Inferencia Estadística

Maestría en Análisis Estadístico y Computación

Capítulo 3: Distribuciones muestrales y métodos de estimación

---

Dra. L. Leticia Ramírez Ramírez

Enero–Mayo, 2020

Contenido

Agradecimientos

En forma de agradecimiento, se enlistan personas que han contribuido de una u otra forma en la construcción de todo el material utilizado, a través de los años:

- Graciela González Farías
- Ulises Márquez
- Víctor Muñiz
- Juan Antonio López
- Sigfrido Iglesias González
- Rodrigo Macías Paéz
- Edgar Jiménez
- Todos los estudiantes que han colaborado con sugerencias y comentarios sobre estas notas.

Estas notas son de uso exclusivo para enseñanza y no pretende la sustitución de los textos y artículos en la bibliografía.

Distribuciones Muestrales

Introducción

Distribución Muestral

Obtener información muestral y reconstruir variables de interés es el objetivo de la inferencia. Para ello definimos lo que es un **estadístico de prueba**

$$Y = h(X_1, X_2, \ldots, X_n),$$

y notamos que depende **sólo** de la muestra y **no** de los parámetros.

Distribución Muestral

Notemos además que: Y el estadístico, es una función de variables aleatorias, por lo que en sí mismo es una variable aleatoria y con los métodos de la sección anterior, sería posible establecer su distribución en forma exacta. A ésta se le llama Distribución Muestral

Ahora, más que obtener cualquier función de la muestra, estamos interesados en aquéllas que nos digan algo "inteligente" sobre los parámetros u otras características de la población.

Definición: Dada una población que tiene parámetro θ (notación general) y $X_1, X_2, \ldots, X_n$ una muestra aleatoria de esta población, entonces un estimador para θ es un estadístico que de alguna manera nos da valores próximos al valor real de θ.

Distribución Muestral

Ejemplos:

- En cualquier población $\overline{X}$ es un estimador de la media de la población.
- En cualquier población $\tilde{X}$ (mediana) es también un estimador de la media de la población.
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ es un estimador de σ^2.
- $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ es otro estimador para σ^2.
- En una población distribuida Uniforme en el intervalo (a, b),
  $\hat{a}_0 = X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ es un estimador para a y
  $\hat{b}_0 = X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$, es un estimador para b.

Distribución Muestral

Ejemplo. (Sólo para fijar ideas): Supongamos que el modelo poblaciónal está dado por:

x	0	1	2	3
f(x)	.2	.4	.3	.1

a) Considera una muestra de tamaño 2: X_1, X_2 (todas las posibles parejas que se pueden formar a partir de los 4 valores que toma la variable X). Construir la distribución de la media muestral:

$$\overline{X} = \frac{X_1 + X_2}{2}$$

Para ello, comenzaremos formando la distribución conjunta de X_1 y X_2.

Distribución Muestral

Notando que son independientes, construir la conjunta es sólo tomar el producto de los valores marginales.

x_2 \ x_1	0	1	2	3	suma
0	.04	.08	.06	.02	.2
1	.08	.16	.12	.04	.4
2	.06	.12	.09	.03	.3
3	.02	.04	.03	.01	.1
suma	.2	.4	.3	.1	1

Distribución Muestral

Lo que falta es calcular todos los posibles valores que se pueden obtener de medias muestrales, basados en las 16 parejas factibles, esto quiere decir que cuando uno realmente va y toma la muestra y calcula el valor medio de las dos observaciones entonces pueden darse los siguientes valores:

valor de $\overline{x}$	0	.5	1	1.5	2	2.5	3	Total
probabilidad	0.04	0.16	0.28	0.28	0.17	0.06	0.01	1

Distribución muestral de $\overline{X}$

No es difícil ver de la función de probabilidad $f(x)$ que la media y varianza poblaciónales son $\mu = 1.3, \sigma^2 = 0.81$, y de la tabla anterior podemos calcular cuáles serían estas mismas características para la distribución de $\overline{X}$:

$$E(\overline{X}) = 0(0.04) + .5(0.16) + 1(0.28) + 1.5(.28) + 2(0.17) + 2.5(0.06) + 3(0.01) = 1.3 = \mu$$

y

$$V(\overline{X}) = E(\overline{X})^2 - [E(\overline{X})]^2 = 0^2(0.04) + .5^2(0.16) + 1^2(0.28) + 1.5^2(0.28) + \\ + 2^2(0.17) + (2.5)^2(0.06) + 3^2(0.01) - (1.3)^2 = .405 = \frac{0.81}{2} = \frac{\sigma^2}{n}$$

Distribución Muestral

En general para conocer los valores medios o de varianza de $\overline{X}$ o de S^2, o cualquier otro estimador que se exprese en términos de sumas o sumas de cuadrados, no es necesario conocer su distribución, sencillamente utilizamos propiedades de valores esperados

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu$$
$$= \frac{n\mu}{n} = \mu \quad .$$

(Por el concepto de muestra aleatoria, misma distribución).

Distribución Muestral

Además,

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

(por el concepto de muestra aleatoria: independencia).

Estos resultados son válidos sin importar si la variable es discreta o continua (i.e. para cualquier distribución muestreada).

Pero la distribución muestral dependerá del comportamiento de donde se obtiene la muestra.

Distribución del Estimador de la Media

Nota que en este sentido $\overline{X}$ es un estimador natural de μ, dado que en promedio $\overline{X}$ toma ese valor (μ), y la variabilidad asociada con los valores de $\overline{X}$ se puede controlar mediante el tamaño de muestra. Esto es, entre más grande sea n, los valores de $\overline{X}$ se agrupan más alrededor de su media, que es μ.

Ya vimos, como un ejemplo en el capítulo anterior, que si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria tomada de una población Normal(μ ,σ) entonces:

$$\overline{X} \sim \mathit{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

Aquí se hereda la distribución, aunque los parámetros hay que ajustarlos.

Distribución del Estimador de la Media

Cuando, por ejemplo, se muestrea de una Poisson o de una Gamma vimos que no era sencillo establecer la distribución de $\overline{X}$

Un resultado que es de capital importancia, debido a la simplicidad que se deriva de éste, en cuanto al establecimiento de distribuciones muestrales es "El Teorema Central de Límite". Existen muchas versiones, la que verás aquí, es la más simple de todas ellas.

Distribución del Estimador de la Media

Teorema (Teorema del Central del Límite -TCL-)

Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una población arbitraria que tiene media μ y varianza σ^2 (y su función generatriz de momentos es $M_X(t)$), entonces la distribución límite cuando $n \to \infty$ de

$$Z_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma}$$

es la distribución normal estándar.

Por distribución límite debemos entender que sólo es un comportamiento aproximado, esto es, calcular probabilidades con la distribución muestral exacta o con la normal estándar, bajo muestras grandes, nos dará cantidades muy similares entre sí.

Distribución del Estimador de la Media

Nota: La población no cambia de distribución, es el comportamiento de sus valores medios el que puede ser modelado en forma aproximada por una Normal Estándar.

¿Qué tan buena es la aproximación? Esto depende de dos factores entrelazados:

i) La forma de la distribución de donde se obtiene la muestra (básicamente, qué tan asimétrica es) y,

ii) el tamaño de la muestra.

Entrelazados porque entre más asimétrica sea la distribución original, los valores medios obtenidos de muestras de esas poblaciónes, más tardarán en comportarse como una Normal. Aquí, "más" significa: más grande debe ser la muestra de donde calculemos el valor medio.

Distribución del Estimador de la Media

Las siguientes gráficas ilustran el razonamiento anterior:

- Se simularon 10,000 valores de un modelo dado.
- Se tomaron 1000 muestras de tamaños $n = 2, 5, 10, 30$ y 60 cada una.
- Se calcularon los valores medios en cada una de las 1000 muestras para cada uno de los diferentes casos (n's).
- Se graficaron los histogramas para visualizar la distribución muestral (estandarizando los valores de cada $\bar{x}$, recordemos que nosotros sabemos cuál es la media y varianza poblaciónal para estas simulaciones).

Distribución del Estimador de la Media

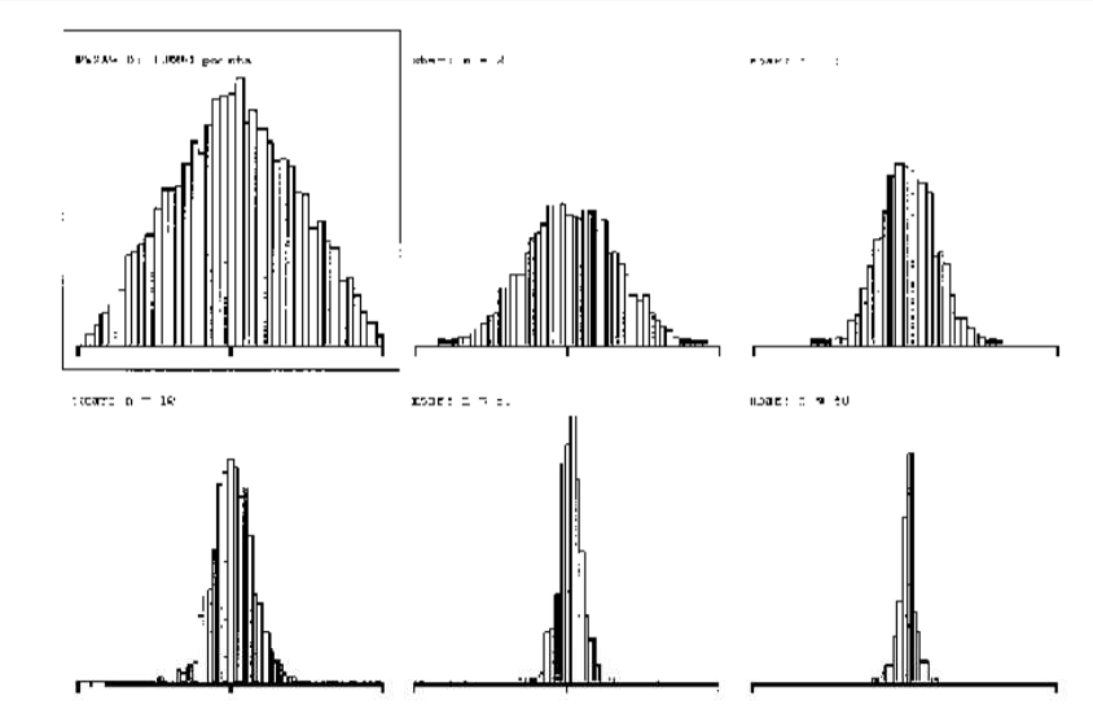


Figure 1: Una población con distribución triangular. Esta distribución es simétrica y "medio parecida" a una Normal.

Distribución del Estimador de la Media

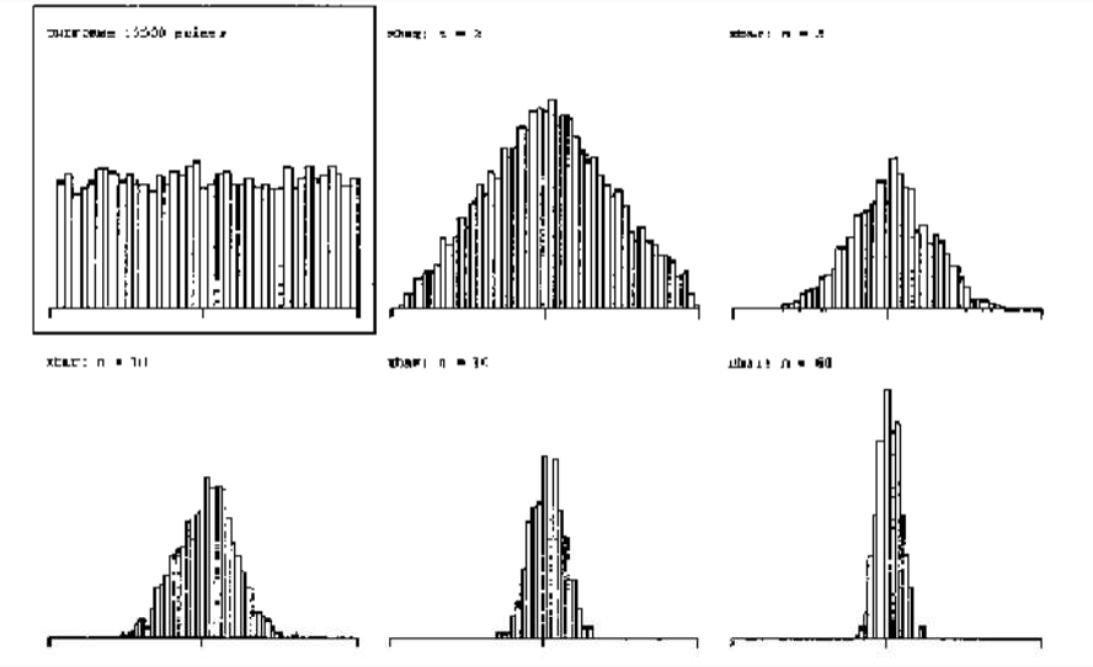


Figure 2: Distribución Uniforme. Esta distribución es simétrica, pero no se parece al comportamiento de una Normal pues las colas de la distribución no "decaen".

Distribución del Estimador de la Media

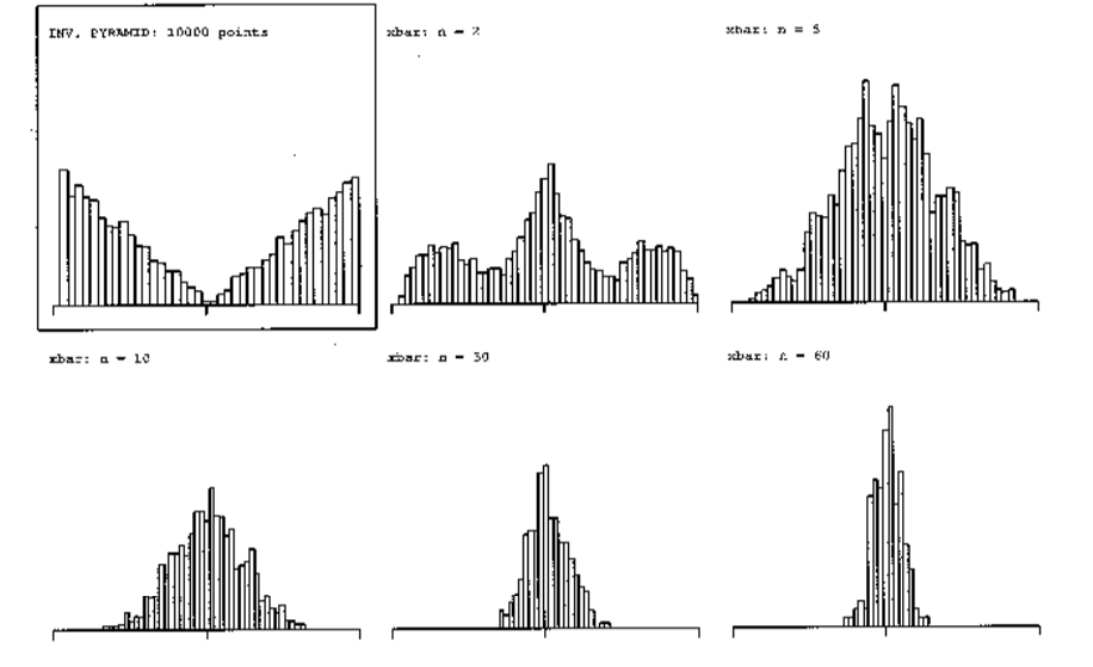


Figure 3: Distribución Triangular inversa. Esta distribución es simétrica pero su comportamiento en el centro y colas de la distribución son los opuestos al de una Normal.

Distribución del Estimador de la Media

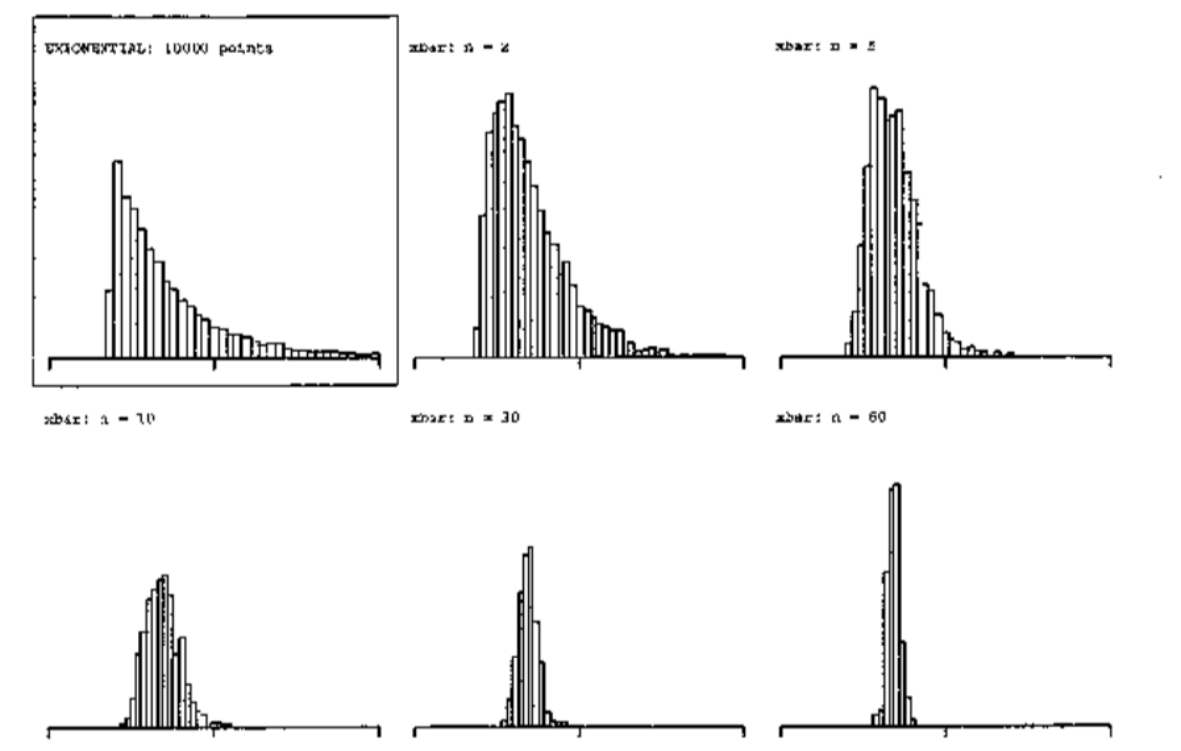


Figure 4: Distribución Exponencial. Esta distribución no es simétrica, ni se parece a la normal.

Distribución del Estimador de la Media

¿Qué conclusiones puedes obtener de estas gráficas?

¿La afirmación: "generalmente se considera como valor $n \geq 30$ que da una "buena aproximación" tiene más sentido? Pues sí, pero no siempre es necesario muestrear tanto para obtener buenos resultados.

Ahora daremos una demostración analítica del teorema, para satisfacer a los fans de la formalidad matemática (opcional)

Demostración: Se utiliza el método de la generatriz de momentos. Probaremos que $M_Z(t) \to e^{t^2/2}$ ya que $e^{t^2/2}$ es la generatriz de momentos de una variable aleatoria con distribución normal estándar.

Distribución del Estimador de la Media *

$$
\begin{aligned}
M_Z(t) = E(e^{Zt}) = E(e^{t(\frac{\sqrt{n}}{\sigma}(\overline{X}-\mu))}) = E(e^{t\frac{\sqrt{n}\bar{X}}{\sigma}} \cdot e^{-t\frac{\sqrt{n}\mu}{\sigma}}) \\
= e^{-t\frac{\sqrt{n}\mu}{\sigma}} E(e^{t\frac{\sqrt{n}\bar{X}}{\sigma}}) = e^{-t\frac{\sqrt{n}\mu}{\sigma}} E(e^{\frac{t}{\sqrt{n}\sigma}\sum_{i=1}^{n} X_i}) \\
= e^{-t\frac{\sqrt{n}\mu}{\sigma}} E(e^{\frac{t}{\sqrt{n}\sigma}X_1} \cdot e^{\frac{t}{\sqrt{n}\sigma}X_2} \cdots e^{\frac{t}{\sqrt{n}\sigma}X_n}),
\end{aligned}
$$

dado que las X_i son independientes

$$
M_Z(t) = e^{-t\frac{\sqrt{n}\mu}{\sigma}} \left[M_{X_1}(\frac{t}{\sigma\sqrt{n}}) \cdot M_{X_2}(\frac{t}{\sigma\sqrt{n}}) \cdots M_{X_n}(\frac{t}{\sigma\sqrt{n}}) \right]
$$

dado que son distribuciones idénticas

Distribución del Estimador de la Media *

$$M_Z(t) = e^{-t\frac{\sqrt{n}\mu}{\sigma}} \left[M_X(\frac{t}{\sigma\sqrt{n}}) \cdot M_X(\frac{t}{\sigma\sqrt{n}}) \cdots M_X(\frac{t}{\sigma\sqrt{n}}) \right]$$
$$= e^{-t\frac{\sqrt{n}\mu}{\sigma}} \left[M_X(\frac{t}{\sigma\sqrt{n}}) \right]^n$$

Ahora expandemos $M_X(\frac{t}{\sigma\sqrt{n}})$ en series de potencias alrededor de $t = 0$ (Taylor):

$$M_X\left(\frac{t}{\sigma\sqrt{n}}\right) = M_X(0) + M_X'(0)\left(\frac{t}{\sigma\sqrt{n}}\right) + \frac{M_X''(0)}{2!}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \frac{M_X'''(0)}{3!}\left(\frac{t}{\sigma\sqrt{n}}\right)^3 + \cdots$$

$$M_X(t) = E(e^{tx})$$
$$M_X(0) = E(e^0) = 1$$

Distribución del Estimador de la Media *

$$\Longrightarrow M_X\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \mu_1'\frac{t}{\sigma\sqrt{n}} + \mu_2'\frac{t^2}{2\sigma^2 n} + \mu_3'\frac{t^3}{6\sigma^3 n^{3/2}} + \cdots$$

Entonces la generatriz de Z esta dada por:

$$M_Z(t) = e^{-t\frac{\sqrt{n}\mu}{\sigma}}\left[1 + \mu_1'\frac{t}{\sigma\sqrt{n}} + \mu_2'\frac{t^2}{2\sigma^2 n} + \mu_3'\frac{t^3}{6\sigma^3 n^{3/2}} + \cdots\right]^n$$

y tomando logaritmos:

$$\ln M_Z(t) = -\frac{t\sqrt{n}}{\sigma}\mu + n\ln\left[1 + \mu_1'\frac{t}{\sigma\sqrt{n}} + \mu_2'\frac{t^2}{2\sigma^2 n} + \mu_3'\frac{t^3}{6\sigma^3 n^{3/2}} + \cdots\right]$$

Distribución del Estimador de la Media *

Sabiendo que $\ln(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \frac{u^4}{4} + \cdots$

$$\begin{aligned}
\ln M_Z(t) =& \frac{t\sqrt{n}}{\sigma}\mu + n\left\{\left[\mu_1'\frac{t}{\sigma\sqrt{n}} + \mu_2'\frac{t^2}{2\sigma^2 n} + \cdots\right] - \frac{1}{2}\left[\mu_1'\frac{t}{\sigma\sqrt{n}} + \mu_2'\frac{t^2}{2\sigma^2 n} + \cdots\right]^2 + \right. \\
& \left. + \frac{1}{3}\left[\mu_1'\frac{t}{\sigma\sqrt{n}} + \mu_2'\frac{t^2}{2\sigma^2 n} + \cdots\right]^3 - \cdots\right\} \\
=& t\left[\frac{-\mu\sqrt{n}}{\sigma} + \frac{n\mu_1'}{\sigma\sqrt{n}}\right] + t^2\left[\frac{n\mu_2'}{2\sigma^2 n} - \frac{n(\mu_1')^2}{2\sigma^2 n}\right] \\
& + t^3\left[\frac{n\mu_3'}{6\sigma^3 n^{3/2}} - \frac{2n\mu_1'\mu_2'}{4\sigma^3 n^{3/2}} + \frac{n(\mu_1')^3}{3\sigma^3 n^{3/2}}\right] + \cdots
\end{aligned}$$

Distribución del Estimador de la Media *

como $\mu_1' = E(X) = \mu$ y $\mu_2' - \mu_1^2 = E(X^2) - [E(X)]^2 = V(X) = \sigma^2$

$$\Rightarrow \ln M_Z(t) = \frac{t^2}{2} + \frac{t^3}{\sigma^3\sqrt{n}}\left[\frac{\mu_3'}{6} - \frac{\mu_1'\mu_2'}{2} + \frac{(\mu_1')^3}{3}\right] + \cdots$$

Cuando $n \to \infty$, en la expresión anterior obtenemos que:

$$\lim_{n\to+\infty} \ln M_Z(t) = \frac{t^2}{2}$$

$$\ln[\lim_{n\to+\infty} M_Z(t)] = \frac{t^2}{2}$$

$$\therefore \lim_{n\to+\infty} M_Z(t) = e^{t^2/2}$$

Distribución del Estimador de la Media

Algunos casos particulares del TCL que son de importancia práctica:

1. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población *Bernoulli*(p), i.e. cada X_i puede tomar solo valores $\{0, 1\}$. Recordemos que para cada X_i, $E(X_i) = p$ y $V(X_i) = pq$. La cantidad

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\#\text{ número de éxitos en la muestra}}{\#\text{total de observaciones}}$$

es la proporción de éxitos en muestra (%) $\overline{X} = \hat{p}$. Además

$$Y = \sum_{i=1}^{n} X_i \sim \textit{Binomial}(n, p)$$

ya que las X_i son una muestra aleatoria i.i.d.

Distribución del Estimador de la Media

El TCL implica

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

Haciendo álgebra llegamos a que

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{n(\overline{X} - \mu)}{n\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

$$\Rightarrow \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim N(0,1), \quad \text{cuando } n \to \infty$$

Distribución del Estimador de la Media

De aquí, haciendo los ajustes para la población Bernoulli, se llega a que

$$Z = \frac{\hat{p} - p}{\sqrt{pq}/\sqrt{n}} = \frac{Y - np}{\sqrt{npq}} \sim N(0,1) \quad \text{cuando } n \to \infty.$$

Este es el denominado Teorema de De-Moivre-Laplace.

Se pueden calcular probabilidades binomiales a través de aproximarlas con valores de probabilidad obtenidos de la distribución Normal. Que tan buena es la aproximación de nuevo depende de lo asimétrica que sea la Binomial considerada, esto es, depende del valor de p: si p es pequeña o muy grande, deberíamos tener muestras relativamente grandes para contrarrestar el efecto del sesgo.

Estudiar esta aproximación, así como la idea de en el libro de Factor de Corrección por continuidad.

Distribución del Estimador de la Media

2. Una aproximación semejante se puede encontrar para una muestra de tamaño n obtenida de una población Poisson. La razón del porque funciona se encuentra en el hecho de que, dado un proceso Poisson (Y) éste se puede subdividir en partes iguales y en cada subdivisión se auto define de nuevo un Proceso Poisson con λ restringido al subintervalo. Así, Y se puede expresar como una suma de variables Poisson i.i.d., y sobre éstas se aplica el TCL.

Distribución del Estimador de la Media

3. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con media μ_1 y varianza σ_1^2. Además, sea $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria de otra población con media μ_2 y varianza σ_2^2, tal que las muestras sean independientes entre sí.

   El TCL implica que:

$$\frac{\overline{X} - \mu_1}{\frac{\sigma_1}{\sqrt{n}}} \to N(0,1) \quad \text{cuando } n \to \infty$$

   y

$$\frac{\overline{Y} - \mu_2}{\frac{\sigma_2}{\sqrt{n}}} \to N(0,1) \quad \text{cuando } n \to \infty.$$

Distribución del Estimador de la Media

En otras palabras $\overline{X}$ y $\overline{Y}$ se pueden aproximar como

$$\overline{X}_{n_1} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \qquad \overline{Y}_{n_2} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

cuando n_1 y n_2 es grande. Esto implica que, bajo las mismas premisas,

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

y

$$\overline{X} + \overline{Y} \sim N\left(\mu_1 + \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Distribución del Estimador de la Media

Luego, cuando n_1 y n_2 son grandes, tenemos la siguiente aproximación

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

Se puede hacer algo similar para $\overline{X} + \overline{Y}$.

Distribución del Estimador de la Media

- Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una población Bernoulli(θ_1) y $Y_1, Y_2, \ldots, Y_n$ es una muestra de otra población Bernoulli(θ_2), tenemos los siguientes estimadores

$$\overline{X} = \hat{\theta}_1, \qquad \bar{Y} = \hat{\theta}_2.$$

Aplicando lo hecho antes llegamos a que

$$\frac{\hat{\theta}_1 - \hat{\theta}_2 - (\theta_1 - \theta_2)}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} \sim N(0,1), \qquad \text{cuando } n_1 \to \infty,\ n_2 \to \infty.$$

Distribución del Estimador de la Media

El TCL nos dice que la distribución de

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

se aproxima a la de una normal para n grande.

Existe un resultado similar cuando lo que interesa es una función de $\overline{X}$, digamos $g(\overline{X})$, método que estudiaremos después.

A continuación aplicaremos el TCL en algunos ejemplos.

Distribución del Estimador de la Media

Ejemplo 1: Treinta componentes electronicos estan conectados de la siguiente forma: Tan pronto como D_1 falle, D_2 empieza a funcionar, y así sucesivamente. Supón que cada componente tiene un tiempo de vida exponencial con parametro $\theta = 10$ horas y que sus tiempos de vida son independientes. Calcular la probabilidad de que el sistema funcione cuando menos 350 horas.

Sea

$$T_1 = \text{Tiempo de vida del componente } D_1$$
$$T_2 = \text{Tiempo de vida del componente } D_2,$$
$$\vdots$$
$$T_{30} = \text{Tiempo de vida del componente } D_{30}.$$

Distribución del Estimador de la Media

Sabemos que $T_i \sim Exp(\theta = 10)$ $i = 1, 2, \ldots, 30$ y que además son independientes entre sí. Definamos T como el tiempo de vida de todo el sistema. Entonces $T = T_1 + T_2 + \ldots + T_{30}$ y por tanto

$$P(T \geq 350) = P\left(\sum_{i=1}^{30} T_i \geq 350\right)$$

$$= P\left(\frac{\sum_{i=1}^{30} T_i}{30} \geq \frac{350}{30}\right) = P(\overline{T} \geq \frac{350}{30}).$$

En este caso tenemos que los primeros momentos están dados por $\mu = \theta$ y$\sigma^2 = \theta^2$. Luego, aplicando el TCL llegamos a que

$$P\left(\frac{\overline{T} - \theta}{\frac{\theta}{\sqrt{n}}} \geq \frac{\frac{350}{30} - 10}{\frac{10}{\sqrt{30}}}\right) \cong P(Z \geq 0.9128)$$

$$= 1 - P(Z < 0.9128) = 1 - 0.8186 = 0.1814.$$

Distribución del Estimador de la Media

Ejemplo 2: Supongamos que $X_1, \ldots, X_{50}$ son i.i.d. $Poisson(\lambda = 0.03)$ y sea $S = \sum_{i=1}^{50} X_i$.

a) Calcular $P(S > 3)$ usando TCL.

$$P(S > 3) = P\left(\sum_{i=1}^{50} X_i > 3\right) = P\left(\frac{\sum_{i=1}^{50} X_i}{50} > \frac{3}{50}\right)$$

$$= P\left(\overline{X} > \frac{3}{50}\right), \quad \mu = \lambda, \quad \sigma^2 = \lambda$$

$$= P\left(\frac{\overline{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} > \frac{\frac{3}{50} - 0.03}{\sqrt{\frac{0.03}{50}}}\right) \cong P(Z > 1.22) = 1 - P(Z < 1.122)$$

$$= 1 - 0.8888 = 0.1112$$

Con corrección por continuidad, con $S > 3.5$ queda 0.0516.

Distribución del Estimador de la Media

b) Calcular $P(S > 3)$ usando la distribucion de la suma de Poisson's.

Tenemos que

$$S \sim Poisson\left(\sum_{i=1}^{50} \lambda_i\right) \Longrightarrow S \sim Poisson(50(0.03)) = Poisson(1.5)$$

Así que

$$\begin{aligned} P(S > 3) &= 1 - P(S \leq 3) \\ &= 1 - [P(S = 0) + P(S = 1) + P(S = 2) + P(S = 3)] \\ &= 0.065. \end{aligned}$$

La aproximación se recomienda más en la medida que λ sea grande (> 5). Esto porque entonces la distribución Poisson es menos asimétrica.

Distribución del Estimador de la Media

Ejercicio 4: La proporcion real de familias en cierta ciudad que viven en casa propia es 0.7. Si se escogen al azar 84 familias de esa ciudad y se les pregunta si viven o no en casa propia, ¿Con qué probabilidad podemos asegurar que el valor que se obtendra de la proporcion muestral caerá entre 0.64 y 0.76?

Denotemos por X_i a la variable de que la i-ésima familia viva en en casa propia (1 ó 0). En este caso tenemos que $X_i \sim Bernoulli(0.7)$. La proporción muestral está dada por

$$\hat{\theta}_{84} = \frac{\sum_{i=1}^{84} X_i}{84}.$$

Distribución del Estimador de la Media

$$P(0.64 < \hat{\theta}_{84} < 0.76) = P\left(\frac{0.64 - 0.7}{\sqrt{\frac{0.7(0.3)}{84}}} < \frac{\hat{\theta}_{84} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} < \frac{0.76 - 0.7}{\sqrt{\frac{0.7(0.3)}{84}}}\right)$$

$$\simeq P(-1.2 < Z < 1.2) = 2P(0 < Z < 1.2)$$

$$= 2(0.3849) = 0.7698$$

De nuevo, esta aproximación puede mejorarse usando la corrección por continuidad y se deja como ejercicio.

Distribución del Estimador de la Media

En resumen, la distribución muestral de $\overline{X}$ puede trabajarse como:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \begin{cases} N(0,1) & \text{si la muestra viene de una población normal} \\ N(0,1) & \text{si la muestra viene de una población arbitraria} \\ & \text{y } n \text{ es grande} \end{cases}.$$

Distribución del Estimador de la Varianza

Otra distribución muestral que es de nuestro interés es la de la varianza muestral S^2 , puesto que S^2 es un estimador natural de σ^2.

Recordemos nuestra definición de la varianza muestral:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left\{(X_i-\mu)-(\overline{X}-\mu)\right\}^2.$$

Entonces,

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}\left\{(X_i-\mu)-(\overline{X}-\mu)\right\}^2 \\ &= \sum_{i=1}^{n}(X_i-\mu)^2 - 2\sum_{i=1}^{n}(X_i-\mu)(\overline{X}-\mu) + n(\overline{X}-\mu)^2 \\ &= \sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2.\end{aligned}$$

Distribución del Estimador de la Varianza

Además,

$$\begin{aligned} E((n-1)S^2) &= E\left\{\sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2\right\} \\ &= \sum_{i=1}^{n} E(X_i-\mu)^2 - nE(\overline{X}-\mu)^2 \\ &= \sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = \sigma^2(n-1), \end{aligned}$$

pues $E(X_i-\mu)^2 = \sigma^2$ y $E(\overline{X}-\mu)^2 = \frac{\sigma^2}{n}$.

Por lo tanto,

$$E(S^2) = \sigma^2.$$

Distribución del Estimador de la Varianza *

Establecer la varianza de S^2 es bastante más complejo, por lo que en principio lo haremos sólo para el caso cuando la muestra aleatoria $X_1, X_2, \ldots, X_n$ haya sido tomada de una población $Normal(\mu, \sigma)$; esto se hará derivando primero la distribución muestral y, posteriormente, sus momentos, media y varianza.

La media ya sabemos que es la varianza de la población, sin importar de dónde hayamos tomado la muestra $E(S^2) = \sigma^2$ siempre.

Sabemos que

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

también sabemos que, cuando la muestra es tomada de una población normal,

$$\frac{(X_i - \mu)}{\sigma} \sim N(0,1), \qquad \frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Distribución del Estimador de la Varianza *

Con la información anterior en mente, consideremos lo siguiente.

Si en la expresión para la varianza multiplicamos ambos lados por $n-1$, restamos y sumamos μ dentro del paréntesis, tenemos

$$(n-1)S^2 = \sum_{i=1}^{n} \left\{(X_i - \mu + \mu - \overline{X})\right\}^2 .$$

Agrupando y elevando al cuadrado llegamos a

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n} \left\{(X_i - \mu) - (\overline{X} - \mu)\right\}^2 = \sum_{i=1}^{n} [(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2] \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) + \sum_{i=1}^{n} (\overline{X} - \mu)^2 .\end{aligned}$$

Distribución del Estimador de la Varianza *

En el segundo término se dejó sólo el término que tiene índice (sobre el que corre la sumatoria), ahora, puesto que

Observemos que

$$\sum_{i=1}^{n} X_i = n\overline{X} \qquad \text{(por definición de media muestra)}$$

$$\sum_{i=1}^{n} \mu = n\mu \qquad \text{(se suma n veces una constante)}$$

$$\sum_{i=1}^{n} (\overline{X} - \mu)^2 = n(\overline{X} - \mu)^2 \qquad \text{(se suma n veces una constante)}$$

Distribución del Estimador de la Varianza *

Usando la observación anterior, podemos escribir que

$$\begin{aligned}(n-1)S^2 &= [\sum_{i=1}^{n}(X_i-\mu)^2] - 2(\overline{X}-\mu)(n\overline{X}-n\mu) + n(\overline{X}-\mu)^2 \\ &= [\sum_{i=1}^{n}(X_i-\mu)^2] - 2n(\overline{X}-\mu)^2 + n(\overline{X}-\mu)^2 \\ &= [\sum_{i=1}^{n}(X_i-\mu)^2] - n(\overline{X}-\mu)^2.\end{aligned}$$

Entonces

$$(n-1)S^2 = [\sum_{i=1}^{n}(X_i-\mu)^2] - \left(\frac{\bar{X}-\mu}{\frac{1}{\sqrt{n}}}\right)^2.$$

Distribución del Estimador de la Varianza *

Dividiendo por σ^2 nos queda

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 - \left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)^2 .$$

El primer término del lado derecho es la suma de n v.a.i. $N(0,1)$ elevadas al cuadrado, y el segundo término es una v.a. $N(0,1)$ también elevada al cuadrado. ¿Qué distribución tiene una suma de variables aleatorias ji-cuadradas independientes con un grado de libertad? Una Ji-cuadrada con n grados de libertad.

Nota: De hecho una suma de l variables aleatorias Ji-cuadradas independientes con n_i grados de libertad $(i = 1, \ldots, l)$ es una variable aleatoria Ji-cuadrada con $m = n_1 + n_2 + \ldots + n_l$ grados de libertad. Verificar.

Distribución del Estimador de la Varianza *

Como tenemos que $\frac{(n-1)S^2}{\sigma^2}$ es igual a la resta de dos variables aleatorias con distribuciones χ^2_n y χ^2_1, no conocemos su distribución.

Aunque en la nota anterior se menciona cual es la distribución de una suma de variables aleatorias Ji-cuadrada independientes, no es tan evidente cuál es la distribución de las variables restadas. En primer lugar no hay garantía de independencia, y en segundo lugar si las variables inmiscuidas fueran independientes, la generatriz de la combinación no es una reconocible.

Distribución del Estimador de la Varianza *

Sin embargo, si vemos la ecuación de la siguiente manera

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2,$$

existe un teorema que garantiza la independencia de los dos términos de la izquierda. Considerando esto podemos calcular la generatriz de la combinación de las dos variables e igualarla a la generatriz de la variable de la derecha.

Si llamamos $U = \frac{(n-1)S^2}{\sigma^2}$, $V = (\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}})^2$ y $W = \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2$, tenemos que $U + V = W$ y entonces

$$M_U(t) \cdot M_V(t) = M_{U+V}(t) = M_W(t).$$

Distribución del Estimador de la Varianza *

Como $V \sim \chi^2_1 \equiv Gamma(\alpha = \frac{1}{2}, \beta = 2)$ y $W \sim \chi^2_n \equiv Gamma(\frac{n}{2}, 2)$, podemos sustituir las generatrices correspondientes

$$M_U(t) \cdot (1-2t)^{-\frac{1}{2}} = (1-2t)^{-\frac{n}{2}}.$$

Despejando $M_U(t)$, se tiene que

$$M_U(t) = \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} = (1-2t)^{-\frac{n-1}{2}},$$

la cual corresponde a la generatriz de una distribución Gamma con $n-1$ grados de libertad. Es decir que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \equiv Gamma(\alpha = \frac{n-1}{2}, \beta = 2).$$

(Nota que cuando formamos el estimador de la varianza como S^2, la suma esta dividida por $n-1$ y los grados de libertad asociados aquí también son $n-1$. Esto no se obtiene por casualidad).

Distribución del Estimador de la Varianza *

Ahora bien, recordemos que anteriormente se vio que si se tiene una v.a. $Y \sim Gamma(\alpha = \frac{\nu}{2}, \beta = 2)$, entonces un múltiplo de Y cumple $aY \sim Gamma(\alpha = \frac{\nu}{2}, \beta^* = a\beta = 2a)$, que es distinta de una χ^2.

Si identificamos a $\frac{(n-1)S^2}{\sigma^2}$ con Y y tomamos $a = \frac{\sigma^2}{n-1}$, entonces $aY = S^2$. Así,

$$Y = \frac{(n-1)S^2}{\sigma^2} \sim Gamma(\alpha = \frac{n-1}{2}, \beta = 2),$$

lo que implica que

$$aY = \frac{\sigma^2}{n-1} \cdot \frac{n-1}{\sigma^2} S^2 = S^2 \sim Gamma(\alpha = \frac{n-1}{2}, \beta = \frac{2\sigma^2}{n-1}).$$

En general preferimos trabajar con una Ji-cuadrada que con una Gamma.

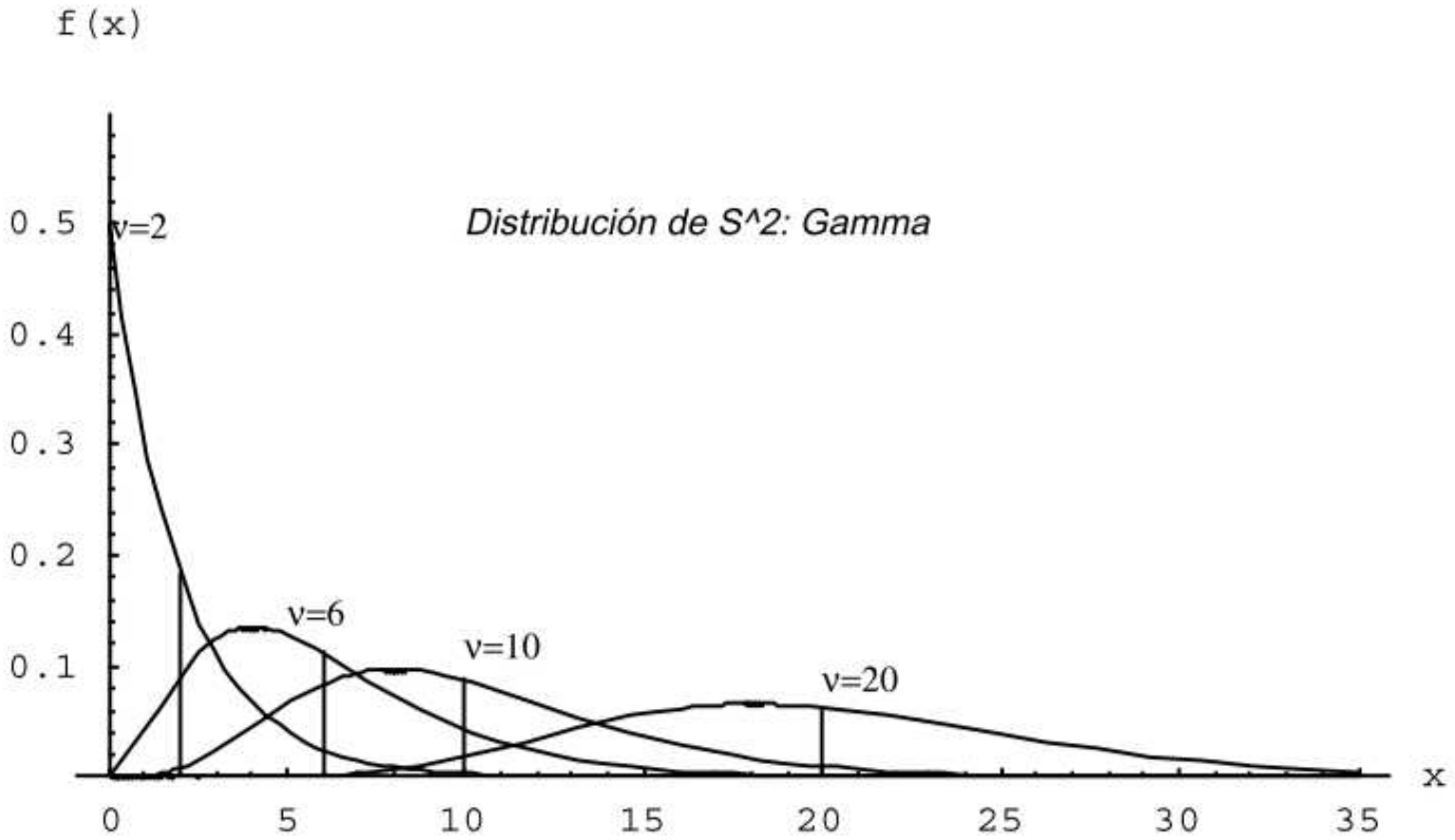
Distribución del Estimador de la Varianza *

Como conocemos que la varianza de una v.a. $Gamma(\alpha, \beta)$ es $\alpha\beta^2$, tenemos que

$$V(S^2) = \underbrace{\frac{n-1}{2}}_{\alpha} \underbrace{\left(\frac{2\sigma^2}{n-1}\right)^2}_{\beta^2} = \frac{2\sigma^4}{n-1}.$$

La siguiente figura muestra en forma esquemática las distribuciones muestrales para los estimadores de la media y de la varianza, obtenidos de muestras de poblaciónes normales.

Distribución del Estimador de la Varianza *



Distribución del Estimador de la Varianza

Observación:

Hemos demostrado que la distribución de una v.a. S^2 es una distribución Gamma, la cual usualmente es sesgada a la derecha. Por ello el valor promedio es "jalado" por la cola "pesada" y existe una mayor posibilidad de obtener un valor alejado del promedio (σ^2); esto es, S^2 generalmente subestima el valor de la varianza.

Sin embargo, si el tamaño de muestra n es grande, la variación de S^2 es menor y la curva tiende a ser más simétrica, más "acampanada", algo como una normal, y la subestimación tiende a desaparecer. Esto concuerda con el TCL ya que S^2, en última instancia, se expresa como un promedio de variables aleatorias independientes, $Y_i = (X_i - \overline{X})^2$.

Distribución del Estimador de la Varianza

El hablar de algo llamado precisión en la estimación implica que deberá existir su contraparte: error en la estimación. Esto depende al menos de dos factores, tamaño de muestra y distribución de la población.

Según nuestro grado de precisión establecido, la forma de la población y los objetivos planteados, podremos fijar un tamaño de muestra. Más adelante estudiaremos algunos casos particulares.

Nota: En el caso de no tener una población normal se tiene la siguiente relación para la varianza de S^2

$$V(S^2) = \sigma^4 \left(\frac{2}{n-1} + \frac{E(X-\mu)^4 - 3\sigma^4}{n\sigma^4} \right) = \sigma^4 \left(\frac{2}{n-1} + \frac{\delta}{n} \right),$$

donde $\delta =$ coeficiente de curtosis.

Distribución del Estimador de la Varianza

Habíamos dicho que si δ era aproximadamente cero, nuestra curva se asemejaba más a la de una normal. En esta fórmula, si $\delta = 0$, la varianza de S^2 se reduce a la varianza correspondiente para una población normal $(\frac{2\sigma^4}{n-1})$.

Denotamos con $\chi^2_{\alpha,\nu}$ al numero real que satisface que

$$P(\chi^2 > \chi^2_{\alpha,\nu}) = \alpha$$

Por ejemplo, si $\alpha = 0.95$ y los grados de libertad son $\nu = 15$, tenemos que $\chi^2_{0.95,15} = 7.261 \Rightarrow P(\chi^2_{15} > 7.26) = 0.95$.

Estimación Puntual

Introducción

Estimación Puntual

Hemos visto que las cantidades como $\overline{X}, S^2$ son estimadores naturales de los parámetros μ, σ^2 y que en vista de su carácter aleatorio tienen asociada una distribución de probabilidad, la cual contiene información acerca del estimador.

Sin embargo los estimadores no son únicos.

Por otro lado podemos estar interesados en estimar parámetros muy diferentes a μ y σ^2. Por ejemplo, los parámetros α, β de la distribución Gamma.

El estudio de tales estimadores es tema de la estimación puntual.

Estimación Puntual

La estimación puntual tiene dos objetivos

- Encontrar estimadores para los parámetros de interés.
- Evaluar la calidad de los estimadores con el fin de seleccionar al más adecuado.

No intentamos obligarte a que cuando tengas un problema, te avoques inmediatamente a usar la teoría para hallar un estimador, más bien intentamos mostrar los estimadores más convenientes, los más usuales, aproximaciones útiles y darte una idea clara de la confianza que puedes tener al elegir el valor (o intervalo) de un cierto estimador como "cercano" al parámetro.

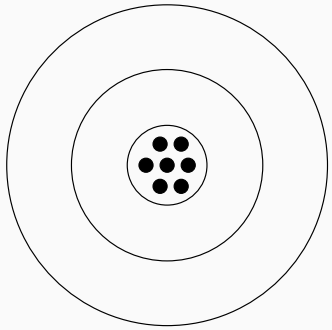
Estimación Puntual

Existen una serie de criterios que ayudan a la elección de un estimador

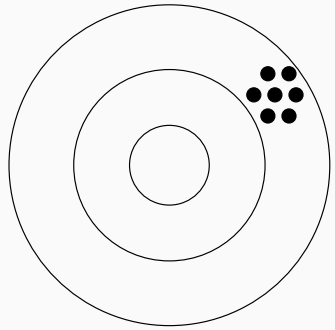
- Insesgamiento.
- Eficiencia.
- Consistencia.
- Suficiencia.
- Invarianza, etc.

Estimación Puntual

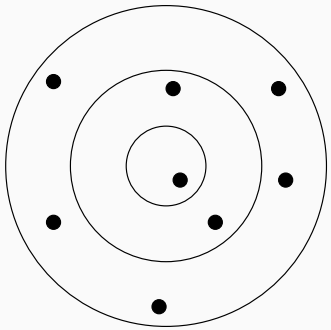
Varianza y sesgo.



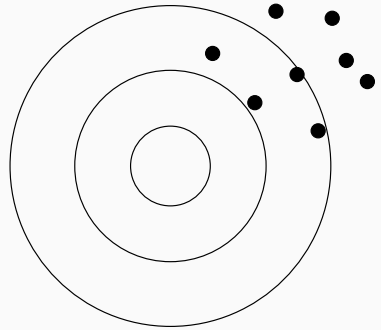
menor varianza, menor sesgo



menor varianza, mayor sesgo



mayor varianza, menor sesgo



mayor varianza, mayor sesgo

Estimación Puntual: Insesgamiento

Al principio del capítulo se dijo que: "...$\overline{X}$ es un estimador natural de μ, dado que en promedio se toma ese valor (μ)..."; este es el concepto de insesgamiento. Como el promedio de una cierta v.a. es igual a su valor esperado, tenemos la siguiente definición.

Definición

Sea $\hat{\theta}$ un estimador del parámetro θ. Diremos que $\hat{\theta}$ es un estimador insesgado para θ si

$$E(\hat{\theta}) = \theta$$

O sea, podemos pedir que "en promedio" el valor del estimador sea igual al valor del parámetro en cuestion.

Estimación Puntual: Ejemplos

1. $\overline{X}$ en una población cualquiera satisface que $E(\overline{X}) = \mu \implies \overline{X}$ es insesgado para μ.
2. Si $X_1, X_2, \ldots, X_n$ es una m.a. de una población con media μ y varianza σ^2 entonces

$$E(X_i) = \mu, \qquad Var(X_i) = \sigma^2;$$

esto es, cada variable en la muestra constituye un estimador insesgado para la media y la varianza.

3. En poblaciónes normales S^2 es un estimador de σ^2, además

$$E(S^2) = E\left(\frac{(n-1)S^2}{\sigma^2} \cdot \frac{\sigma^2}{n-1}\right) = \frac{\sigma^2}{n-1} E\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2,$$

lo cual implica que S^2 es un estimador insegado para σ^2.

Estimación Puntual: Ejemplos

4. Otro estimador para σ^2 en poblaciónes normales es $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Tenemos que,

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}(n-1)S^2 = \frac{n-1}{n}S^2,$$

lo cual implica que

$$E(S_n^2) = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}E(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Por lo tanto, S_n^2 no es un estimador insesgado para σ^2. En este caso se dice que S_n^2 subestima en promedio el valor de σ^2; se dice que hay un sesgo.

Definición

Si un estimador $\hat{\theta}$ no es insesgado para θ, se dice que es sesgado y se define el sesgo de $\hat{\theta}$ como

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Cuando el sesgo es positivo se dice que el estimador está sesgado a la derecha y si es negativo se dice que está sesgado a la izquierda.

Ejercicio: En el ejemplo 4 anterior calcula el sesgo.

Es posible, sin embargo, que para algunos estimadores sesgados, cuando n (el tamaño de muestra) aumenta, el sesgo disminuya. Tenemos por lo tanto la siguiente definición.

Definición

Sea $\hat{\theta}$ un estimador para θ, diremos que $\hat{\theta}$ es asintoticamente insesgado para θ si

$$\lim_{n \to \infty} E(\hat{\theta}) = \theta$$

donde n es el tamano de la muestra.

Ejemplo: S_n^2 es asintóticamente insesgado en una población normal porque

$$\lim_{n\to\infty} E[S_n^2] = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2.$$

Como ya dijimos, existen múltiples estimadores insesgados para un mismo parámetro en una población dada. La pregunta que surge ahora es **de qué forma decidir cuál será más conveniente utilizar**.

Un método consiste en comparar las variabilidades de cada estimador y escoger el que tenga la menor variabilidad, es decir, la menor varianza. Veamos un ejemplo; posteriormente escribiremos la definición formal.

Ejemplo: En una población normal, se toma una muestra aleatoria $X_1, X_2, \ldots, X_n$. Sabemos que S^2 es insesgado para σ^2; mostraremos que $\hat{\sigma}^2 = \frac{1}{2}(X_1 - X_n)^2$, también es un estimador insesgado para σ^2.

Tenemos que $X_i \sim N(\mu, \sigma^2)$, lo que implica que $X_1 - X_n \sim N(0, 2\sigma^2)$. Estandarizando obtenemos que $\frac{X_1 - X_n}{\sqrt{2}\sigma} \sim N(0,1)$. Luego,

$$\frac{(X_1 - X_n)^2}{2\sigma^2} \sim \chi^2_1.$$

Esto implica que

$$E\left(\frac{(X_1 - X_n)^2}{2\sigma^2}\right) = 1,$$

y por tanto

$$E(\hat{\sigma}^2) = E\left(\frac{(X_1 - X_n)^2}{2}\right) = \sigma^2,$$

es decir que $\hat{\sigma}^2$ es insesgado.

Ahora calculemos la varianza de $\hat{\sigma}^2$.

Sabemos que $V(\frac{(X_1-X_n)^2}{2\sigma^2}) = 2$, por ser una variable χ^2 con 1 grado de libertad. Entonces,

$$V\left(\frac{(X_1 - X_n)^2}{2}\right) = 2\sigma^4,$$

lo que implica que

$$V(\hat{\sigma}^2) = 2\sigma^4 \geq \frac{2\sigma^4}{n-1} = V(S^2).$$

Entonces se prefiere S^2 sobre $\hat{\sigma}^2$.

Definición

Si $\hat{\theta}_1$ y $\hat{\theta}_2$ son dos estimadores insesgados de θ, se dice que $\hat{\theta}_1$ es más eficiente (más preciso) que $\hat{\theta}_2$ si

$$V(\hat{\theta}_1) < V(\hat{\theta}_2)$$

Ejemplo: Supongamos que se quiere estudiar el tiempo de reacción de una determinada substancia química. Por falta de más información, se asumió que dichos tiempos siguen un comportamiento uniforme en el intervalo $(0, \theta)$ y lo que interesa entonces es, al menos, hacer una buena estimación de θ.

Se puede demostrar (ver el ejercicio siguiente) que cuando $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una distribución uniforme en $(0, \theta)$ el estimador

$$\hat{\theta}_1 = \frac{n+1}{n} \max(X_1, X_2, \ldots, X_n)$$

es un estimador insesgado para θ. Este no es el único estimador insesgado de θ. Por ejemplo, tenemos que

$$E(X_i) = \frac{a+b}{2} = \frac{0+\theta}{2} = \frac{\theta}{2}$$

$$\Longrightarrow E(\overline{X}) = \frac{\theta}{2}, \qquad E(2\overline{X}) = \theta.$$

Por lo tanto $\hat{\theta}_2 = 2\overline{X}$ también es un estimador insesgado para θ.

La anterior es una práctica común para deducir un estimador insesgado, pero depende del resultado del valor esperado.

¿Cuál de los dos estimadores tiene menor varianza? O dicho de otra forma, ¿cuál estimador escogeríamos? Calculemos la varianza de cada estimador y el que tenga la menor será "el agraciado".

Tenemos que

$$V(\hat{\theta}_1) = \left(\frac{n+1}{2}\right)^2 V(X_{(n)})$$

donde $X_{(n)}$ es el estadístico de orden n. Te darás cuenta que nos falta $V(X_{(n)})$. Para calcular la varianza del estadístico de orden n recordemos que

$$f_{X_{(n)}}(x) = n[F_X(x)]^{n-1} f_X(x).$$

Se dejará como ejercicio demostrar que la función de densidad de $X_{(n)}$ y su varianza son

$$f_{X_{(n)}}(x) = n\frac{x^{n-1}}{\theta^n} \quad 0 < x < \theta$$

y

$$V(X_{(n)}) = \frac{n}{(n+1)^2(n+2)}\theta^2,$$

respectivamente.

Entonces,

$$V(\hat{\theta}_1) = \left(\frac{n+1}{n}\right)^2 \left[\frac{n}{(n+1)^2(n+2)}\theta^2\right] = \frac{\theta^2}{n(n+2)}.$$

Por otra parte, recordando que $V(aX) = a^2 V(X)$, y que $V(\overline{X}) = \frac{\sigma^2}{n}$, tenemos que

$$\begin{aligned} V(\hat{\theta}_2) &= V(2\bar{X}) = 2^2 V(\overline{X}) \\ &= 4\frac{\frac{(\theta-0)^2}{12}}{n} = \frac{\theta^2}{3n}, \end{aligned}$$

donde hemos usado el hecho de que la varianza de una f.d.p. uniforme en (a, b) es $\frac{(b-a)^2}{12}$. ¿Cuál tiene la menor varianza?

Si dividimos las dos varianzas obtenemos

$$\frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)} = \frac{\frac{\theta^2}{n(n+2)}}{\frac{\theta^2}{3n}} = \frac{3}{n+2},$$

aquí podemos observar que, si $n > 1$, la varianza de $\hat{\theta}_1$ es menor que la de $\hat{\theta}_2$. Por ejemplo, si $n = 4$, tendríamos

$$\frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)} = \frac{3}{n+2} = \frac{3}{6} = \frac{1}{2},$$

es decir,

$$V(\hat{\theta}_1) = \frac{1}{2} V(\hat{\theta}_2).$$

Así, el estimador insesgado $\hat{\theta}_1 = \frac{n+1}{n} \max(X_1, X_2, \ldots, X_n)$ tiene menor varianza que el estimador insesgado $\hat{\theta}_2 = 2\overline{X}$.

Si te preguntas porqué dividimos las dos varianzas, es por el hecho de que si dividimos dos cantidades y el resultado es mayor que 1, significará que el numerador es mayor que el denominador, y si el resultado es menor que 1, significará que el numerador es menor que el denominador.

No importa a quién escribas en el numerador o denominador sino la interpretación del cociente en sí.

Ejemplo: Sea una m.a. $X_1, X_2, \ldots, X_n$ de una población normal. La media muestral $\overline{X}$ y la mediana muestral $\widetilde{X}$ son estimadores insesgados para μ, y sus varianzas respectivas son

$$V(\overline{X}) = \frac{\sigma^2}{n}, \quad V(\widetilde{X}) \approx \frac{\pi\sigma^2}{2n},$$

y por lo tanto

$$\frac{V(\widetilde{X})}{V(\overline{X})} = 1.57.$$

Esto es, la media es 57% más eficiente que la mediana. Si $n = 100$ se necesita una muestra de tamaño $n = 157$ para que la mediana fuese tan precisa para la estimación de μ como lo es $\overline{X}$.

El concepto de eficiencia puede conducirnos en forma natural a pensar que si encontramos el de menor varianza de todos los posibles estimadores insesgados tendríamos un estimador "de alto rendimiento", éste será llamado **Estimador Insesgado de Mínima Varianza (EIMV)**.

Antes de ver la teoría pertinente, los siguientes ejemplos nos mostrarán diversos estimadores de la media μ y un experimento que podríamos llamar "experimento truncado".

Estimación Puntual: Estimadores para μ

Existen diversos estimadores para la media μ , y aunque el más usual es $\overline{X}$, la elección final depende fuertemente de la distribución que está siendo muestreada. A continuación te presentamos algunos estimadores para μ a partir de una m.a. $X_1, X_2, \ldots, X_n$.

1. La media muestral $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. La mediana muestral

$$\widetilde{X} = \begin{cases} \text{El valor central si n es impar} \\ \\ \text{El promedio de los datos centrales si n es par} \end{cases} .$$

3. El promedio de los datos extremos $\overline{X}_e = \frac{X_{(1)}+X_{(n)}}{2}$.
4. $\overline{X}_{tr(10)}$ el promedio de los datos al descartar el 10% inferior y el 10% superior de los mismos (media ajustada).

Estimación Puntual: Estimadores para μ

Nota: El subíndice en el último estimador es por "trimmed".

No hay respuesta a la pregunta: **¿cuál de estos estimadores es el más cercano al verdadero valor de μ?** Pero sí a la pregunta: **¿cuál estimador tenderá a producir estimaciones más cercanas al verdadero valor?** Veamos el siguiente ejemplo.

Ejemplo: Supóngase que se desea estimar la conductividad térmica promedio μ de un cierto material. Usando técnicas de medición típicas, se obtiene una m.a. $X_1, X_2, \ldots, X_n$ de n mediciones de conductividad térmica. Se piensa que la distribución poblaciónal es alguna de las siguientes tres familias.

Estimación Puntual

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad -\infty < x < \infty;$$

$$f(x) = \frac{1}{\pi[1 + (x-\mu)^2]} \qquad -\infty < x < \infty;$$

$$f(x) = \begin{cases} \frac{1}{2c} & -c \leq x - \mu \leq c \\ 0 & \text{otra parte} \end{cases}.$$

La primera f.d.p. es la distribución normal, la segunda es la distribución de Cauchy, y la tercera es una distribución uniforme. Las tres distribuciones son simétricas alrededor de μ. La distribución de Cauchy es acampanada con colas mucho más pesadas (hay más probabilidad para valores alejados) que la de la curva normal. La distribución uniforme no tiene colas.

Estimación Puntual

Haremos un análisis de la conveniencia de usar distintos estimadores para μ dependiendo de qué distribución (población) proviene la muestra.
El mejor estimador de μ depende fuertemente de la distribución (población) que está siendo muestreada. En particular,

- Si la muestra aleatoria proviene de una distribución normal, entonces $\overline{X}$ es el mejor de los cuatro estimadores, ya que tiene la mínima varianza de entre todos los estimadores insesgados.
- Si la muestra aleatoria viene de una distribución Cauchy, entonces $\overline{X}$ y $\overline{X}_e$ son pésimos estimadores para μ, mientras que $\widetilde{X}$ es muy bueno (el EIMV no es conocido). $\overline{X}$ es malo ya que es muy sensible a observaciones "outlying" y las colas pesadas de la distribución de Cauchy hacen que tales observaciones probablemente aparezcan en una muestra.(**Nota**: en este caso μ no tiene la misma interpretación, dado que la esperanza matemática, no existe).

Estimación Puntual

- Si la distribución que corresponde es uniforme, el mejor estimador es $\overline{X}_e$; este estimador es grandemente influenciado por observaciones "outlying", pero la falta de colas en la uniforme hace tales observaciones imposibles.
- La media ajustada ("trimmed") no es la mejor en ninguna de estas tres situaciones, pero trabaja razonablemente bien en cada una de ellas. Esto es, $\overline{X}_{tr(10)}$ "no pierde mucho" aún cuando se compara con los mejores estimadores en cada una de las tres situaciones.

Estimación Puntual

Recientes investigaciones en estadística han establecido que cuando se estima μ para una distribución continua, una media ajustada con proporción de ajuste de 10% ó 20% (de cada extremo de la muestra) produce estimaciones razonables sobre un amplio rango de modelos posibles. Por ello, una media ajustada con pequeño porcentaje de ajuste se dice que es un estimador robusto.

Existen situaciones en las que la elección del estimador no se dá entre diferentes estimadores producidos por la misma muestra sino entre estimadores basados en dos experimentos diferentes.

Ejemplo: Supóngase que un cierto tipo de componente tiene una distribución de su tiempo de vida $exp(\theta)$. Una muestra de tamaño n de tales componentes es seleccionada y cada una es puesta en operación. Si el experimento es continuado hasta que todos los n tiempos de vida, $X_1, X_2, \ldots, X_n$, son observados, entonces $\overline{X}$ es un estimador insesgado de $\mu = \theta$.

En algunos experimentos, sin embargo las componentes son dejadas en operación hasta que **falla la r-ésima**, donde $r < n$. Sean Y_1 el tiempo de la primera falla (el mínimo tiempo de vida de entre las n componentes), Y_2 el tiempo al cual la segunda falla ocurre (el segundo tiempo de falla más pequeño), y así sucesivamente.

Estimación Puntual

Ya que el experimento termina al tiempo Y_r, tenemos que $T_r =$ **"el tiempo de vida total acumulado de las componentes a la terminación del experimento"** es

$$T_r = \sum_{i=1}^{r} Y_i + (n-r)Y_r.$$

El primer término del lado derecho representa la suma de los tiempos de vida de las componentes que fallaron, y el segundo término representa la suma de los tiempos de las $(n-r)$ restantes, recuerda que se terminó el experimento y por ende las componentes restantes vivieron el tiempo máximo de la r-ésima que falló, esto es Y_r.

Calcular el valor esperado de la variable aleatoria T_r y a partir del resultado proponer un estimador insesgado para μ.

Nota que los Y_i son los estadísticos de orden i de los tiempos de vida. Una forma de proceder puede ser obteniendo el valor esperado de la fórmula anterior, sin embargo necesitamos conocer las densidades de los estadísticos de orden, hasta el de orden r inclusive. Esto sería algo extenuante ya que la distribución estaría cambiando para cada valor de la variable Y_i, por ello se prefiere una forma alternativa de escribir el tiempo T_r.

El estadístico de orden 1 para una m.a. de tamaño n de una población exponencial, tiene distribución exponencial con parámetro $\frac{\theta}{n}$ (verificar). Es posible escribir una expresión para T_r en términos de estadísticos de primer orden y utilizando propiedades de valor esperado encontrar $E(T_r)$.

Estimación Puntual

Es importante ver que **cuando falla una componente las restantes siguen teniendo una distribución exponencial con el mismo parámetro θ, a pesar de saber que han vivido por lo menos el tiempo de las componentes que fallaron**(Propiedad de pérdida de memoria). Entonces:

- las n componentes duran hasta lo que dura el estadístico de primer orden, Y_1, de las n componentes, cada una $Exp(\theta)$.
- las $(n-1)$ componentes restantes duran un tiempo adicional de $Y_2 - Y_1$, que representa el nuevo estadístico de orden uno de las $(n-1)$ componentes restantes, cada una $Exp(\theta)$.
- las $(n-2)$ componentes restantes duran un tiempo adicional de $Y_3 - Y_2$ que representa el nuevo estadístico de orden uno de las $(n-2)$ componentes restantes, cada una $Exp(\theta)$.
- y así sucesivamente.

Por lo tanto, otra expresión para T_r es

$$T_r = nY_1+(n-1)(Y_2-Y_1)+(n-2)(Y_3-Y_2)+\cdots+[n-(r-1)](Y_r-Y_{r-1})$$

donde se están sumando los mínimos tiempos cada vez. Observa que

$$\begin{aligned}
E(Y_1) &= \frac{\theta}{n}, && Y_1 = \text{est. de orden 1 de las } n \text{ comp., cada una } exp(\theta)\\
E(Y_2-Y_1) &= \frac{\theta}{n-1}, && Y_2-Y_1 = \text{est. orden 1 de las } n-1 \text{ comp. restantes}\\
E(Y_3-Y_2) &= \frac{\theta}{n-2}, && Y_3-Y_2 = \text{est. orden 1 de las } n-2 \text{ comp. restantes}\\
&\vdots\\
E(Y_r-Y_{r-1}) &= \frac{\theta}{[n-(r-1)]}.
\end{aligned}$$

Así,

$$\begin{aligned} E(T_r) &= E[nY_1 + (n-1)(Y_2 - Y_1) + (n-2)(Y_3 - Y_2) + \cdots + [n-(r-1)](Y_r - Y_{r-1})] \\ &= \underbrace{n(\frac{\theta}{n})}_{\text{1er término}} + \underbrace{(n-1)(\frac{\theta}{n-1})}_{\text{2do término}} + \cdots + \underbrace{[n-(r-1)](\frac{\theta}{[n-(r-1)]})}_{\text{r-ésimo término}} \\ &= \theta + \theta + \cdots + \theta \\ &= r\theta. \end{aligned}$$

Por lo tanto,

$$E(T_r) = r\theta,$$

de donde un estimador insesgado utilizando este procedimiento sería

$$\hat{\theta} = \frac{T_r}{r}.$$

Estimación Puntual

Ya que se está eliminando la información de la duración final de las componentes restantes, a este tipo de procedimiento se le llama **experimento de datos censurados**.

Este tipo de datos es muy común en los estudios de confiabilidad y/o sobrevivencia.

Como un ejemplo numérico, considera 20 componentes puestas a prueba en la cual se ha fijado $r = 10$. Se encuentra que los tiempos de falla de las primeras 10 (en las unidades respectivas) son 11, 15, 29, 33, 35, 40, 47, 55, 58, y 72; de donde el valor estimado para μ es

$$\hat{\mu} = \frac{11 + 15 + \cdots + 72 + 10(72)}{10} = 111.5.$$

Por simplicidad se usa la primera expresión para T_r.

La ventaja del experimento con censura es que termina más rapidamente que el experimento sin censura. Una desventaja es que la varianza de $\hat{\theta} = \frac{T_r}{r}$ en el experimento con censura es mayor que la varianza de $\overline{X}$ en el experimento sin censura.

Estimadores Insesgados de Mínima Varianza

Estimadores Insesgados de Mínima Varianza

Habíamos comentado el hecho de considerar a un estimador insesgado como más eficiente que otro comparando sus varianzas respectivas y que esto conducía a preguntarse si existiría una forma de saber cuando un estimador insesgado tenía la menor varianza posible. En algunos casos es factible establecer este estimador.

Definición

Sea $\hat{\theta}$ un estimador de θ, diremos que $\hat{\theta}$ es un estimador insesgado de mínima varianza (EIMV) si:

1. $\hat{\theta}$ es insesgado,
2. Para cualquier otro estimador insesgado $\widetilde{\theta}$, se satisface que $V(\hat{\theta}) \leq V(\widetilde{\theta})$.

Estimadores Insesgados de Mínima Varianza

Teorema (Desigualdad de Cramer-Rao)

Sea $\hat{\theta}$ un estimador insesgado para θ entonces:

$$V(\hat{\theta}) \geq \frac{1}{nE\left[(\frac{\partial}{\partial\theta}\log f(x))^2\right]},$$

donde $f(x)$ es la funcion de probabilidad o densidad de la poblacion y n el tamano de la muestra.

Si por algún método encontramos un estimador insesgado cuya varianza sea igual al valor

$$\frac{1}{nE\left[(\frac{\partial}{\partial\theta}\log f(x))^2\right]}$$

éste será el EIMV.

Estimadores Insesgados de Mínima Varianza

Al número $\frac{1}{nE\left[(\frac{\partial}{\partial\theta}\log f(x))^2\right]}$ se le llama la Cota de Cramer-Rao y al número $nE\left[(\frac{\partial}{\partial\theta}\log f(x))^2\right]$ se le llama la información proporcionada por la muestra.

Teorema

$$nE\left[(\frac{\partial}{\partial\theta}\log f(x))^2\right] = -nE\left[\frac{\partial^2}{\partial\theta^2}\log f(x)\right].$$

La demostración se deja como ejercicio opcional.

Estimadores Insesgados de Mínima Varianza

Ejemplo

En una poblacion $Exp(\theta)$, $\overline{X}$ es el EIMV.

*Poblacion exponencial media $=\theta$ varianza $=\theta^2$

$$
\begin{aligned}
E(\overline{X}) &= \text{Media de la población} = \theta \\
V(\overline{X}) &= \frac{\text{varianza pob.}}{n} = \frac{\theta^2}{n} \\
f(x) &= \frac{1}{\theta} e^{-\frac{x}{\theta}} \\
\log f(x) &= \log\left[\frac{1}{\theta} e^{-\frac{x}{\theta}}\right] = -\log\theta - \frac{x}{\theta} \\
\frac{\partial}{\partial\theta} \log f(x) &= -\frac{1}{\theta} - (-\frac{x}{\theta^2}) = \frac{x}{\theta^2} - \frac{1}{\theta} = \frac{x-\theta}{\theta^2} \\
E\left[(\frac{\partial}{\partial\theta} \log f(x))^2\right] &= E\left[(\frac{X-\theta}{\theta^2})^2\right] = E\left[\frac{(X-\theta)^2}{\theta^4}\right] = \frac{1}{\theta^4} \underbrace{E[(X-\theta)^2]}_{V(X)} \\
&= \frac{1}{\theta^4} V(X) = \frac{\theta^2}{\theta^4} = \frac{1}{\theta^2}
\end{aligned}
$$

Ejemplo (Cont...)

La cota de Cramer-Rao queda:

$$\frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\log f(x)\right)^2\right]} = \frac{1}{n\left(\frac{1}{\theta^2}\right)} = \frac{\theta^2}{n}$$

Como $V(\overline{X})$ =cota Cramér-Rao, entonces $\overline{X}$ es EIMV para θ en una poblacion exponencial.

Estimadores Insesgados de Mínima Varianza

Ejemplo

En una poblacion Poisson (λ), $\overline{X}$ es el EIMV para λ.

*Poblacion Poisson; Media=λ , Varianza=λ

$$E(\overline{X}) = \lambda, \quad Var(\overline{X}) = \frac{\lambda}{n}.$$

Calculamos la cota Cramer-Rao:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \implies \log f(x) = \log\left[\frac{\lambda^x e^{-\lambda}}{x!}\right] = x\log\lambda - \lambda - \log(x!)$$

$$\frac{\partial}{\partial\lambda}\log f(x) = \frac{\partial}{\partial\lambda}[x\log\lambda - \lambda - \log(x!)] = \frac{x}{\lambda} - 1 = \frac{x-\lambda}{\lambda}$$

$$E\left[(\frac{\partial}{\partial\lambda}\log f(x))^2\right] = E\left[(\frac{X-\lambda}{\lambda})^2\right] = \frac{1}{\lambda^2}\underbrace{E[(X-\lambda)^2]}_{V(X)} = \frac{V(X)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

$$\implies \frac{1}{nE\left[(\frac{\partial}{\partial\lambda}\log f(x))^2\right]} = \frac{1}{n(\frac{1}{\lambda})} = \frac{\lambda}{n} = V(\overline{X}).$$

Por lo tanto, $\overline{X}$ es EIMV para λ en una población Poisson.

Estimadores no insesgados (sesgados)

Estimadores no insesgados

También podemos comparar estimadores no insesgados y escoger el más eficiente en el sentido llamado error cuadrático medio:

Definición (Error Cuadrático Medio)

Sea $\hat{\theta}$ un estimador de θ, definimos el error cuadrático medio de $\hat{\theta}$ como

$$ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

Estimadores no insesgados

Observaciones:

1. Si $\hat{\theta}$ es insesgado, $ECM(\hat{\theta}) = V(\hat{\theta})$.
2. Para cualquier $\hat{\theta}$, $ECM(\hat{\theta}) = V(\hat{\theta}) + [Sesgo(\hat{\theta})]^2$. La demostración de éste hecho es la siguiente:

$$\begin{aligned} ECM(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2 + 2[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] + [E(\hat{\theta}) - \theta]^2\} \\ &= V(\hat{\theta}) + 2[E(\hat{\theta}) - \theta]E[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]^2 \\ &= V(\hat{\theta}) + 2[E(\hat{\theta}) - \theta]\{E(\hat{\theta}) - E[E(\hat{\theta})]\} + [E(\hat{\theta}) - \theta]^2 \\ &= V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2, \end{aligned}$$

por lo tanto,

$$ECM(\hat{\theta}) = V(\hat{\theta}) + [Sesgo(\hat{\theta})]^2.$$

El error cuadrático medio es entonces usado para determinar eficiencias relativas de estimadores no necesariamente insesgados.

Estimadores no insesgados

Ejemplo

En una poblacion normal, tanto S^2 como S_n^2 son estimadores de σ^2. ¿Cuál es más eficiente usando el criterio de ECM?

$$E(S^2) = \sigma^2, \qquad V(S^2) = \frac{2\sigma^4}{n-1}$$

$$\Rightarrow ECM(S^2) = \frac{2\sigma^4}{n-1} \text{ por ser insesgado.}$$

$$E(S_n^2) = \frac{n-1}{n}\sigma^2,$$

$$V(S_n^2) = V\left(\frac{n-1}{n}S^2\right) = \left[\frac{n-1}{n}\right]^2 V(S^2)$$

$$= \frac{(n-1)^2}{n^2} \cdot \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2},$$

Ejemplo (Cont.)

$$
\begin{aligned}
Sesgo(S_n^2) &= E(S_n^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{(n-1)\sigma^2 - n\sigma^2}{n} = -\frac{\sigma^2}{n}, \\
\Rightarrow ECM(S_n^2) &= V(S_n^2) + [Sesgo(S_n^2)]^2 \\
&= \frac{2(n-1)\sigma^4}{n^2} + \left[-\frac{\sigma^2}{n}\right]^2 = \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} \\
&= \frac{2(n-1)\sigma^4 + \sigma^4}{n^2} \\
&= \frac{(2n-1)\sigma^4}{n^2}.
\end{aligned}
$$

Ejemplo (Cont...)

Eficiencia (según el error cuadrático medio)

$$\begin{aligned}\frac{ECM(S^2)}{ECM(S_n^2)} &= \frac{\frac{2\sigma^4}{n-1}}{\frac{(2n-1)\sigma^4}{n^2}} \\ &= \frac{2n^2\sigma^4}{(2n-1)(n-1)\sigma^4} \\ &= \frac{2n^2}{(2n-1)(n-1)} \\ &= \left(\frac{2n}{2n-1}\right)\left(\frac{n}{n-1}\right) > 1.\end{aligned}$$

Por lo tanto, $ECM(S_n^2) < ECM(S^2)$; es decir, S_n^2 es mas eficiente que S^2 por el criterio de ECM.

Estimadores no insesgados

Notas:

1. Aún cuando S^2 es un estimador insesgado para σ^2,$S = \sqrt{S^2}$ *no* es un estimador insesgado de σ.
2. En general el error cuadrático medio nos permite calcular la eficiencia relativa de dos estimadores cualesquiera insegados o sesgados.

Mencionaremos dos propiedades más, entiende la definición y trata de digerir los principales resultados que se desprenden de éstas, los cuales encontrarás subrayados.

Otras Propiedades de Estimadores Puntuales

Consistencia

Un estimador se dice que es consistente si para n grande, el estimador toma con una alta probabilidad valores cercanos al parámetro que estima.

Definición

$\hat{\theta}$ es un estimador consistente para θ si para cualquier constante $c > 0$

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < c) = 1 \text{ ó, equivalentemente, } \lim_{n\to\infty} P(|\hat{\theta} - \theta| > c) = 0.$$

Consistencia

Ejemplo

Considerar una muestra aleatoria de tamaño n de una poblacion uniforme $[0, \theta]$ y sea $\hat{\theta} = X_{(n)}$ un estimador de θ.

La función de densidad de la población

$$f(x) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{en otra parte} \end{cases}.$$

La función de distribución acumulada

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{\theta} & 0 < x < \theta \\ 1 & x > \theta \end{cases}.$$

Ejemplo (Cont...)

La densidad del n-ésimo estadístico de orden es

$$f_{X_{(n)}}(y) = n\frac{y^{n-1}}{\theta^n} \qquad 0 < y < \theta$$

$$\Rightarrow P(|\hat{\theta} - \theta| < c) = P(-c < \hat{\theta} - \theta < c) = P(\theta - c < \hat{\theta} < \theta + c)$$

$$= P(\theta - c < X_{(n)} < \theta + c) = \int_{\theta-c}^{\theta} \frac{ny^{n-1}}{\theta^n} dy = \frac{y^n}{\theta^n}\Big|_{\theta-c}^{\theta} = 1 - \frac{(\theta-c)^n}{\theta^n}$$

$$\lim_{n\to\infty}\left[1 - \frac{(\theta-c)^n}{\theta^n}\right] = 1 - \lim_{n\to\infty}\left(\frac{(\theta-c)}{\theta}\right)^n$$

$$\text{como } \frac{\theta-c}{\theta} < 1 \implies \lim_{n\to\infty}\left(\frac{(\theta-c)}{\theta}\right)^n = 0$$

y por lo tanto

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < c) = 1,$$

o en otras palabras, $\hat{\theta}$ es consistente para θ.

Consistencia

Teorema (Condición suficiente pero no necesaria)

Un estimador $\hat{\theta}$ de θ que satisface que:

$$\lim_{n\to\infty} ECM(\hat{\theta}) = 0$$

es un estimador consistente de θ.

Observaciones

1. Si $\hat{\theta}$ es insesgado y su varianza es tal que $\lim_{n\to\infty} V(\hat{\theta}) = 0$ entonces $\hat{\theta}$ es consistente.
2. Para el caso en que $ECM(\hat{\theta})$ no tienda a cero cuando $n \to \infty$, no se puede concluir nada con este teorema, es decir, $\hat{\theta}$ puede o no ser consistente. Habría que usar un metodo alternativo (la definición) para verificar la consistencia del estimador.

Consistencia

Ejemplo

Considera una poblacion Normal $N(\mu, \sigma^2)$ y una muestra aleatoria de tamano n, *entonces* S^2 *es un estimador consistente de* σ^2.

$$E(S^2) = \sigma^2 \qquad V(S^2) = \frac{2\sigma^4}{n-1}$$
$$ECM(S^2) = V(S^2) + [Sesgo(S^2)]^2 = \frac{2\sigma^4}{n-1} + 0$$
$$\lim_{n\to\infty} ECM(S^2) = \lim_{n\to\infty} \frac{2\sigma^4}{n-1} = 0,$$

entonces, S^2 es un estimador consistente de σ^2

Consistencia

Ejemplo

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población cualquiera con media μ y una varianza σ^2. Entonces $\overline{X}$ es un estimador consistente de μ. (Ley Débil de los Grandes Números).

$$E(\overline{X}) = \mu \qquad V(\overline{X}) = \frac{\sigma^2}{n}$$
$$ECM(\overline{X}) = V(\overline{X}) + [Sesgo(\overline{X})]^2 = \frac{\sigma^2}{n} + 0$$
$$\Rightarrow \lim_{n\to\infty} ECM(\overline{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0,$$

Entonces $\overline{X}$ es un estimador consistente de μ.

Esto es, $\overline{X}$ siempre es un estimador consistente.

Suficiencia

Se dice que un estimador $\hat{\theta}$ es suficiente para θ, si engloba o contiene toda la información que proporciona la muestra referente al parámetro θ, de tal forma que los valores individuales en la muestra pueden ser desechados para propositos de estimación de θ. En poblaciones normales, $\overline{X}$ y S^2 son estimadores suficientes para μ y σ^2.

Métodos para construir estimadores

Método de Momentos

Método de momentos

Este es un método relativamente simple para encontrar estimadores. La única propiedad que puede garantizarse de estos estimadores es la consistencia. Pero puede tener otras, uno debería verificar caso por caso.

Recordemos la definición del momento de orden r centrado en el origen de una variable aleatoria X:

$$\mu'_r = E(X^r)$$

Definición

El momento de orden r de una muestra, denotado por m'_r, se define como:

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

Método de momentos

El método de momentos consiste en formar un sistema de ecuaciones igualando los momentos muestrales con los momentos poblacionales y resolviendo con respecto a los parámetros de la poblacion. A la solución de ese sistema de ecuaciones se le llama los estimadores de momentos de los parámetros.

Al final no debemos olvidar reemplazar el símbolo del parámetro θ por el símbolo de estimador $\hat{\theta}$.

Este procedimiento es muy usado en la estimación de componentes de varianza, en la corrección del sesgo de muchos estimadores, en la propuesta de metodologías en situaciones complejas y como valores iniciales en otros procedimientos de estimación que son de carácter recursivo.

Método de momentos

Ejemplo

Considerar una población uniforme $[0, \theta]$. Encontrar el estimador de momentos para θ basandose en una muestra aleatoria de tamano n.

$$\mu_1' = E(X) = \frac{\theta}{2}, \qquad m_1' = \frac{1}{n}\sum_{i=0}^{n} x_i = \overline{x}.$$

Formamos la ecuación $\mu_1' = m_1'$, lo que produce $\frac{\theta}{2} = \overline{x}$. Resolviendo θ para encontramos $\theta = 2\overline{x}$

$$\Longrightarrow \text{el estimador de momentos para } \theta \text{ es } \hat{\theta} = 2\overline{X}.$$

Método de momentos

Ejemplo

Considerar una población Normal $N(\mu, \sigma^2)$. Basándose en una muestra aleatoria de tamano n, encontrar los estimadores de momentos para μ y σ^2.

$$\mu_1' = E(X) = \mu, \quad \mu_2' = E(X^2) = V(X) + [E(X)]^2 = \sigma^2 + \mu^2$$

porque $V(X) = E(X^2) - [E(X)]^2$

$$\Rightarrow m_1' = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}, \qquad m_2' = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Ejemplo (Cont...)

Formamos el sistema:

$$\left.\begin{array}{c} \mu = \overline{x} \\ \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{array}\right\} \Longrightarrow \left\{\begin{array}{c} \mu_1' = m_1' \\ \mu_2' = m_2' \end{array}\right.$$

$$\left\{\begin{array}{c} \mu = \overline{x} \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \end{array}\right.$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = S_n^2,$$

por lo tanto, los estimadores de momentos para μ y σ^2 son

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma}^2 = S_n^2.$$

Nota: $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

Método de Máxima Verosimilitud

Método de Máxima Verosimilitud

Este método consiste en encontrar el valor del parámetro que favorece más los valores que se obtuvieron en la muestra. En otras palabras, dados los valores en la muestra buscamos los valores de los parámetros de la población que más posibilidades tengan de representar a la población que generó a la muestra. (En palabras coloquiales, sería como obtener una muestra de sangre y en base a su composición determinar quién es más factible que sea el papá!).

Se conoce la distribución de la población (por ejemplo una del catálogo) pero falta especificar sus parámetros. Al tomar la muestra se obtendrán, en principio, valores que estén "favorecidos" con probabilidades grandes. Veamos un poco de nomenclatura que nos ayudará a poner en claro estas ideas.

Método de Máxima Verosimilitud

Supongamos que se tiene una población con parámetro θ (un número real) y que $x_1, x_2, \ldots, x_n$ son los valores de una muestra aleatoria de tamaño n, definimos la verosimilitud de la muestra como la función de densidad o de probabilidad conjunta de las variables $X_1, X_2, \ldots, X_n$ evaluada en el punto $x_1, x_2, \ldots, x_n$, se denota por:

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta).$$

Aquí se vuelve muy importante el uso de las letras mayúsculas y minúsculas para denotar a la v. a. y sus valores (respectivamente). Nota que $x_1, x_2, \ldots, x_n$ son los valores observados de la muestra. En este sentido, la función de verosimilitud, depende sólo del valor de θ.

Método de Máxima Verosimilitud

Este método consiste en maximizar esta función con respecto a θ. Al valor donde ocurre el máximo se le llama estimador de máxima verosimilitud para θ. La clave acerca del método es que no se toman como variables cada una de las x's, sino que ahora es el parámetro de la población de donde provienen las observaciones lo que consideramos como variable; dado que las x's ya tomaron un valor específico $x_1, x_2, \ldots, x_n$, en nuestra muestra, inspeccionaremos sobre los valores posibles del parámetro. Por esto, será factible usar en la mayoría de los casos las técnicas clásicas de maximización de funciones continuas.

En muchos casos es más sencillo maximizar el logaritmo natural de la función en lugar de la función directamente.Como el logaritmo natural es una función creciente, el punto donde ambas funciones alacanza el máximo es el mismo.

El estimador de máxima verosimilitud se denota como $\hat{\theta}_{MV}$.

Método de Máxima Verosimilitud

Propiedades:

- Todos los estimadores de máxima verosimilitud son estimadores **suficientes** –cuando esta clase de estimadores existen para la familia de distribuciones considerada.
- **Consistentes**.
- **Asintóticamente insesgados**.
- **Asintóticamente normales**, el TLC es válido para ellos:

$$\frac{\hat{\theta}_{MV} - \theta}{\sqrt{V(\hat{\theta}_{MV})}} \sim N(0,1), \text{ cuando n es grande.}$$

- Son **invariantes ante transformaciones continuas**; es decir, si encontramos $\hat{\theta}_{MV}$,la función $g(\hat{\theta}_{MV})$ será un estimador de máxima verosimilitud para $g(\theta)$. Este resultado se puede generalizar para funciones en general.

Método de Máxima Verosimilitud

Pasos para encontrar el estimador de maxima verosimilitud:

1. Encontrar la verosimilitud $L(\theta)$.
2. Sacar el logaritmo natural de $\ell(\theta) = \log(L(\theta))$ (No obligatorio).
3. Maximizar la verosimilitud $L(\theta)$, o bien su logaritmo $\ell(\theta)$, con respecto a θ.

Nota: En muchos casos, el proceso de maximización genera un sistema de ecuaciones no lineales que solo puede ser resuelto por métodos numéricos.

Método de Máxima Verosimilitud

Ejemplo

Supongamos que tenemos una poblacion Poisson (λ). Encontrar el estimador de máxima verosimilitud para λ basándose en una muestra aleatoria de tamaño n .

La función de probabilidad de la población es $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$. Así, la función de verosimilitud es

$$L(\lambda) = \left(\frac{\lambda^{x_1} e^{-\lambda}}{x_1!}\right) \cdot \left(\frac{\lambda^{x_2} e^{-\lambda}}{x_2!}\right) \cdots \left(\frac{\lambda^{x_n} e^{-\lambda}}{x_n!}\right) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!},$$

y por tanto,

$$\log L(\lambda) = (\sum_{i=1}^{n} x_i) \log \lambda - n\lambda - \log \left[\prod_{i=1}^{n} x_i!\right].$$

Ejemplo (Cont...)

Derivando e igualando a cero para obtener puntos críticos, obtenemos

$$\frac{d}{d\lambda}\log(L(\lambda)) = \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\lambda} - n.$$

Así que,

$$\left(\sum_{i=1}^{n} x_i\right)\frac{1}{\lambda} - n = 0 \quad \Longrightarrow \quad \text{el punto crítico es } \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Ejemplo (Cont...)

Verificamos que efectivamente es un máximo por el método de la segunda derivada:

$$\frac{d^2}{d\theta^2}\log(L(\lambda)) = -\left(\sum_{i=1}^{n} x_i\right)\frac{1}{\lambda^2},$$

evaluando en el punto crítico $\hat{\lambda} = \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)$, obtenemos

$$\frac{d^2}{d\theta^2}\log(L(\lambda))\Big|_{\hat{\lambda}} = -\left(\sum_{i=1}^{n} x_i\right)\cdot\frac{1}{\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)} = -\frac{n^2}{(\sum_{i=1}^{n} x_i)} < 0.$$

Esto implica que el punto $\hat{\lambda}$ nos da un maximo de la funcion, por lo tanto $\hat{\lambda} = \overline{X}$ es el estimador de maxima verosimilitud para λ.

Método de Máxima Verosimilitud

Ejemplo

Supóngase que el tiempo de espera del transporte colectivo de Juan López se distribuye uniformemente en el intervalo $[0, \theta]$. Como regularmente se aburre parado en la esquina de su casa, un día decidió tomar de cuando en cuando los tiempos que pasaba en tan peculiar lugar, planeando para ello la recolección de n de ellos en días seleccionados al azar. Denotemos estos resultados por: $x_1, x_2, \ldots x_n$.

Encontrar el estimador de máxima verosimilitud para θ basándose en los valores observados de la m. a. de tamaño n.

Aquí, X es el tiempo de espera del transporte.

Encontremos el valor del parámetro θ que favorecería más los valores que se obtuvieron en la muestra.

Ejemplo (Cont...)

La funcion de densidad es $f(x) = \frac{1}{\theta}$, $0 < x < \theta$. Así,

$$L(\theta) \equiv L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdots \frac{1}{\theta}, \qquad 0 < x_1 < \theta, \ldots, 0 < x_n < \theta,$$

la cual puede ser escrita, y así cumplir cada desigualdad, como

$$L(\theta) = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdots \frac{1}{\theta} = \frac{1}{\theta^n},$$

para $\theta > x_{(n)}$, donde $x_{(n)} = max(x_i)$. Entonces,

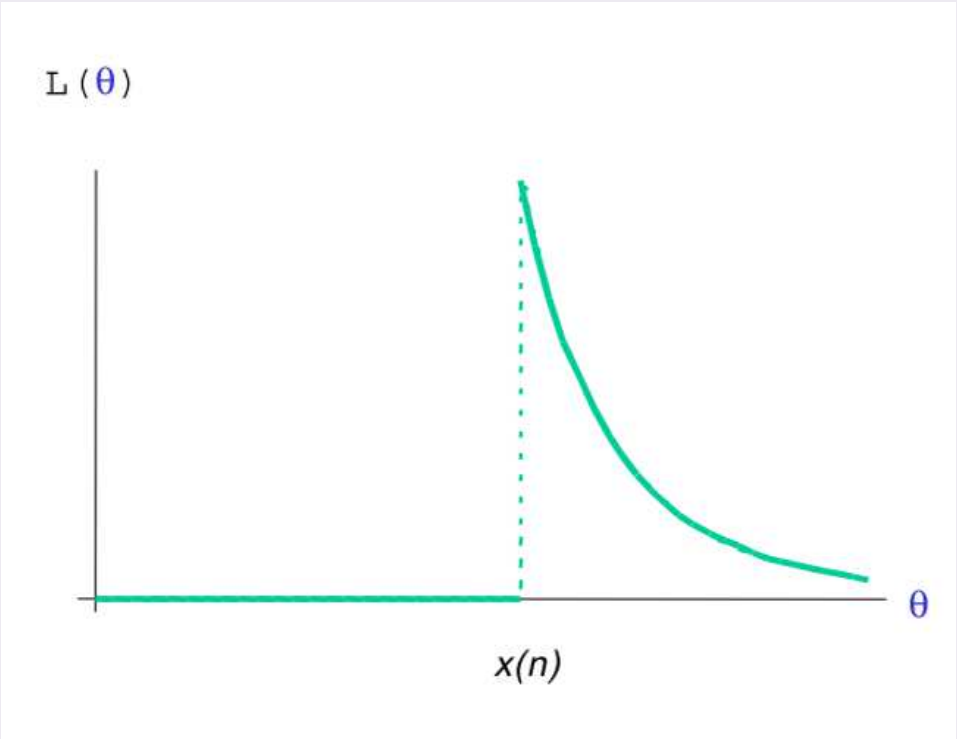
$$L(\theta) = \frac{1}{\theta^n}, \quad \theta > x_{(n)},$$

$$\log L(\theta) = -n \log \theta, \quad \theta > x_{(n)}.$$

Ejemplo (Cont...)

$$\frac{d \log L(\theta)}{d\theta} = -\frac{n}{\theta} \quad \Rightarrow \quad -\frac{n}{\theta} = 0 \text{ ?????}$$

No funciona el método de la derivada.



Ejemplo (Cont...)

Por inspección (ver gráfica) de la función $L(\theta) = \frac{1}{\theta^n}$ (se hace cero antes del valor de x_n) y sabiendo que $\theta > x_{(n)}$, vemos que es maximizada para el menor valor que pueda tomar θ; es decir, cuando $\hat{\theta} = X_{(n)}$; por lo tanto el estimador de máxima verosimilitud para θ es

$$\hat{\theta} = X_{(n)}$$

Entonces, si los tiempos de espera observados por Juan fueron 7.5, 5, 11, 10.5, 4, y 2.3 minutos, el estimador de máxima verosimilitud es $\hat{\theta}$ =11 minutos.

La diferencia entre los casos de los ejemplos previos, radica en el hecho de que en este segundo caso, los valores de la variable aleatoria, dependen del valor del parámetro y esto dificulta el proceso de maximización.

Método de Máxima Verosimilitud

Ejemplo

Un método usado muy a menudo para estimar el tamaño de una población animal (peces, conejos, etc.) consiste en realizar un experimento de captura/recaptura. En este experimento, una muestra inicial de k animales es capturada, cada uno de estos animales es etiquetado y entonces liberado para que se reintegre a su hábitat. Después de permitir que pase un tiempo razonable para que los individuos etiquetados se mezclen con la población, otra muestra de tamaño n es capturada.

Sea X el número de animales etiquetados en la segunda muestra. Usar el valor observado x para estimar el tamaño poblacional N.

Ejemplo (Cont...)

Se están realizando:

- n observaciones, animales capturados, de un conjunto total de N posibles.
- la probabilidad de éxito, animal capturado con etiqueta, en cada observación cambia "paso a paso".
- se conoce k, cuantos elementos tienen etiqueta (éxitos) en nuestra población.

Nuestra población es finita (N) y no hay independencia entre observaciones; esto es, el resultado de cada observación será afectado por los resultados anteriores.

Se realiza un muestreo sin reemplazo. Los valores de la v.a. dependen de los valores de k, n y N.

Ejemplo (Cont...)

¿Qué ley de comportamiento, distribución de probabilidad, tiene nuestra v.a.?

De las anteriores consideraciones y consultando nuestro catálogo podemos afirmar que la ley de comportamiento asociada con nuestra v.a. es la distribución hipergeométrica (ver notas anteriores)

$$f(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad max\{0, n-(N-k)\} \leq x \leq min\{n, k\}.$$

Esto es, $X \sim Hiper(N, k, n)$.

Ejemplo (Cont...)

Nos interesa estimar el tamaño de la población N. De todos los valores posibles del parametro $\theta = N$, se escogerá aquel que haga que la probabilidad de que en la muestra se obtenga el valor que ya se obtuvo, sea máxima

Se observó entonces $X = x$, así

$$L(\theta) = \frac{\binom{k}{x}\binom{\theta-k}{n-x}}{\binom{\theta}{n}}.$$

Observación: la variable no es x sino que ahora la variable es el parámetro de la población de donde proviene la observación puesto que X ya tomó un valor específico x.

Ejemplo (Cont...)

Es clara la dificultad para hallar el valor de θ que maximiza la función de verosimilitud, en vista de los factoriales que intervienen. Para vencer esta dificultad, observa que al efectuar la división $\frac{n!}{(n-1)!}$, el factorial desparece: $\frac{n!}{(n-1)!} = n$; y que si dividimos $\frac{\binom{n}{x}}{\binom{n-1}{x}}$ obtenemos: $\frac{\frac{n!}{(n-x)!x!}}{\frac{(n-1)!}{(n-1-x)!x!}} = \frac{n}{n-x}$.

Así, podemos considerar la siguiente división con el fin de eliminar los factoriales:

$$\frac{L(\theta)}{L(\theta-1)} = \frac{\frac{\binom{k}{x}\binom{\theta-k}{n-x}}{\binom{\theta}{n}}}{\frac{\binom{k}{x}\binom{\theta-1-k}{n-x}}{\binom{\theta-1}{n}}} = \frac{(\theta-k)(\theta-n)}{\theta(\theta-k-n+x)} = \frac{\theta^2-\theta n-k\theta+kn}{\theta^2-\theta n-k\theta+\theta x}$$

Ejemplo (Cont...)

Entonces si la función de verosimilitud es evaluada en el valor máximo de θ y dividida entre la función de verosimilitud evaluada en otro valor de θ (en $\theta - 1$), el cociente debería ser mayor que 1:

$$\frac{L(\theta)}{L(\theta-1)} = \frac{\theta^2 - \theta n - k\theta + kn}{\theta^2 - \theta n - k\theta + \theta x} > 1$$

y para que esto se cumpla, $\frac{kn}{\theta x}$ debe ser mayor que 1:

$$\frac{kn}{\theta x} > 1, \Longrightarrow \frac{n}{x} > \frac{\theta}{k}$$

de donde

$$\theta < \frac{nk}{x}.$$

Ejemplo (Cont...)

Ahora, θ es un entero positivo, de donde, si $\frac{nk}{x}$ da un número con decimales, θ deberá tomar el valor entero inmediatamente anterior. Por lo tanto,

$$\hat{\theta} = \text{el entero inferior más cercano a la proporción } \frac{nk}{x}.$$

Observa que esto concuerda con nuestro tratamiento en la distribución hipergeométrica, en el sentido de que la proporción $\frac{n}{x}$ (muestral), nos da información de la proporción $\frac{\theta}{k}$(poblacional).

Ejemplo (Cont...)

Para fijar ideas supóngamos que $k = 200$ peces son tomados de un lago etiquetados y regresados al lago. Posteriormente $n = 100$ peces son recapturados; de entre los 100 hay $x = 11$ etiquetados. De esta información encontramos que

$$\begin{aligned}\hat{\theta} &= \text{el entero inferior más cercano a } \frac{nk}{x} \\ &= \text{el entero inferior más cercano a } \frac{(200)(100)}{11} = 1818.\end{aligned}$$

Método de Máxima Verosimilitud

Ejemplo

Un asistente del laboratorio X del Campus Monterrey, se encuentra en la etapa de recolección de información para tratar de solventar algunas hipótesis que plantea en su trabajo de tesis. Sus experimentos consisten en registrar los tiempos de vida de ciertas componentes.

Al tiempo $t = 0$, veinte componentes idénticas son sometidas a una prueba. La distribución de tiempos de vida de cada una es exponencial con parámetro λ (esto lo sabe por las características globales de las componentes y basándose en información obtenida en su revisión bibliográfica).

El estudiante se fastidia de estar esperando la finalización del experimento, dado que dura muchas horas y abandona la prueba sin monitorearla.

Ejemplo

A su regreso, el edificio esta cerrado y no logra convencer al guardia de que debe entrar así que para cuando logra el acceso han transcurrido 24 horas. Finaliza inmediatamente la prueba después de notar que $y = 15$ de las 20 componentes aún están en operación (han fallado 5).

Obtener el estimador de máxima verosimilitud de λ.

Solución. Si conocieramos las duraciones de cada componente (los resultados de una m.a.) podríamos usar el EMV de la distribución exponencial (ver uno de los ejemplos anteiores). Sin embargo, no tenemos esa información.

Ejemplo (Cont...)

Si recuerdas hay una relación ente la distribución exponencial y la distribución Poisson, en particular hay una relación entre sus parámetros. Pero, de nuevo, no tenemos la información del proceso Poisson correspondiente.

Lo que tenemos es la información de un experimento Binomial: # de pruebas fijo, componentes idénticas de donde probabilidad de "éxito" (falla de componente) es constante bajo la condición de 24 horas de prueba, y hay independencia.

También nos damos cuenta que la probabilidad de "éxito" ($=p$) está relacionada con la distribución exponencial, ya que $p =$probabilidad de que una componente siga funcionando después de 24 horas, y esto se calcula mediante la distribución exponencial.

Ejemplo (Cont...)

De las consideraciones anteriores vemos que es factible atacar el problema mediante la distribución binomial y de ahí conectar con la distribución exponencial.

Aquí, $X =$ el tiempo de vida de una componente, y se sabe que $X \sim Exp(\lambda)$. Se nos da la información de que se probaron 20 y de cuántas continúan funcionando. Observa que esta información se puede representar como los resultados de un experimento binomial (o de 20 repeticiones de experimentos Bernoulli), esto es, Y=el número de componentes que sobreviven 24 horas, de donde $Y \sim Bin(n = 20, p =?)$.

No conocemos p, sin embargo, $p =$probabilidad de que una cualquiera de las componentes dure las 24 horas, esto equivale a $P(X \geq 24)$.

Ejemplo (Cont...)

Es decir,

$$p = 1 - P(X \leq 24) = 1 - F(24) = e^{-\frac{24}{\lambda}}.$$

Observa que si encontramos un estimador para p, encontraremos un estimador para λ (por la propiedad de los estimadores de máxima verosimilitud de que si $\hat{\theta}$ es un estimador EMV para θ,$g(\hat{\theta})$ será un EMV para $g(\theta)$, donde $g(\theta)$ es una función uno a uno). Busquemos pues el EMV para p.

Ejemplo (Cont...)

Podemos ver el experimento como una m. a. de experimentos Bernoulli, donde cada resultado es el valor de una v.a. $Bernoulli(p)$, así

$$L(p) = \prod_{i=1}^{20} (1-p)^{1-x_i} p^{x_i} = (1-p)^{(1-x_1)} p^{x_1} \cdots (1-p)^{(1-x_{20})} p^{x_{20}},$$

la cual se puede escribir como

$$L(p) = (1-p)^{\sum_{i=1}^{20}(1-x_i)} p^{\sum_{i=1}^{20} x_i} = (1-p)^{(20-\sum_{i=1}^{20} x_i)} p^{\sum_{i=1}^{20} x_i},$$

$$\log[L(p)] = (20 - \sum_{i=1}^{20} x_i) \log(1-p) + (\sum_{i=1}^{20} x_i) \log p,$$

$$\frac{d}{dp}\{\log[L(p)]\} = \frac{(20 - \sum_{i=1}^{20} x_i)}{(1-p)}(-1) + \frac{(\sum_{i=1}^{20} x_i)}{p} = 0.$$

Ejemplo (Cont...)

Despejando a p

$$p = \frac{1}{20}\sum_{i=1}^{20} x_i = \overline{x}.$$

Para comprobar que corresponde a un máximo sacamos la segunda derivada de $\log[L(p)]$

$$\frac{d^2}{dp^2}\{\log[L(p)]\}\Big|_{p=\overline{x}} = \frac{(20 - \sum_{i=1}^{20} x_i)}{(1-p)^2}(-1) - \frac{(\sum_{i=1}^{20} x_i)}{p^2}\Big|_{p=\overline{x}}.$$

Ejemplo (Cont...)

Evaluando en $p = \overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i$, obtenemos

$$\frac{d^2}{dp^2}\{\log[L(p)]\}\Big|_{p=\overline{x}} = \frac{(20 - \sum_{i=1}^{20} x_i)}{(1 - \frac{1}{20}\sum_{i=1}^{20} x_i)^2}(-1) - \frac{(\sum_{i=1}^{20} x_i)}{(\frac{1}{20}\sum_{i=1}^{20} x_i)^2}$$

$$= -20\left[\frac{1}{1 - \frac{1}{20}\sum_{i=1}^{20} x_i} + \frac{1}{\frac{1}{20}\sum_{i=1}^{20} x_i}\right] < 0.$$

Por lo tanto

$$\hat{p} = \overline{X}$$

es el estimador de máxima verosimilitud de p.

Ejemplo (Cont...)

Ahora, sabemos que $p = e^{-\frac{24}{\lambda}}$, entonces

$$\lambda = -\frac{24}{\log(p)},$$

esto es,

$$\hat{\lambda} = -\frac{24}{\log(p)} = -\frac{24}{\log(\frac{15}{20})} = 83.42$$

es el EMV para λ.

Ejemplo

Suponer una población Exponencial(θ). Obtener el estimador de máxima verosimilitud para θ, basándose en una muestra aleatoria de tamaño n.

La función de densidad está dada por $f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ para $x > 0$. Así que, la verosimilitud es

$$\begin{aligned} L(\theta) &= \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_2}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_3}{\theta}} \cdots \frac{1}{\theta} e^{-\frac{x_n}{\theta}}, \quad x_i > 0 \\ &= \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^{n} x_i} \end{aligned}$$

$$\Longrightarrow \log L(\theta) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i.$$

Ejemplo (Cont...)

Derivando con respecto a θ obtenemos

$$\frac{d}{d\theta}\log L(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}$$

Igualando ahora a cero para obtener puntos críticos:

$$\begin{aligned} -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} &= 0 \\ \Rightarrow \quad \frac{-n\theta + \sum_{i=1}^{n} x_i}{\theta^2} &= 0, \end{aligned}$$

de donde se encuentra el punto críctico

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

Ejemplo (Cont...)

Además, se tiene que

$$\frac{d^2}{d\theta^2}\log L(\theta) = \frac{n}{\theta^2} - \frac{2\sum_{i=1}^n x_i}{\theta^3} = \frac{n\theta - 2\sum_{i=1}^n x_i}{\theta^3}$$

Notemos que

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} \quad \Longrightarrow \quad \left.\frac{d^2}{d\theta^2}\log L(\theta)\right|_{\hat{\theta}} = \frac{n\frac{\sum_{i=1}^n x_i}{n} - 2\sum_{i=1}^n x_i}{(\frac{\sum_{i=1}^n x_i}{n})^3}$$

$$= \frac{-n^3}{\left(\sum_{i=1}^n x_i\right)^2} < 0.$$

$\therefore \hat{\theta} = \overline{X}$ es el estimador de máxima verosimilitud para θ.

Método de Máxima Verosimilitud

El método de máxima verosimilitud puede ser extendido para el caso en que haya más de un parámetro en la población, lo único que hay que hacer es maximizar la verosimilitud con respecto a todos los parámetros en forma simultánea.

Ejemplo

Suponer una población normal(μ, σ^2) . Encontrar los estimadores de máxima verosimilitud para μ y σ^2basándose en una muestra aleatoria de tamaño n.

La función de densidad de cada X_i es $f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2}$. Por lo tanto, la funcion de verosimilitud es

Ejemplo (Cont...)

$$L(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_1-\mu}{\sigma})^2} \cdots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_n-\mu}{\sigma})^2}$$

$$= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\Longrightarrow$$

$$\log L(\mu,\sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Longrightarrow$$

$$\frac{\partial}{\partial\mu}\log L(\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$\Longrightarrow$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0; \sum_{i=1}^{n}x_i - n\mu = 0 \;\Rightarrow\; \mu = \frac{\sum_{i=1}^{n}x_i}{n} = \overline{x}$$

Ejemplo (Cont...)

Ahora,

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0,$$

entonces

$$\frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{n}{2\sigma^2},$$

$$\sum_{i=1}^{n} (x_i - \mu)^2 = n\sigma^2,$$

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

Ejemplo (Cont...)

Por lo tanto, los estimadores para μ y σ^2 son:

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Faltaría verificar que este punto efectivamente nos da un máximo de la función pero para poderlo hacer necesitamos resultados más elaborados de cálculo avanzado. Omitiremos esa parte aquí.

$\therefore$ Los estimadores de máxima verosimilitud para μ y σ^2 son:

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Ejemplo (Cont...)

Es claro además, que por las propiedades de invarianza, el estimador de MV para σ es sencillamente

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Método de Máxima Verosimilitud

Ejercicio. Considera una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una f.d.p. Exponencial Recorrida

$$f(x; \lambda, \theta) = \begin{cases} \lambda e^{-\lambda(x-\theta)} & x \geq \theta \\ 0 & \text{en otra parte} \end{cases}.$$

Tomando $\theta = 0$ se obtiene la distribución exponencial considerada previamente (θ : *threshold*).

En el flujo vehicular, "Tiempo Pico" es el tiempo transcurrido entre el tiempo que un automóvil acaba de pasar por un punto fijo y el instante en el que el siguiente automóvil comienza a pasar por ese mismo punto. Suponiendo que la v.a. $X =$*el "tiempo pico" para dos autos consecutivos tomados al azar en una carretera durante periodos de "horas pico"* tiene la distribución exponencial recorrida, haz lo siguiente.

Método de Máxima Verosimilitud

1. Hallar los estimadores de máxima verosimilitud de θ y λ.
2. Al hacer 10 observaciones de tiempos pico, se obtuvieron los siguientes valores (en segundos):

   3.11, 0.64, 2.55, 2.20, 5.44, 3.42, 10.39, 8.93, 17.82, y 1.30.

   Calcular los estimados de θ y λ.

Existe otro método de estimación que se conoce como Mínimos Cuadrados, el cual juega un papel primordial cuando trabajamos con modelos lineales, en donde el valor medio de una variable aleatoria es modelado como una función lineal de otros factores (típicamente no considerados aleatorios). Esta metodología es discutida a detalle en el curso de Estadística Multivariada.

Estimación por Intervalos

Intervalos de Verosimilitud

Más allá de estimadores puntuales

Usando la función de verosimilitud podemos obtener el estimador máximo verosímil, pero esta función también da información sobre el "ranking" de otros valores del parámetro θ, así podemos decir si un valor θ_0 "es casi" tan bueno como $\hat{\theta}$ o si este valor θ_0 es muy inverosímil, dada las observaciones que se tienen.

Para poder hablar de la verosimilitud de un conjunto de valores de parámetros, conviene estandarizar $L(\theta) = L(\theta, \boldsymbol{x}) = f(\boldsymbol{x}, \theta)$. Definimos la verosimilitd relativa como

$$\Lambda(\theta) = \frac{L(\theta, \boldsymbol{x})}{L(\hat{\theta}, \boldsymbol{x})}$$

y la log verosimilitud relativa como

$$r(\theta) = log(L(\theta, \boldsymbol{x})) - log(L(\hat{\theta}, \boldsymbol{x}))$$

donde $\hat{\theta}$ es el estimador máximo verosímil (EMV).

Intervalos (regiones) de verosimilitud

Definición

La región al $1-\alpha\%$ de verosimilitud es el conjunto de valores de θ tal que tienen verosimilitud relativa mayor o igual a $1-\alpha$. Esto es

$$\begin{aligned} R_\theta(1-\alpha) &= \{\theta \in \Theta | \Lambda(\theta) \geq 1-\alpha\} \\ &= \{\theta \in \Theta | r(\theta) \geq log(1-\alpha)\} \end{aligned}$$

Ejemplo

Sea $x_1, \ldots, x_n$ una m.a. $\exp(\theta)$. Calculamos su verosimilitud y verosimilitud relativa.

$$L(\theta) = \left(\frac{1}{\theta}\right)^n \exp\left\{\sum_{i=1}^{n} x_i/\theta\right\}$$

$$\hat{\theta} = \sum_{i=1}^{n} x_i/n$$

$$\Lambda(\theta) = \left(\frac{\sum x_i}{n\theta}\right)^n \exp\left\{n - \sum x_i/\theta\right\}$$

$$r(\theta) = n\left[log(\sum x_i) - log(n) - log(\theta)\right] + n - \sum x_i/\theta$$

Intervalos de Confianza

Inclusión de la Variabilidad

En la estimación de un parámetro para una población dada, además de sólo asignarle un valor único (obtenido del estimador), es factible incluir de alguna forma la información correspondiente a la variabilidad del estimador.

Por ejemplo, si queremos estimar μ, sabemos que podemos usar el valor de $\hat{\theta} = \overline{X}$, pero también sabemos que $V(\overline{X}) = \frac{\sigma^2}{n}$. ¿Cómo incluir esta información?

Una primera forma de incluir esta información es:

$$\hat{\theta} \pm \sqrt{V(\hat{\theta})}.$$

Inclusión de la Variabilidad

- con signo menos tenemos el extremo izquierdo de un intervalo localizado a una distancia de una desviación estándar del valor del estimador,
- con el signo más tenemos el extremo derecho de un intervalo localizado a una distancia de una desviación estándar del valor del estimador.

Ejemplo: Si en un problema obtenemos $\overline{X} = 3$ y $V(\overline{X}) = \frac{\sigma^2}{n} = 4$, entonces se podría reportar el intervalo

$$3 \pm \sqrt{4}.$$

Sin embargo, si $V(\overline{X}) = 0.4$, entonces el intervalo

$$3 \pm \sqrt{0.4}$$

nos da más información que el primer intervalo (recuerda la desigualdad de Chebyshev, por ejemplo).

Inclusión de la Variabilidad: Ejemplo

La desigualdad de Chevyshev:

Para cuaquier variable aleatoria X con primer y segundo momento finitos, se tiene la siguiente desigualdad

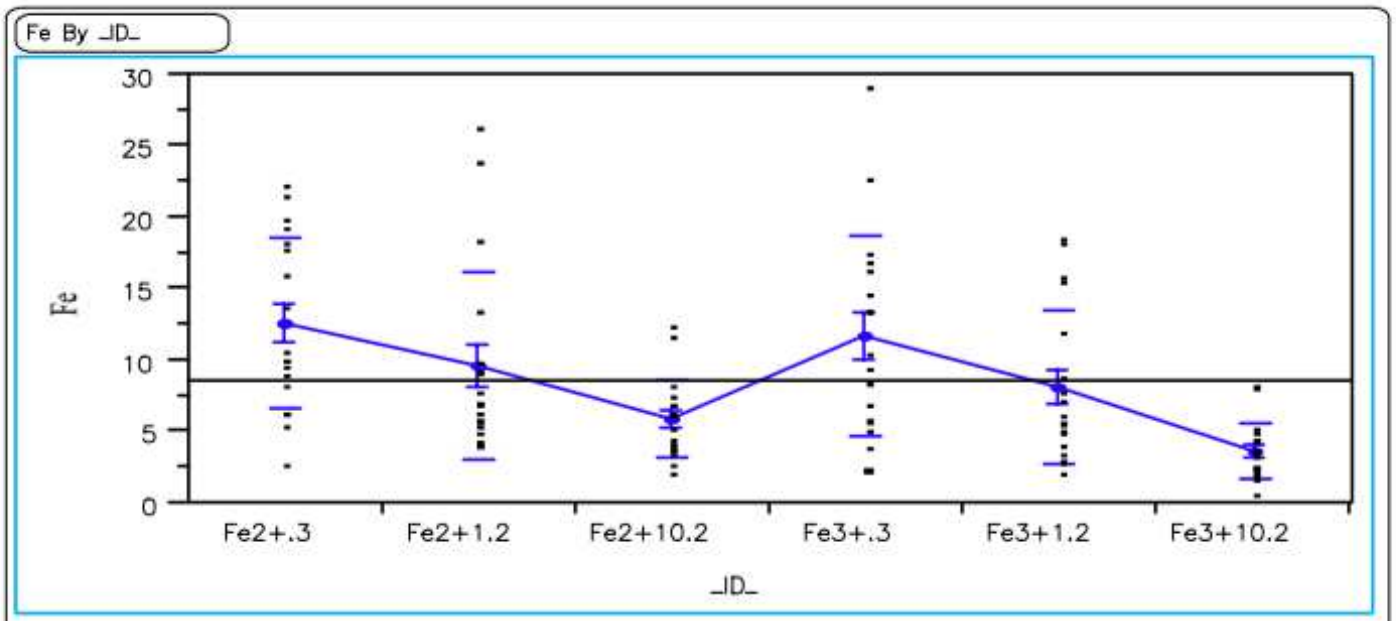
$$P(|X - \mu| < \epsilon) > 1 - \frac{\sigma^2}{\epsilon^2}$$

donde $E(X) = \mu$, $Var(X) = \sigma^2$ y $\epsilon > 0$.

Entonces para un ϵ fijo, entre más pequeña sea la varianza de X, más grande es la probabilidad de que X esté alrededor de su media μ

Inclusión de la Variabilidad

En la práctica es común ver gráficas como:



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
Fe2+.3	18	12.6394	6.08209	1.4336
Fe2+1.2	18	9.6322	6.69122	1.5771
Fe2+10.2	18	5.9367	2.80678	0.6616
Fe3+.3	18	11.7500	7.02815	1.6566
Fe3+1.2	18	8.2039	5.44739	1.2840
Fe3+10.2	18	3.6989	2.03087	0.4787

Inclusión de la Variabilidad

De estas gráficas podemos construir algunas conclusiones preliminares. Por ejemplo, parece ser que los valores medios de fierro varían de acuerdo a las concentraciones y al tipo de Fe empledo. Así también algunos casos presentan menor variabilidad que otros.

Sin embargo, esto no nos permite más que describir la situación. La pregunta obligada es: ¿Son estas diferencias reales?

Recordemos que existe una distribución asociada con estas mediciones. En este caso estamos hablando de comparar el comportamiento de 6 poblaciones. Más adelante discutiremos algunos aspectos experimentales que nos permitirán establecer otras "propiedades" de nuestras poblaciones, en particular, el concepto de homogeneidad e independencia. Por ahora, veamos una forma "mejor" de trabajar con la variabilidad: el concepto de **Intervalo de Confianza**.

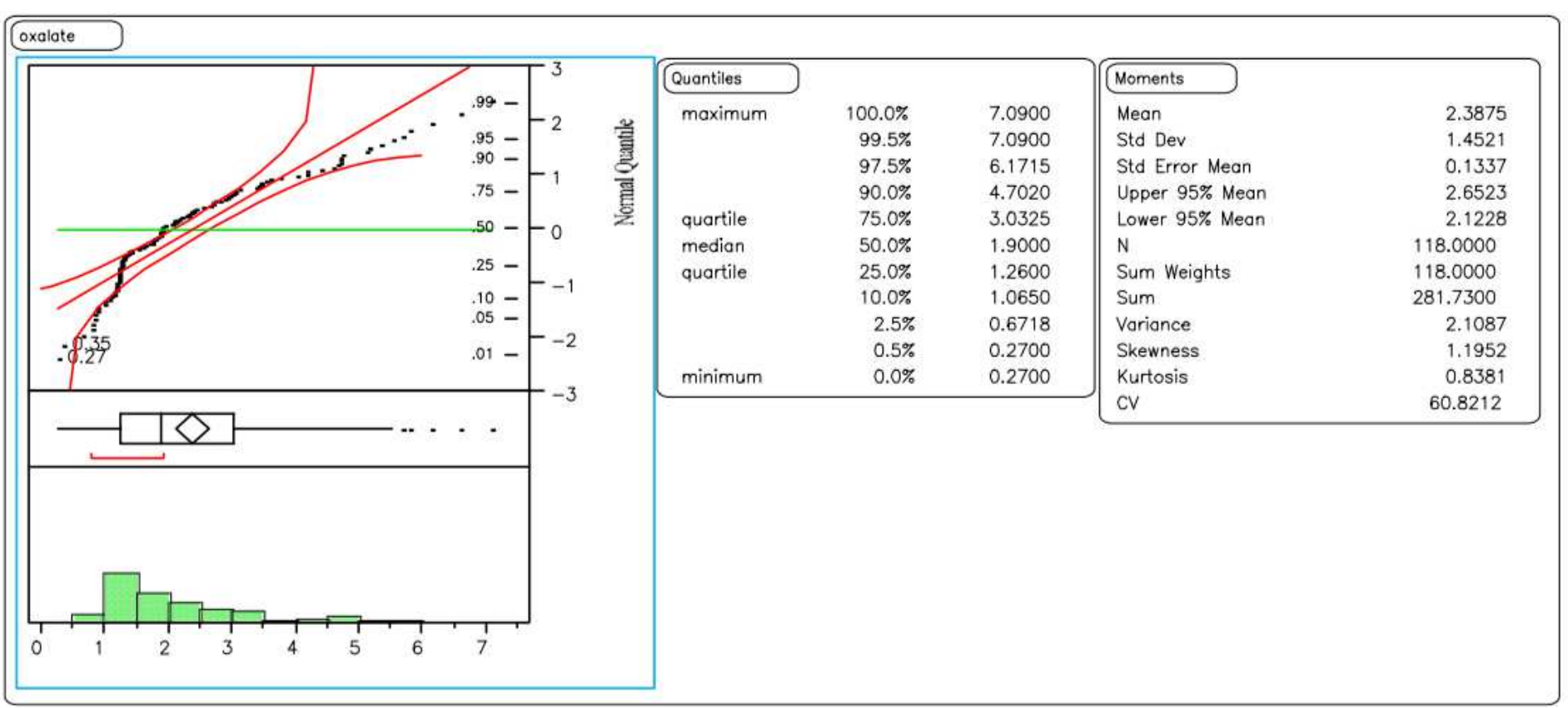
Inclusión de la Variabilidad

Veamos un ejemplo.

Ejemplo: En una rama de la industria alimentaria, se realizan en forma rutinaria mediciones del contenido de calcio en comida para animales (mascotas). El método estándar utiliza precipitación de oxalato de calcio seguida de tritanio; es una técnica que consume tiempo. Los resultados de 118 muestras (Heckman 1960), se dan en el archivo calcio.txt (columna 1). Ahora, con lo que hemos discutido hasta aquí, ¿qué podemos decir a partir del comportamiento de estos valores muestrales?

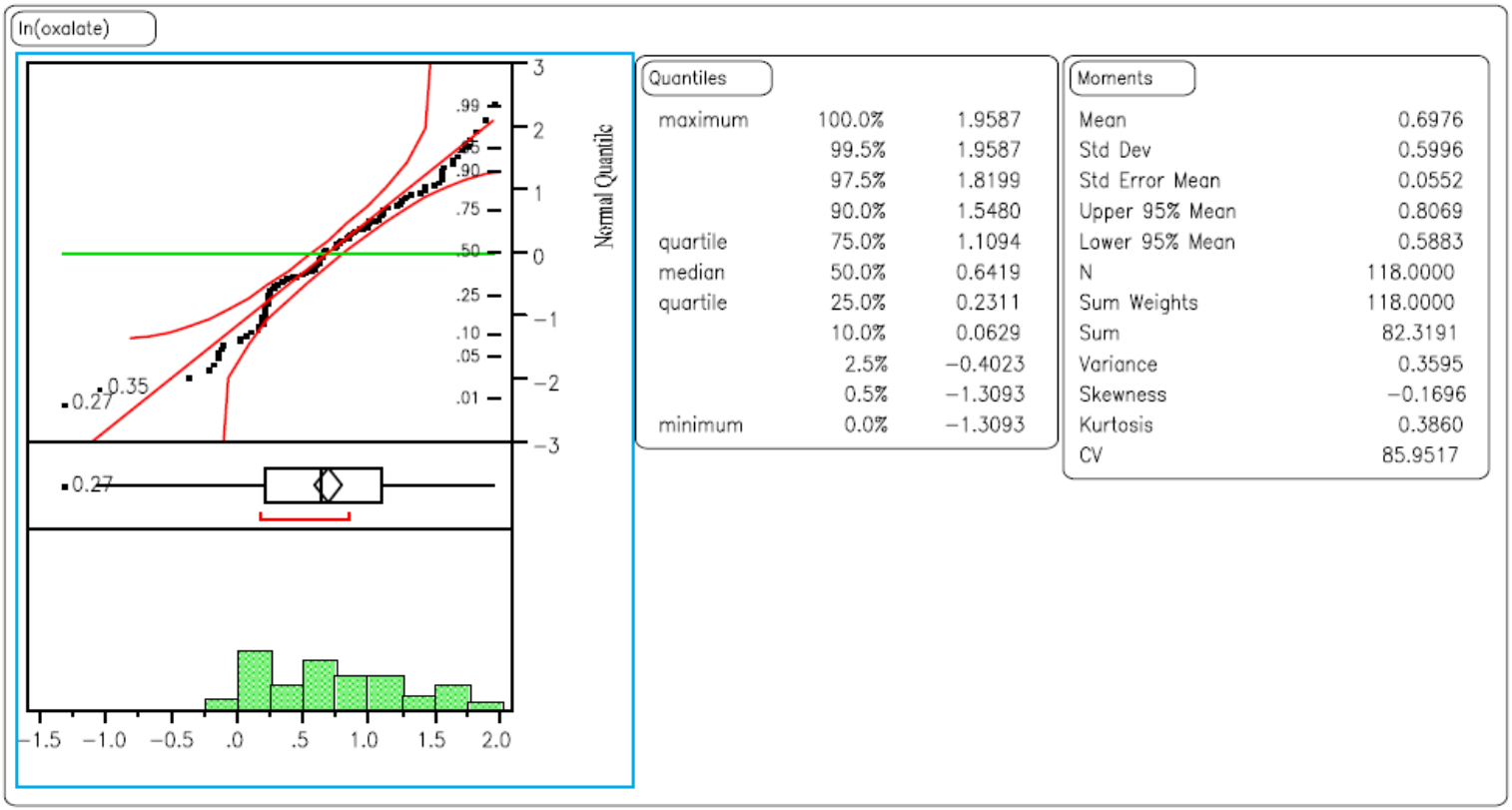
Sabemos cómo ajustar gráficamente un modelo probabilístico, y conocemos, al menos en algunos casos, lo que esperamos de nuestros valores muestrales a través de caracterizar las distribuciones muestrales. Veamos que pasa exactamemte para este ejemplo:

Inclusión de la Variabilidad



Inclusión de la Variabilidad

De las graficas anteriores y de los valores muestrales de simetría, curtosis, el desplazamiento entre la media y la mediana y los posibles valores extremos, es claro que nuestros datos no parecen provenir de una población con distribución Normal. Sin embargo, considerando una transformación de ellos, digamos sus logaritmos, observamos los siguientes cambios:



Quantiles		
maximum	100.0%	1.9587
	99.5%	1.9587
	97.5%	1.8199
	90.0%	1.5480
quartile	75.0%	1.1094
median	50.0%	0.6419
quartile	25.0%	0.2311
	10.0%	0.0629
	2.5%	−0.4023
	0.5%	−1.3093
minimum	0.0%	−1.3093

Moments	
Mean	0.6976
Std Dev	0.5996
Std Error Mean	0.0552
Upper 95% Mean	0.8069
Lower 95% Mean	0.5883
N	118.0000
Sum Weights	118.0000
Sum	82.3191
Variance	0.3595
Skewness	−0.1696
Kurtosis	0.3860
CV	85.9517

Inclusión de la Variabilidad

De estos resultados incluso nos atreveríamos a decir que la muestra posiblemente fue obtenida de una población (valores del logaritmo del contenido de calcio) normal, con la posible presencia de dos valores atípicos.

Por otra parte, como la muestra es en particular "grande", podríamos obtener el comportamiento aproximado de cantidades tales como la media muestral y la varianza muestral, usando el TLC. Para fines de ilustración pensaremos que la distribución de la población es normal (y por ende, la distribución de los valores de calcio, sería lognormal).

Inclusión de la Variabilidad

Bajo las condiciones anteriores, ¿qué nos dicen los valores de $\overline{x} = 0.6976$ y de $s^2 = 0.3595$? Estos son los estimadores puntuales de μ y σ^2 en la población. ¿Qué más podemos decir sobre μ y σ^2?

Nos gustaría no sólo describir la variabilidad de nuestros estimadores puntuales, sino aprovechar el hecho de que sabemos que la distribución de los valores de $\hat{X}$ es $Normal(\mu, \frac{\sigma}{\sqrt{n}})$ y que la distribución de S^2 es $\frac{n-1}{\sigma^2}\chi^2$.

Bootstrap y jacknife

Sobre las distribuciones Empíricas

Cuantiles

Las gráficas Quantile-Quantile (QQ) son útiles para comparar funciones de distribución.

Definición

Sea X una v.a. con función de distribución (fd) F. La fd inversa o función de cuantiles es

$$F^{-1}(q) = \inf \{x : F(x) \geq q\},$$

para que $q \in [0, 1]$.

Si F es estrictamente creciente, entonces el $p-$ésimo cuantil es el único número real x_p tal que

$$F(x_p) = p.$$

Equivalentemente

$$x_p = F^{-1}(p).$$

Cuantiles

Sea X una v.a. continua con fd F_X y definamos $Y = aX + b$, con $a, b \in \mathbb{R}$. Denotemos con F_X^{-1} la función de cuantiles de X. Observemos lo siguiente

$$\begin{aligned} F_Y(y) = p &\Leftrightarrow P(aX + b \leq y) = P(X \leq (y-b)/a) = p \\ &\Leftrightarrow F_X((y-b)/a) = p, \end{aligned}$$

de donde obtenemos

$$F_Y^{-1}(p) = aF_X^{-1}(p) + b.$$

Esta relación es clave para encontrar percentiles para la trasformación lineal de variables aleatorias. Un caso importante es cuando se estandariza una variable con distribución $N(\mu, \sigma^2)$ para poder obtener la distribución acumulada o percentiles a partir de tablas.

Función de distribución empírica

Un primer problema de inferencia consiste en la estimación no paramétrica de la función de distribución empírica (fde) de F. Sea $X_1, \ldots, X_n \sim F$ una muestra independiente e idénticamente distribuida (iid), donde F es una fd sobre la recta real. Estimaremos F con la fde que se define a continuación.

Definición

La fde $\hat{F}_n$ es la función de distribución acumulada que asigna masa $1/n$ en cada punto X_i. Formalmente,

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} 1_{\{X_i \leq x\}}}{n},$$

donde

$$1_{\{X_i \leq x\}} = \begin{cases} 1 & \text{si } X_i \leq x \\ 0 & \text{si } X_i > x \end{cases}$$

Función de distribución empírica

Teorema

En cualquier valor fijo x,

$$\begin{aligned} E(\hat{F}_n(x)) &= F(x) \\ Var(\hat{F}_n(x)) &= \frac{F(x)(1-F(x))}{n} \to 0 \\ \hat{F}_n(x) &\xrightarrow{P} F(x). \end{aligned}$$

Teorema (Glivenko-Cantelli)

Sean $X_1, \ldots, X_n \sim F$. Entonces

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{P} 0$$

Función de distribución empírica

Las gráficas Q-Q son una herramienta visual extremadamente útil para establecer cualitativamente el ajuste de un conjunto de datos a una distribución teórica.

Supongamos que la hipótesis de que X sigue cierta distribución F es de interés. Dada una muestra $X_1, \dots X_n$, y usando el teorema de Glivenko-Cantelli, tenemos un buen estimador de la distribución de estos datos:

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} 1_{\{X_i \leq x\}}}{n} \xrightarrow{P} F(x).$$

Función de distribución empírica

Entonces, graficando

$$F(x) \text{ vs } \hat{F}_n(x)$$

para x apropiados, podremos determinar visualmente la validez de la hipótesis.

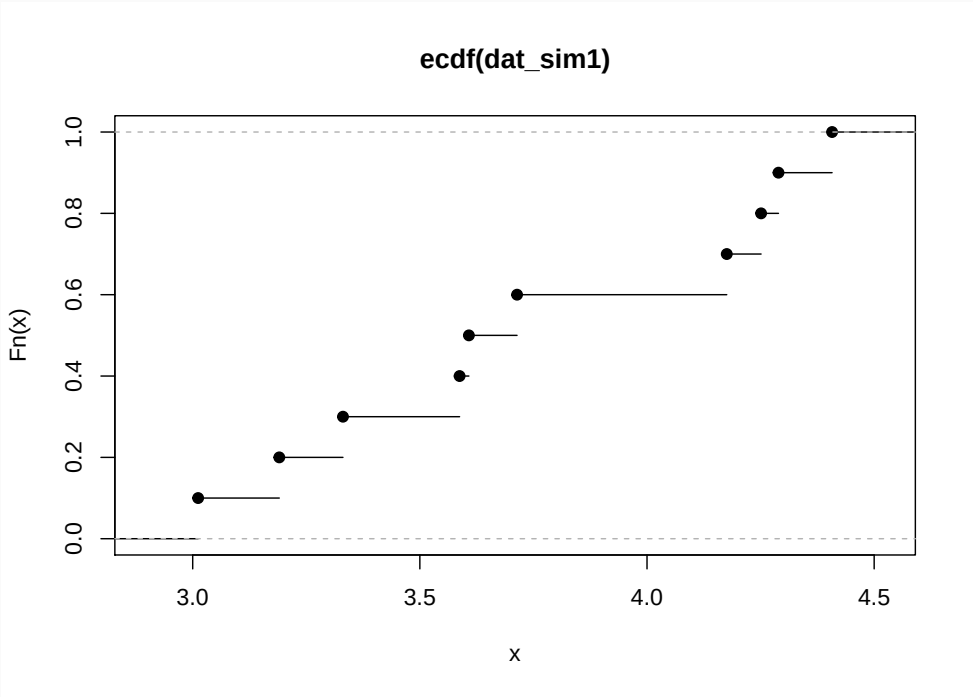
Si $F(x)$ corresponde a una función de distribución continua, entonces casi seguramente los valores muestreados no se repetirán. En este caso si $x_{(i)}$ denota el $i-$ésimo estadístico de orden, se tiene que

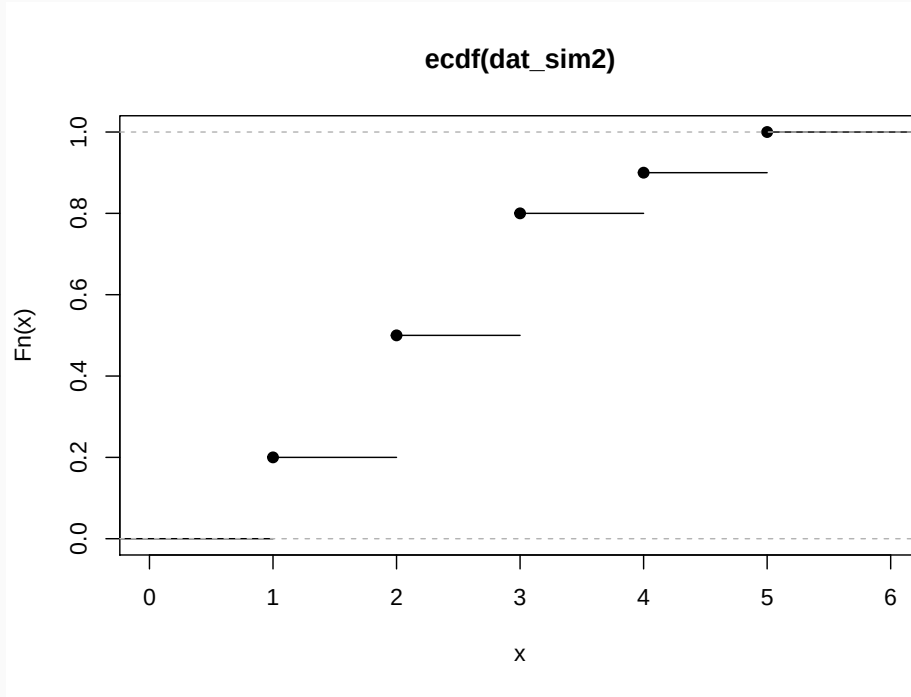
$$\hat{F}_n(x_{(i)}) = \frac{\sum_{i=1}^{n} 1_{\{x_i \leq x_{(i)}\}}}{n} = \frac{i}{n},$$

es decir que cada salto de la fdp empírica es de longitud $1/n$ sobre cada x_i.

Función de distribución empírica: Ejemplo

```
plot(ecdf(datsim1))
datsim2<-rpois(10,3) distribución discreta
plot(ecdf(datsim2))
```





Cuantil de una Función de distribución empírica

Sin importar si F es discreta o continua, el número de observaciones que se tiene en la m.a. es finito, lo cual induce una función de distribución escalonada, como en los ejemplos anteriores. Al igual que en el caso discreto, entonces se define el p-ésimo cuantil empírico como

$$\hat{F}_n^{-1}(q) = \inf\left\{x : \hat{F}_n(x) \geq q\right\},$$

con $q \in (0, 1)$.

Esta es la forma más sencilla de definir los percentiles, pero no es la única. En R se calculan los cuantiles con varios tipos de suavizamiento usando la función "quantile". "Type 1" corresponde a la definición anterior. Por ejemplo

```
quantile(datsim2,prob=0.2,type=1)
1
quantile(datsim2,prob=0.4,type=1)
2
```

Q-Q plot empírica contra teórica

La gráfica permite visualizar una distribución candidata para los datos. Se grafican los cuantiles para cada valor q. Por ejemplo se pueden hacer tablas como la siguiente

x_i	rango$= i$	$p_i = \frac{i}{n+1}$	$z_{p_i} = q_i$ tal que $P(Z \leq z_{p_i}) = p_i$
31	1	0.0625	-1.53412
32	2	0.125	-1.15035
33.2	3	0.1875	-0.88715
34.5	4	0.25	-0.67449
35	5	0.3125	-0.48878
35.5	6	0.375	-0.31864
36.5	7	0.4375	-0.15731
37.2	8	0.5	0
37.5	9	0.5625	0.157311
37.8	10	0.625	0.318639
38	11	0.6875	0.488776
38.3	12	0.75	0.67449
38.5	13	0.8125	0.887147

Q-Q plot empírica contra teórica

En R se puede utilizar la gráfica qqplot y qqnorm (para comparar con la ditribución Normal). Por ejemplo

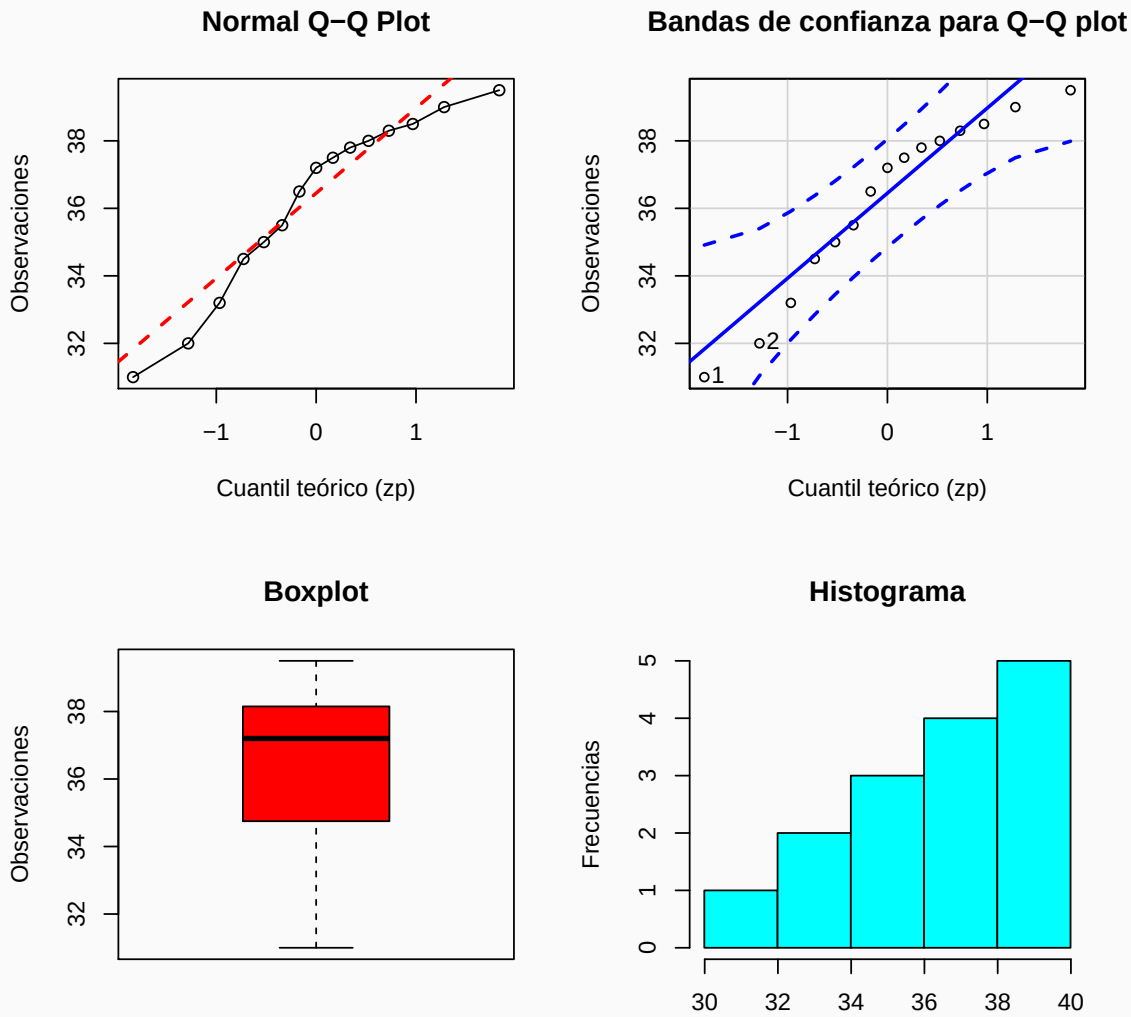
```
x1<-rnorm(100)
qqnorm(x1)
qqline(x1,distribution=qnorm)

x2<-runif(100)
qqnorm(x2)
qqline(x2,distribution=qnorm)

qqplot(x2,runif(1000))
qqline(x2,distribution=qunif)
```

Función de distribución empírica

Otra gráfica auxiliar para describir algunas características de los datos es el diagrama de cajas (que en R se implementa con el comando "boxplot")



Jackknife y Bootstrap

Jackknife y Bootstrap

Estas técnicas de remuestreo permiten utilizar la muestra que se tiene para principalemente poder

- estimar la variabilidad y la desviación estándar del estimador que nos interesa;
- estimar su sesgo y proponer una modificación que sea insesgada;
- estimar la tasa de error de una regla de predicción basada en datos.

Estas técnicas requieren de poder computacional para las estimaciones y aunque bajo ciertas condiciones cumplen con propiedades asíntoticas, no se obtienen expresiones cerradas de los estimadores de sesgo y desviación estándar.

Jackknife

En estadística, Jackknife es una técnica de remuestreo que es especialmente útil para estimar varianzas y sesgos.

El estimador Jackknife de un parámetro se encuentra al sistemáticamente dejar fuera cada una de las observaciones de un conjunto de datos y calcular el estimado y al final encontrar el promedio de estos cálculos.

Dada una muestra de tamaño n, el estimador Jackknife se encuentra al sumar los estimados de cada sub-muestra de tamaño $(n-1)$.

Jackknife es anterior a otras técnicas de remuestreo como bootstrap.

La técnica Jackknife fue desarrollada por Quenouille (1949) y expandida por John Tukey (1958), quién propuso el nombre Jackknife.

Jackknife

Sea $X_1, \ldots, X_n$ una m.a. de F (esto es, son i.i.d. con distribución F) y sea $T_n = T(X_1, \ldots, X_n)$ un estimador de alguna cantidad θ y denotemos su sesgo por $\text{bias}(T_n) = E(T_n) - \theta$.

Llamaremos a la submuestra

$$X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$$

la ***i*-ésima muestra Jackknife** de $(X_1, \ldots, X_n)$.

Denotemos por $T_{(i)}$ al estadístico T_n con la i-ésima observación removida, es decir

$$T_{(i)} = T(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Llamaremos a $T_{(i)}$ la ***i*-ésima replica de Jackknife**. En otras palabras, $T_{(i)}$ es el resultado de evaluar la i-ésima muestra Jackknife en el estadístico.

Notemos que entonces tenemos n diferentes $T_{(i)}$, que pueden o no coincidir algunas en valor, pero cada una es calculada de un diferente subconjunto de la muestra original.

Jackknife: Estimación de sesgo y corrección

El **estimador Jackknife del sesgo** se define como

$$b_{Jack} = (n-1)(\hat{T}_{(\cdot)} - T_n)$$

donde $\hat{T}_{(\cdot)} = \sum_i T_{(i)}/n$.

El **estimador corregido del sesgo** se obtiene restando el sesgo estimado al estimador original. Ésto es

$$T_{Jack} = T_n - b_{Jack}. \tag{1}$$

Jackknife: Estimación de sesgo y corrección

¿Por qué se define b_{Jack} de esta manera? **Para muchos estadísticos (estimadores)** se puede mostrar que

$$\text{bias}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right)$$

para algunos a y b.

Por ejemplo, sea $\sigma^2 = V(X_i)$ y $\hat{\sigma}^2 = n^{-1}\sum_i (X_i - \overline{X})^2$. Entonces,

$$\text{bias}(\hat{\sigma}^2) = -\frac{\sigma^2}{n},$$

es decir que $\hat{\sigma}^2$ tiene la forma anterior con $a = -\sigma^2$ y $b = 0$.

Jackknife: Estimación de sesgo y corrección *

Cuando el sesgo tiene la forma descrita anteriormente, se cumple que

$$\text{bias}(T_{(i)}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

Se sigue que el sesgo de $\hat{T}_{(\cdot)}$ también tiene dicha forma. Por tanto,

$$\begin{aligned} E(b_{Jack}) &= (n-1)(\hat{T}_{(\cdot)} - T_n) \\ &= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\ &= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\ &= \text{bias}(T_n) + O\left(\frac{1}{n^2}\right), \end{aligned}$$

lo cual muestra que el b_{Jack} estima el sesgo hasta un orden de $O(n^{-2})$.

Jackknife: Estimación de sesgo y corrección *

Por un cálculo similar,

$$\text{bias}(T_{Jack}) = \frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right)$$

así que el sesgo de T_{Jack} es un orden de magnitud menor que el de T_n.

T_{Jack} también puede escribirse como

$$T_{Jack} = \frac{\sum_{i=1}^{n} \tilde{T}_i}{n} \tag{2}$$

donde

$$\tilde{T}_i = nT_n - (n-1)T_{(i)}$$

son los llamados **pseudo-valores**.

Jackknife: T_{Jack}

Es facil corroborar que las expresiones del estimador corregido por sesgo, T_{Jack} en (1) y (2) son equivalentes.

$$
\begin{aligned}
T_{Jack} &= T_n - b_{jack} = T_n - (n-1)(\hat{T}_{(\cdot)} - T_n) \\
&= nT_n - (n-1)\hat{T}_{(\cdot)} \\
&= \frac{\sum_{i=1}^{n} T_n}{n} - \frac{(n-1)\sum_{i_1}^{n} T_{(i)}}{n} \\
&= \frac{\sum_{i=1}^{n}[nT_n - (n-1)T_{(i)}]}{n} \\
&= \frac{\sum_{i=1}^{n} \tilde{T}_i}{n}
\end{aligned}
$$

Jackknife: Estimación de la varianza

En el caso particular del estadístico $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ como un estimador de $E(X)$.

Tenemos una medida de la precisión del estimador $\overline{x}$ dada por

$$\hat{\sigma}(\overline{x}) = \left(\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{1/2}. \tag{3}$$

Problema: Esta fórmula para la precisión de $\overline{x}$ no se extiende en forma obvia para otros estimadores T_n, e.g. la mediana.

Jackknife: Estimación de la varianza

El **estimador jackknife de** $\mathrm{var}\,(T_n)$ es

$$v_{Jack} = \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{(i)} - \hat{T}_{(\cdot)} \right)^2$$

y el **estimador jackknife del error estándar** es

$$\widehat{\mathrm{se}}_{Jack} = \sqrt{v_{Jack}}.$$

Jackknife: Estimación de la varianza

Similarmente al caso del sesgo, también podemos obtener v_{Jack} usando los pseudo-valores. En este caso, el estimador Jackknife de $V(T_n)$ es

$$v_{Jack} = \frac{\tilde{s}^2}{n}$$

donde

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n}(\tilde{T}_i - \frac{1}{n}\sum_{j=1}^{n}\tilde{T}_j)^2}{n-1}$$

es la varianza muestral de los pseudo-valores. Bajo condiciones adecuadas en T, se puede demostrar que v_{Jack} estima consistentemente $V(T_n)$.

Por ejemplo, si T es una función suave de la media muestral entonces la consistencia se cumple.

Jackknife: Limites y consistencia *

Se puede demostrar que bajo ciertas condiciones de T, v_{jack} es un estimador consistente de var(T_n), en el sentido de que $v_{Jack}/\,\text{var}\,(\mathrm{T_n}) \xrightarrow{\mathrm{P}} 1$. En particular para un estimador T_n que es función de la media $\overline{X}_n$, tenemos el siguiene teorema.

Teorema

Sea $\mu = E(X_1)$ y $\sigma^2 = V(X_1) < \infty$ y supongamos que $T_n = g(\overline{X}_n)$ donde g tiene una derivada continua no-nula en μ. Entonces

$$\frac{(T_n - g(\mu))}{\sigma_n} \xrightarrow{d} N(0,1)$$

donde $\sigma_n^2 = n^{-1}(g'(\mu))^2\sigma^2$.

El Jackknife es consistente significando

$$\frac{v_{Jack}}{\sigma_n^2} \xrightarrow{c.s.} 1.$$

Jackknife: Limites y consistencia *

En contraste del boostrap jacknife no produce estimadores consistentes para los cuantiles, pero la la mediana se tiene un limite en distribución.

Teorema

Si $T(F) = F^{-1}(p)$ es el p-ésimo cuantil, entonces el estimador Jackknife de la varianza es inconsistente.

Para la mediana ($p = 1/2$) tenemos que

$$v_{Jack}/\sigma_n^2 \xrightarrow{d} (\chi_2^2/2)^2$$

donde σ_n^2 es la varianza asintótica de la mediana muestral.

Jackknife: Ejemplo 1

Sea $T_n = \overline{X}_n$. Sabemos que es unestadístico insesgado y tenemos una expresión para estimar su varianza (3), pero exploremos sus estimadores tipo Jackknife.

Es fácil ver que $\tilde{T}_i = X_i$. Por tanto usando (2) tenemos $T_{Jack} = T_n$, su sesgo $b_{Jack} = 0$ y el estimador Jackknife de la desviación estándar de $\overline{X}_n$

$$\hat{\sigma}_{Jack} = \left(\frac{n-1}{n} \sum_{i=1}^{n} (\overline{X}_{(i)} - \overline{X}_{(\cdot)})^2 \right)^{1/2}.$$

Jackknife: Ejemplo 2

Consideremos los siguientes 20 datos de Manly (2007, "Randomization, Bootstrap and Monte Carlo Methods in Biology. 3rd ed").

3.56,	0.69,	0.10,	1.84,	3.93,	1.25,	0.18,	1.13,	0.27,	0.50,
0.67,	0.01,	0.61,	0.82,	1.70,	0.39,	0.11,	1.20,	1.21,	0.72.

Se desea estimar la desviación estándar de la población (σ) corrigiendo con su sesgo estimado. Manly (2007) estima la raíz cuadrada del MLE de la varianza poniendo n en lugar de $n-1$ en el denominador, es decir

$$T_n = \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = 1.03,$$

y no con la desviación estándar muestral la cual tiene $(n-1) = 19$ grados de libertad.

Jackknife: Ejemplo 2

Calculamos T_n (el estimador sin corregir)

```
dat<-c(3.56, 0.69, 0.10, 1.84, 3.93, 1.25, 0.18, 1.13, 0.27, 0.50,
0.67, 0.01, 0.61, 0.82, 1.70, 0.39, 0.11, 1.20, 1.21, 0.72)
n<-length(dat)
Tn<-sqrt(var(dat)*(n-1)/n)
Tn
[1] 1.032848
```

Hacemos una matriz donde cada columna i corresponde al vector de datos menos su i-ésima observacion.

```
jdat<-matrix(dat,nrow=n,ncol=n)
diag(jdat)<-NA
jdat<-matrix(jdat[which(!is.na(jdat))],nrow=n-1,ncol=n)
```

Jackknife: Ejemplo 2

```
Ti<-apply(jdat,2,FUN=function(x){sqrt(var(x)*(length(x)-1)/length(x))})
Ti
[1] 0.8788364 1.0563893 1.0360975 1.0430060 0.8134128 1.0585751 1.0399595
[8] 1.0594885 1.0438813 1.0519008 1.0560070 1.0313246 1.0547329 1.0583612
[15] 1.0483872 1.0484218 1.0365998 1.0590473 1.0589633 1.0569234
Tdot<-mean(Ti)
Tdot
[1] 1.029516

tildeTi<-n*Tn-(n-1)*Ti
Tjack<-mean(tildeTi)
Tjack
[1] 1.096158
```

Entonces el sesgo estimado de T_n es $\hat{\text{bias}}(T_n) = T_n - T_{Jack} = -0.06331$

Jackknife

Ejercicio: Supongamos que queremos aplicar Jackknife en otras 3 situaciones:

1. Supongamos que queremos estimar $\theta = \mu$. ¿Qué pasaría si usamos la media muestral ($\hat{\theta} = \overline{x}$)?
2. Supongamos que queremos estimar $\theta = \sigma$. ¿Qué pasaría la desviación estándar muestral ($\hat{\theta} = s$)?
3. Supongamos que queremos estimar $\theta = \sigma^2$. ¿Qué pasaría si usamos la varianza muestral estándar ($\hat{\theta} = s^2$)?

Jackknife

Hay algunas cosas importantes que observar en estos resultados,

1. Sabemos que $\overline{x}$ y s^2 son estimadores insesgados de μ y σ. Aplicar el método de Jackknife a un estimador insesgado no tiene efecto. Es decir, $\hat{\theta}_{Jack} = \hat{\theta}$ y el sesgo estimado será 0.
2. Los errores estándar pueden usarse para calcular intervalos de confianza para θ usando los estimador corregidos por sesgo ($\hat{\theta}_{Jags} = T_{Jags}$) de $\hat{\theta} = T_n$ y el estimador de su error estándar ($\text{se}(\hat{\theta}_{Jack})$).

   $$\hat{\theta}_{Jack} \pm 1.960 \cdot \text{se}(\hat{\theta}_{Jack}).$$

   Observación:

   $$\hat{\theta}_{Jack} \pm t^* \cdot \text{se}(\hat{\theta}_{Jack})$$

   donde t^* es el valor crítico $1 - \alpha/2$ de una distribución t con $n - 1$ grados de libertad.

Jackknife: Limitaciones

- El método de Jackknife puede fallar si el estadístico $\hat{\theta}$ no es suave. La suavidad implica que cambios relativamente pequeños los datos provocarán sólo un pequeño cambio en el estadístico.
- La mediana muestra es un ejemplo de un estadístico que no es suave.
- Por ejemplo, volviendo a los datos de Manly (2007), los valores ordenados son

0.01	0.10	0.11	0.18	0.27	0.39	0.50	0.61
0.67	0.69	0.72	0.82	1.13	1.20	1.21	1.25
1.70	1.84	3.56	3.93				

  Notemos que hay dos estimadores Jackknife:
  - $1/2$ de los estimadores de Jackknife son iguales a 0.72 (al borrar alguno de los 10 primeros valores).
  - $1/2$ de los estimadores de Jackknife son iguales a 0.69 (al borrar alguno de los 10 últimos valores).

Jackknife: Limitaciones

- Por lo tanto, el método de Jackknife no es bueno para estimar percentiles (como la mediana), o cuándo se usa un estimador no suave.
- Esto no será el caso cuando usemos el método de estimación bootstrap.

Bootstrap

Considera una m.a. $X_1, \ldots, X_n$ de alguna población y un estadístico $\hat{\theta} = T_n = T(X_1, \ldots, X_n)$ calculado a partir de la muestra que es el estimador de algún parámetro θ de la población. Por ejemplo, θ puede ser la media de la población y $\hat{\theta}$ la media muestral.

Para poder entender que tan bien $\hat{\theta}$ estima θ, es necesario conocer la varianza de la v.a. $\hat{\theta}$. Sin embargo, para poder calcularla, uno necesita saber algo de la población.

Existe un método general para estimar la varianza de $\hat{\theta}$ a partir de la muestra. Este método es conocido como **bootstrap**.

La idea de bootstrap es tratar a la muestra como una nueva población.

Bootstrap

Pensemos a nuestra m.a. $X_1, \ldots, X_n$ de v.a.i.i.d. como una **población finita** y consideremos el experimento de **tomar una m.a. con reemplazo y ordenada de tamaño n de esta población finita**. Denotaremos esta nueva muestra por $X_1^*, \ldots, X_n^*$ y la llamaremos muestra bootstrap.

Así, $X_1^*, \ldots, X_n^*$ son v.a. independientes, cada una de las cuales puede tomar el valor X_i con probabilidad $1/n$. Es decir,

$$P(X_i^* = X_j) = \frac{1}{n}.$$

A partir de la muestra bootstrap podemos calcular la v.a. correspondiente $\hat{\theta}^*$.

Ejemplo: Si θ es la media de la población y $\hat{\theta}$ la media muestral de la m.a. $X_1, \ldots, X_n$, entonces $\hat{\theta}^*$ es la media muestral de $X_1^*, \ldots, X_n^*$.

Bootstrap

La idea es usar la varianza de la variable $\hat{\theta}^*$ (con una muestra $X_1, \ldots, X_n$ fija) como un estimador de la varianza de $\hat{\theta}$. Esta varianza es llamada el **estimador bootstrap ideal**.

¿Cómo podemos calcular el estimador bootstrap ideal?

En principio, es simple. Existe un número finito n^n de muestras bootstrap, cada una con probabilidad $1/n^n$. Podríamos hacer lo siguiente:

1. Para cada una de tales muestras bootstrap, calcula el valor de $\hat{\theta}^*$.
2. Calcula la media de estos n^n números.
3. Entonces calcula la media del cuadrado de las desviaciones de su media. Este número es el estimador bootstrap ideal.

El problema con esto es, por supuesto, que n^n es un número enorme. Enumerar todas las muestras bootstrap de esta forma es impráctico, incluso para n's pequeños.

Bootstrap

El bootstrap es un método para estimar la varianza y la distribución de un estadístico $T_n = g(X_1, \ldots, X_n)$.

También se utiliza al bootstrap para construir intervalos de confianza.

Denotemos por $V_F(T_n)$ a la varianza de T_n. Añadimos el subíndice F para enfatizar que la varianza es una función de la función de ditribución F.

Si conocemos F podríamos, al menos en principio, calcular la varianza.

Por ejemplo, si $T_n = \sum_i X_i/n$, entonces

$$V_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - (\int x dF(x))^2}{n},$$

la cuál claramente está en función de F.

Bootstrap

Denotemos por $\hat{F}_n$ a la distribución empírica de la m.a. $X_1, \ldots, X_n$.

Con bootstrap estimamos $V_F(T_n)$ a tráves de $V_{\hat{F}_n}(T_n)$. En otras palabras, utilizamos un estimador "plug-in" de la varianza.

Como $V_{\hat{F}_n}(T_n)$ puede ser muy difícil de calcular, lo aproximamos con un estimado simulado denotado por v_{boot}. Específicamente, seguimos los siguientes pasos.

Algoritmo (Estimación de la varianza por bootstrap)

1. Obten una muestra $X_1^*, \ldots, X_n^* \sim \hat{F}_n$.
2. Calcula $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repite los pasos 1 y 2, B veces para obtener $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Pongamos
$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2$$

Por la Ley de Grandes Números,

$$v_{boot} \xrightarrow{c.s.} V_{\hat{F}_n}(T_n)$$

cuando $B \to \infty$.

El error estándar estimado de T_n es $\hat{se}_{boot} = \sqrt{v_{boot}}$.

Bootstrap

El siguiente diagrama ilustra la idea del bootstrap:

$$\begin{array}{lccccc} \text{Mundo real:} & F & \Rightarrow & X_1, \ldots, X_n & \Rightarrow & T_n = g(X_1, \ldots, X_n) \\ & n \Uparrow & & & & B \Uparrow \\ \text{Mundo bootstrap:} & \hat{F}_n & \Rightarrow & X_1^*, \ldots, X_n^* & \Rightarrow & T_n^* = g(X_1^*, \ldots, X_n^*) \end{array}$$

$$V_{F_n}(T_n) \overset{O(1/\sqrt{n})}{\approx} V_{\hat{F}_n}(T_n) \overset{O(1/\sqrt{B})}{\approx} v_{boot}.$$

¿Cómo simulamos de $\hat{F}_n$?

Bootstrap

Ya que $\hat{F}_n$ da probabilidad $1/n$ a cada observación de los datos originales, muestrar aleatoriamente n observaciones de $\hat{F}_n$ es lo mismo que obtener una muestra con reemplazamiento de tamaño n a partir de los datos originales.

Por lo tanto, el paso 1 del algoritmo anterior puede reemplazarse por

1. Obten una muestra $X_1^*, \ldots, X_n^*$ con reemplazo de $X_1, \ldots, X_n$.

Bootstrap: Estimación de Función de distribución de T_n

Bootstrap puede usarse para aproximar no sólo la varianza sino la función de distribución acumulada de un estadístico T_n.

Sea $G_n(t) = P(T_n \leq t)$ la función de distribución de T_n.

La aproximación bootstrap a G_n está dada por

$$\tilde{G}_n^*(t) = \frac{1}{B} \sum_{b=1}^{B} I(T_{n,b}^* \leq t).$$

Con esta expresión entonces podemos obener estimación de probabilidades y cuantiles. Por ejemplo, podemos estimar

$$P(0 < T_n \leq 10) = \tilde{G}_n^*(10) - \tilde{G}_n^*(0)$$

o estimar el valor t tal que

$$P(T \leq t) > 0.95 = (\tilde{G}_n^*)^{-1}(0.95)$$

Bootstrap: Nota sobre el caso paramétrico

Hasta ahora, hemos estimado F de forma no-paramétrica. Existe también un bootstrap paramétrico. Si F_θ depende de un parámetro θ y $\hat{\theta}$ es un estimado de θ, entonces podemos simplemente muestrear de $F_{\hat{\theta}}$ en lugar de F_n.

Esto es igual de exacto, pero mucho más simple que el método delta.

Bootstrap: Intervalos de confianza

Existen diversas maneras de contruir intervalos de confianza bootstrap. Varían en la facilidad de calcular y en la exactitud.

Entre estas formas encontrarmos:

- Intervalo normal.
- Intervalo pivotal.
- Intervalo basado en percentiles.
- Intervalo pivotal studentizado.

Los primeros tres se presentan a continuación.

Bootstrap: Intervalo normal

El intervalo bootstrap más sencillo es el intervalo normal

$$T_n \pm z_{\alpha/2} \hat{\text{se}}_{boot}$$

donde $\hat{\text{se}}_{boot}$ es el estimador bootstrap del error estándar.

Este intervalo no es muy exacto a menos de que la distribución de T_n sea cercana a la normal.

Bootstrap: Intervalo pivotal

Sea $\theta = T(F)$ y $\hat{\theta}_n = T(\hat{F}_n)$ y definamos el pivote $R_n = \hat{\theta}_n - \theta$.

Denotemos por $H(r)$ a la función de distribución del pivote R_n,

$$H(r) = P_F(R_n \leq r).$$

Sea $C_n^* = (a, b)$ donde

$$a = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right), \qquad b = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right);$$

así

$$\begin{aligned} P(a \leq \theta \leq b) &= P(\hat{\theta}_n - b \leq R_n \leq \hat{\theta}_n - a) = H(\hat{\theta}_n - a) - H(\hat{\theta}_n - b) \\ &= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \end{aligned}$$

Bootstrap: Intervalo pivotal

Por tanto C_n^* es un intervalo de confianza exacto para θ, de confiabilidad $1-\alpha$. Desafortunadamente, a y b dependen de la distribución desconocida H.

Sin embargo, podemos tomar el estimado bootstrap de H

$$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^* \leq r),$$

donde $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$.

Denotemos por r_β^* al cuantil muestral β de $(R_{n,1}^*, \ldots, R_{n,B}^*)$ y por θ_β^* al cuantil muestral β de $(\theta_{n,1}^*, \ldots, \theta_{n,B}^*)$. Notemos que $r_\beta^* = \theta_\beta^* - \hat{\theta}_n$.

Bootstrap: Intervalo pivotal

Se sigue que un intervalo de confianza aproximado de $1-\alpha$ para θ es $C_n = (\hat{a}, \hat{b})$ donde

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1}\left(1-\frac{\alpha}{2}\right) = \hat{\theta}_n - r^*_{1-\alpha/2} = 2\hat{\theta}_n - \theta^*_{1-\alpha/2},$$

$$\hat{b} = \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r^*_{\alpha/2} = 2\hat{\theta}_n - \theta^*_{\alpha/2}.$$

En resumen, el intervalo de confianza bootstrap pivotal es

$$C_n = (2\hat{\theta}_n - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta}_n - \hat{\theta}^*_{\alpha/2}).$$

Típicamente este es un intervalo de confianza puntual y asintótico.

Bootstrap: Intervalos basados en percentiles

El intervalo bootstrap de percentiles esta definido por

$$(T^*_{\alpha/2}, T^*_{1-\alpha/2}).$$

Bootstrap: Ejemplo

Se muestrea (con remplazo) 10 alumnos de posgrado de un grupo que asisten a una clase de estadística. Para este ejercicio se sabe que en el grupo de 30 alumnos, el promedio de edad es 39.17. Las edades muestreadas son:

$$30, \quad 31, \quad 32, \quad 34, \quad 37, \quad 37, \quad 37, \quad 44, \quad 49, \quad 53$$

A continuación obtenemos los intervalos Pivotal y basados en percentiles al 95% de confianza.

```
muestra<-c(31, 37, 49, 34, 37, 30, 44, 53, 32, 37)
n<-length(muestra)
B<-1000
bsam<-matrix(sample(muestra,size=n*B,replace=TRUE), nrows=n, ncols=B)
bxbar<-apply(bsam,2,FUN=mean)
head(bxbar)
[1] 39.5 43.1 35.9 37.1 36.9 38.7
```

Bootstrap: Ejemplo

Intervalo Pivotal

```
bRnb<-bxbar-mean(muestra)
HatH<-ecdf(bRnb) # Función de distribución empirica
plot(HatH)
# El calculo de los cuantiles se puede hacer directamente con bRnb o bxbar u
# Intervalo:
c(2*mean(muestra)-quantile(bxbar,probs=0.975),
(2*mean(muestra)-quantile(bxbar,probs=0.025)))
# Equivalentemente
2*mean(muestra)-quantile(bxbar,probs=c(0.975,0.025))
```

Origina el intervalo $(33.4, 42.7)$ que contiene al verdadero valor.

Ejercicio: Repetir esto 1000 veces y obtener la proporción de veces que los intervalos obtenidos contienen al verdadero valor de 39.17.

Bootstrap: Ejemplo

Intervalo Basado en percentiles

Estos intervalos se basan en los percentiles de los B estimadores bootstrap

```
quantile(bxbar,probs=c(0.025, 0.975))
2.5% 97.5%
34.1  43.4
```

Origina el intervalo $(34.1, 43.4)$ que contiene al verdadero valor.

Ejercicio: Repetir esto 1000 veces y obtener la proporción de veces que los intervalos obtenidos contienen al verdadero valor de 39.17.

Bootstrap

Como se ha comentado anteriormente, bootstrap es superior que Jackknife para estimar cuantiles y bajo ciertas condiciones de regularidad se puede demostrar la validez del bootstrap en la inferencia de una gran variedad de parámetros. Aquí no las estudiaremos pero se debe de tener en cuenta que el bootstrap no es infalible. Existen condiciones en la que los resultados pueden ser muy inadecuados.

También se puede demostrar que el estimador bootstrap de la varianza es consistente bajo algunas condiciones de T. En general, las condiciones de consistencia para el bootstrap son más débiles que las necesarias para el jackknife.

Estimación no paramétrica

Introducción

Dada una secuencia de variables aleatorias independientes e identicamente distribuidas $x_1, x_2, \ldots x_n$ con función de densidad común $f(x)$,

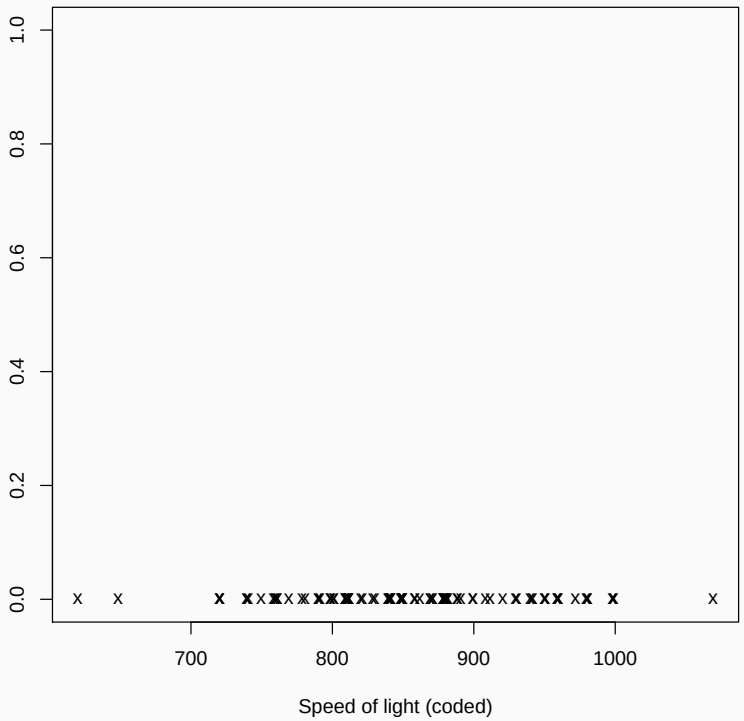
¿Cómo podemos estimar $f(x)$?

(Parzen, E. *On estimation of a Probability Density Function and Mode*, 1962)

Introducción

Un esquema tradicional de análisis de datos.

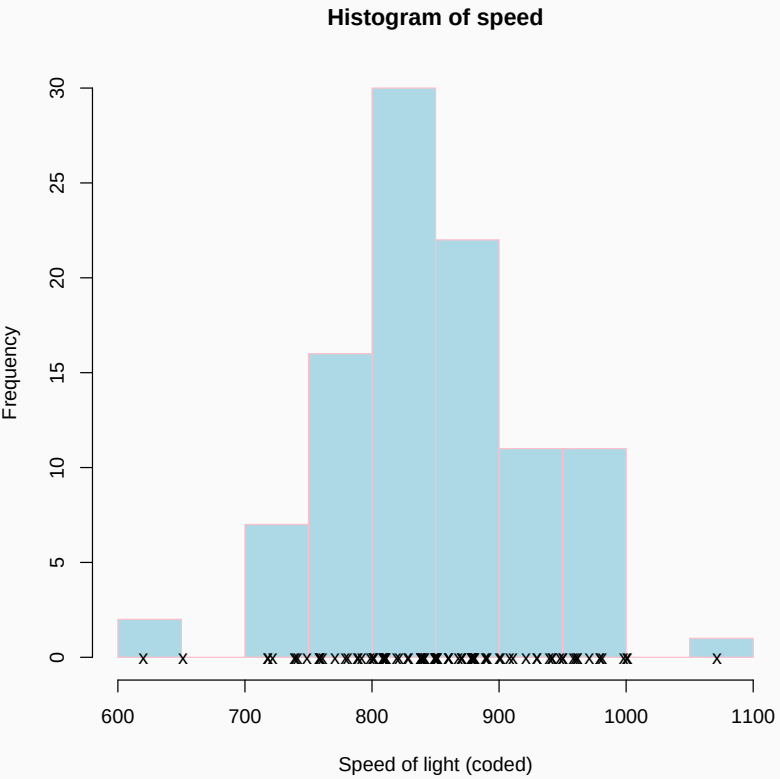
Ejemplo: Estimación de la velocidad de la luz experimentalmente (Michaelson & Morley, 1986)



Introducción

Un esquema tradicional de análisis de datos.

Ejemplo: Estimación de la velocidad de la luz experimentalmente (Michaelson & Morley, 1986)

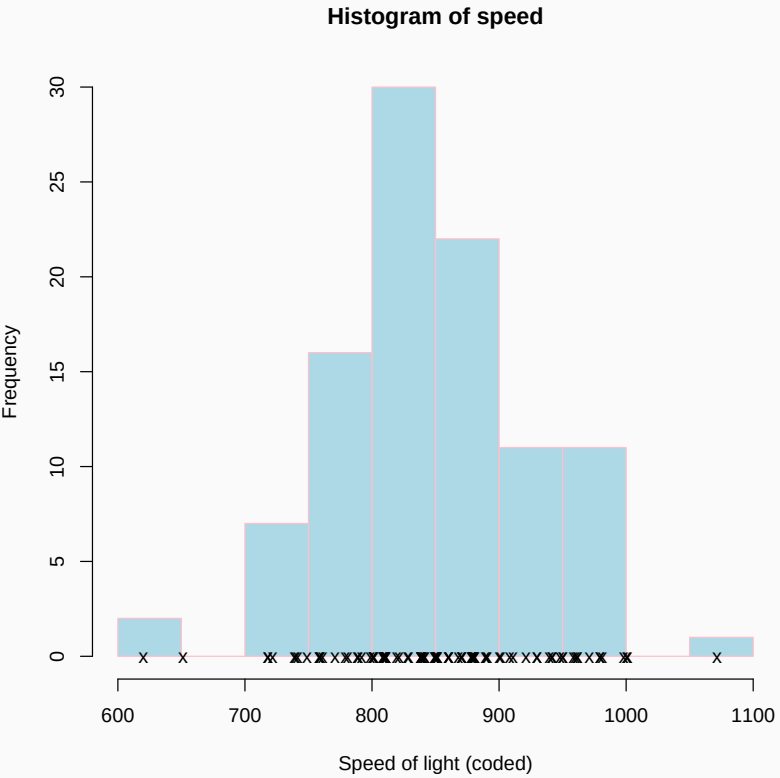


- Estimamos su densidad

Introducción

Un esquema tradicional de análisis de datos.

Ejemplo: Estimación de la velocidad de la luz experimentalmente (Michaelson & Morley, 1986)

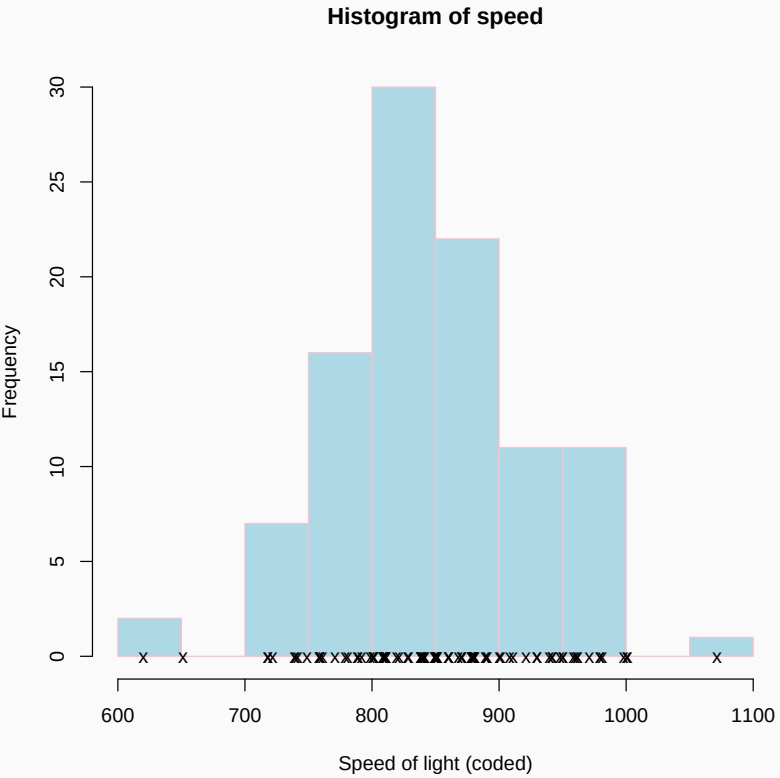


- Estimamos su densidad
- ¿Qué modelo es el adecuado?

Introducción

Un esquema tradicional de análisis de datos.

Ejemplo: Estimación de la velocidad de la luz experimentalmente (Michaelson & Morley, 1986)

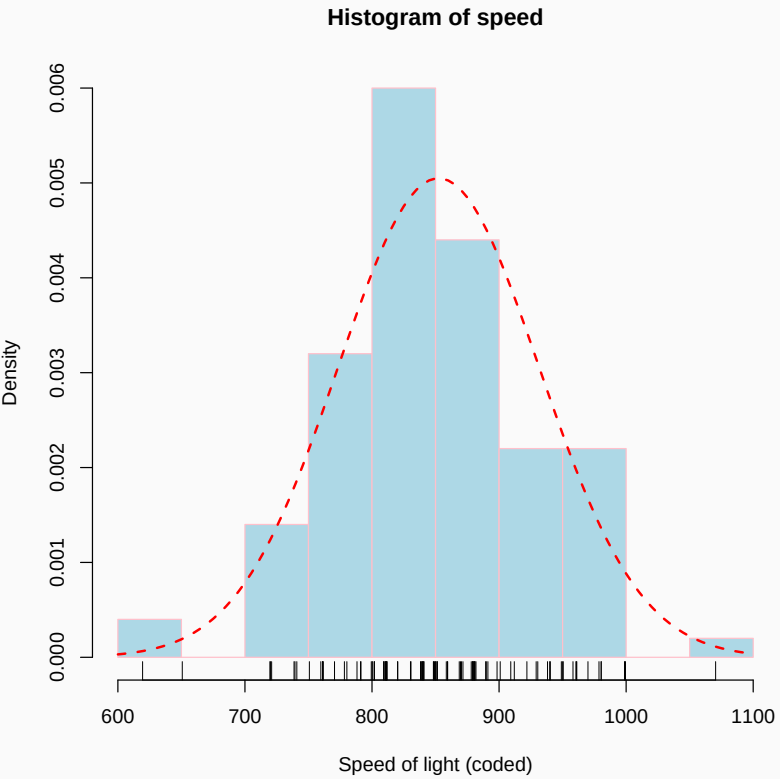


- Estimamos su densidad
- ¿Qué modelo es el adecuado?
- Esto dependerá de varios factores, por ejemplo, la naturaleza de los datos, el conocimiento apriori del fenómeno estudiado o la experiencia del analista.

Introducción

Un esquema tradicional de análisis de datos.

Ejemplo: Estimación de la velocidad de la luz experimentalmente (Michaelson & Morley, 1986)

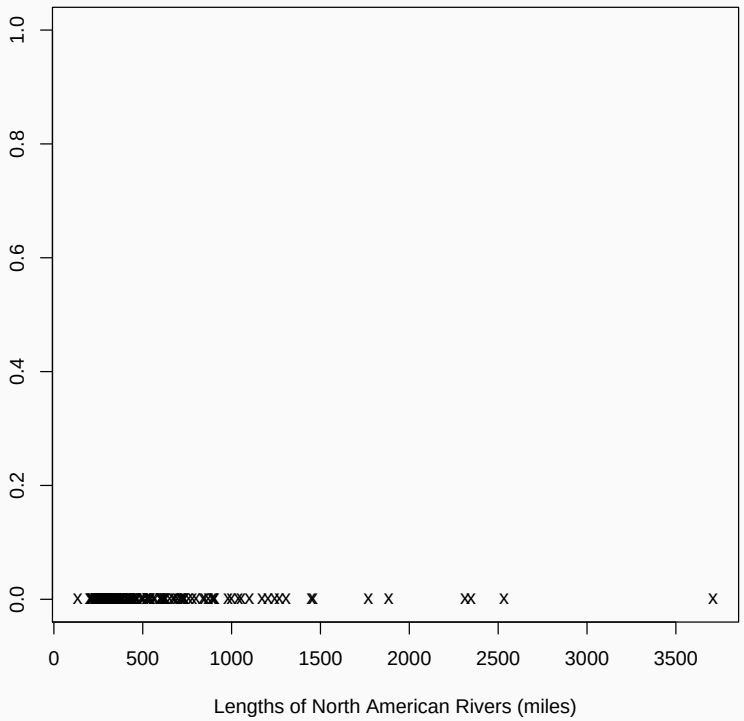


- Ajustemos un modelo normal, los parámetros $\theta = (\mu, \sigma)$ los estimamos de la muestra.

Introducción

Un esquema tradicional de análisis de datos.

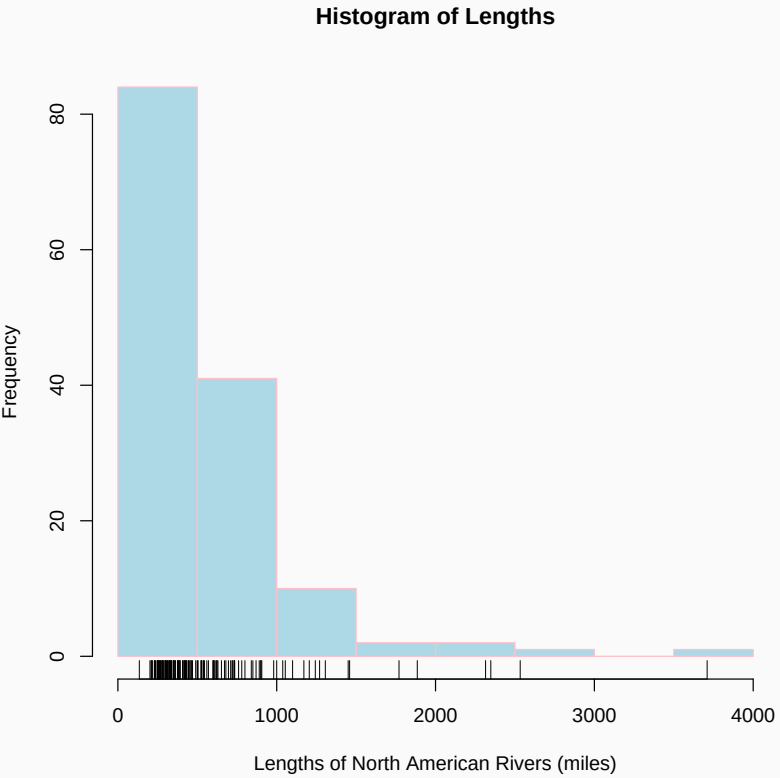
Ejemplo: Longitud de ríos de Norteamérica según la US Geological Survery (1975).



Introducción

Un esquema tradicional de análisis de datos.

Ejemplo: Longitud de ríos de Norteamérica según la US Geological Survery (1975).

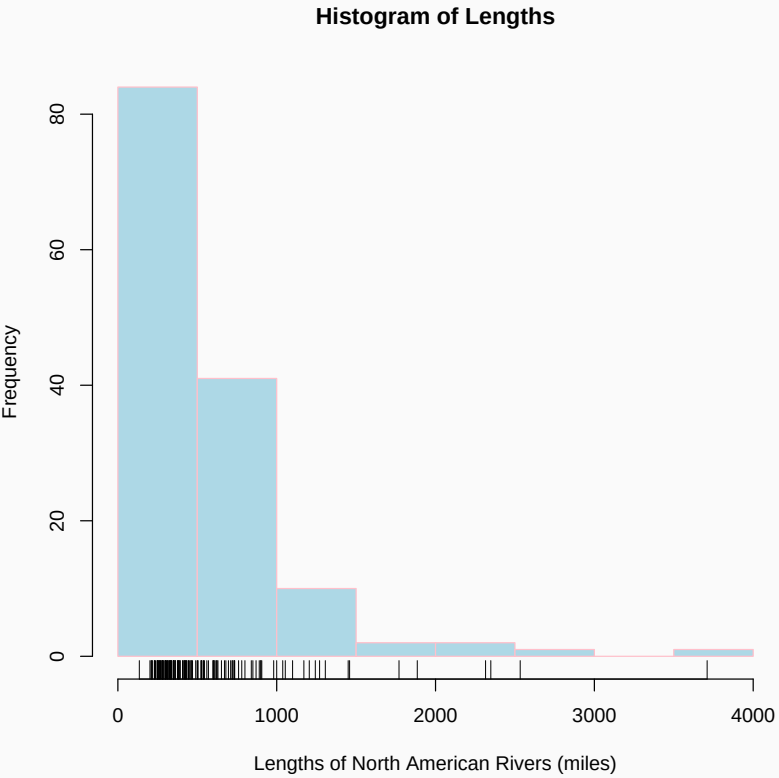


- Estimamos su densidad

Introducción

Un esquema tradicional de análisis de datos.

Ejemplo: Longitud de ríos de Norteamérica según la US Geological Survery (1975).

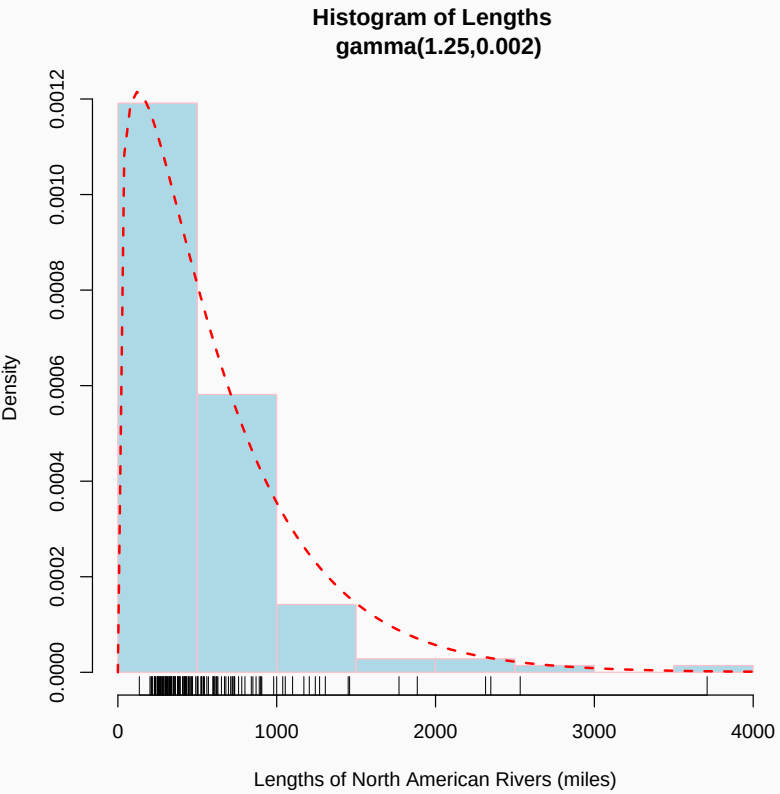


- Estimamos su densidad
- ¿Qué modelo parece apropiado?

Introducción

Un esquema tradicional de análisis de datos.

Ejemplo: Longitud de ríos de Norteamérica según la US Geological Survery (1975).



- Ajustamos un modelo Gamma(α, λ), con α el parámetro de forma y λ el parámetro de escala.

Introducción

- Aquí, todo el conocimiento que obtengamos del fenómeno bajo estudio proviene de **EL modelo** que escojamos, que a su vez nos proveé **LA función de densidad** de los datos.

Introducción

- Aquí, todo el conocimiento que obtengamos del fenómeno bajo estudio proviene de **EL modelo** que escojamos, que a su vez nos proveé **LA función de densidad** de los datos.
- La modelación estadística tradicional escoge algún modelo paramétrico conocido para $f(x)$, por ejemplo, Normal, Gamma, Exponencial, etc...

Introducción

- Aquí, todo el conocimiento que obtengamos del fenómeno bajo estudio proviene de **EL modelo** que escojamos, que a su vez nos proveé **LA función de densidad** de los datos.
- La modelación estadística tradicional escoge algún modelo paramétrico conocido para $f(x)$, por ejemplo, Normal, Gamma, Exponencial, etc...
- Esto tiene mucho sentido para fenómenos muy estudiados y analizados (estudios de confiabilidad, por ejemplo)

Introducción

- Aquí, todo el conocimiento que obtengamos del fenómeno bajo estudio proviene de **EL modelo** que escojamos, que a su vez nos proveé **LA función de densidad** de los datos.
- La modelación estadística tradicional escoge algún modelo paramétrico conocido para $f(x)$, por ejemplo, Normal, Gamma, Exponencial, etc...
- Esto tiene mucho sentido para fenómenos muy estudiados y analizados (estudios de confiabilidad, por ejemplo)
- Sin embargo, para fenómenos o datos más complejos, no siempre es conveniente o válido suponer una distribución de antemano. Verás muchos ejemplos más adelante.

EDNP

En la Estimación de Densidad No Paramétrica (EDNP), no hacemos supuestos distribucionales sobre los datos.

Dado un conjunto de datos $\mathbf{X} \in \mathbb{R}^d$ (por simplicidad, empezaremos consideraremos el caso univariado):

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) \sim f(\mathbf{x}),$$

queremos estimar **UNA** distribución de densidad $\hat{f}(\mathbf{x})$ que aproxime a $f(\mathbf{x})$ tal que

$$\hat{f}(\mathbf{x}) \geq 0$$

$$\int_{\mathbb{R}} \hat{f}(\mathbf{x}) d\mathbf{x} = 1,$$

(*bona fide* density).

EDNP

¿Qué nos gustaría de $\hat{f}(\mathbf{x})$?

- Que se parezcan: $E\hat{f}(\mathbf{x}) = f(\mathbf{x})$.
  Si $\hat{f}(\mathbf{x})$ es una estimación basada en una muestra de tamaño n, ésta característica nos asegura que

$$E\hat{f}(\mathbf{x}) \to f(\mathbf{x}) \text{ cuando } n \to \infty$$

  (insesgado)
- Que converja a la vedadera distribución:

$$\hat{f}(\mathbf{x}) \xrightarrow{P} f(\mathbf{x})$$

  (consistente)

Para esto, debemos definir una forma de medir la diferencia entre ambas distribuciones...

El histograma.

Es quizá el método no paramétrico más usado para estimar y visualizar una densidad. El método es muy simple.

Supongamos que $\mathbf{x} \in [a, b]$

- Crea una partición fija de M celdas disjuntas $T_0, T_1, \ldots, T_{M-1}$ que comprendan el intervalo $[a, b]$, cada una con un ancho h.
- La densidad se estima mediante:

$$\hat{f}(x) = \frac{1}{nh} \sum_{m=0}^{M-1} N_m I_{T_m}(x),$$

donde

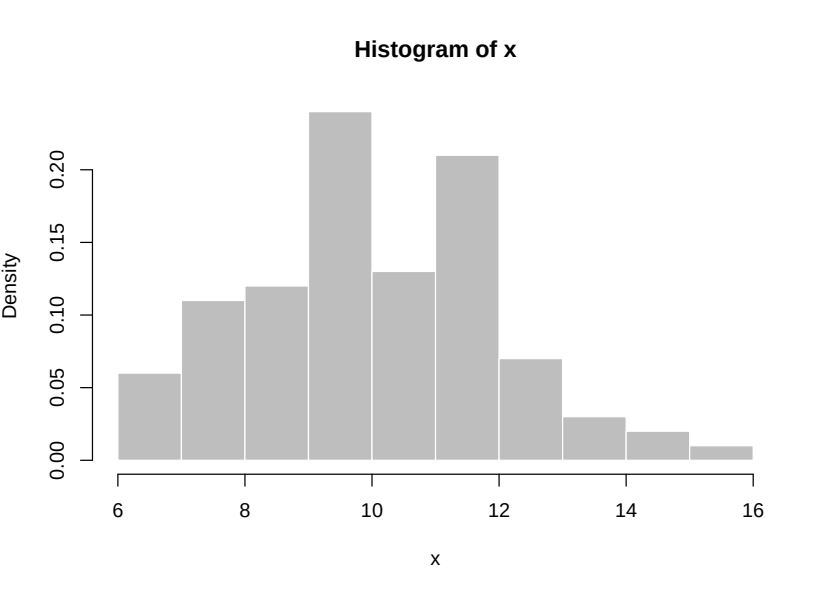
- I_{T_m} es la función indicadora del intervalo m,
- $N_m = \sum_{i=1}^{n} I_{T_m}(x_i)$, es decir, el número de valores que caen en la celda T_m

EDNP: El histograma

Las desventajas del histograma como un estimador de la densidad han sido mencionadas por varios autores. Por ejemplo:

- Celdas fijas
- Discontinuidades en las fronteras de las celdas
- Selección del origen del histograma

Ejemplo




```
x<-rnom(1000,mean=10,sd=2)
hist(x,freq=FALSE,col="grey",b
```

Estimación usando un Kernel ("Kernel Density Estimation", KDE).

Es el método mas popular. Para el caso univariado, el KDE está dado por

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad x \in \mathbb{R}, h > 0. \tag{4}$$

K es la función Kernel, y h es el **ancho de banda**, que determina la suavidad de la estimación.

Bajo condiciones no muy restrictivas (h debe decrecer cuando n aumenta), puede mostrarse que KDE converge en probabilidad a la verdadera densidad.

KDE

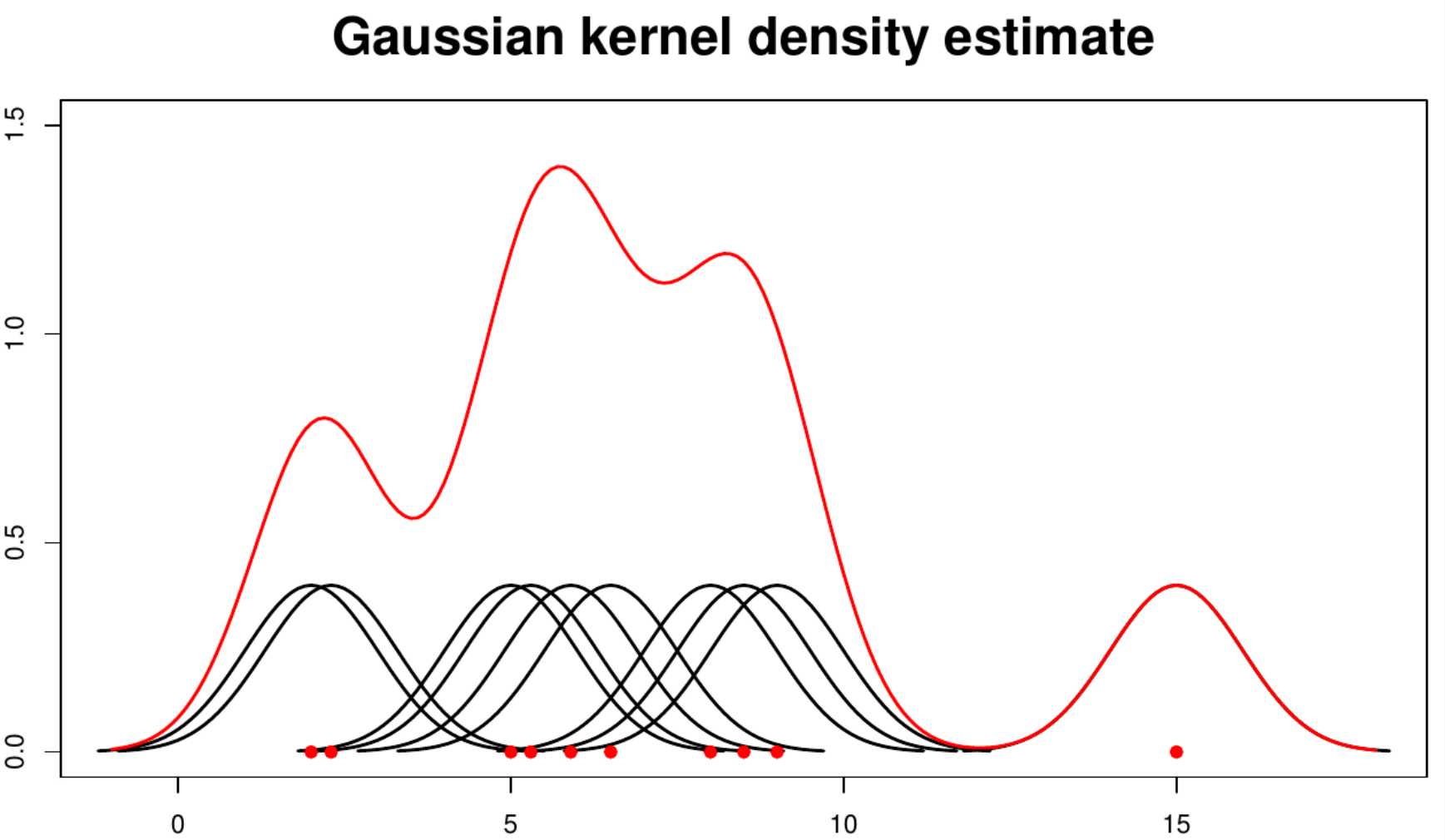
$K(y)$ se selecciona como función de densidad simétrica alrededor del 0, para que el "peso" de cada observación se distribuya de forma igual a la derecha y a la izquierda. Se utiliza n distribuciones K en (4), cada una centrada en las observaciones x_i's y con una vecindad de "influencia" definida por el ancho de banda h.

Como K es una función de densidad, se puede verificar fácilmente que (4) es una función de densidad.

(i) Como $K(y)$ es no negativa para todo $y \in \mathbb{R}$ entonces $\hat{f}(x)$ en (4) es no negativa

(ii) $K(y)$ se centra y escala considerando el cambio de varible $x = yh + x_i$. Por el Teorema de Cambio de Variable, $K(x) = K((x - x_i)/h))/h$. Entonces (4) es la mezcla de n densidades con constantes de mezcla $1/n$ para cada una (as´ se tiene que la suma de los pesos es 1). Entonces $\int_{-\infty}^{\infty} \hat{f}(x) = (n)^{-1} \sum \int K(y)dy = 1$.

EDNP: KDE



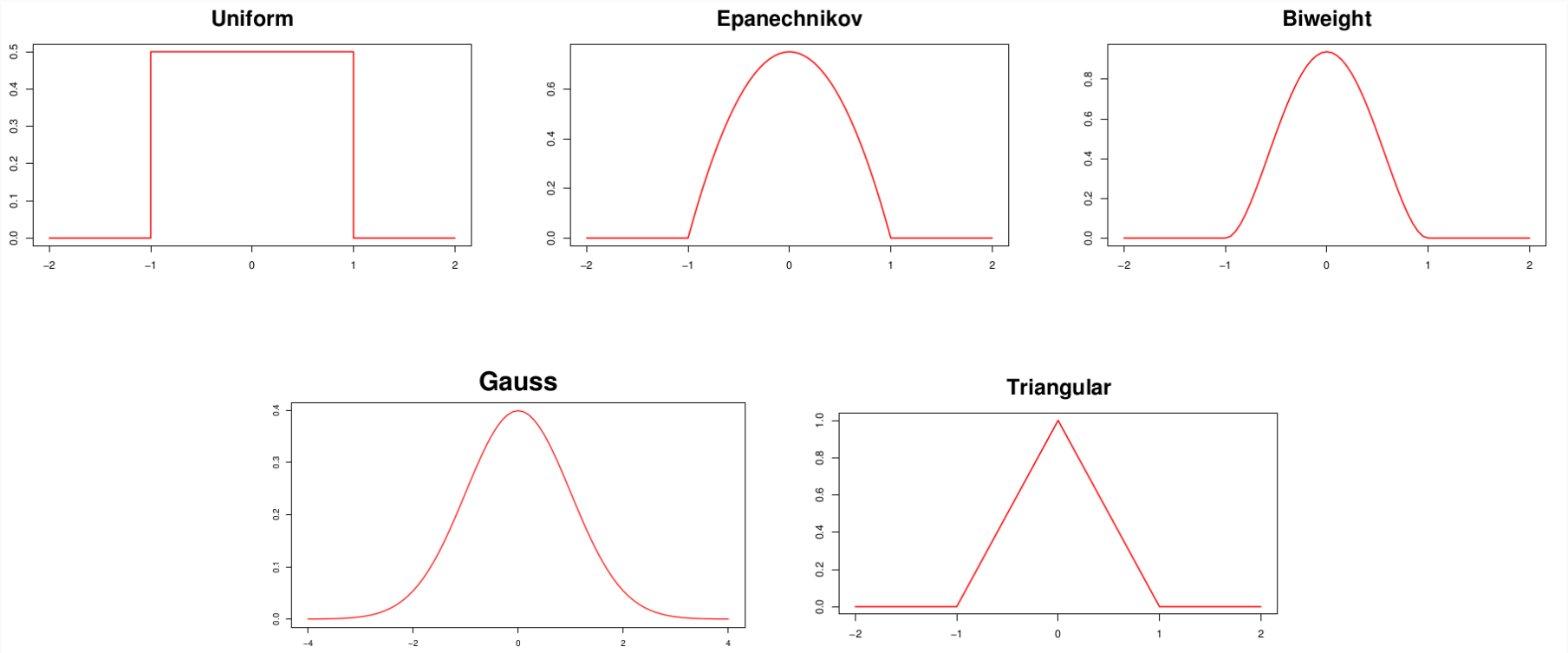
KDE

Dado un kernel K y un ancho de banda h, la KDE es *única* para un conjunto de datos específico, entonces, **no depende** de la selección del origen, como pasa con los histogramas.

El Kernel K

- Puede ser una función de densidad también, generalmente se selecciona una función unimodal y simétrica.
- El centro del kernel se coloca sobre cada dato x_i
- La influencia de cada dato se propaga en su vecindad
- La contribución de cada punto se suma para la estimación total

KDE



KDE

Más importante que el Kernel, es la elección del ancho de banda.

El ancho de banda h

- Es un factor de escala
- Controla la suavidad o rugosidad de la estimación
- Introducimos un concepto importante: **sobreestimación**
- Esto a su vez, lleva a otro concepto aún mas importante: **Bias-Variance tradeoff**
- Veamos un ejemplo

KDE: Elección de h

Cómo elegimos h

- A "ojo" (¿qué quieres ver?)
- Dos criterios:
  - Normal scale rule (también conocida como "rule of thumb")
  - Validación cruzada (CV)

KDE: Elección de h *

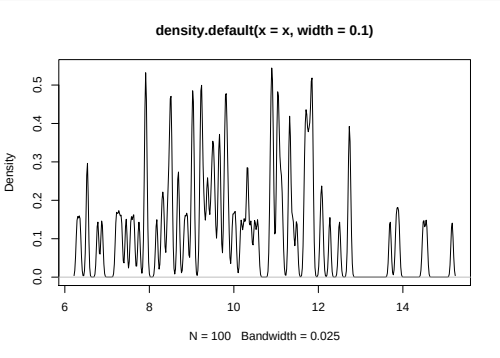
Cómo elegimos h

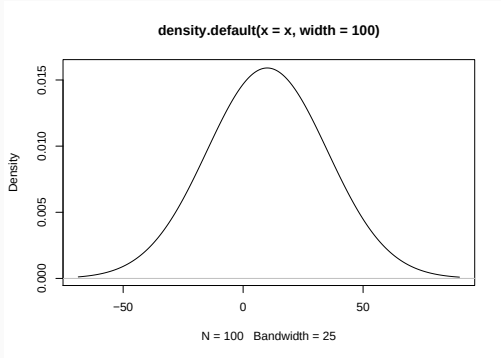
- Introducimos un criterio: MISE (mean integrated squared error). Ver notas.
- Normal scale rule (también conocida como "rule of thumb")
  - Asume que f es Normal, y calcula h óptima que minimice MISE según este supuesto.
  - Puede mostrarse que, si usamos un kernel Gaussiano, el h óptimo bajo este esquema es
  $$h^{ROT} = 1.06sn^{-1/5},$$
  con s es la estimación de σ.
  - La opción por default en `R`
  - Bien para un primer vistazo, pero tiende a suavizar de más cuando f es multimodal o claramente no-Gaussiana.
- Validación cruzada: utiliza el criterio *leave-one-out* CV.

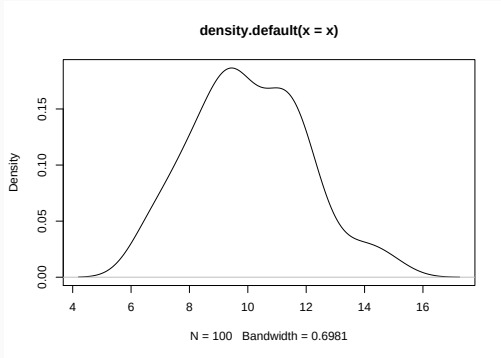
KDE

Retomamos el ejemplo anterior (pag 244)

```
?density
plot(density(x,width=0.1)) #kernel normal es el default
plot(density(x,width=100))
plot(density(x))
```








```
#otro kernels se pueden especificar con la opción ‘‘kernel’’
plot(density(x,width=0.1,kernel="triangular"))
```

Ejemplo

Sea $Y \in \{0, 1\}$ una variable aleatoria binaria que indica si una persona tiene *coronary heart disease (CHD)* (1) o no la tiene (0), y X una variable aleatoria contínua que representa su medición de *systolic blood pressure (SBP)*.

1. Estima $P(X = x \mid Y = 0)$ y $P(X = x \mid Y = 1)$ usando estimaciones basadas en un kernel Gaussiano.
2. Realiza una estimación de la probabilidad posterior de que un paciente tenga la enfermedad CHD en base a su medición de SBP, es decir:

$$P(Y = 1 \mid X = x) = \frac{P(X = x \mid Y = 1)\, P(Y = 1)}{P(X = x \mid Y = 1)\, P(Y = 1) + P(X = x \mid Y = 0)\, P(Y = 0)}.$$

   Utiliza las estimaciones usadas en el inciso anterior y distintos valores de ancho de banda del kernel. ¿Qué valor tomarías para la apriori $P(Y = 0)$ y $P(Y = 1)$? ¿Qué efecto tiene el parámetro del ancho de banda en los resultados?

EDNP: KDE *

KDE multivariado.

La extensión al caso multivariado es sencilla:

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n} K(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)), \quad \mathbf{x} \in \mathbb{R}^d,$$

donde $\mathbf{H}$ es una matriz de $d \times d$ no singular que generaliza el ancho de banda h, y K es una función con media $\mathbf{0}$ e integra 1.

Estimación bayesiana

Apéndice. Maximizar la verosimilitud

Apéndice. Maximizar la verosimilitud

La maximización se realiza numéricamente. Tenemos una gran variedad de métodos:

- Basados en derivadas: Newton, Quasi-Newton, etc...
- Estocásticos: recocido simulado
- Heurísticas: algoritmos genéticos, evolutivos, etc...

Apéndice. Maximizar la verosimilitud

Para un problema de optimización sin restricciones, consideraremos resolver lo siguiente:

$$\min_{\mathbf{x}} f(\mathbf{x}),$$

con $\mathbf{x} \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$.

¿Cómo lo resolvemos? o **¿Cómo sabemos si una solución es óptima?**

Hay toda una teoría de optimización, por el momento diremos que $\mathbf{x}^*$ es una solución (global) si

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \quad \forall \mathbf{x}.$$

Muchas veces, el óptimo global no se conoce, así que nos conformaremos con óptimos locales $\mathbf{x}^*$, tales que

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{N},$$

donde $\mathcal{N}$ es una vecindad de $\mathbf{x}^*$.

Apéndice. Maximizar la verosimilitud

La optimización es un proceso iterativo, donde buscamos **direcciones** $\mathbf{p} \in \mathbb{R}^d$ que minimicen gradualmente nuestra función objetivo hasta un punto estacionario donde (esperamos), se encuentra el óptimo. En cada iteración, daremos un paso α_t en dirección de $\mathbf{p}$, es decir:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \mathbf{p}_t.$$

¿Cómo escoger la dirección de descenso? La expansión de Taylor nos da una pista:

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})'\mathbf{p} + \frac{1}{2}\mathbf{p}'\nabla^2 f(\mathbf{x} + c\mathbf{p})\mathbf{p}.$$

Apéndice. Maximizar la verosimilitud

El método de Newton y sus derivados, usan como dirección de descenso:

$$\mathbf{p}_t = -\mathbf{B}_k^{-1} \nabla f(\mathbf{x})_k,$$

donde $\mathbf{B}_k = \nabla^2 f(\mathbf{x}_k)$, o alguna aproximación (Quasi-Newton).

Escoger el tamaño de paso α_t es más complicado. Hay varias propuestas (que verás en su momento), que en general, son una "negociación" entre encontrar la solución (descender lo suficiente) y que sea computacionalmente eficiente (tiempo de ejecución y exactitud).

Gradiente y Hessiano, tendrán que calcularse muchas veces, numéricamente. Métodos para calcularlos los verás (e implementarás) el próximo semestre.

Apéndice. Maximizar la verosimilitud: En R

Tenemos varias opciones como la libreria `maxLik` (`https://cran.r-project.org/web/packages/maxLik/index.html`)

```
library(maxLik)
?maxLik
Maximum likelihood estimation

Description:

This is the main interface for the ''maxLik'' package, and the
function that performs Maximum Likelihood estimation.  It is a
wrapper for different optimizers returning an object of class
"maxLik".  Corresponding methods handle the likelihood-specific
properties of the estimates, including standard errors.

Usage:
maxLik(logLik, grad = NULL, hess = NULL, start, method,
constraints=NULL, ...)
```

Apéndice. Maximizar la verosimilitud

```
maxLik                  package:maxLik                  R Documentation

Maximum likelihood estimation

Arguments:

logLik: log-likelihood function.  Must have the parameter vector as
the first argument.  Must return either a single
log-likelihood value, or a numeric vector where each
component is log-likelihood of the corresponding individual
observation.

grad: gradient of log-likelihood.  Must have the parameter vector
as the first argument.  Must return either a single gradient
vector with length equal to the number of parameters, or a
matrix where each row is the gradient vector of the
corresponding individual observation.  If NULL, numeric
gradient will be used.

hess: hessian of log-likelihood.  Must have the parameter vector as
the first argument.  Must return a square matrix.  If NULL,
numeric Hessian will be used.

start: numeric vector, initial value of parameters.  If it has
names, these will also be used for naming the results.

method: maximisation method, currently either "NR" (for
Newton-Raphson), "BFGS" (for
Broyden-Fletcher-Goldfarb-Shanno), "BFGSR" (for the BFGS
```

Apéndice. Maximizar la verosimilitud

```
maxLik                  package:maxLik                  R Documentation

Maximum likelihood estimation

Arguments:
hess: hessian of log-likelihood.  Must have the parameter vector as
the first argument.  Must return a square matrix.  If NULL,
numeric Hessian will be used.

start: numeric vector, initial value of parameters.  If it has
names, these will also be used for naming the results.

method: maximisation method, currently either "NR" (for
Newton-Raphson), "BFGS" (for
Broyden-Fletcher-Goldfarb-Shanno), "BFGSR" (for the BFGS
algorithm implemented in R), "BHHH" (for
Berndt-Hall-Hall-Hausman), "SANN" (for Simulated ANNealing),
"CG" (for Conjugate Gradients), or "NM" (for Nelder-Mead).
Lower-case letters (such as "nr" for Newton-Raphson) are
allowed.  If missing, a suitable method is selected
automatically.
```

Apéndice. Maximizar la verosimilitud

```
> ## ejemplo: exponencial theta=2
> N <- 100
> t <- rexp(N, 2)
> loglik <- function(theta) log(theta) - theta*t
> gradlik <- function(theta) 1/theta - t
> hesslik <- function(theta) -N/theta^2
> ## Estimate with numeric gradient and hessian
> a <- maxLik(loglik, start=1, control=list(printLevel=2))
----- Initial parameters: -----
fcn value: -45.56026
parameter initial gradient free
[1,]         1         54.43974    1
Condition number of the (active) hessian: 1
-----Iteration 1 -----
-----Iteration 2 -----
-----Iteration 3 -----
-----Iteration 4 -----
-----Iteration 5 -----
gradient close to zero
5  iterations
estimate: 2.194895

estimate: 2.062606    Function value: -21.38656
```

Apéndice. Maximizar la verosimilitud

```
> summary( a )
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 5 iterations
Return code 1: gradient close to zero
Log-Likelihood: -21.38656
1  free parameters
Estimates:
Estimate Std. error t value Pr(> t)
[1,]   2.1949     0.2195   9.999  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

Apéndice. Maximizar la verosimilitud

```
> ## Estimate with analytic gradient and hessian
> a <- maxLik(loglik, gradlik, hesslik, start=1)
> summary( a )
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 5 iterations
Return code 1: gradient close to zero
Log-Likelihood: -27.60296
1  free parameters
Estimates:
Estimate Std. error t value Pr(> t)
[1,]   2.0626     0.2063      10  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

Apéndice. Maximizar la verosimilitud

```
> ## ejemplo: gaussiana
> loglik <- function(param) {
+     mu <- param[1]
+     sigma <- param[2]
+     ll <- -0.5*N*log(2*pi) - N*log(sigma) - sum(0.5*(x - mu)^2/sigma^2)
+     ll
+ }
> N <- 100
> x <- rnorm(N, 1, 2) # mean=1, stdd=2
> res <- maxLik(loglik, start=c(0,1))
```

Apéndice. Maximizar la verosimilitud

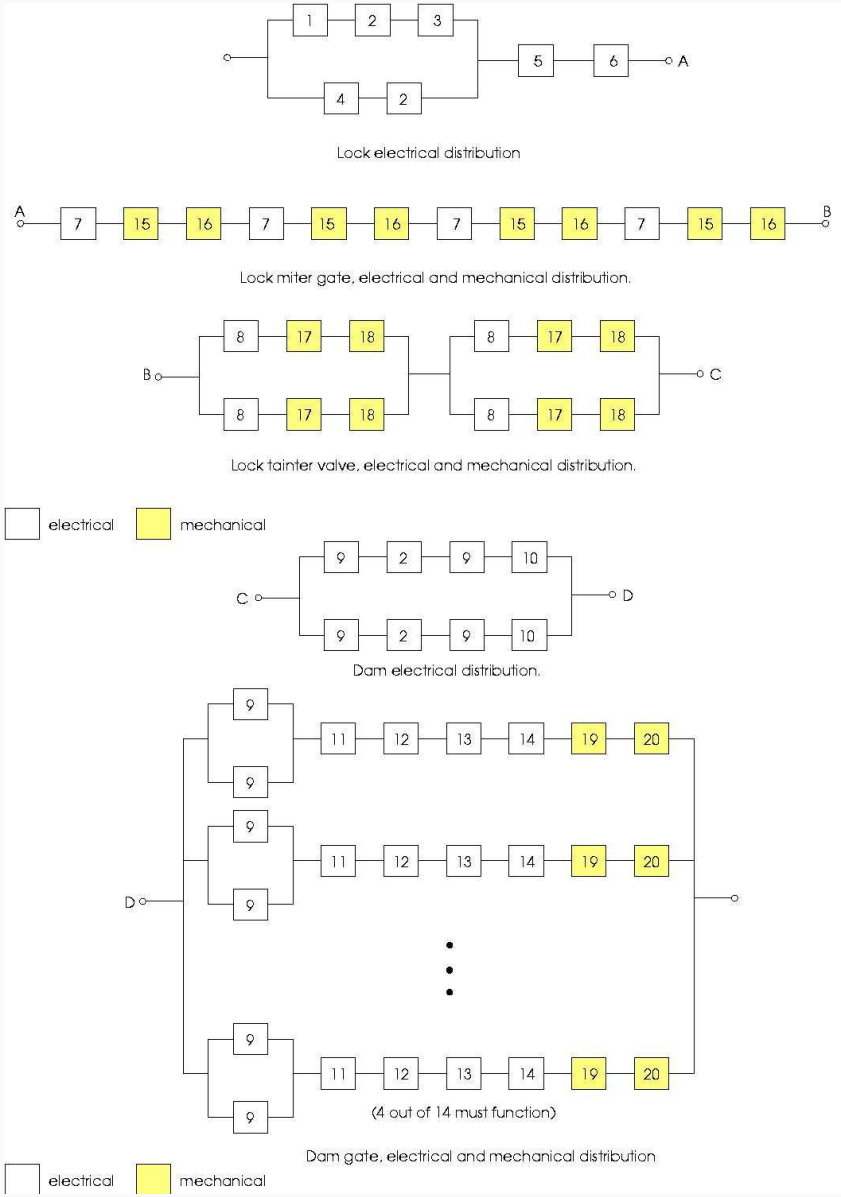
```
> ## ejemplo: gaussiana
> summary( res )
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 7 iterations
Return code 2: successive function values within tolerance limit
Log-Likelihood: -207.0033
2  free parameters
Estimates:
Estimate Std. error t value  Pr(> t)
[1,]   1.4605     0.1917   7.617 2.59e-14 ***
[2,]   1.9176     0.1357  14.136  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
--------------------------------------------
```

Apéndice. Maximizar la verosimilitud

Maximizar la verosimilitud puede ser muy complejo.

Piensa por ejemplo, en un problema de Confiabilidad de Sistemas.

Apéndice. Maximizar la verosimilitud



Apéndice. Maximizar la verosimilitud

Una opción: Heurísticas de optimización, por ejemplo, algoritmos evolutivos.

- Son algoritmos evolutivos probabilísticos que mantienen un conjunto (*población*) de *individuos* $P(t) = \{x_1^t, \ldots, x_n^t\}$ en cada iteración t y cada individuo representa una solución potencial al problema en cuestión.
- Están "inspirados" en un modelo de evolución biológica natural. Estos modelan el proceso colectivo de aprendizaje dentro de una *población* de *individuos*, cada uno de los cuales representa un punto de búsqueda en el espacio de soluciones potenciales a un problema dado.

Vista general

begin

$t \leftarrow 0$ *initialize* $P(t)$

evaluate $P(t)$

while (not termination-condition) do

begin

$t \leftarrow t + 1$

select $P(t)$ *from* $P(t-1)$

alter $P(t)$

evaluate $P(t)$

end

end

Apéndice. Maximizar la verosimilitud

- Algoritmo Genético (AG). Programa Evolutivo donde cada individuo es representado mediante una cadena de bits de longitud fija.

Algoritmo Genético Simple

begin

$t \leftarrow 0$ *initialize* $P(t)$

evaluate $P(t)$

while (not termination-condition) do

begin

$P'(t) \leftarrow$ *select from* $P(t)$ *(selection operator)*

$P''(t) \leftarrow$ *crossover* $P'(t)$ *(crossover operator)*

$P'''(t) \leftarrow$ *mutate* $P''(t)$ *(mutation operator)*

$P(t+1) \leftarrow P'''(t)$

evaluate $P(t+1)$

$t \leftarrow t+1$

end

end

Apéndice. Maximizar la verosimilitud

Estrategias evolutivas.

- Programas evolutivos donde se usa una representación en punto flotante
- Existen versiones individuales y poblacionales
- Los hijos (o hijo) generados son evaluados y comparados contra sus padres y el mejor de ellos sobrevive para formar parte (como un nuevo padre) en la siguiente generación

Ejemplo: algoritmo EE-$(\mu + \lambda)$.

$\mu =$ número de padres

$\lambda =$ número de hijos

begin

$t \leftarrow 0$ *initialize* $P(0) := \{a_1(0), \ldots, a_\mu(0)\}$

evaluate $P(t) := \{\Phi(a_1(0)), \ldots, \Phi(a_\mu(0))\}$ *where* $\Phi(a_k(0)) = f(x_k(0))$ $k = 1, \ldots, \mu$

while (not termination-condition) do

$a'_k(t) \leftarrow c'(P(t))$ $k = 1, \ldots, \lambda$ *(crossover)*

$a''_k(t) \leftarrow m'_{\tau,\tau',\beta}(a'_k(t))$ $k = 1, \ldots, \lambda$ *(mutation)*

$P''(t) \leftarrow \{a''_1(t), \ldots, a''_\lambda(t)\}$

evaluate $P''(t) := \{\Phi(a''_1(t)), \ldots, \Phi(a''_\lambda(t))\}$ *where* $\Phi(a''_k(t)) = f(x''_k(t))$ $k = 1, \ldots, \lambda$

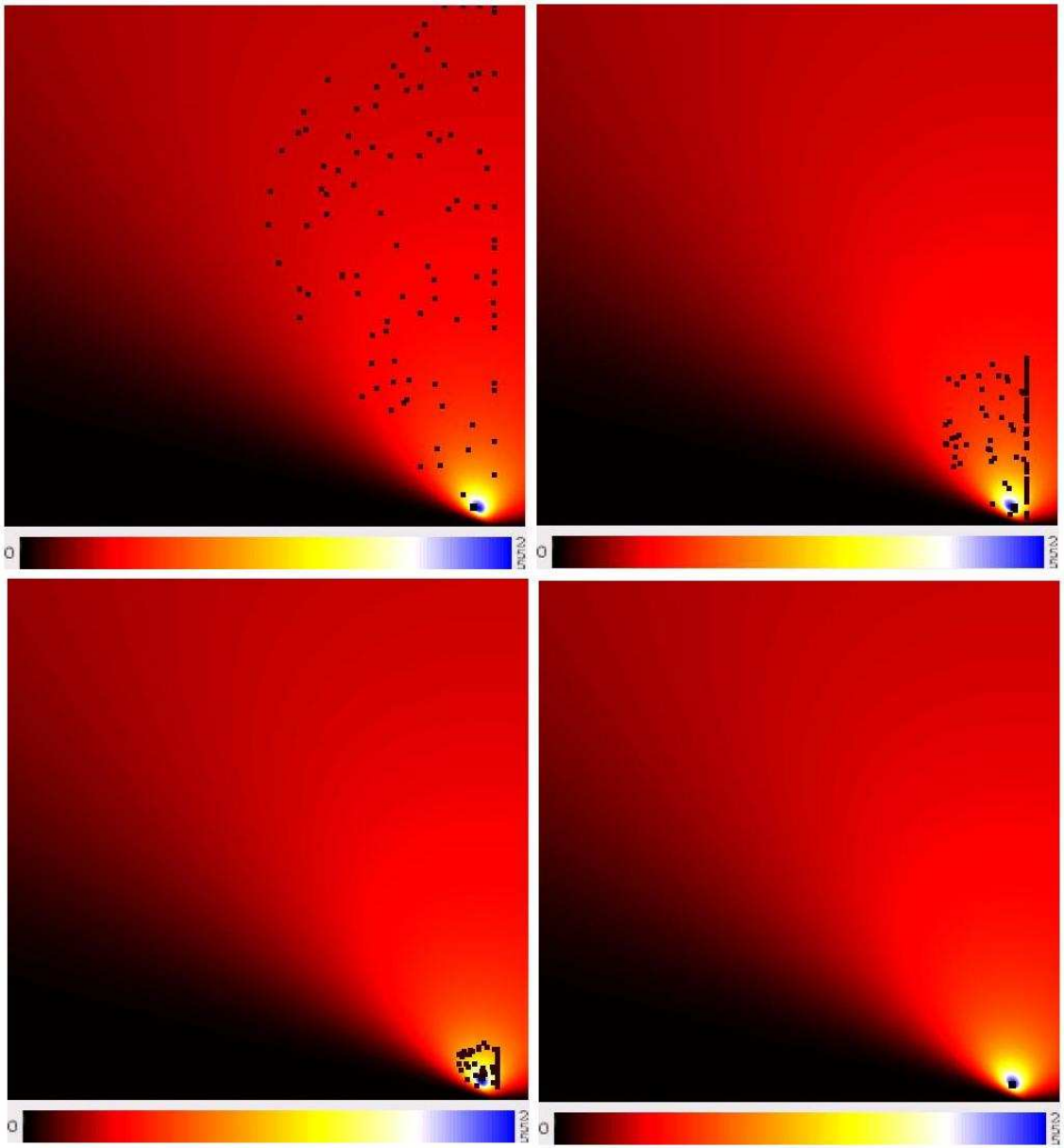
$P(t+1) := s_{(\mu+lambda)}(P(t) \bigcup P''(t))$ *selection*

$t \leftarrow t + 1$

end

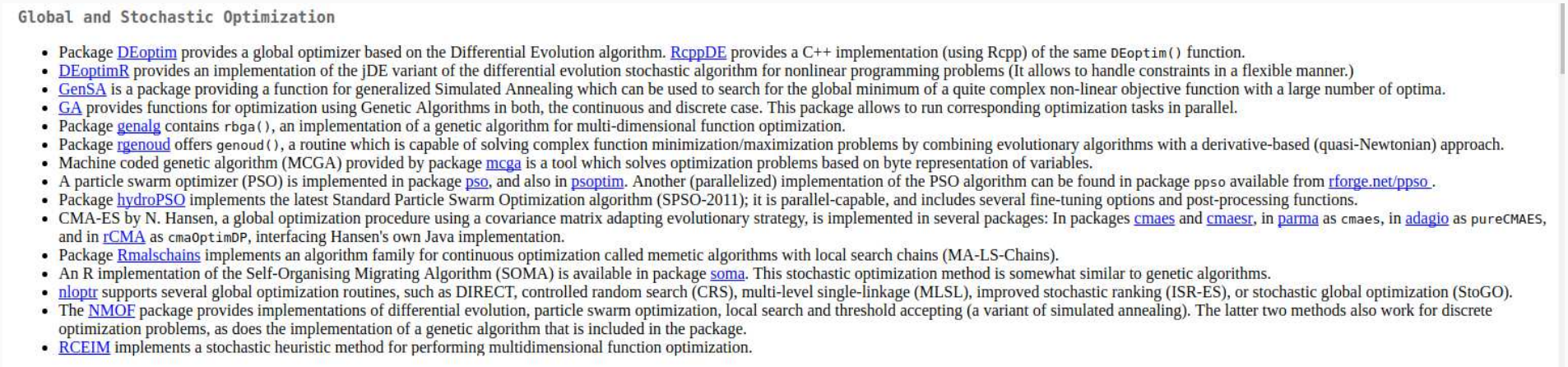
end

Apéndice. Maximizar la verosimilitud



Apéndice. Maximizar la verosimilitud

En R, puedes recurrir a https://cran.r-project.org/, -->packages-->CRAN Task Views-->Optimization

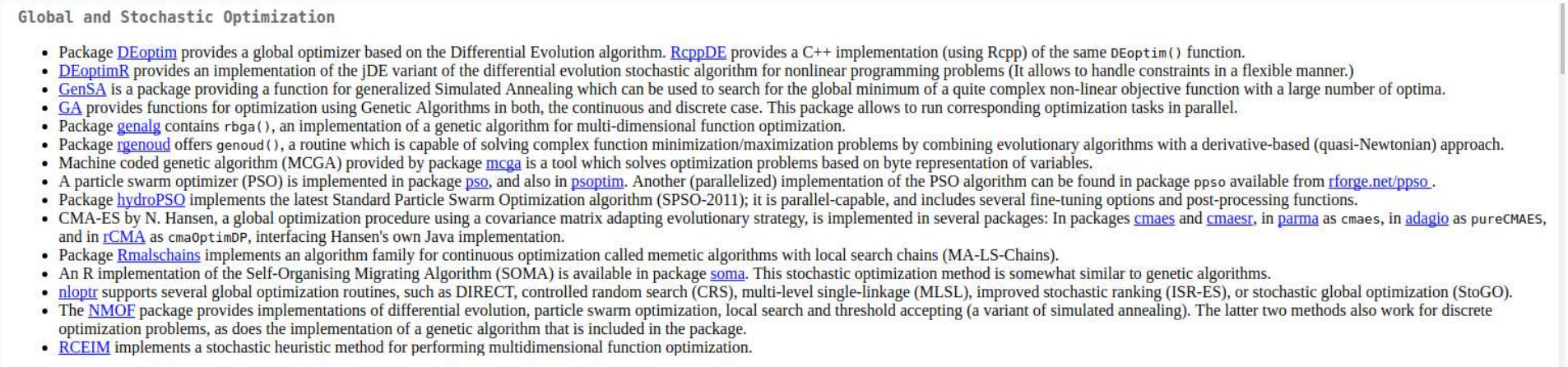


Global and Stochastic Optimization

- Package DEoptim provides a global optimizer based on the Differential Evolution algorithm. RcppDE provides a C++ implementation (using Rcpp) of the same `DEoptim()` function.
- DEoptimR provides an implementation of the jDE variant of the differential evolution stochastic algorithm for nonlinear programming problems (It allows to handle constraints in a flexible manner.)
- GenSA is a package providing a function for generalized Simulated Annealing which can be used to search for the global minimum of a quite complex non-linear objective function with a large number of optima.
- GA provides functions for optimization using Genetic Algorithms in both, the continuous and discrete case. This package allows to run corresponding optimization tasks in parallel.
- Package genalg contains `rbga()`, an implementation of a genetic algorithm for multi-dimensional function optimization.
- Package rgenoud offers `genoud()`, a routine which is capable of solving complex function minimization/maximization problems by combining evolutionary algorithms with a derivative-based (quasi-Newtonian) approach.
- Machine coded genetic algorithm (MCGA) provided by package mcga is a tool which solves optimization problems based on byte representation of variables.
- A particle swarm optimizer (PSO) is implemented in package pso, and also in psoptim. Another (parallelized) implementation of the PSO algorithm can be found in package `ppso` available from rforge.net/ppso .
- Package hydroPSO implements the latest Standard Particle Swarm Optimization algorithm (SPSO-2011); it is parallel-capable, and includes several fine-tuning options and post-processing functions.
- CMA-ES by N. Hansen, a global optimization procedure using a covariance matrix adapting evolutionary strategy, is implemented in several packages: In packages cmaes and cmaesr, in parma as `cmaes`, in adagio as `pureCMAES`, and in rCMA as `cmaOptimDP`, interfacing Hansen's own Java implementation.
- Package Rmalschains implements an algorithm family for continuous optimization called memetic algorithms with local search chains (MA-LS-Chains).
- An R implementation of the Self-Organising Migrating Algorithm (SOMA) is available in package soma. This stochastic optimization method is somewhat similar to genetic algorithms.
- nloptr supports several global optimization routines, such as DIRECT, controlled random search (CRS), multi-level single-linkage (MLSL), improved stochastic ranking (ISR-ES), or stochastic global optimization (StoGO).
- The NMOF package provides implementations of differential evolution, particle swarm optimization, local search and threshold accepting (a variant of simulated annealing). The latter two methods also work for discrete optimization problems, as does the implementation of a genetic algorithm that is included in the package.
- RCEIM implements a stochastic heuristic method for performing multidimensional function optimization.

Apéndice. Maximizar la verosimilitud

En `R`, puedes recurrir a `https://cran.r-project.org/`, `-->packages-->CRAN Task Views-->Optimization`



Global and Stochastic Optimization

- Package DEoptim provides a global optimizer based on the Differential Evolution algorithm. RcppDE provides a C++ implementation (using Rcpp) of the same `DEoptim()` function.
- DEoptimR provides an implementation of the jDE variant of the differential evolution stochastic algorithm for nonlinear programming problems (It allows to handle constraints in a flexible manner.)
- GenSA is a package providing a function for generalized Simulated Annealing which can be used to search for the global minimum of a quite complex non-linear objective function with a large number of optima.
- GA provides functions for optimization using Genetic Algorithms in both, the continuous and discrete case. This package allows to run corresponding optimization tasks in parallel.
- Package genalg contains `rbga()`, an implementation of a genetic algorithm for multi-dimensional function optimization.
- Package rgenoud offers `genoud()`, a routine which is capable of solving complex function minimization/maximization problems by combining evolutionary algorithms with a derivative-based (quasi-Newtonian) approach.
- Machine coded genetic algorithm (MCGA) provided by package mcga is a tool which solves optimization problems based on byte representation of variables.
- A particle swarm optimizer (PSO) is implemented in package pso, and also in psoptim. Another (parallelized) implementation of the PSO algorithm can be found in package `ppso` available from rforge.net/ppso .
- Package hydroPSO implements the latest Standard Particle Swarm Optimization algorithm (SPSO-2011); it is parallel-capable, and includes several fine-tuning options and post-processing functions.
- CMA-ES by N. Hansen, a global optimization procedure using a covariance matrix adapting evolutionary strategy, is implemented in several packages: In packages cmaes and cmaesr, in parma as `cmaes`, in adagio as `pureCMAES`, and in rCMA as `cmaOptimDP`, interfacing Hansen's own Java implementation.
- Package Rmalschains implements an algorithm family for continuous optimization called memetic algorithms with local search chains (MA-LS-Chains).
- An R implementation of the Self-Organising Migrating Algorithm (SOMA) is available in package soma. This stochastic optimization method is somewhat similar to genetic algorithms.
- nloptr supports several global optimization routines, such as DIRECT, controlled random search (CRS), multi-level single-linkage (MLSL), improved stochastic ranking (ISR-ES), or stochastic global optimization (StoGO).
- The NMOF package provides implementations of differential evolution, particle swarm optimization, local search and threshold accepting (a variant of simulated annealing). The latter two methods also work for discrete optimization problems, as does the implementation of a genetic algorithm that is included in the package.
- RCEIM implements a stochastic heuristic method for performing multidimensional function optimization.

Aprenderás más en los siguientes semestres!