

se tiene que $T = \sum X_i$ es un estadístico suficiente para θ.

Sea $Y_i = \begin{cases} 1 & \text{si } X_i = 1 \\ 0 & \text{si } X_i \neq 1 \end{cases}$.

El objetivo es estimar p donde p es la probabilidad de que $X_i = 1$ (solo llegue un cliente por hora). Un estimador de p (MLE) es

$$\delta(x) = \frac{\sum Y_i}{n}$$

¿Es el óptimo?

Calculamos

$$\mathbb{E}[\delta(x)|T] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i|T)$$

Vea que

$$\begin{aligned}\mathbb{E}[Y_i|T=t]=\mathbb{P}(X_i=1|T=t)&=\frac{\mathbb{P}(X_i=1,T=t)}{\mathbb{P}(T=t)}\\&=\frac{\mathbb{P}(X_i=1,\sum_{j\neq i}X_j=t-1)}{\mathbb{P}(T=t)}\\&=\frac{\mathbb{P}(X_i=1)\mathbb{P}(\sum_{j\neq i}X_j=t-1)}{\mathbb{P}(T=t)}=\Delta\end{aligned}$$

- $\mathbb{P}(X_i=1)=\theta e^{-\theta}$
- $\mathbb{P}(\sum_{j\neq i}X_j=t-1)=e^{-(n-1)\theta}\dfrac{((n-1)\theta)^{t-1}}{(t-1)!}$
- $\mathbb{P}(T=t)=e^{-n\theta}\dfrac{(n\theta)^t}{t!}$

Entonces,

$$\Delta=\frac{\theta e^{-n\theta}\dfrac{((n-1)\theta)^{t-1}}{(t-1)!}}{e^{-n\theta}\dfrac{(n\theta)^t}{t!}}=\frac{t}{n}\left(1-\frac{t}{n}\right)^{t-1}=G\left(\frac{t}{n}\right)$$

es el estadístico con MSE mínimo.

Capítulo 12

Distribución muestral de un estadístico

Capítulo 13

Distribución muestral

Definición. Suponga que $X_1, \ldots, X_n$ es una muestra con parámetro θ con parámetro θ (desconocido). Sea $T = r(X_1, \ldots, X_n, \theta)$. La distribución de T dado θ se llama **distribución muestral**.

Ejemplo. Si $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. El MLE de μ es

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La distribución muestral del estadístico $\bar{X}_n$ es

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- $\mathbb{E}[\bar{X}_n] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\mathbb{E}[X_i] = \dfrac{1}{n}\cdot n\mathbb{E}[X_1] = \mu.$
- $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\right) = \dfrac{1}{n^2}\cdot n \cdot \mathrm{Var}(X_1) = \dfrac{\sigma^2}{n}.$

Ejemplo. X_i : tiempo de vida de un aparato. $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathrm{Exp}(\theta)$. La previa de θ es $\Gamma(1, 2)$. Solamente observamos $n = 3$. La posterior sería

$$\theta | X \sim \Gamma(1 + 3, 2 + \sum_{i=1}^{3} X_i).$$

El estimador bayesiano, bajo pérdida cuadrática, es

$$\mathbb{E}[\theta|X] = \frac{4}{2+\sum X_i} = \hat{\theta}$$

Problema: estimar $\mathbb{P}(|\hat{\theta}-\theta| < 0{,}1)$.

Vea que $P(|\hat{\theta}-\theta| < 0{,}1) = \mathbb{E}[P(|\hat{\theta}-\theta| < 0{,}1|\theta)]$

Sea

$$\begin{aligned} F(t|\theta) = \mathbb{P}(\hat{\theta} \leq t|\theta) &= \mathbb{P}\left(\frac{4}{2+T} \leq t \middle| \theta\right) \\ &= \mathbb{P}\left(2+T \geq \frac{4}{t} \middle| \theta\right) \\ &= \mathbb{P}\left(T \geq \frac{4}{t} - 2 \middle| \theta\right) \end{aligned}$$

Nota. Suma de exponenciales es una gamma.

Entonces $T \sim \Gamma(3,\theta)$, por lo que $F(t|\theta) = 1 - G_{\Gamma(3,0)}\left(\frac{4}{t}-2\right)$.

De esta manera,

$$\begin{aligned} \mathbb{P}[|\hat{\theta}-\theta| < 0{,}1|\theta] &= [-0{,}1+\theta < \hat{\theta} < 0{,}1+\theta|\theta] \\ &= G_{\Gamma(3,0)}\left(\frac{4}{0{,}1+\theta}-2\right) + G_{\Gamma(3,0)}\left(\frac{4}{-0{,}1+\theta}-2\right) \end{aligned}$$

y se toma la esperanza. Otra solución es cambiar la probabilidad de forma que no dependa de θ.

$$\mathbb{P}\left(\left| \underbrace{\frac{\hat{\theta}_{MLE}}{\theta} - 1}_{\text{Cambio relativo}} \right| < 0{,}1 \middle| \theta\right) = \mathbb{P}\left(\left|\frac{3}{\theta T} - 1\right| < 0{,}1 \middle| \theta\right) = \Delta$$

Si $T \sim \Gamma(3,0) \implies \theta T \sim \Gamma(3,1)$.

Por lo tanto,

$$\Delta = \mathbb{P}\left(0{,}9 < \frac{3}{\theta T} < 1{,}1 \middle| \theta\right) = \mathbb{P}\left(\frac{3}{1{,}1} < \theta T < \frac{3}{0{,}9}\right) = 13,4\,\%$$

Capítulo 14

Distribución χ^2

Definición. Para $m > 0$ definimos

$$\chi^2_m \sim \Gamma\left(\frac{m}{2}, \frac{1}{2}\right)$$

la distribución **chi-cuadrado** con m grados de libertad.

Propiedades:

- $\mathbb{E}[X] = m$.
- $\mathrm{Var}(X) = 2m$.
- Para $X_i \sim \chi^2_{m_i}$, $i = 1, \ldots, k$, independientes, entonces

$$\sum_{i=1}^{k} X_i \sim \chi^2_{\sum m_i}$$

- Si $X \sim N(0,1) \implies Y = X^2 \sim \chi^2_1$.
- Si $X_i \overset{i.i.d}{\sim} N(0,1) \implies \sum_{i=1}^{m} X_i^2 = \chi^2_m$.

Ejemplo. Si $X_1, \ldots, X_n \sim N(\mu, \sigma^2) \implies Z = \dfrac{X_i - \mu}{\sigma} \sim N(0,1)\ \forall i$.

Entonces

$$\sum Z_i^2 \sim \chi^2_n \implies \sum \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n \quad (*)$$

Además, si μ es conocido y σ^2 desconocido, entonces

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Su prueba queda como ejercicio.

De esta manera, observe que, de $(*)$,

$$\frac{n}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_n^2$$

La principal limitación es que μ es conocida. Asuma que también es desconocida. ¿Cuál es la distribución muestral de $(\bar{X}_n, \hat{\sigma}^2)$?

Teorema. Bajo las condiciones anteriores,

1) $\bar{X}_n$ y $\hat{\sigma}_n$ son independientes aunque $\hat{\sigma}_n$ es función de $\bar{X}_n$.

2) La distribución muestral de $\bar{X}_n$ es $N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

3) $n\dfrac{\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n} \dfrac{(X_i - \mu)^2}{\sigma^2} \sim \chi_{n-1}^2$.

De álgebra lineal, recuerde que una matriz $A_{n\times n}$es ortogonal si cumple que $A^{-1} = A$, $\det(A) = 1$. Si $X, Y \in \mathbb{R}^n$, $AX = Y$, A ortogonal, entonces

$$\|Y\|_2^2 = \|X\|_2^2 \quad (\Delta\Delta)$$

Teorema. Si $X_1, \dots, X_n \sim N(0, 1)$, A es ortogonal $n \times n$ y $Y = AX$ donde $X = (X_1, \dots, X_n)^T$ entonces $Y_1, \dots, Y_n \sim N(0, 1)$.

Prueba. Ver 8.3.1.

Si $X_1, \dots, X_n \sim N(0, 1)$, use Gram-Schmidt con vector inicial $u = \dfrac{1}{\sqrt{n}} 1_{n\times 1}$.

Generamos $A = \begin{bmatrix} u \\ \vdots \end{bmatrix}$. Defina $Y = AX$. Entonces

$$Y_1 = uX = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i = \sqrt{n}\bar{X}_n.$$

Por la propiedad $(\Delta\Delta)$, $\sum\limits_{i=1}^n Y_i^2 = \sum\limits_{i=1}^n X_i^2$. Entonces,

$$\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Como Y_1^2 y $\sum_{i=2}^n Y_i^2$ son independientes, entonces $\bar{X}_n$ y $\dfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ son independientes.

Note que $\sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$ ya que $Y_i \overset{i.i.d}{\sim} N(0,1)$.

Si $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, tome $Z_i = \dfrac{X_i - \mu}{\sigma}$ y repita todo lo anterior.

Ejemplo. $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ (μ, σ desconocidos). Los MLE son

$$\hat{\mu} = \bar{X}_n, \quad \hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2\right]^{\frac{1}{2}}.$$

Encuentre n tal que

$$p = \mathbb{P}\left[|\hat{\mu} - \mu| < \frac{6}{5}, |\hat{\sigma} - \sigma| < \frac{6}{5}\right] \geq \frac{1}{2}.$$

Por independencia de $\bar{X}_n$ y $\hat{\sigma}_n^2$,

$$p = \mathbb{P}\left[|\hat{\mu} - \mu| < \frac{\sigma}{5}\right]\mathbb{P}\left[|\hat{\sigma} - \sigma| < \frac{\sigma}{5}\right]$$

Por un lado,

$$\mathbb{P}\left[|\hat{\mu} - \mu| < \frac{6}{5}\right] = \mathbb{P}\left[-\frac{\sqrt{n}}{5} \leq \underbrace{\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma}}_{N(0,1)} < \frac{\sqrt{n}}{5}\right] = \Phi\left(\frac{\sqrt{n}}{5}\right) - \Phi\left(-\frac{\sqrt{n}}{5}\right).$$

Además,

$$
\begin{aligned}
\mathbb{P}\left[|\hat{\sigma}-\sigma|<\frac{\sigma}{5}\right] &= \mathbb{P}\left[\frac{4}{5}\frac{\hat{\sigma}}{\sigma}<\frac{6}{5}\right] \\
&= \mathbb{P}\left[0{,}64n\frac{n\hat{\sigma}}{\sigma}<1{,}44n\right] \\
&= F_{\chi^2_{n-1}}(1{,}44n)-F_{\chi^2_{n-1}}(0{,}64n)
\end{aligned}
$$

Estime n de manera que

$$\left[1-2\Phi\left(-\frac{\sqrt{n}}{5}\right)\right][F_{\chi^2_{n-1}}(1{,}44n)-F_{\chi^2_{n-1}}(0{,}64n)]\geq\frac{1}{2}.$$

Se resuelve numéricamente, y si $n=21$ se cumple.

14.1. Distribución t

Definición. Sea Y y Z dos variables independientes tal que $Y\sim\chi^2_m$ y $Z\sim N(0,1)$. Si

$$X:=\frac{Z}{\sqrt{\frac{Y}{m}}},$$

tiene una distribución t **de Student** con m grados de libertad. Tiene como densidad

$$f_X(x)=\frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\Gamma\left(\frac{m}{2}\right)}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}},\quad x\in\mathbb{R}.$$

Propiedades:

1) f_X es simétrica.
2) La media de X no existe si $m\leq 1$. Si la media existe, es 0.
3) Las colas de una t de Student son más pesadas que una $N(0,1)$.
4) Si m es entero, los primeros $m-1$ momentos de X existen y no hay momentos de orden superior.

5) Si $m > 2$, $\mathrm{Var}\,(X) = \dfrac{m}{m-2}$.

6) Si $m = 1$, $X \sim$ Cauchy.

7) **Ejercicio**: $f_x(x) \xrightarrow[m\to\infty]{} \Phi(x)$ (sirve como aproximación). La discrepancia de ambas está en la cola y se disipa cuando m es grande.

Recuerde que, por el teorema 8.3.1, $\bar{X}_n$ y $Y = \dfrac{n\hat{\sigma}^2}{\sigma}$ son independientes, con $\bar{X}_n \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ y $Y \sim \chi^2_{n-1}$. Además,

$$Z = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim N(0,1).$$

Sea

$$T = \frac{Z}{\sqrt{\dfrac{Y}{n-1}}} = \frac{\sqrt{n}\dfrac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\dfrac{\frac{n\hat{\sigma}^2}{\sigma^2}}{n-1}}} = \frac{\bar{X}_n - \mu}{\sqrt{\dfrac{\hat{\sigma}}{n-1}}}$$

el cual no depende de σ.

Teorema. Si $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, defina

$$\sigma' = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right]^{\frac{1}{2}}.$$

Entonces

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma'} \sim t_{n-1}$$

Nota. $\sigma' = \left(\dfrac{n}{n-1}\right)^{\frac{1}{2}} \hat{\sigma}$ (si n es grande, $\sigma' = \hat{\sigma}$).

Prueba. Sean

$$S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad Z = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}.$$

Dado que $Y = \dfrac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$, entonces

$$
\begin{aligned}
U = \frac{Z}{\sqrt{\dfrac{Y}{n-1}}} &= \frac{\dfrac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)}{\sqrt{\dfrac{S_n^2}{\sigma^2(n-1)}}} \\
&= \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\dfrac{S_n^2}{n-1}}} \\
&= \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma'} \sim t_{n-1}.
\end{aligned}
$$