Estimación de parámetros

Reconocimiento de Patrones

Reveles Martínez, R.

Universidad Autónoma de Zacatecas,
Unidad Académica de Ingeniería Eléctrica,
Maestría en Ciencias del Procesamiento de la Información

17 de Abril del 2021

Contenido

Distribución de probabilidad

La distribución de probabilidad de una variable aleatoria (X) es una lista de todos los valores posibles de X, junto con sus probabilidades correspondientes.

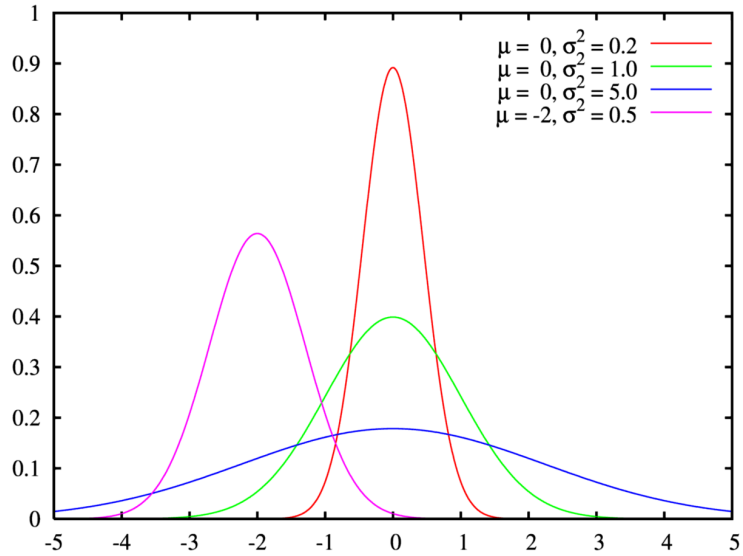

Figura 1: Distribuciones de probabilidad

La distribución de probabilidad de variables continuas se representan mediante una notación funcional f(x) y se les conoce como **funciones de densidad.**

Función de densidad de probabilidad

Definición

Una función de densidad de probabilidad (p.d.f) describe el comportamiento de una variable aleatoria al proporcionar las probabilidades de ocurrencia de ciertos valores. Estas pdf's tienen formas muy complejas y no hay expresiones analíticas que las expliquen. En otras ocasiones, resulta conveniente aproximarlas por una expresión cerrada. Esto es conveniente para propósitos de derivar propiedades y caracterizar un conjunto de datos. Hay que seleccionar apropiadamente los parámetros de las pdf's de tal forma que describan adecuadamente los datos.

Estimación de parámetros

La estimación de parámetros consiste en cómo utilizar las distribuciones de probabilidad para modelar incertidumbre debido al proceso de observación y la incertidumbre en los parámetros de nuestros predictores.

Estimación de parámetros

Sea $\boldsymbol{x} = \boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{...}, \boldsymbol{x_n}$ n variables aleatorias que representan observaciones $\boldsymbol{x_1} = x_1$,$\boldsymbol{x_2} = x_2$,..., $\boldsymbol{x_n} = x_n$, y supongamos que estamos interesados en estimar un parámetro no aleatorio desconocido θ, con base en estos datos. Por ejemplo suponga que estas variables aleatorias tienen una distribución normal común $N(\mu, \sigma^2)$ donde se desconoce μ. Supongamos que los datos medidos tienen algo que ver con este parámetro $\boldsymbol{\theta}$. Más precisamente, asumimos que la función de densidad de probabilidad conjunta de $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{...}, \boldsymbol{x_n}$ dada por $f_x(x_1, x_2, ..., x_n; \theta)$ depende de θ. Estimamos el parámetro desconocido θ en función de estas muestras de datos. El problema de la estimación es seleccionar un estadístico unidimensional $\boldsymbol{T}(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{...}, \boldsymbol{x_n})$ que mejor estime θ en el sentido óptimo. El método de **Máxima Verosimilitud (MLE)** se utiliza a menudo para esto.

Estimación de parámetros

Estimación de Máxima Verosimilitud

La idea detrás de la estimación de máxima verosimilitud (MLE) es definir una función de parámetros que nos permita encontrar un modelo que se ajuste bien a los datos. El problema de estimación se centra en la función de Verosimilitud. Dicha función incluye los datos representados por una variable aleatoria x y para una familia de densidades de probabilidad $p(x|\theta)$ parametrizada por θ.

Método de Máxima Verosimilitud

El principio de MLE asume que un conjunto de datos muestrales es representativo de una población $f_x(x_1, x_2, ..., x_n; \theta)$ y elige el valor para θ que más probablemente causó que ocurrieran los datos observados, es decir, una vez que se dan las observaciones $x_1, x_2, ..., x_n$, $f_x(x_1, x_2, ..., x_n; \theta)$ solo es una función de θ, y el valor de θ que maximiza la p.d.f es el valor más probable para θ y se elige como su estimación MLE $\hat{\theta}_{ML}(\mathrm{x})$. Dado $\boldsymbol{x}_1 = x_1$,$\boldsymbol{x}_2 = x_2$,..., $\boldsymbol{x}_{\boldsymbol{n}} = x_n$, la f.d.p. conjunta $f_x(x_1, x_2, ..., x_n; \theta)$ se define como la función de Verosimilitud, y la estimación de MLE se puede determinar a partir de la ecuación de Verosimilitud como sigue:

$$\mathcal{L}(x_1, x_2, ..., x_n; \theta) \triangleq log\ f_x(x_1, x_2, ..., x_n; \theta) \tag{1}$$

Método de máxima verosimilitud

Si $\mathcal{L}(x_1, x_2, ..., x_n; \theta)$ es diferenciable y el supremo $\hat{\theta}_{ML}$ existe, entonces se debe satisfacer la siguiente ecuación:

$$\frac{\partial \ \log \ f_x(x_1, x_2, ..., x_n; \theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}_{ML}} \tag{2}$$

Ejemplo 1

Estimación de parámetros para una distribución normal

Suponga que un conjunto de datos se distribuye de manera normal. Su función de densidad de probabilidad $\mathcal{L}(x_1, x_2, ..., x_n; \theta)$ es diferenciable y el supremo $\hat{\theta}_{ML}$ existe y cumpliéndose la ecuación 2.

Ejemplo 1

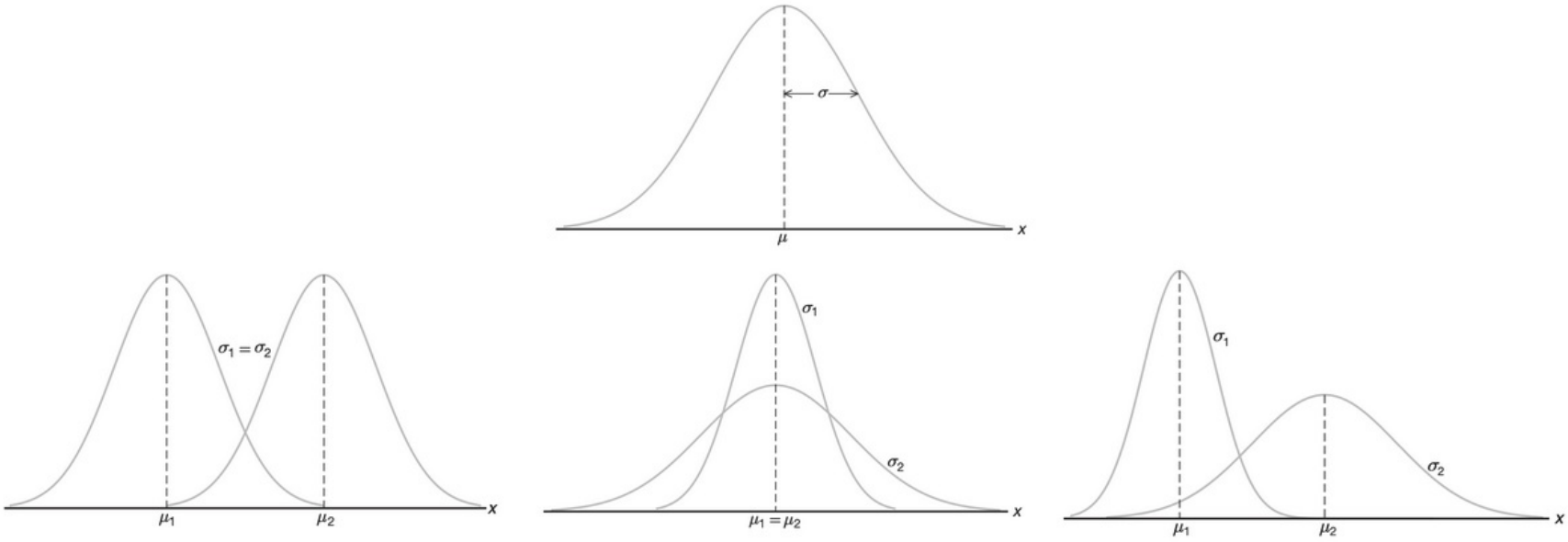

Figura 2: Familia de distribuciones curva normal

FDP Distribución normal

$$P(x|\theta) = P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3}$$

Verosimilitud vs μ

Podemos identificar en el gráfico de verosimilitud dónde la pendiente de la curva es cero en la siguiente figura.

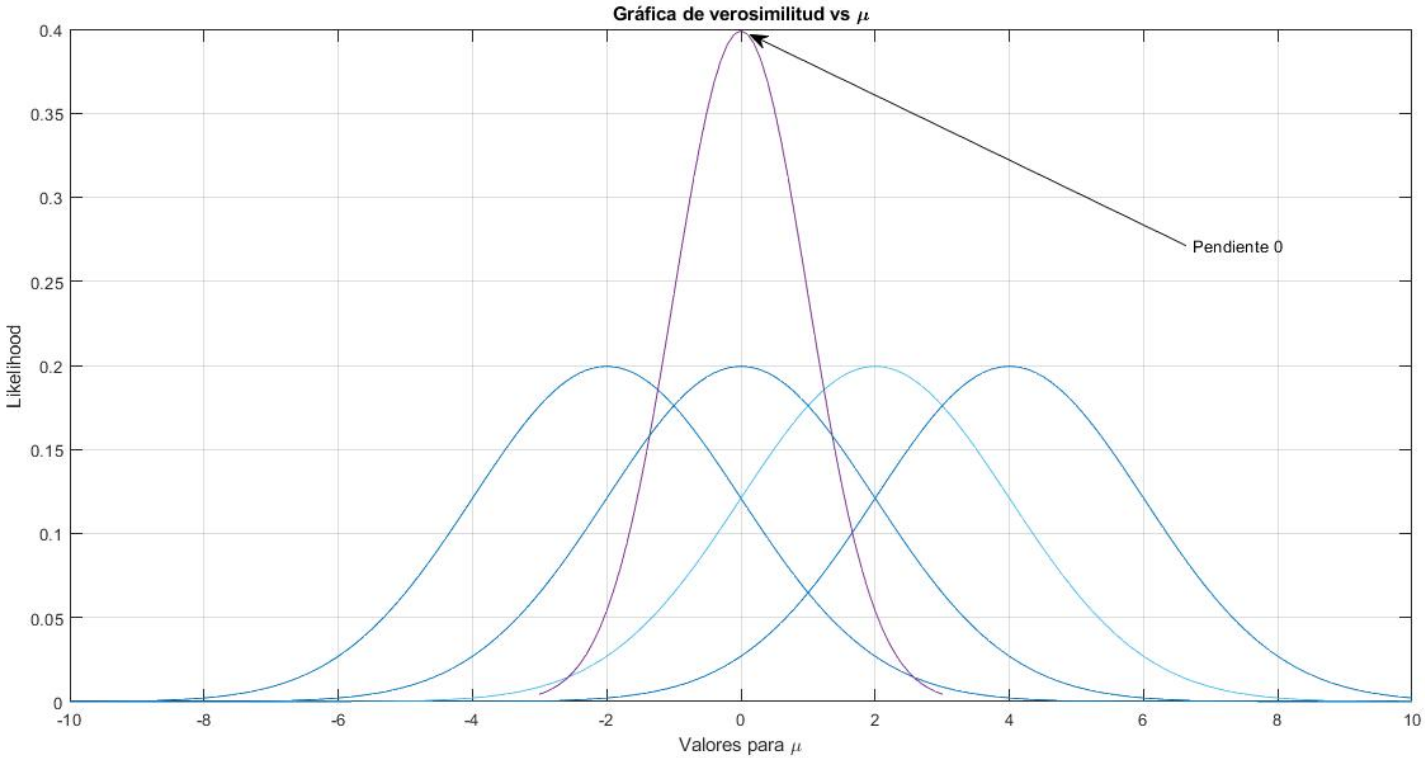

Figura 3: Verosimilitud vs μ

Verosimilitud vs σ

Podemos identificar en el gráfico de verosimilitud dónde la pendiente de la curva es cero en la siguiente figura.

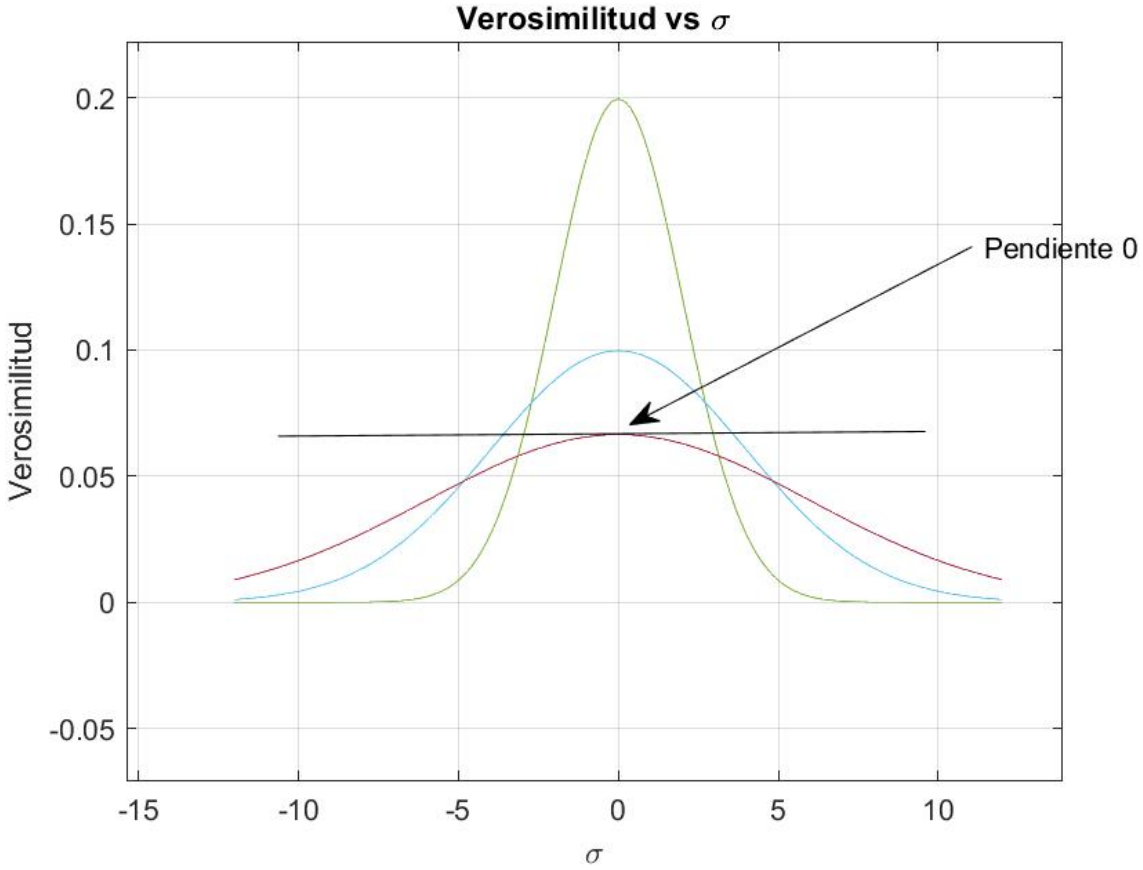

Figura 4: Verosimilitud vs σ

Verosimilitud para n datos

Para un conjunto de datos con n muestras, podemos calcular dichas curvas de verosimilitud de la siguiente manera:

Método MLE para n datos

$$\mathcal{L}(x_1, x_2, ..., x_n; \theta) = \mathcal{L}(x_1, x_2, ..., x_n; \mu, \sigma) = \mathcal{L}(x_1|\mu, \sigma) \times ... \times \mathcal{L}(x_n|\mu, \sigma) \quad (4)$$

Tenemos la función de verosimilitud sin ningún valor especifico para μ y para σ, el cual es igual al producto de las funciones de verosimilitud para las N mediciones individuales.

Máxima Verosimilitud para n datos

$$\mathcal{L}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{5}$$

$$\mathcal{L}(x_1|\mu,\sigma) \times \mathcal{L}(x_2|\mu,\sigma) \times ... \times \mathcal{L}(x_n|\mu,\sigma) \tag{6}$$

$$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_2-\mu)^2}{2\sigma^2}} \times ... \times \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_n-\mu)^2}{2\sigma^2}} \tag{7}$$

Máxima Verosimilitud para n datos

Ahora lo que tenemos que hacer es tomar dos derivadas de esta ecuación para determinar la máxima verosimilitud para μ y para σ.

Una derivada con respecto a μ, cuando tratamos a σ como constante, y podemos encontrar la estimación de máxima verosimilitud para μ al encontrar donde esta derivada es igual a cero. De manera similar para la verosimilitud de σ.

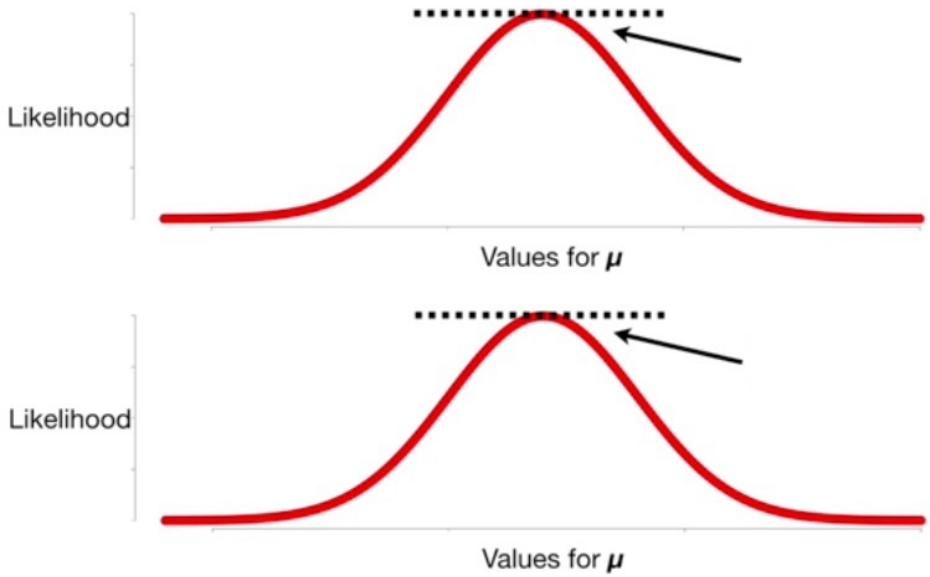

Figura 5: Máxima verosimilitud

Máxima Verosimilitud para n datos

Antes de determinar derivadas, por simplicidad matemática es conveniente, remontarnos a la ecuación 2, para poder determinar las derivadas de manera más fácil. En la función de verosimilitud y en el logaritmo ambas alcanzan los mismos valores para μ y σ.

Máxima Verosimilitud para n datos

$$\ln\left[\mathcal{L}(x|\mu,\sigma)\right] \tag{8}$$

$$ln\left[\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_2-\mu)^2}{2\sigma^2}} \times \ldots \times \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}\right] \tag{9}$$

Máxima Verosimilitud para n datos

Utilizando las leyes de los logaritmos $\ln\ (AB) = \ln\ A + \ln\ B$

$$ln\left[\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\right] + ln\left[\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}\right] + ... + ln\left[\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}\right] \tag{10}$$

Tomemos solo el primer termino de la ecuación 10 y volvamos a aplicar la misma ley del logaritmo del producto.

$$ln\left[\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\right] = ln\left[\frac{1}{\sqrt{2\pi}\sigma}\right] + ln\left[e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\right] \tag{11}$$

Máxima Verosimilitud para n datos

$$ln\left[(2\pi\sigma^2)^{-\frac{1}{2}}\right] + ln\left[e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\right] \tag{12}$$

$$ln\left[(2\pi\sigma^2)^{-\frac{1}{2}}\right] + \frac{-(x_1-\mu)^2}{2\sigma^2} \tag{13}$$

$$-\frac{1}{2}\, ln\left[(2\pi\sigma^2)\right] + \frac{-(x_1-\mu)^2}{2\sigma^2} \tag{14}$$

$$-\frac{1}{2}\, ln\left[2\pi\right] - \frac{1}{2}\, ln\left[\sigma^2\right] - \frac{(x_1-\mu)^2}{2\sigma^2} \tag{15}$$

$$-\frac{1}{2}\, ln\left[2\pi\right] - \frac{2}{2}\, ln\left[\sigma\right] - \frac{(x_1-\mu)^2}{2\sigma^2} \tag{16}$$

Máxima Verosimilitud para n datos

$$-\frac{1}{2}\ ln\left[2\pi\right]-\ \ ln\left[\sigma\right]-\frac{(x_1-\mu)^2}{2\sigma^2} \tag{17}$$

Regresemos a la ecuación 10 y sustituyendo la ecuación 17 tenemos lo siguiente:

$$-\frac{1}{2}\ ln\left[2\pi\right]-\ \ ln\left[\sigma\right]-\frac{(x_1-\mu)^2}{2\sigma^2}-\ldots--\frac{1}{2}\ ln\left[2\pi\right]-\ \ ln\left[\sigma\right]-\frac{(x_n-\mu)^2}{2\sigma^2} \tag{18}$$

Máxima Verosimilitud para n datos

Para simplificar la ecuación 18, recordamos que tenemos n observaciones o datos y obtendremos el algoritmo de verosimilitud al cual derivaremos para estimar los parámetros μ y σ.

$$-\frac{n}{2}\, ln\,[2\pi] - \quad n\; ln\,[\sigma] - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_n-\mu)^2}{2\sigma^2} \tag{19}$$

Derivación del algoritmo de verosimilitud

Dado que tenemos que

$$\frac{\partial}{\partial \mu} \ln \left[\mathcal{L}(x_1, ..., x_n | \theta) \right] = 0 \tag{20}$$

Empezaremos tomando la derivada con respecto μ. Esta derivada es la función de pendiente para el logaritmo de la curva de probabilidad y la usaremos para encontrar el valor pico.

Retomando la ecuación 19

$$-\frac{n}{2}\ ln\left[2\pi\right] - \quad n\ ln\left[\sigma\right] - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_n-\mu)^2}{2\sigma^2}$$

Observamos que los dos primeros términos no contienen μ, por lo que sus derivadas son cero. Para los términos restantes, tomaremos solo un termino subsecuente.

$$\frac{\partial}{\partial\mu}\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right]$$

Aplicando regla de la cadena para derivar mediante el uso de la fórmula $\frac{d}{dx}u^n = nu^{n-1}\frac{du}{dx}$ así como $\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{v\,\frac{du}{dx} - u\,\frac{du}{dx}}{v^2}$.

$$\frac{\partial}{\partial \mu}\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] =$$

$$n = 2; n-1 = 1;$$

$$u = -(x-\mu)^2; \frac{du}{d\mu} = -2(x-\mu)(-1) = 2(x-\mu);$$

$$v = 2\sigma^2; \ \frac{dv}{d\mu} = 0;$$

$$\frac{\partial}{\partial \mu}\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] = \frac{2\sigma^2 2(x-\mu) + (x-\mu)^2(0)}{(2\sigma^2)^2}$$

$$\frac{\partial}{\partial \mu}\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right]=\frac{4\sigma^2(x-\mu)}{(4\sigma^4)}=\frac{(x-\mu)}{\sigma^2}$$

En conclusión tenemos entonces que

$$\frac{\partial}{\partial \mu}\ln\left[\mathcal{L}(x_1,...,x_n|\theta)\right]=0+0+\frac{(x_1-\mu)}{\sigma^2}+...+\frac{(x_n-\mu)}{\sigma^2}$$

Reagrupando las x_i y factorizando el termino común σ^2

$$\frac{\partial}{\partial \mu} ln\left[\mathcal{L}(x_1,...,x_n|\theta)\right]=\frac{1}{\sigma^2}\left[(x_1+x_2+...+x_n)-n\mu\right]$$

Una vez encontrada la ecuación de la derivada, determinaremos o estimaremos la máxima verosimilitud para μ. Tenemos que resolver donde la derivada es 0, por que la pendiente es 0 en el pico de la curva.

$$\frac{\partial}{\partial \mu} ln\left[\mathcal{L}(x_1, ..., x_n|\theta)\right] = \frac{1}{\sigma^2}\left[(x_1 + x_2 + ... + x_n) - n\mu\right] = 0$$

$$\frac{1}{\sigma^2}\left[(x_1 + x_2 + ... + x_n) - n\mu\right] = 0$$

$$\left[(x_1 + x_2 + ... + x_n) - n\mu\right] = 0$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\sum_{i=1}^{n} x_i = n\mu$$

$$n$$

Concluimos que la estimación de máxima verosimilitud para μ es la media de las observaciones.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Para estimar la máxima verosimilitud para σ debemos emplear la misma metodología. Ahora derivar la ecuación 19 respecto a σ, igualarla a cero y despejar.

$$\frac{\partial}{\partial \sigma} \ln\left[\mathcal{L}(x_1, ..., x_n|\theta)\right] = 0 \tag{21}$$

$$\frac{\partial}{\partial \mu} ln\left[L(x_1, ..., x_n|\theta)\right] = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\left[(x_1 - \mu)^2 + ... + (x_n - \mu)^2\right]$$

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\left[(x_1 - \mu)^2 + ... + (x_n - \mu)^2\right] = 0$$

$$\frac{1}{\sigma}\left[-n + \frac{1}{\sigma^2}\left[(x_1 - \mu)^2 + ... + (x_n - \mu)^2\right]\right] = 0$$

$$n = \frac{1}{\sigma^2}\left[(x_1 - \mu)^2 + ... + (x_n - \mu)^2\right]$$

$$\sigma^2 = \frac{1}{n}\left[(x_1 - \mu)^2 + ... + (x_n - \mu)^2\right]$$

$$\sigma = \sqrt{\frac{1}{n}\left[(x_1 - \mu)^2 + ... + (x_n - \mu)^2\right]}$$

Concluimos que la estimación de máxima verosimilitud es la desviación estándar de las mediciones, y este parámetro determina el ancho de la curva de distribución.

Bibliografía

Walpole R., Myers R., Myers S., Ye K. Probabilidad y Estadística para ingeniería y ciencias. Novena Edición. Pearson, Prentice Hall. 2012