Inferencia Estadística

Dra. Graciela González Farías
Dr. Ulises Márquez

Maestría en Cómputo Estadístico.

CIMAT Monterrey.

Agradecimientos

En forma de agradecimiento, se enlistan personas que han contribuido de una u otra forma en la construcción de estas notas a través de los años:

- 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 involucrados.

Temario

1. Variables aleatorias y distribuciones de probabilidad.
   a) Distribuciones de probabilidad de variables aleatorias discretas.
   b) Procesos de Poisson.
   c) Distribuciones de probabilidad de variables aleatorias continuas.
   d) Métodos gráficos para la identificación de distribuciones.
   e) Estimación de densidades.
   f) Distribuciones de probabilidad de vectores aleatorios.
   g) Esperanzas condicionales y regresión.
   h) Modelos jerárquicos, compuestos y mezclas de variables aleatorias.
   i) Transformaciones de variables aleatorias.
   j) Simulación de variables aleatorias.
   k) Convergencia de variables aleatorias y el Teorema del Límite Central.

Temario

2. Distribuciones muestrales y métodos de estimación.
   a) Propiedades de los estimadores.
   b) Estimadores no insesgados
   c) Distribuciones muestrales.
   d) Principio de máxima verosimilitud.
   e) Estimación puntual.
   f) Bootstrap y jackknife.

Temario

3. Pruebas de Hipótesis e intervalos de confianza.
   a) Definición de conceptos.
   b) Potencia de la prueba.
   c) Pruebas para dos poblaciones normales independientes.
   d) Pruebas para medias en muestras pareadas.
   e) Pruebas básicas de varianzas.
   f) Pruebas para proporciones.
   g) Conceptos de estimación bayesiana.
   h) Temas optativos de modelos para presentaciones finales, por ejemplo:
      1. Pruebas no-paramétricas clásicas.
      2. Pruebas de permutaciones.
      3. Estimación no paramétrica (suavizadores y splines).
      4. Modelos gráficos probabilístas.
      5. Entre muchos otros.

Evaluación y acreditación

- Dos exámenes parciales, 18 de septiembre y 5 de noviembre: 15 %, cada uno.
- Evaluación de las tareas (de 2 tipos) y actividades en clase y asistencia: 40 %.
- Un examen final, consistente en una exposición donde se entrega un reporte y se hace una presentación de 1/2 hora. La presentación debe incluir antecedentes, metodología, un ejemplo práctico y compartir el código. Deberán entregar a los instructores y a sus compañeros el resumen. Adicionalmente, deberán dejar un ejercicio sobre el tema a sus compañeros que calificarán en forma honesta: 30 %.

Las tareas tienen una frecuencia quincenal e incluyen TODOS los ejercicios dejados en las notas y requerirán en general el uso de recursos computacionales.

Textos

- **Larry Wasserman (2004) . All of Statistics, A concise course in Statistical Inference. Springer.**
- F.M. Dekking, C. Kraaikamp, H.P. Lopuhaa L.E. Meester (2005). A Modern Introduction to Probability and Statistics, Understanding Why and How. Springer text in Statistics.
- John A. Rice (1995). Mathematical Statistics and Data Analysis, Second Edition. Duxbury Press.
- Casella & Berger. (2002). Statistical Inference, Second Edition . Duxbury Press.
- Richard J. Larsen and Morris L. Marx (2011). An Introduction to Mathematical Statistics and its Applications. Fifth Edition. Prentice Hall.

Introducción: Modos de convergencia

- Ya que la estadística tiene que ver con recolectar información, naturalmente estamos interesados en saber que pasa cuando juntamos más y más datos.
- ¿Qué podemos decir del compartamiento límite de una sucesión de variables aleatorias $X_1, X_2, \ldots$?
- El estudio del comportamiento de sucesiones de variables aleatorias es uno de los aspectos más importantes de la teoría de la probabilidad. Esta parte de la probabilidad recibe los nombres de Teoría límite, Teoría asintótica o Teoría de grandes muestras.

Introducción: Modos de convergencia

Recordemos de Cálculo que una sucesión $\{x_n\}$ converge a un límite x si

$$\forall \epsilon > 0,\ \exists N > 0 \text{ tal que } |x_n - x| < \epsilon\ \forall n > N.$$

Veamos con un ejemplo porque la definición anterior es insuficiente para trabajar con variables aleatorias.

Si $x_n = x \in \mathbb{R}$, obtenemos que $\lim_{n\to\infty} x_n = x$. Consideremos una versión probabilistica de esto: Supongamos que $\{X_n\}$ es una sucesión de v.a. aleatorias independientes con distribución $N(0,1)$. ¿$X_n \to X \sim N(0,1)$?

Lo anterior no puede ser cierto pues $P(X_n = X) = 0$ para todo $n \in \mathbb{N}$. Lo anterior deja claro que en probabilidad necesitamos algo más sútil para establecer convergencias.

Introducción: Modos de convergencia

Tomemos otro ejemplo. Consideremos $X_1, X_2, \ldots$ donde $X_i \sim N(0, 1/i)$.

X_n está muy concentrada alrededor de 0 para n grande, por lo que nos gustaría decir $X_n \to 0$. Sin embargo, $P(X_n = 0) = 0$ para todo n.

Estudiaremos cinco conceptos de convergencia de variables aleatorias: convergencia en probabilidad, convergencia en distribución, convergencia en L_2, convergencia casi segura y convergencia en L_1.

La convergencia en L_1 y L_2 son casos particulares de la convergencia L_p. Debido a que queda fuera de los objetivos del curso, no estudiaremos está convergencia. Sin embargo las ideas son similares a los casos L_1 y L_2.

Tipos de convergencia

Definición

Sea X_n una sucesión de v.a. y sea X una v.a. Denotemos por F_n a la función de distribución de X_n y por F a la función de distribución de X.

1. *X_n converge a X en probabilidad, denotado por $X_n \xrightarrow{P} X$, si para cada $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

2. *X_n converge a X en distribución, denotado por $X_n \xrightarrow{d} X$, si*

$$\lim_{n\to\infty} F_n(t) = F(t)$$

para todo t donde F es continua.

3. *X_n converge a X en media cuadrática (o en L_2), denotado por $X_n \xrightarrow{L_2} X$, si*

$$\lim_{n\to\infty} E\left[(X_n - X)^2\right] = 0.$$

Tipos de convergencia

Ejemplo (Convergencia en distribución)

Sea $X_n \sim N(0, 1/n)$. *Tenemos que*

$$F_n(t) \equiv P(X_n < t) = P(\sqrt{n}X_n < \sqrt{n}t) = P(Z < \sqrt{n}t).$$

Luego,

$$\lim_{n\to\infty} F_n(t) = \begin{cases} 0 & si\ t < 0, \\ 1 & si\ t > 0. \end{cases}$$

¿Conocemos alguna función de distribución parecida? Consideremos la v.a. constante $Y = 0$. *Se tiene que la función de distribución de* Y *está dada por*

$$F_Y(t) = \begin{cases} 0 & si\ t < 0, \\ 1 & si\ t \geq 0. \end{cases}$$

Debido a que F_Y *no es continua en* 0, *podemos escribir que* $X_n \xrightarrow{d} Y \equiv 0$.

Tipos de convergencia

Notemos que en el ejemplo anterior $F_n(0) = 1/2$ para todo n y que $F_n(0) = 1/2 \neq 1 = F_Y(0)$. Es decir que la convergencia falla en $t = 0$; sin embargo esto no importa porque F_Y no es continua en $t = 0$.

Ejemplo (Convergencia en probabilidad)

Como antes, sea $X_n \sim N(0, 1/n)$. *La desigualdad de Markov implica que, para* $\epsilon > 0$,

$$P(|X_n| > \epsilon) = P(|X_n|^2 > \epsilon^2) \leq \frac{E(X_n^2)}{\epsilon^2} = \frac{(1/n)}{\epsilon^2} \to 0$$

cuando $n \to \infty$. *Así que* $X_n \xrightarrow{P} Y \equiv 0$.

Tipos de convergencia

El siguiente resultado establece relaciones entre los distintos tipos de convergencia.

Teorema

Se tienen las siguientes relaciones:

a) $X_n \xrightarrow{L_2} X$ *implica que* $X_n \xrightarrow{P} X$.

b) $X_n \xrightarrow{P} X$ *implica que* $X_n \xrightarrow{d} X$.

c) *Si* $X_n \xrightarrow{d} X$ *y si* $P(X = c) = 1$ *para algún número real* c, *entonces* $X_n \xrightarrow{P} X$.

En general ninguna de las implicaciones inversas se cumplen, a excepción del caso especial en c).

Tipos de convergencia

Demostración

a) Supongamos que $X_n \xrightarrow{L_2} X$. *Sea* $\epsilon > 0$ *fijo. De la desigualdad de Markov se sigue que*

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) \leq \frac{E\left[|X_n - X|^2\right]}{\epsilon^2} \to 0.$$

b) Sea $\epsilon > 0$ *y* x *un punto de continuidad de* F. *Entonces*

$$\begin{aligned} F_n(x) &= P(X_n \leq x) \\ &= P(X_n \leq x, X \leq x + \epsilon) + P(X_n \leq x, X > x + \epsilon) \\ &\leq P(X \leq x + \epsilon) + P(|X_n - X| > \epsilon). \end{aligned}$$

Tipos de convergencia

Demostración (Continuación)

También

$$\begin{aligned} F(x-\epsilon) &= P(X \leq x-\epsilon) \\ &= P(X \leq x-\epsilon, X_n \leq x) + P(X \leq x-\epsilon, X_n > x) \\ &\leq F_n(x) + P(|X_n - X| > \epsilon). \end{aligned}$$

Por lo tanto,

$$F(x-\epsilon) - P(|X_n - X| > \epsilon) \leq F_n(x) \leq F(x+\epsilon) - P(|X_n - X| > \epsilon).$$

Tomando el límite es $n \to \infty$, concluimos que

$$F(x-\epsilon) \leq \lim_{n\to\infty} F_n(x) \leq F(x+\epsilon).$$

Tipos de convergencia

Demostración (Continuación)

Debido a que F es continua en x, al hacer $\epsilon \downarrow 0$, llegamos a que

$$\lim_{n\to\infty} F_n(x) = F(x).$$

c) Sea $\epsilon > 0$ fijo. Entonces,

$$\begin{aligned} P(|X_n - c| > \epsilon) &= P(X_n < c - \epsilon) + P(X_n > c + \epsilon) \\ &\leq P(X_n \leq c - \epsilon) + P(X_n > c + \epsilon) \\ &= F_n(c - \epsilon) + 1 - F_n(c + \epsilon) \\ &\to F(c - \epsilon) + 1 - F(c + \epsilon) \\ &= 0 + 1 - 1 = 0. \end{aligned}$$

Ahora demostraremos que las implicaciones inversas no se cumplen.

Tipos de convergencia

Demostración (Continuación)

Convergencia en probabilidad no implica convergencia en L_2. *Sea* $U \sim Uniforme(0,1)$ *y* $X_n = \sqrt{n}1_{(0,1/n)}(U)$. *Entonces, si* $\epsilon > 0$,

$$P(|X_n| > \epsilon) = P(\sqrt{n}1_{(0,1/n)}(U) > \epsilon) = P(0 \leq U < 1/n) = 1/n \to 0.$$

Por lo que, $X_n \xrightarrow{P} 0$. *Sin embargo* $E(X_n^2) = n\int_0^{1/n} = 1$ *para todo n, lo que implica que* X_n *no converge en* L_2.

Tipos de convergencia

Demostración (Continuación)

Convergencia en distribución no implica convergencia en probabilidad. *Sea* $X \sim N(0,1)$ *y* $X_n = -X$ *para* $n \in \mathbb{N}$. *Así,* $X_n \sim N(0,1)$. *Como* X_n *tiene la misma distribución que* X *para todo* n, *luego* $\lim F_n(x) = F(x)$ *para todo* x. *Por lo tanto* $X_n \xrightarrow{d} X$. *Pero, si* $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) = P(|2X| > \epsilon) = P(|X| > \epsilon/2) \neq 0.$$

Así que X_n *no converge en probabilidad.*

Tipos de convergencia

Uno podría pensar que si $X_n \xrightarrow{P} b$, para b constante, entonces $E(X_n) \to b$. Sin embargo esto no se cumple.

Sea X_n una variable aleatoria definida por $P(X_n = n^2) = 1/n$ y $P(X_n = 0) = 1 - 1/n$. Notemos que

$$P(|X_n| < \epsilon) = P(X_n = 0) = 1 - (1/n) \to 1,$$

para $\epsilon > 0$ pequeño. Así que $X_n \xrightarrow{P} 0$. Sin embargo,

$$E(X_n) = n^2(1/n) + 0 \cdot (1 - 1/n) = n.$$

Así, $E(X_n) \to \infty$.

Tipos de convergencia

Algunas propiedades de convergencia se conservan bajo transformaciones.

Teorema

Sean X_n, X, Y_n, Y v.a. Sea g una función continua y $c \in \mathbb{R}$.

a) *Si* $X_n \xrightarrow{P} X$ *y* $Y_n \xrightarrow{P} Y$ *entonces* $X_n + Y_n \xrightarrow{P} X + Y$.

b) *Si* $X_n \xrightarrow{L_2} X$ *y* $Y_n \xrightarrow{L_2} Y$ *entonces* $X_n + Y_n \xrightarrow{L_2} X + Y$.

c) *Si* $X_n \xrightarrow{d} X$ *y* $Y_n \xrightarrow{d} c$ *entonces* $X_n + Y_n \xrightarrow{d} X + c$.

d) *Si* $X_n \xrightarrow{P} X$ *y* $Y_n \xrightarrow{P} Y$ *entonces* $X_n Y_n \xrightarrow{P} XY$.

e) *Si* $X_n \xrightarrow{d} X$ *y* $Y_n \xrightarrow{d} c$ *entonces* $X_n Y_n \xrightarrow{d} cX$.

f) *Si* $X_n \xrightarrow{P} X$ *entonces* $g(X_n) \xrightarrow{P} g(X)$.

g) *Si* $X_n \xrightarrow{d} X$ *entonces* $g(X_n) \xrightarrow{d} g(X)$.

Tipos de convergencia

- Los incisos c) y e) son conocidos como el Teorema de Slutzky.
- $X_n \xrightarrow{d} X$ y $Y_n \xrightarrow{d} Y$ en general no implican que $X_n + Y_n \xrightarrow{d} X + Y$.

Existen otros tipos de convergencia que estudiaremos: la convergencia casi segura (c.s.) y la convergencia en L_1.

Definición

Diremos que X_n converge casi seguramente a X, denotado por $X_n \xrightarrow{c.s.} X$, si

$$P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1.$$

Además diremos que X_n converge en L_1 a X, denotado por $X_n \xrightarrow{L_1} X$ si

$$E\left[|X_n - X|\right] \to 0$$

cuando $n \to \infty$.

Tipos de convergencia

A continuación establecemos relaciones entre los nuevos tipos de convergencia introducidos y los introducidos al principio de las diapositivas.

Teorema

Sean $\{X_n\}$ una sucesión de variables aleatorias y X una variable aleatoria. Entonces,

a) $X_n \xrightarrow{c.s.} X$ *implica que* $X_n \xrightarrow{P} X$.

b) $X_n \xrightarrow{L_2} X$ *implica que* $X_n \xrightarrow{L_1} X$.

c) $X_n \xrightarrow{L_1} X$ *implica que* $X_n \xrightarrow{P} X$.

Tipos de convergencia

- Convergencia en probabilidad no implica convergencia casi segura (Borel-Cantelli).
- Convergencia en probabilidad no implica convergencia L_1.
- Convergencia en probabilidad de una sucesión uniformemente integrable implica convergencia en L_1.
- Convergencia casi segura no implica convergencia en L_1. Tomemos X_n de tal manera que $P(X_n = 0) = 1 - 1/n^2$ y $P(X_n = 2^n) = 1/n^2$. Claramente $X_n \xrightarrow{c.s.} 0$ pero $E(X_n)$ no converge.

Ley de los Grandes Números

A continuación estudiaremos un resultado muy importante de Teoría de la Probabilidad: **La Ley de Grandes Números**.

Este Teorema nos dice que la media de una muestra grande es cercana a la media de la distribución.

La primera versión y prueba formal se debe a Jacob Bernoulli. Bernoulli dio una prueba rigurosa para el caso en el que las variables aleatorias son independientes e identicamente distribuidas con distribución *Bernoulli*$(1/2)$. Le tomó 20 años poderlo demostrar.

La designación de Ley de los Grandes Números se debe a Poisson.

Ley de los Grandes Números

Teorema (Ley Débil General de los Grandes Números)

Supongamos que $\{X_n, n \geq 1\}$ *son variables aleatorias independientes y definamos* $S_n = \sum_{j=1}^{n} X_j$. *Si*

1. $\sum_{j=1}^{n} P(|X_n| > n) \to 0$,
2. $\frac{1}{n^2} \sum_{j=1}^{n} E[|X_j| 1_{\{|X_j| \leq n\}}] \to 0$;

entonces si definimos

$$a_n = \sum_{j=1}^{n} E[X_j 1_{\{|X_j| \leq n\}}],$$

obtenemos que

$$\frac{S_n - a_n}{n} \xrightarrow{P} 0.$$

Ley de los Grandes Números

Un caso particular importante del resultado anterior es la llamada Ley Débil de los Grandes Números de Khintchin.

Teorema (Ley Débil de los Grandes Números de Khintchin)

Supongamos que $\{X_n, n \geq 1\}$ *son v.a.i.i.d. con* $E(|X_1|) < \infty$ *y* $E(X_1) = \mu$. *Entonces,*

$$\frac{S_n}{n} \xrightarrow{P} \mu.$$

Observemos que no estamos requiriendo nada de la varianza en el resultado anterior. Si suponemos que $\text{Var}(X_1) < \infty$, la desigualdad de Chebyshev implica que

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\text{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2},$$

que tiende a 0 cuando $n \to \infty$. Esto constituye una demostración de un caso particular de LDGNK.

Ley de los Grandes Números

Ejemplo: Consideremos el lanzamiento de una moneda con probabilidad p de obtener cara. Denotemos por X_i el resultado de un lazamiento (0 ó 1). Sabemos que $E(X_1) = p$. La fracción de caras después de n lanzamientos es $\overline{X}_n$.

De acuerdo a la LDGNK, $\overline{X}_n$ converge a p en probabilidad. Sin embargo, esto no quiere decir que converja numéricamente (ver LFGNK) a p. Significa que cuando n es grande, la distribución de $\overline{X}_n$ está concentrada alrededor de p.

Supongamos que $p = 1/2$. ¿Qué tan grande debe de ser n para que $P(0.4 \leq \overline{X}_n \leq 0.6) \geq 0.7$?

Ley de los Grandes Números

Notemos que $E(\overline{X}_n) = p = 1/2$ y que $\text{Var}(\overline{X}_n) = p(1-p)/n = 1/(4n)$.
De la desigualdad de Chebyshev

$$\begin{aligned} P(0.4 \leq \overline{X}_n \leq 0.6) &= P(|\overline{X}_n - \mu| \leq 0.1) \\ &= 1 - P(|\overline{X}_n - \mu| > 0.1) \\ &\geq 1 - \frac{1}{4n(0.1)^2} \\ &= 1 - \frac{25}{n} \end{aligned}$$

La expresión anterior será mayor que 0.7 si $n = 84$.

Ley de los Grandes Números

Teorema (Ley Fuerte de Grandes Números de Kolmogorov)

Sea $\{X_n, n \geq 1\}$ *una sucesión de v.a.i.i.d. y pongamos* $S_n = \sum_{i=1}^{n} X_i$. *Existe* $c \in \mathbb{R}$ *tal que*

$$\overline{X}_n \xrightarrow{c.s} c$$

ssi $E(|X_1|) < \infty$, *en cuyo caso* $c = E(X_1)$.

Teorema (Corolario de la LFGNK)

Si $\{X_n\}$ *es una sucesión de v.a.i.i.d., entonces*

$$E(|X_1|) < \infty \quad \textit{implica que } \overline{X}_n \xrightarrow{c.s} \mu = E(X_1),$$

y

$$E(X_1^2) < \infty \quad \textit{implica que } \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2 \xrightarrow{c.s.} \sigma^2 = Var(X_1).$$

Ley de los Grandes Números

Una consecuencia muy importante de la LFGNK es el llamado Teorema de Glivenko-Cantelli. Este Teorema nos dice que la función distribución empirica es una aproximación uniforme de la verdadera función de distribución.

Teorema (Glivenko-Cantelli)

Sea $X_1, \ldots, X_n$ una muestra de variables aleatorias con función de distribución F. Además, sea F_n la función de distribución empírica de $X_1, \ldots, X_n$. Definamos

$$D_n = \sup_x |F_n(x) - F(x)|.$$

Entonces

$$D_n \xrightarrow{c.s.} 0,$$

cuando $n \to \infty$.

Distribución del Estimador de la Media

Teorema del Límite Central (TLC) Clásico: Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria (i.e. independientes e identicamente distribuidas) de una población arbitraria que tiene media μ y varianza σ^2, entonces cuando $n \to \infty$

$$Z_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

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.

Teorema del Límite Central

A continuación veremos algunas versiones alternativas o más generales del TLC. Omitiremos las demostraciones pues están quedan fuera de los objetivos del curso.

Estos teoremas nos darán condiciones sobre cuando podemos aproximar la distribución de la media muestral por una normal estándar.

Teorema del Límite Central

Sean $\{X_n, n \geq 1\}$ independientes (pero no necesariamente idénticamente distribuidas) y supongamos que X_k tiene distribución F_k y que $E(X_k) = 0$ y $\text{Var}(X_k) = \sigma_k^2$. Definamos

$$s_n^2 = \sum_{j=1}^{n} \sigma_j^2 = \text{Var}(\sum_{j=1}^{n} X_j).$$

Diremos que $\{X_k\}$ satisface la condición de Lindeberg si, para todo $t > 0$, cuando $n \to \infty$

$$\frac{1}{s_n^2} \sum_{j=1}^{n} E(X_j^2 1_{\{|X_j/s_n| > t\}}) \to 0.$$

Teorema del Límite Central

- La condición de Lindeberg nos dice que para cada k, la mayoría de la masa de X_k está centrada en un intervalo alrededor de la media $(= 0)$ y que este intervalo es pequeño relativo a s_n.
- La condición de Lindeberg implica que

$$\sup_{k \leq n} \frac{\sigma_k^2}{s_n^2} \to 0$$

cuando $n \to \infty$.

- De la desigualdad de Chebyshev, el límite anterior implica que

$$\max_{k \leq n} P(|X_k|/s_n > \epsilon) \to 0.$$

Esta condición es típica en teoremas del límite central y asegura de que no existe un sumando que domina y que cada sumando contribuye un pequeña porción del total.

Teorema del Límite Central

Teorema (TLC de Lindeberg-Feller)

Bajo la notación usada arriba, la condición de Lindeberg implica

$$\frac{S_n}{s_n} \equiv \frac{\sum_{i=1}^n X_i}{s_n} \xrightarrow{d} N(0,1).$$

El Teorema anterior puede extenderse de la siguiente manera.

Teorema (Condición de Liapunov)

Sea $\{X_k, k \geq 1\}$ *una sucesión de v.a.i. satisfaciendo* $E(X_k) = 0$, $Var(X_k) = \sigma_k^2 < \infty$ *y* $s_n^2 = \sum_{j=1}^n \sigma_k^2$. *Si para algún* $\delta > 0$

$$\frac{\sum_{k=1}^n E\left[|X_k|^{\delta+2}\right]}{s_n^{\delta+2}} \to 0,$$

entonces la condición de Lindeberg se cumple y, por lo tanto, también el TLC.

Teorema del Límite Central

Ejemplo (Récords)

Sea $\{X_n, n \geq 1\}$ *una sucesión de v.a. i.i.d. continuas. Decimos que* X_n *es un récord de la sucesión si*

$$X_n > \text{máx}\{X_1, \ldots, X_{n-1}\}$$

y definamos el evento

$$A_k = \{X_k \text{ es un récord}\}.$$

Un resultado debido a Renyi dice que los eventos $\{A_k, k \geq 1\}$ *son independientes y que*

$$P(A_j) = \frac{1}{j} \qquad j \geq 2.$$

Teorema del Límite Central

Ejemplo (Récords)

Definamos

$$1_k = 1_{A_k}, \qquad \mu_n = \sum_{j=1}^{n} 1_j.$$

Notemos que

$$E(1_k) = \frac{1}{k}, \qquad Var(1_k) = \frac{1}{k} - \frac{1}{k^2},$$

de donde

$$s_n^2 = Var(\mu_n) = \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k^2} \right) = \sum_{k=1}^{n} \frac{1}{k} - \sum_{k=1}^{n} \frac{1}{k^2} \sim \log n.$$

Así que

$$s_n^3 \sim (\log n)^{3/2}.$$

Teorema del Límite Central

Ejemplo (Récords)

Por otra parte,

$$E|1_k - E(1_k)|^3 = E\left|1_k - \frac{1}{k}\right|^3 = \left|1 - \frac{1}{k}\right|^3 \frac{1}{k} + \frac{1}{k^3}\left(1 - \frac{1}{k}\right) \leq \frac{1}{k} + \frac{1}{k^3},$$

y por lo tanto

$$\frac{\sum_{k=1}^n E|1_k - E(1_k)|^3}{s_n^3} \leq \frac{\sum_{k=1}^n \left(\frac{1}{k} + \frac{1}{k^3}\right)}{(\log n)^{3/2}} \sim \frac{\log n}{(\log n)^{3/2}} \to 0.$$

Esto implica que $\{X_k\}$ satisface la condición de Liapunov.

Teorema del Límite Central

Ejemplo (Récords)

Entonces tenemos que

$$\frac{\mu_n - E(\mu_n)}{\sqrt{Var(\mu_n)}} \xrightarrow{d} N(0,1).$$

Notemos que

$$\sqrt{Var(\mu_n)} = s_n \sim \sqrt{\log n}$$

y que

$$\frac{E(\mu_n) - \log n}{\sqrt{\log n}} = \frac{\sum_{k=1}^{n} \frac{1}{k} - \log n}{\sqrt{\log n}} \sim \frac{\gamma}{\sqrt{\log n}} \to 0,$$

donde γ es la constante de Euler.

Teorema del Límite Central

Ejemplo (Récords)

Lo que finalmente da como resultado (Convergence to Types Theorem)

$$\frac{\mu_n - \log n}{\sqrt{\log n}} \xrightarrow{d} N(0,1).$$

Si quisieramos darle formalidad a está ultima parte, una forma de hacerlo es aplicando el "Convergence to Types Theorem".

Teorema del Límite Central

Es natural preguntarse que tan precisa es la aproximación normal del TLC. EL siguiente Teorema responde esta inquietud.

Teorema (Desigualdad de Berry-Essèen)

Sean $\{X_n, n \geq 1\}$ *i.i.d. y supongamos que* $E(X_k) = 0$, $E(X_1^2) = \sigma^2 < \infty$ *y* $E(|X_1|^3) < \infty$. *Entonces,*

$$\sup_x \left| P\left(\frac{\sqrt{n} \cdot \overline{X}_n}{\sigma} \leq z \right) - \Phi(z) \right| \leq \frac{33}{4} \frac{E\left[|X_1|^3\right]}{\sqrt{n}\sigma^3}.$$

Teorema del Límite Central

Decimos que una sucesión $\{X_n, n \geq 1\}$ es estrictamente estacionaria si, para cada k, la distribución conjunta de

$$(X_{n+1}, \ldots, X_{n+k})$$

es independiente de n para $n = 0, 1, \ldots$. Por otra parte, llamamos a la sucesión m-dependiente si para cada t,

$$(X_1, \ldots, X_{t-1}), \qquad (X_{t+m+1}, X_{t+m+2}, \ldots)$$

son independientes. Así, si las variables son lo suficientemente lejanas son independientes.

Un ejemplo de un sucesión estacionaria y m-dependiente es el modelo de series de tiempo llamado medias móviles de orden m: Sean $\{Z_n\}$ v.a.i.i.d. y, para constantes dadas $c_1, c_2, \ldots, c_m$, definamos el proceso

$$X_t := \sum_{i=1}^{m} c_i Z_{t-j}, \qquad t = 0, 1, \ldots.$$

Teorema del Límite Central

Teorema (Hoeffding and Robbins)

Supongamos que $\{X_n, n \geq 1\}$ *es una sucesión estrictamente estacionaria y m-dependiente con* $E(X_1) = 0$ *y*

$$Cov(X_t, X_{t+h}) = E(X_t X_{t+h}) =: \gamma(h).$$

Supongamos que

$$v_n := \gamma(0) + 2\sum_{k=1}^{n} \gamma(k) \neq 0.$$

Entonces

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \xrightarrow{d} N(0, v_m)$$

y

$$nVar(\overline{X}_n) \to v_m.$$

Teorema del Límite Central

Finalmente, estudiaremos un TLC multivariado.

Teorema (Teorema del Límite Central Multivariado)

Supongamos que $X_1, X_2, \ldots, X_n$ *son vectores aleatorios i.i.d. donde*

$$X_i = (X_{1i}, X_{2i}, \ldots, X_{ki})'$$

con media

$$\mu = (\mu_1, \mu_2, \ldots, \mu_k)' = (E(X_{1i}), E(X_{2i}), \ldots, E(X_{ki}))'$$

y matriz de varianza Σ. *Sea*

$$\overline{X} = \left(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k\right)'$$

donde $\overline{X}_j = n^{-1} \sum_{i=1}^{n} X_{ji}$. *Entonces,*

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} N(0, \Sigma).$$

Distribuciones Asociadas al Muestreo y Estimación Puntual

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), $\hat{a}_0 = X_{(1)} = \text{mín}\{X_1, X_2, \ldots, X_n\}$ es un estimador para a y $\hat{b}_0 = X_{(n)} = \text{máx}\{X_1, X_2, \ldots, X_n\}$, 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

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

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

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

Distribución Muestral

Además,

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

(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

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:

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

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 $\overline{X}$ (lo hicimos sólo para la suma y concluimos que $\overline{X}$ es un múltiplo de la distribución de la suma).

- 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 de Límite Central". Existen muchas versiones, la que verás aquí, es la más simple de todas ellas. Para otros casos ver [2], y algunos serán discutidos más adelante.

Distribución del Estimador de la Media

Teorema del Límite Central (TLC): 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 poblaciones, 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

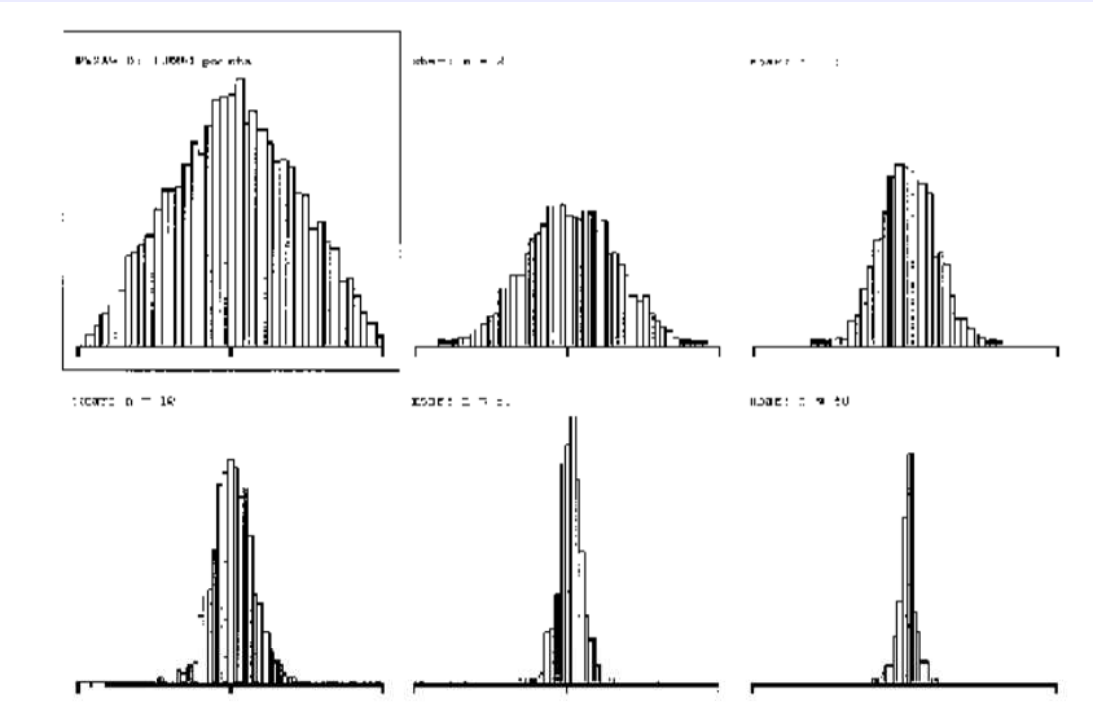


Figura: Una población con distribución triangular. Esta distribución es simétrica y "medio parecida" a una Normal.

Distribución del Estimador de la Media

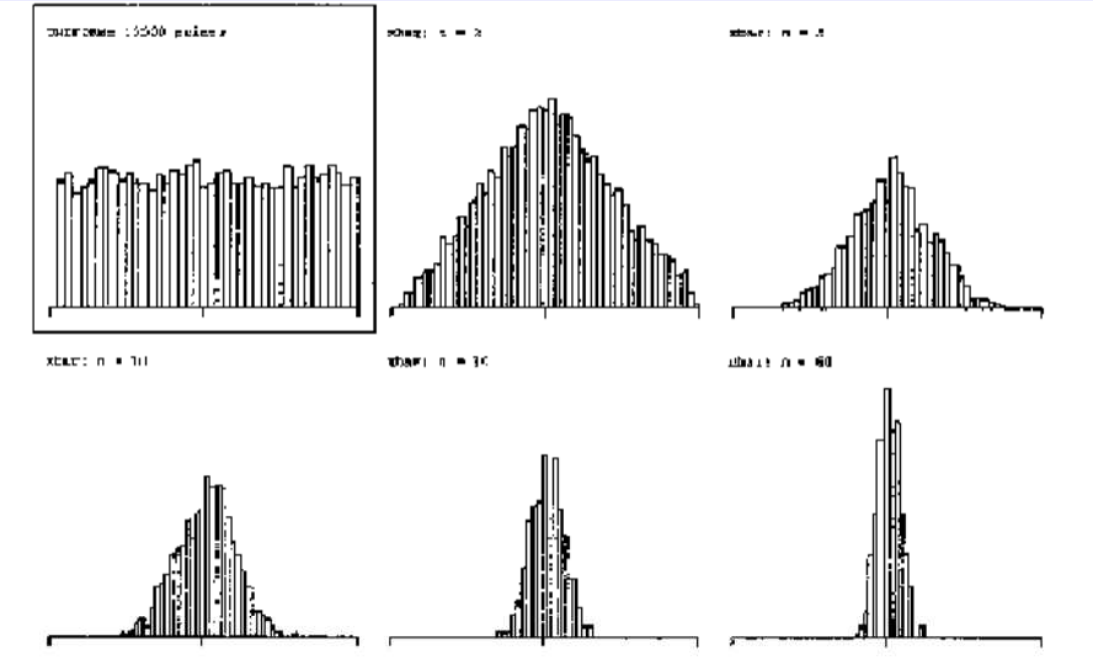


Figura: Distribución Uniforme. Esta distribución es simétrica, pero no se parece al comportamiento de una Normal pues las colas de la distribución no "decaen".

Distribución del Estimador de la Media

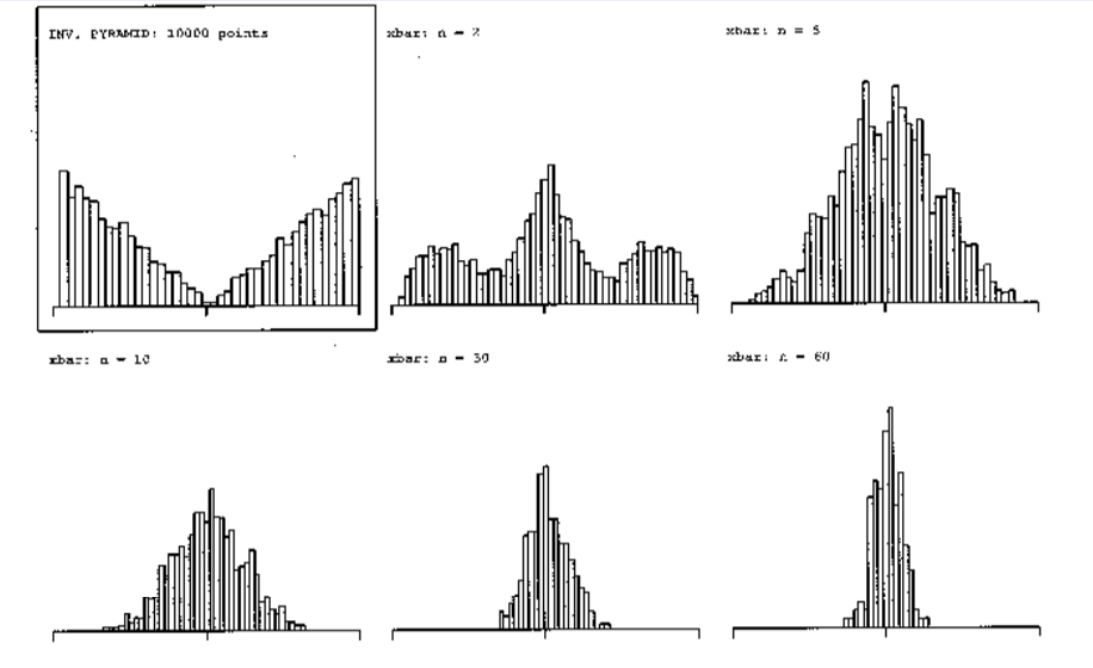


Figura: Distribución Triangular inversa. Esta distribución es simétrica pero su comportamiento en el centro y colas de la distribución son los opuestos al de una Normal.

Distribución del Estimador de la Media

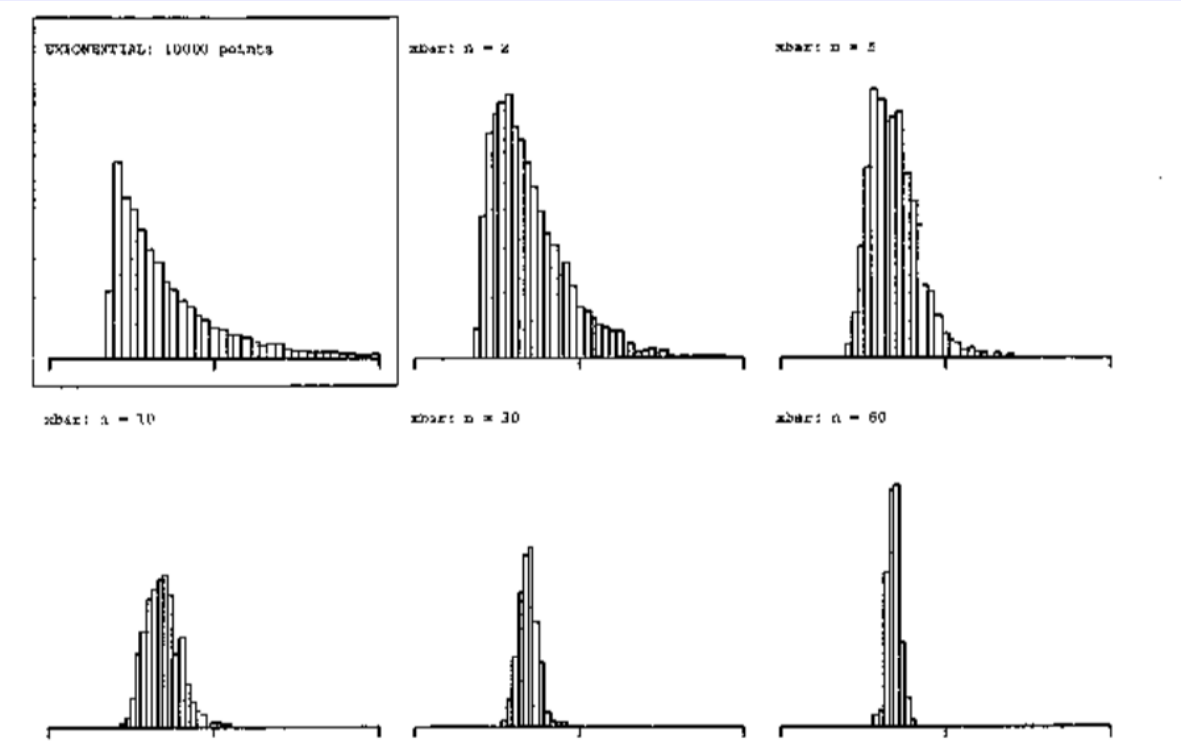


Figura: Distribución Exponencial. Esta distribución no es simétrica, ni se parece a la normal.

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

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

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 TLC 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 TLC 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, Texto[0].

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 TLC.

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 TLC 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

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

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

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

y

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

Distribución del Estimador de la Media

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

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

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

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

Distribución del Estimador de la Media

El TLC 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 TLC 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 TLC 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: Suponegamos 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 TLC.

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

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

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

$$= 1 - 0.8888 = 0.1112$$

Con correccion 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

$$\hat{\theta}_{84} = (84)^{-1} \sum_{i=1}^{84} X_i.$$

Distribución del Estimador de la Media

$$
\begin{aligned}
P(0.64 < \hat{\theta}_{84} < 0.76) &= P\left(\frac{0.64 - 0.7}{\sqrt{\frac{0.7(0.3)}{84}}} < \frac{\hat{\theta}_{84} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} < \frac{0.76 - 0.7}{\sqrt{\frac{0.7(0.3)}{84}}}\right) \\
&\simeq P(-1.2 < Z < 1.2) = 2P(0 < Z < 1.2) \\
&= 2(0.3849) = 0.7698
\end{aligned}
$$

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:

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

Distribución del Estimador de la Varianza

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

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

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

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.

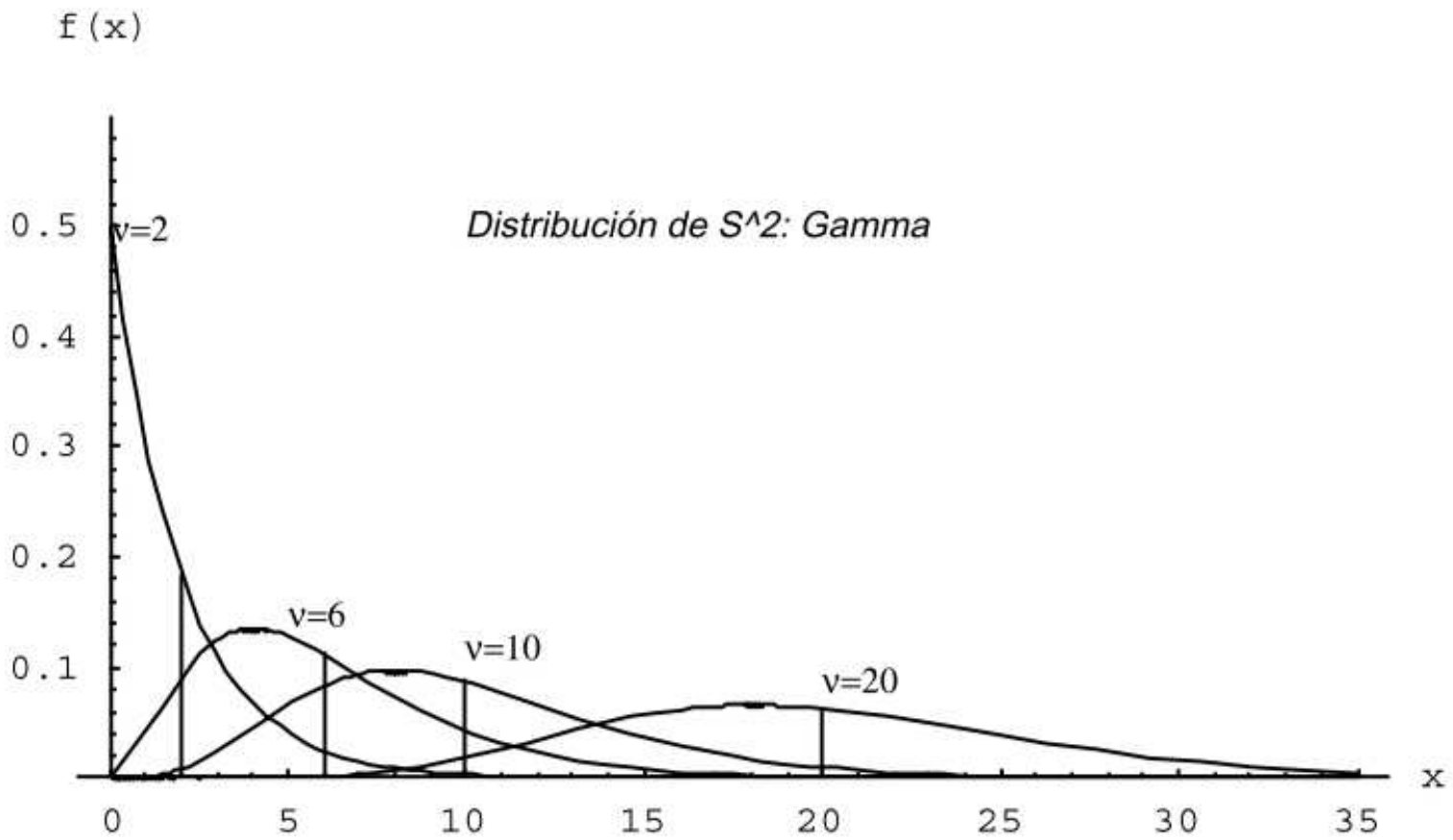
Distribución del Estimador de la Varianza

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 poblaciones 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 T.L.C. 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

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.

También mencionamos que $\overline{X}$ y S^2 no son los únicos estimadores para μ y σ^2, e incluso que pueden ser otros los parámetros en los que inicialmente estemos interesados estimar, 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.

Nos enfocaremos a algunos comentarios importantes sobre el segundo objetivo y dejaremos el resto como material para el apéndice al final del capítulo.

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.

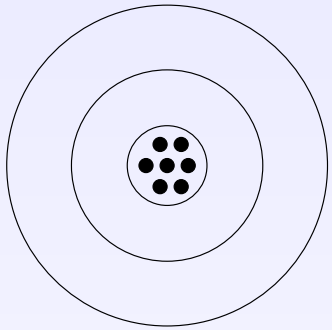
Estimación Puntual

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

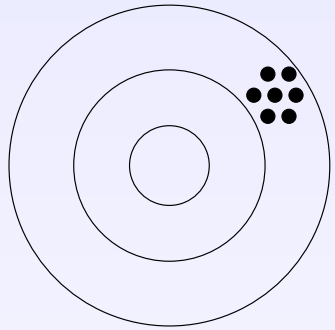
- Insesgamiento.
- Eficiencia.
- Consistencia.
- Suficiencia.
- Invarianza, etc.

Estimación Puntual

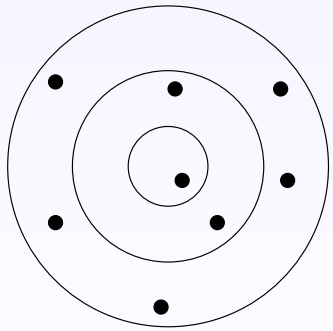
Varianza y sesgo.



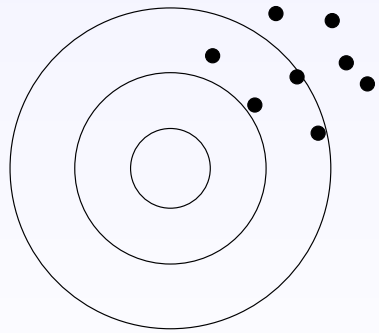
menor varianza, menor sesgo



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 poblaciones 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

4. Otro estimador para σ^2 en poblaciones 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.

Estimación Puntual

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.

Estimación Puntual

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.

Estimación Puntual

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.

Estimación Puntual

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.

Estimación Puntual

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 $\hat{\sigma}^2$.

Estimación Puntual

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.

Estimación Puntual

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

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

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.

Estimación Puntual

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

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

Estimación Puntual

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,

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

Estimación Puntual

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

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

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?

Estimación Puntual

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

Estimación Puntual

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.

Estimación Puntual

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 $\overline{X}$.

Estimación Puntual

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

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 "outlier" y las colas pesadas de la distribución de Cauchy hacen que tales observaciones probablemente aparezcan en una muestra.
  (**Nota**: en esta distribución, 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 "outlier", 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.

Estimación Puntual

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.

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

Estimación Puntual

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 recalcar el hecho de 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**. Esta es la propiedad de pérdida de memoria.

Con lo anterior en mente vemos que:

- 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.

Estimación Puntual

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

$$E(Y_1) = \frac{\theta}{n}, \quad Y_1 = \text{est. de orden 1 de las } n \text{ comp., cada una } exp(\theta)$$

$$E(Y_2 - Y_1) = \frac{\theta}{n-1}, \quad Y_2 - Y_1 = \text{est. orden 1 de las } n-1 \text{ comp. restantes}$$

$$E(Y_3 - Y_2) = \frac{\theta}{n-2}, \quad Y_3 - Y_2 = \text{est. orden 1 de las } n-2 \text{ comp. restantes}$$

$$\vdots$$

$$E(Y_r - Y_{r-1}) = \frac{\theta}{[n-(r-1)]}.$$

Estimación Puntual

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 supervivencia.

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.

Estimación Puntual

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 Cramér-Rao)

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

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

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

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

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

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 Cramér-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 población Exp(θ) , $\overline{X}$ es el EIMV.
**población exponencial media $=\theta$ varianza $=\theta^2$*

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

Estimadores Insesgados de Mínima Varianza

Ejemplo (Cont...)

La cota de Cramér-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 población exponencial.*

Estimadores Insesgados de Mínima Varianza

Ejemplo

En una población Poisson (λ), $\overline{X}$ es el EIMV para λ.
**población Poisson; Media=λ , Varianza=λ*

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

Calculamos la cota Cramér-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

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 población normal, tanto S^2 como S_n^2 son estimadores de σ^2. ¿Cual es mas eficiente usando el criterio de ECM?

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

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

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

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

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

Estimadores no insesgados

Ejemplo (Cont.)

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

$$\Rightarrow ECM(S_n^2) = V(S_n^2) + [Sesgo(S_n^2)]^2$$

$$= \frac{2(n-1)\sigma^4}{n^2} + \left[-\frac{\sigma^2}{n}\right]^2 = \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2}$$

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

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

Estimadores no insesgados

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 población uniforme $[0, \theta]$ *y sea* $\hat{\theta} = X_{(n)}$ *un estimador de* θ.

La función de densidad de la población

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

La función de distribución acumulada

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

Consistencia

Ejemplo (Cont...)

La densidad del n-esimo estadístico de orden es

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

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

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

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

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 población 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

Definición

Sea $X = (X_1, \ldots, X_n)$ una muestra aleatoria. Un estadístico $t = T(X)$ se dice suficiente para el parámetro subyacente θ si la distribución condicional de X, dado el estadístico $t = T(X)$, no depende del parámetro θ.

Teorema (Factorización de Fisher)

Sea $X = (X_1, \ldots, X_n)$ una muestra aleatoria. Si $f_\theta(x)$ es la función de densidad de X, entonces T es suficiente para θ si y solo si se pueden encontrar funciones no-negativas g_θ y h tal que

$$f_\theta(x) = h(x) \cdot g_\theta(T(x)),$$

es decir, la densidad f_θ se puede factorizar en el producto de dos funciones en el que un factor, h, no depende de θ y el otro factor, el cual depende de θ, depende de X solo a través de $T(x)$.

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

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 población. 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* $\theta = 2\overline{x}$

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

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.

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

$$\text{porque } V(X) = E(X^2) - [E(X)]^2$$

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

Método de momentos

Ejemplo (Cont...)

Formamos el sistema:

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

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

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

El método de máxima verosimilitud 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 valore posibles del parámetro. Por esto, será factible usar en la mayoría de los casos las técnicas clásicas de maximización de funciones continuas.

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

En muchos casos es más sencillo maximizar el logaritmo natural de la función en lugar de la función directamente ,y como la función logaritmo natural es una función creciente, el punto donde se toma el máximo sera el mismo que el de la función original.

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:

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

- 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 $L(\theta)$ (No obligatorio).
3. Maximizar la verosimilitud, o bien el logaritmo ($\log L(\theta)$), 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 población 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,

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

Método de Máxima Verosimilitud

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,

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

Método de Máxima Verosimilitud

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.*

Método de Máxima Verosimilitud

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,*

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

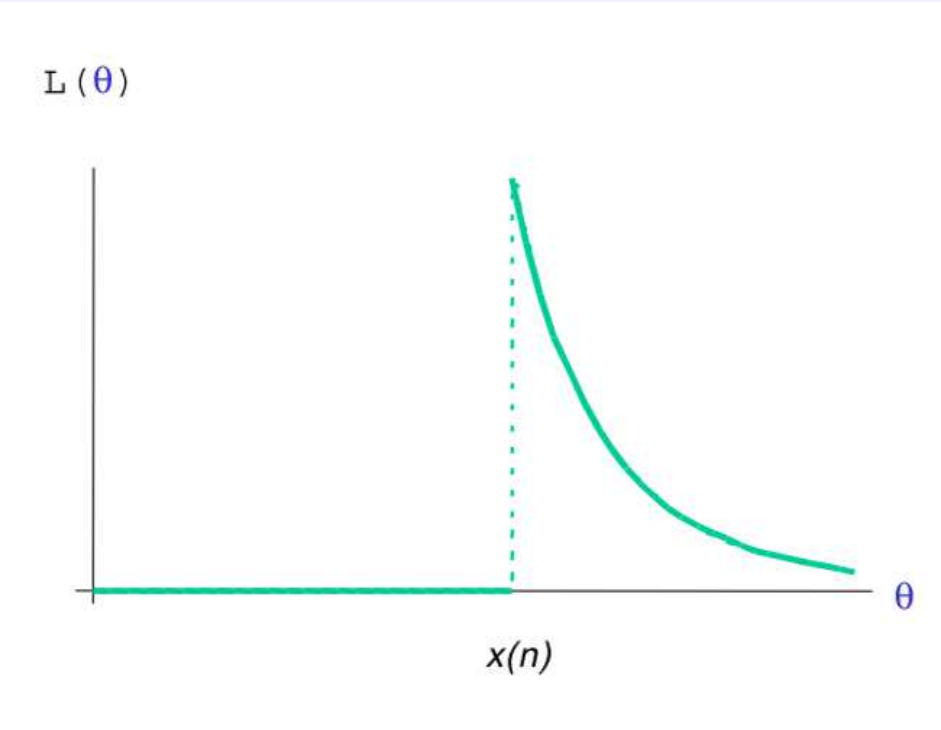
$$\log L(\theta) = -n \log \theta, \quad \theta > x_{(n)}.$$

Método de Máxima Verosimilitud

Ejemplo (Cont...)

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

No funciona el método de la derivada.



Método de Máxima Verosimilitud

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.

Método de Máxima Verosimilitud

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.

Método de Máxima Verosimilitud

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

Método de Máxima Verosimilitud

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.*

Método de Máxima Verosimilitud

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:

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

Método de Máxima Verosimilitud

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

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

Método de Máxima Verosimilitud

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,

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

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).

Método de Máxima Verosimilitud

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*

$$\hat{\theta} = \textit{el entero inferior más cercano a } \frac{nk}{x}$$

$$= \textit{el entero inferior más cercano a } \frac{(200)(100)}{11} = 1818.$$

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.

Método de Máxima Verosimilitud

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.*

Método de Máxima Verosimilitud

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.

Método de Máxima Verosimilitud

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 $P(X \geq 24)$.

Método de Máxima Verosimilitud

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.

Método de Máxima Verosimilitud

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

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

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

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

Método de Máxima Verosimilitud

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 $\log[L(p)]$

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

Método de Máxima Verosimilitud

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.

Método de Máxima Verosimilitud

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.

Método de Máxima Verosimilitud

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*

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

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

Método de Máxima Verosimilitud

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

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

Método de Máxima Verosimilitud

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

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

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

$\therefore \hat{\theta} = \overline{X}$ *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^2 *basándose 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*

Método de Máxima Verosimilitud

Ejemplo (Cont...)

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

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

$$\Longrightarrow$$

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

$$\Longrightarrow$$

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

$$\Longrightarrow$$

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

Método de Máxima Verosimilitud

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

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

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

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

Método de Máxima Verosimilitud

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:

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

Método de Máxima Verosimilitud

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.

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 (que verás en el próximo semestre), 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. Una de ellas es la libreria maxLik (https://cran.r-project.org/web/packages/maxLik/index.html)

```
library(maxLik)
?maxLik
maxLik                    package:maxLik                    R Documentation

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
          algorithm implemented in R), "BHHH" (for
          Berndt-Hall-Hall-Hausman), "SANN" (for Simulated ANNealing),
```

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

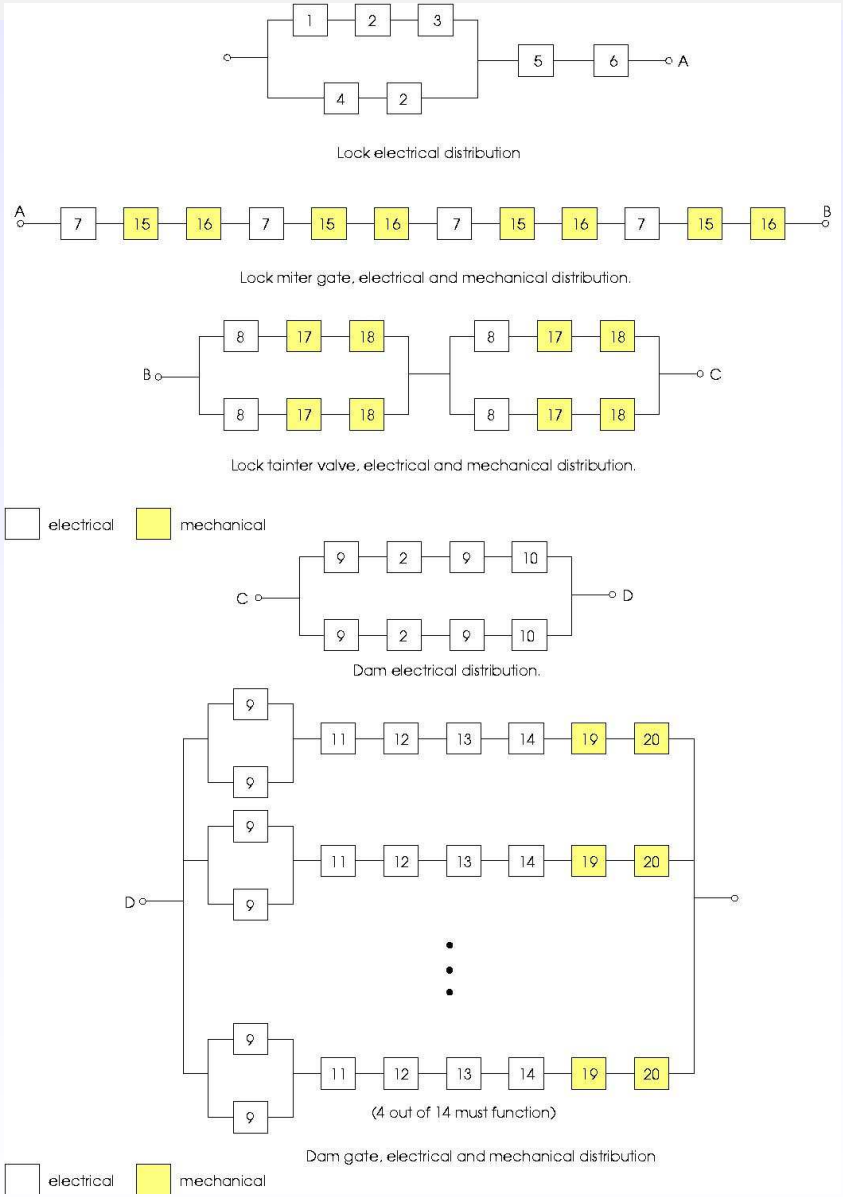
```
> ## ejemplo: gaussiana
> summary( res )
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 7 iterations
Return code 2: successive function values within tolerance limit
Log-Likelihood: -207.0033
2  free parameters
Estimates:
     Estimate Std. 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.

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.

Apéndice. Maximizar la verosimilitud

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

Apéndice. Maximizar la verosimilitud

Ejemplo: algoritmo EE-$(\mu + \lambda)$.

μ = número de padres
λ = número de hijos
begin

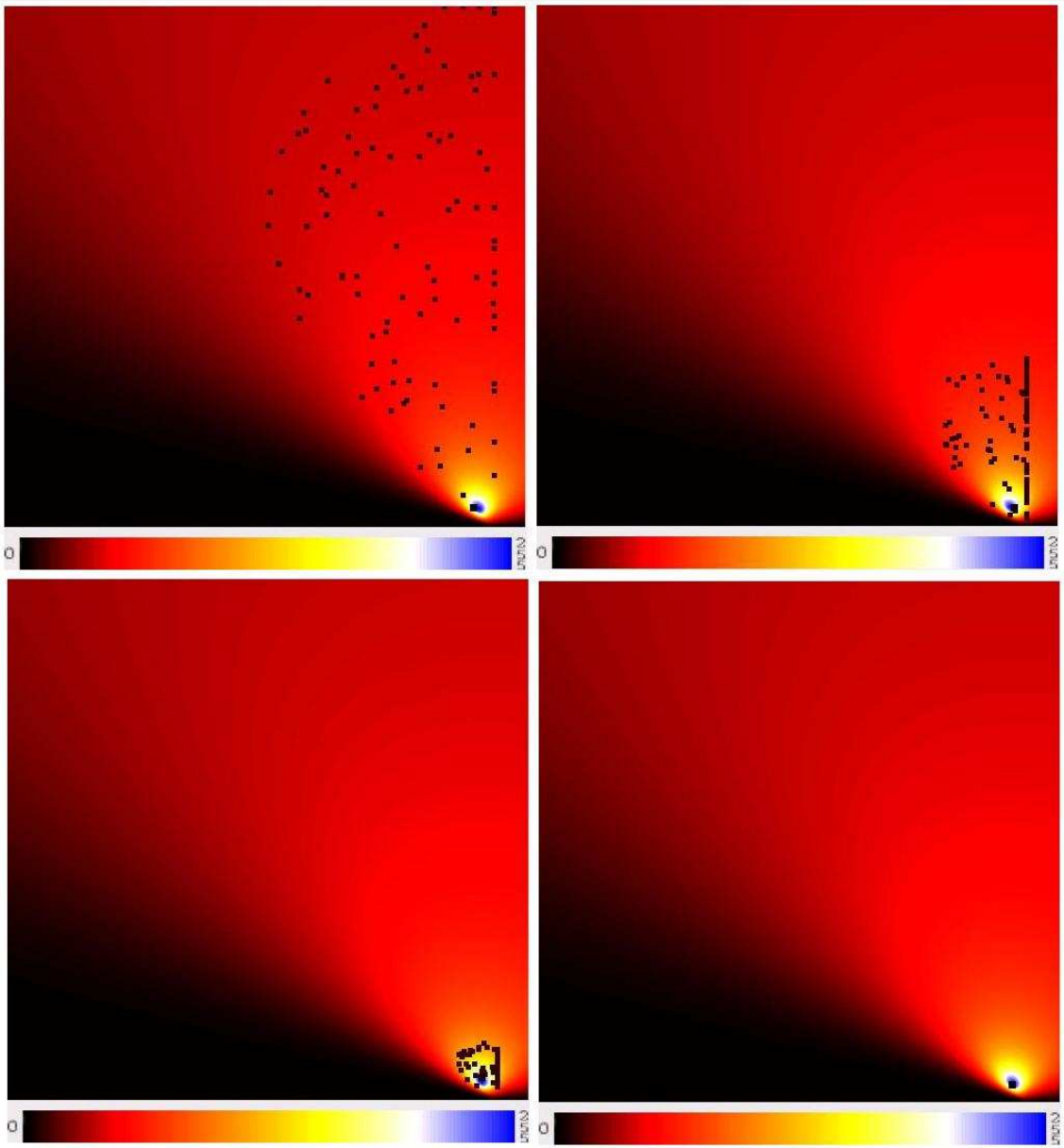
$t \leftarrow 0$ *initialize* $P(0) := \{a_1(0), \ldots, a_\mu(0)\}$
evaluate $P(t) := \{\Phi(a_1(0)), \ldots, \Phi(a_\mu(0))\}$ *where* $\Phi(a_k(0)) = f(x_k(0))$ $k = 1, \ldots, \mu$
while (not termination-condition) do

$a'_k(t) \leftarrow c'(P(t))$ $k = 1, \ldots, \lambda$ *(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

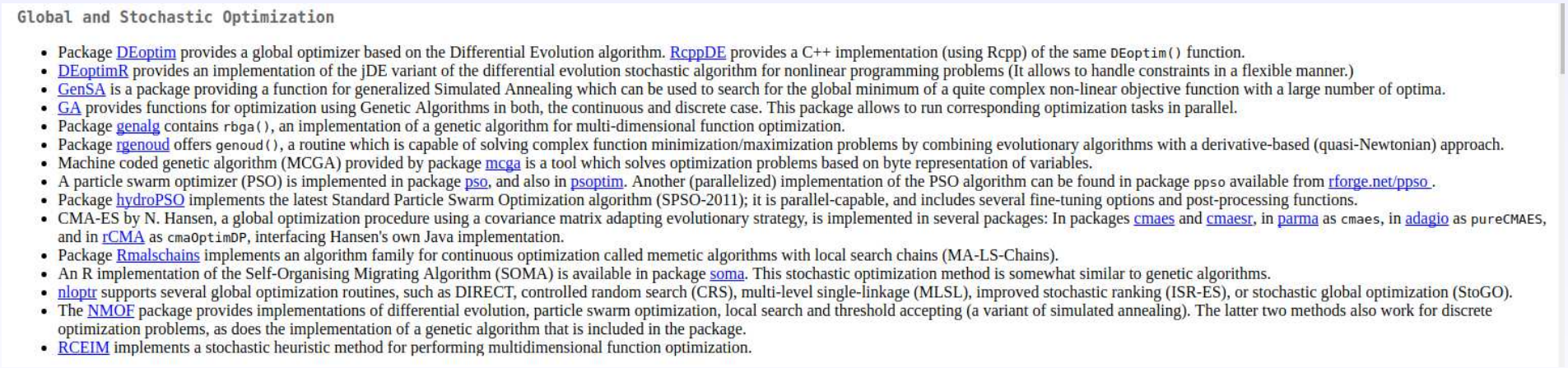
end

Apéndice. Maximizar la verosimilitud



Apéndice. Maximizar la verosimilitud

En R, puedes recurrir a `https://cran.r-project.org/`, `-->packages-->CRAN Task Views-->Optimization`



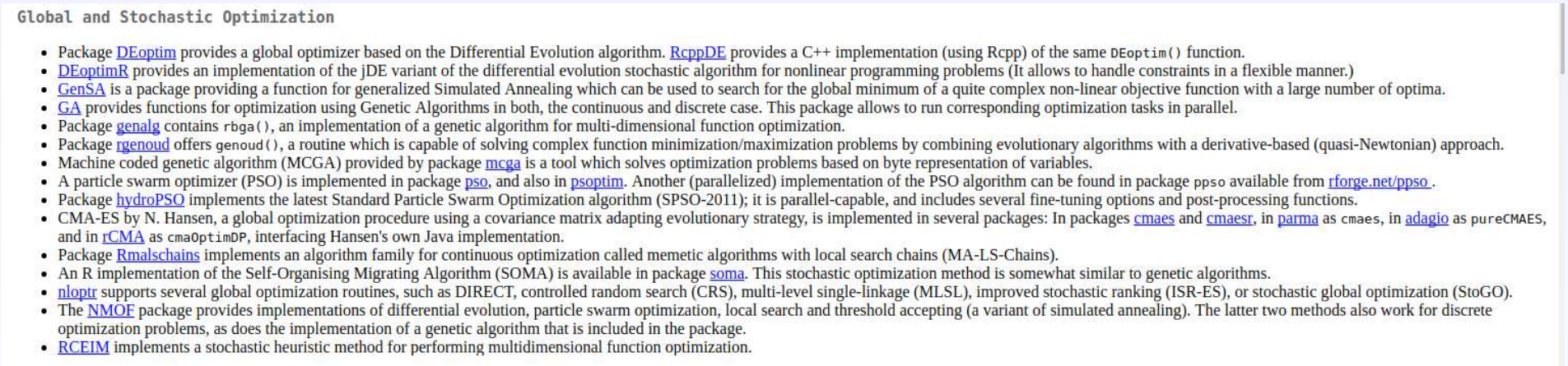
Global and Stochastic Optimization

- 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.

Aprenderás más en los siguientes semestres!

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!