# Análisis de datos con Métodos Bayesianos

Instituto Bayesiano para la Investigación en Desarrollo
(BayesGroup.org)

Rolando Gonzales
rgonzales@bayesgroup.org

16 de octubre de 2015

GRUPO BAYES Instituto Bayesiano para la Investigación en Desarrollo

Fundación Bayes

## Contenido del curso

Este es un curso corto e introductorio sobre Análisis de datos con métodos Bayesianos.

- Un poco de historia como introducción al curso.
- Homogeneización.
- Introducción a la inferencia Bayesiana.
- Introducción a la econometría Bayesiana.
- Aplicaciones.

Ejercicios prácticos con datos del anuario estadístico del municipio de La Paz 2013 y otros datos relevantes.

# Inferencia

- Hasta el momento hablamos abstractamente de "eventos", en lugar de datos convencionales.
- La Inferencia implica realizar afirmaciones probabilísticas acerca de cantidades desconocidas utilizando los datos disponibles (y otra información disponible).
- Sea $\theta$ un parámetro que es el objetivo del análisis, y $\mathbf{D}$ los datos que tenemos disponibles. El objetivo de la inferencia es obtener una afirmación probabilística acerca de $\theta$ dados los datos $\mathbf{D}$: $\mathbb{P}(\theta|\mathbf{D})$.

# Inferencia

- La inferencia comienza (usualmente) especificando un modelo paramétrico para los datos a partir de una distribución de probabilidad, $\mathbb{P}(\mathbf{D}|\theta)$.
- Si formalizamos nuestro conocimiento previo sobre $\theta$ con $\mathbb{P}(\theta)$, antes de observar los datos, es posible inferir $\mathbf{D}$: $\mathbb{P}(\theta|\mathbf{D})$ con,

$$\mathbb{P}(\theta|\mathbf{D}) \propto \mathbb{P}(\theta)\mathbb{P}(\mathbf{D}|\theta).$$

  En inferencia bayesiana, $\mathbb{P}(\theta|\mathbf{D})$ se denomina la probabilidad *a posteriori* o posterior, dada la evidencia de los datos contenida en $\mathbb{P}(\mathbf{D}|\theta)$ y el prior, o distribución de probabilidad *a priori*, $\mathbb{P}(\theta)$.

# Estimación

La estimación es el proceso de extraer información acerca del valor de cierto parámetro poblacional a partir de la muestra $x_1, ..., x_n$, utilizando algún estadígrafo como estimador, calculado con los datos $x_1, ..., x_n$.

**Estadígrafo.**

*Considérese que se obtiene una muestra de tamaño $n$ de una población. Un estadígrafo es una función que se obtiene utilizando como datos las observaciones de la muestra $x_1, ..., x_n$.*

# Estimación puntual e intervalica

Los Bayesianos describimos $\mathbb{P}(\theta|\mathbf{D})$ a través estadígrafos de resumen como medias, modas, fractiles, intervalos de credibilidad (probabilidades entre regiones) y gráficos. Toda la información que se necesita para hacer inferencia se obtiene al calcular $\mathbb{P}(\theta|\mathbf{D})$.

Una medida común de estimación puntual es,

$$\mathbb{E}[\theta|\mathbf{D}] = \int_{-\infty}^{\infty} \theta \mathbb{P}(\theta|\mathbf{D}) d\theta.$$

# Estimación puntual e intervalica

El problema con la estimación puntual de los momentos posteriores es que puede no describir apropiadamente las características de la distribución de $\mathbb{P}(\theta|\mathbf{D})$, por lo que es necesario complementar estas medidas puntuales con medidas interválicas, i.e. intervalos de credibilidad Bayesianos.

## Intervalos de credibilidad

*Sea $\mathcal{C}$ un subconjunto del espacio paramétrico $\Theta$, tal que un $100(1-\alpha)\,\%$ intervalo de credibilidad de colas iguales cumple la condición,*

$$1-\alpha = \int_{\mathcal{C}} \mathbb{P}(\theta|\mathbf{D})d\theta.$$

## Estimación puntual e intervalica

Los intervalos de credibilidad no son iguales a los intervalos de confianza frecuentistas (los intervalos de confianza frecuentistas cubren al "verdadero" parámetro a través de una $1-\alpha$ proporción de las replicaciones en el experimento, en promedio).

Los intervalos de credibilidad se generalizan a conjuntos de credibilidad para distribuciones multidimensionales.

El conjuto $\mathcal{C}$ puede definirse en diferentes maneras para cubrir diferentes partes de $\Theta$ y aún cumplir la definición de intervalos de credibilidad.

# Estimación puntual e intervalica

Una decisión sencilla y común es escoger $\alpha$ para que la $\frac{\alpha}{2}$ densidad quede igualmente distribuida en las colas derecha e izquierda de la distribución, afuera del intervalo de credibilidad:

$$\frac{\alpha}{2} = \int_0^L \mathbb{P}(\theta|\mathbf{D})d\theta,$$

$$\frac{\alpha}{2} = \int_H^\infty \mathbb{P}(\theta|\mathbf{D})d\theta.$$

# Estimación puntual del primer momento: enfoque frecuentista, máxima verosimilitud

- La distribución de probabilidad de una variable contiene un parámetro $\theta$ que se quiere estimar
- El parámetro puede tomar valores de un conjunto llamado espacio paramétrico $\Theta$
- La información sobre $\theta$ es la proporcionada por una muestra aleatoria de tamaño $n$, $\mathbf{x} = (x_1, ..., x_n)$. Sea $\mathbb{P}(\mathbf{x}|\theta)$ la probabilidad de extraer la muestra $\mathbf{x}$ cuando $\theta$ toma un valor en $\Theta$, i.e. una muestra concreta corresponde a un valor concreto de $\theta$. Supóngase sin embargo que, dada una muestra $\mathbf{x} = (x_1, ..., x_n)$, lo que se requiere es recuperar el valor más plausible de $\theta$ que podría haber generado la muestra, $\mathbb{P}(\mathbf{x}|\theta)$ es inadecuado y es necesario definir un nuevo concepto: la verosomilitud (*likelihood*).

# Estimación puntual del primer momento: enfoque frecuentista, máxima verosimilitud

- La función $\mathcal{L}(\theta|\mathbf{x})$ –denominada función de verosimilitud– es una cantidad matemática que expresa la verosomilitud, plausibilidad, de que el parámetro $\theta$ tome un valor concreto en base a la información contenida en la muestra $\mathbf{x} = (x_1, ..., x_n)$
- En la función de verosimilitud $\mathcal{L}(\theta|\mathbf{x})$, la muestra permanece constante, y $\theta$ varía en el espacio paramétrico $\Theta$. Situación contraria a $\mathbb{P}(\mathbf{x}|\theta)$, ya que en este caso, $\theta$ permanece constante y lo que se compara son los posibles sucesos $\mathbf{x}_1, \mathbf{x}_{2,}...$ para el mismo valor de $\theta$.

# Estimación puntual del primer momento: enfoque frecuentista, máxima verosimilitud

Sea una variable aleatoria $\xi$ con una función de densidad $f(x;\theta)$. Si se toma una muestra de tamaño $n$, $x_1, ..., x_n$, la función de densidad conjunta de la muestra será,

$$f(x_1, ..., x_n; \theta) = f(x_1; \theta) \cdots f(x_n; \theta)$$

esta función de densidad conjunta será la función de verosimilitud muestral,

$$\mathcal{L}(\theta|\mathbf{x}) := f(x_1; \theta) \cdots f(x_n; \theta)$$

La elección del estimador $\widehat{\theta}$ entre los posibles valores que puede tomar el parámetro $\theta$ se sgue el criterio de elegir aquel $\widehat{\theta}$ tal que,

$$\mathcal{L}(\widehat{\theta}|\mathbf{x}) = \max_{\theta \in \Theta} \mathcal{L}(\theta|\mathbf{x})$$

# Estimación puntual del primer momento: enfoque frecuentista, máxima verosimilitud

Ya que en general la función de verosimilitud es complicada, para mayor sencillez se calcula el valor de $\theta$ en su logaritmo,

$$\ln \mathcal{L}(\widehat{\theta}|\mathbf{x}) = \max_{\theta\in\Theta} \ln \mathcal{L}(\theta|\mathbf{x})$$

$$\ln \mathcal{L}(\theta|\mathbf{x}) = \ln f(x_1;\theta) + \cdots + \ln f(x_n;\theta) = \sum\nolimits_{i=1}^{n} \ln f(x_i;\theta)$$

Una vez hallado el logaritmo se deriva respecto a $\theta$,

$$\frac{\partial \ln \mathcal{L}(\theta|\mathbf{x})}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial\theta} = 0$$

La solución $\widehat{\theta} = \widehat{\theta}(\mathbf{x})$, únicamente función de los elementos muestrales, será el estimador máximo-verosímil del parámetro de interés.

# Estimación puntual del primer momento: enfoque frecuentista, máxima verosimilitud

Sean una muestra i.i.d. $x := \{x_t\}_{t=1}^{m} = x_1, ..., x_m$, si $x \sim \mathcal{N}(\mu, \sigma^2)$, la función de verosimilitud $\mathcal{L}\left(\mu, \sigma^2\right)$ será,

$$
\begin{aligned}
\mathcal{L}\left(\mu, \sigma^2\right) &= f(x_1, ..., x_m | \mu, \sigma^2) \\
&= \prod_{i=1}^{m} f(x_i | \mu, \sigma^2) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{m} (x_i - \mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

# Estimación puntual del primer momento: enfoque frecuentista, máxima verosimilitud

$\mathcal{L}\left(\mu,\sigma^2\right)$ puede escribirse en función de $\overline{x}$,

$$\mathcal{L}\left(\mu,\sigma^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{m}\left(x_i-\overline{x}\right)^2 + m\left(\overline{x}-\mu\right)^2}{2\sigma^2}\right).$$

Si se aplica logaritmos a $\mathcal{L}\left(\mu,\sigma^2\right)$ y se maximiza derivando respecto al parámetro de interés $\mu$ e igualando a cero:

$$\frac{\partial}{\partial\mu}\log\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{m}\left(x_i-\overline{x}\right)^2 + m\left(\overline{x}-\mu\right)^2}{2\sigma^2}\right) = 0$$

se tendrá el estimador máximo verosimil,

$$\widehat{\mu} = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

# Estimación puntual del primer momento: enfoque Bayesiano

Sea nuevamente una muestra i.i.d. $x_1, ..., x_m$ con $x \sim \mathcal{N}(\mu, \sigma^2)$. Por simpleza, asúmase que se conoce $\sigma_0^2$ y el parámetro de interés a estimar es $\mu$.

En inferencia Bayesiana, los parámetros no son cantidades fijas, sino variables aleatorias que siguen una distribución de probabilidad.

Si $\mu$ sigue una distribución gaussiana, se tiene el prior,

$$\mu | m, s^2 \sim \mathcal{N}(m, s^2) = (2\pi s^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2s^2}(\mu - m)^2\right].$$

# Estimación puntual del primer momento: enfoque Bayesiano

Por el Teorema de Bayes,

$$\begin{aligned}
\mathbb{P}(\mu|x) &\propto \mathbb{P}(\mu)\mathbb{P}(x|\mu) \\
&\propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma_0^2}(x_i-\mu)^2\right] \exp\left[-\frac{1}{2s^2}(\mu-m)^2\right] \\
&= \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{s^2}(\mu-m)^2\right)\right].
\end{aligned}$$

# Estimación puntual del primer momento: enfoque Bayesiano

Después de expandir los cuadrados y reordenar,

$$\mathbb{P}(\mu|x) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{n}{\sigma_0^2}\right)\left(\mu-\frac{\frac{m}{s^2}+\frac{n\overline{x}}{\sigma_0^2}}{\frac{1}{s^2}+\frac{n}{\sigma_0^2}}\right)^2\right].$$

Por lo que el estimador Bayesiano posterior de $\mu$ es,

$$\hat{\mu}=\frac{\frac{m}{s^2}+\frac{n\overline{x}}{\sigma_0^2}}{\frac{1}{s^2}+\frac{n}{\sigma_0^2}}.$$

# Importancia de los datos vs. la información prior

Reordenando el estimador Bayesiano,

$$\begin{aligned}
\hat{\mu} &= \frac{\frac{1}{s^2}m + \frac{n}{\sigma_0^2}\overline{x}}{\frac{1}{s^2} + \frac{n}{\sigma_0^2}} \\
&= \left(\frac{\frac{1}{s^2}}{\frac{1}{s^2} + \frac{n}{\sigma_0^2}}\right) m + \left(\frac{\frac{n}{\sigma_0^2}}{\frac{1}{s^2} + \frac{n}{\sigma_0^2}}\right) \overline{x}
\end{aligned}$$

Con un cambio de variable,

$$\omega_p = \left(\frac{\frac{1}{s^2}}{\frac{1}{s^2} + \frac{n}{\sigma_0^2}}\right), \qquad \omega_d = \left(\frac{\frac{n}{\sigma_0^2}}{\frac{1}{s^2} + \frac{n}{\sigma_0^2}}\right),$$

# Importancia de los datos vs. la información prior

El estimador Bayesiano puede expresarse como,

$$\hat{\mu} = \omega_p m + \omega_d \overline{x},$$

siendo $\omega_p$ el peso (importancia) de la información prior y $\omega_d$ el peso (importancia) de los datos en el estimador.

Nótese que si no se tiene mucha confianza en la información prior, puede definirse un prior no-informativo (i.e. un prior difuso, dominado por la función de verosimilitud) si $s^2 \to \infty^+$.
En este caso $s^2 \to \infty^+$,

$$\hat{\mu} \approx \overline{x}.$$

# ¿Qué sucede cuando $n \to \infty$ ?

Multipliquemos la anterior expresión por $\frac{\sigma_0^2}{n}$, entonces,

$$\hat{\mu} = \frac{\frac{m\sigma_0^2}{ns^2} + \overline{x}}{\frac{\sigma_0^2}{ns^2} + 1}.$$

Si $n \to \infty$ (cuando la muestra de datos es grande),

$$\lim_{n\to\infty} \hat{\mu} = \lim_{n\to\infty} \frac{\frac{m\sigma_0^2}{ns^2} + \overline{x}}{\frac{\sigma_0^2}{ns^2} + 1} = \overline{x}.$$

Este resultado se denomina a veces el Teorema del Límite Central Bayesiano.