Capítulo 7

Análisis estadístico de datos simulados

7.1. Introducción

Para llevar adelante una simulación de una situación real, debemos conocer algo sobre las fuentes de aleatoriedad de esta situación. Cada fuente de aleatoriedad se corresponderá en alguna medida a una variable aleatoria con cierta distribución de probabilidad, y cuanto mejor esté seleccionada esta distribución más adecuada será la simulación.

En la tabla 7.1 se ilustran algunos ejemplos de sistemas a simular y sus correspondientes fuentes de aleatoriedad.

Tipo de sistema	**Fuente de aleatoriedad**
Fabricación	Tiempos de procesamiento. Tiempos de falla de una máquina. Tiempos de reparación de máquinas
Defensa	Tiempos de arribo y carga útil de aviones o misiles. Errores de lanzamiento.
Comunicaciones	Tiempos entre llegadas de mensajes. Longitudes de mensajes.
Transporte	Tiempo de embarque Tiempos entre arribos de pasajeros.

Cuadro 7.1: Sistemas y fuentes de aleatoriedad

Así, para simular un sistema real es necesario:

- Representar cada fuente de aleatoriedad de acuerdo a una distribución de probabilidad.
- Elegir adecuadamente la distribución, para no afectar los resultados de la simulación.

7.2. Selección de la distribución

Para elegir una distribución es necesario trabajar con datos obtenidos del sistema real a simular. Estos datos pueden luego ser usados a) directamente, b) realizando el muestreo a partir de la distribución *empírica* de los datos o c) utilizando técnicas de inferencia estadística.

Si se utilizan los datos **directamente**, entonces sólo se podrán reproducir datos históricos y resulta una información insuficiente para realizar buenas simulaciones del modelo. De todos modos, los datos son importantes para validar el modelo existente con el modelo simulado.

La **distribución empírica** permite reproducir datos intermedios a los datos observados, lo cual es algo deseable fundamentalmente si se tienen datos de tipo continuo. Esta técnica es recomendable en los casos en que no se puedan ajustar los datos a una distribución teórica.

Por otro lado, las técnicas de **inferencia estadística** tienen varias ventajas con respecto al uso de la distribución empírica. Por un lado, esta última puede tener irregularidades si hay poca información mientras que las distribuciones teóricas tienen una forma más suave. Además pueden simularse datos aún fuera del rango de los datos observados. No es necesario almacenar los datos observados ni las correspondientes probabilidades acumuladas. Por otra parte, en ciertos casos puede ser necesario imponer un determinado tipo de distribución por la naturaleza misma del modelo, y en ese caso se pueden modificar fácilmente los parámetros de la distribución elegida. Las desventajas que puede tener la selección de una distribución teórica es que no se encuentre una distribución adecuada, y que se puedan generar valores extremos no deseados.

Para seleccionar una distribución, se analizan ciertos parámetros que indican la distribución particular dentro de una familia. Por ejemplo, si es una distribución normal se necesita determinar μ y σ. Si es exponencial, se debe determinar λ. Los parámetros de una distribución pueden ser de posición, de escala o de forma, de acuerdo a qué características de la distribución determinan.

En el caso en que no se pueda hallar una distribución teórica adecuada que ajuste a los datos observados, o simplemente porque se prefiere simular a partir de las observaciones, se suele utilizar la **distribución empírica**. Esto es, la distribución de los datos de acuerdo a la muestra que se ha observado.

Así, si los datos observados son $X_1, X_2, \ldots, X_n$, la distribución empírica asigna una función de masa de probabilidad empírica a cada x dada por

$$p_e(x) = \frac{\#\{i \mid X_i = x,\, 1 \leq i \leq n\}}{n}.$$

En particular, la función de distribución acumulada está dada por:

$$F_e(x) = \frac{\#\{i \mid X_i \leq x,\, 1 \leq i \leq n\}}{n}.$$

Por ejemplo, si los datos observados son $X_1 = 3.2$, $X_2 = 4.3$, $X_3 = -2.0$, $X_4 = 1.6$, $X_5 = 0$,

entonces

$$X_{(1)} = -2.0, \quad X_{(2)} = 0, \qquad X_{(3)} = 1.6, \qquad X_{(4)} = 3.2 \qquad X_{(5)} = 4.3,$$

y

$$F_e(x) = \begin{cases} 0 & x < -2.0 \\ \frac{1}{5} & -2.0 \leq x < 0 \\ \frac{2}{5} & 0 \leq x < 1.6 \\ \frac{3}{5} & 1.6 \leq x < 3.2 \\ \frac{4}{5} & 3.2 \leq x < 4.3 \\ 1 & x \geq 4.3 \end{cases}$$

Ahora bien, si se sabe que los datos observados provienen de una variable aleatoria continua, entonces es conveniente suavizar F_e para que también resulte continua. Una posibilidad es definir $F(x) = F_e(x)$ en los puntos observados, y unir con una poligonal en los puntos intermedios a las observaciones. Esto es, en primer lugar se ordenan los valores en forma creciente, denotando $X_{(i)}$ a la observación que ocupa el i-ésimo lugar en el ordenamiento:

$$X_{(1)} < X_{(2)} < \ldots, < X_{(n)}.$$

La distribución propuesta es:

$$F_{el}(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i-1}{n-1} + \frac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & X_{(i)} \leq x \leq X_{(i+1)} \\ 1 & x \geq X_{(n)}. \end{cases}$$

La Figura 7.1 ilustra ambas distribuciones empíricas para los datos del ejemplo. Así, será po-

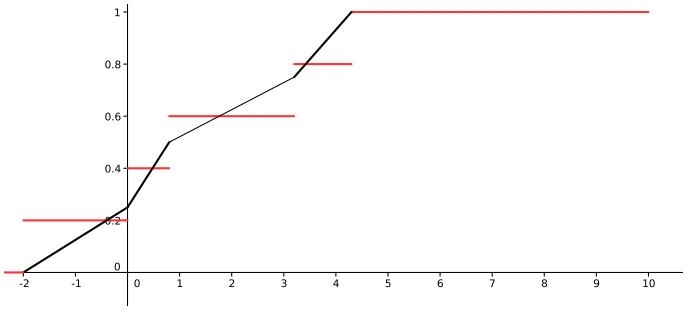

Figura 7.1: Distribuciones empiricas

sible simular cualquier valor entre $X_{(1)}$ y $X_{(n)}$, y se simulará con mayor frecuencia en los intervalos donde han ocurrido más observaciones.

Por último, si lo que se conoce es una agrupación de los datos en distintos intervalos:

$$[a_1, a_2),\ [a_2, a_3),\ \ldots, [a_{k-1}, a_k),$$

es decir, un histograma de los datos, se puede hacer una distribución empírica que aproxime a la frecuencia acumulada de las observaciones. Esto es, si n_j es la cantidad de observaciones en el intervalo $[a_j, a_{j+1})$, entonces

$$n = n_1 + n_2 + \cdots + n_k,$$

y se define la distribución empírica **lineal** G, donde

$$G(a_1) = 0, \qquad G(a_j) = \frac{1}{n}\,(n_1 + n_2 + \cdots + n_{j-1}), \quad 2 \leq j \leq k+1$$

y

$$G(x) = \begin{cases} 0 & x < a_1 \\ G(a_j) + \frac{G(a_{j+1}) - G(a_j)}{a_{j+1} - a_j}(x - a_j) & a_j < x < a_{j+1},\ 1 \leq j \leq k \\ 1 & x \geq a_{k+1} \end{cases}.$$

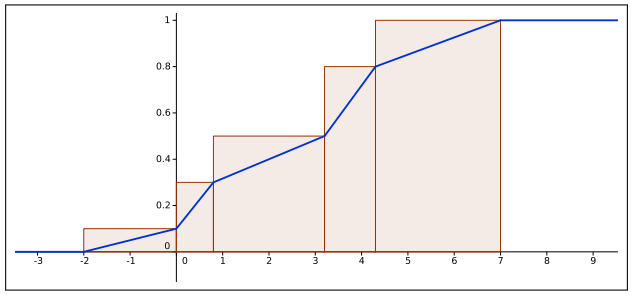

Figura 7.2: Distribución empírica a partir de datos agrupados

7.2.1. Algunas medidas estadísticas

A la hora de seleccionar una determinada distribución teórica de probabilidad para llevar adelante una simulación, es importante conocer algunos valores estadísticos que tienen las distribuciones teóricas y compararlos con los que se obtienen a partir de una muestra.

Por ejemplo, es importante conocer el rango de la variable, su media, su variabilidad, su simetría o tendencia central, entre otras. Ahora bien, estos valores están bien definidos para una distribución teórica pero son desconocidos para una distribución de la cual sólo se conoce una muestra. Entonces, para estimar estos valores, se utilizan los **estadísticos muestrales**. Más

específicamente, un estadístico muestral es una variable aleatoria definida a partir de los valores de una muestra. Por ejemplo, la **media muestral** $X(n)$ es el estadístico definido por:

$$X(n) = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

y que suele utilizarse para estimar la media o valor esperado de la distribución de los datos. La **varianza muestral** $S^2(n)$ es el estadístico definido por

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

y es un estimador no sesgado para la varianza.

Así, si se tiene una muestra de datos y a partir de ella se calculan los valores:

$$\overline{x} = X(n), \qquad \overline{s}^2 = S^2(n),$$

y se pretende analizar su ajuste a una distribución normal, lo más razonable sería considerar la normal $N(\mu, \sigma)$ con $\mu = \overline{x}$ y $\sigma = \sqrt{\overline{s^2}}$.

La Tabla 7.2 muestra algunas de los estimadores y medidas estadísticas que suelen ser útiles para decidir la elección de una distribución teórica a partir de una muestra de datos observados. En cada caso, se considera que la muestra es de tamaño n, los valores observados son

$$X_1, X_2, \ldots, X_n$$

y ordenados se denotan

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}.$$

Función	Estimador muestral	Estima
Min, Max	$X_{(1)}$, $X_{(n)}$	rango
Media μ	$\overline{X}(n)$	Tendencia central
Mediana	$\hat{m} = \begin{cases} X_{(n+1)/2} \\ \frac{1}{2}(X_{n/2} + X_{(n/2+1)}) \end{cases}$	Tendencia central.
Varianza σ^2	$S^2(n)$	Variabilidad
c.v.=$\frac{\sigma}{\mu}$	$\hat{cv}(n) = \frac{\sqrt{S^2(n)}}{\overline{X}(n)}$	Variabilidad
τ	$\hat{\tau} = \frac{S^2(n)}{\overline{X}(n)}$	Variabilidad
Asimetría $\nu = \frac{E[(X-\mu)^3]}{(\sigma^2)^{3/2}}$	$\hat{\nu}(n) = \frac{\sum_i (X_i - \overline{X}(n))^3/n}{[S^2(n)]^{3/2}}$	Simetría

Cuadro 7.2: Tabla de estimadores

Por ejemplo, si una distribución es simétrica, su media y su mediana son iguales. Luego si la media y la mediana muestral son muy diferentes, no se debería elegir una distribución normal para la simulación.

Por otro lado, para la distribución es exponencial el coeficiente de variación es 1: $c.v. = \sigma/\mu = 1$. Es decir, el coeficiente de variación estimado a partir de la muestra debería ser un valor próximo a 1 para decidirse por una distribución exponencial.

Los **histogramas** también son herramientas útiles para seleccionar una distribución, y ciertos tests estadísticos como el test χ-cuadrado se basa justamente en la comparación del histograma de frecuencias observadas y esperadas para determinar cuán buen ajuste hay de la distribución teórica a la distribución de los datos.

Para realizar un histograma, el rango de valores obtenidos en los datos se divide en k intervalos adyacentes disjuntos $[a_1, a_2), [a_2, a_3), \ldots, [a_k, a_{k+1})$ de igual amplitud Δ, se considera h_j la proporción de datos que caen en el intervalo $[a_j, a_{j+1})$, y el histograma se define por la función

$$h(x) = \begin{cases} 0 & x < a_1 \\ h_j & a_j \leq x < a_{j+1} \\ 1 & x \geq a_{k+1} \end{cases}.$$

Notemos que si f es la densidad real de los datos, entonces

$$P(a_j \leq x < a_{j+1}) = \int_{a_j}^{a_{j+1}} f(x)\, dx = f(y)\, \Delta$$

para algún $y \in [a_j, a_{j+1})$.

Así, al ser los intervalos de igual amplitud, las áreas de los rectángulos o barras del histograma son proporcionales a la frecuencia relativa de los datos en el correspondiente intervalo. Luego tiene sentido superponer al histograma normalizado la función de densidad o de probabilidad de masa según corresponda, y comparar ambos gráficos. El histograma normalizado se obtiene dividiendo h_j por Δ para lograr un área total igual a 1.

Otras herramientas estadísticas son los diagramas de caja y q-cuantiles, que permiten hacer una análisis comparativo entre la muestra observada y la distribución teórica.

Los diagramas de caja determinan los cuartiles de la muestra, es decir, los valores donde se ubican el 25 %, 50 % y 75 % de los datos, con una representación gráfica en forma de caja que permite visualizar además la simetría o tendencia central de los datos.

Los q-cuantiles expresan otros cuantiles. Por ejemplo, si $q = 10$, el q-cuantil determina el valor hasta el cual se acumula el 10 % de los datos.

7.3. Estimación de parámetros

Supongamos que hemos obtenido una muestra de n datos, y queremos inferir de qué distribución provienen y qué parámetros corresponden a esa distribución. Por ejemplo, si consideramos que provienen de una distribución exponencial, ¿cómo determinamos el parámetro λ?

En la práctica, no será posible conocer estos parámetros con exactitud si lo que se conoce es sólo una muestra. Pero existen ciertos métodos para **estimar** estos parámetros.

Definición 7.1. Dada una muestra de n datos observados, se llama **estimador** $\hat{\theta}$ del parámetro θ a cualquier función de los datos observados.

Por ejemplo, si se toma una muestra de tamaño n, X_1, X_2, ..., X_n, los siguientes son estimadores:

$$\hat{\theta}_1 = \frac{X_1 + X_n}{2}, \qquad \hat{\theta}_2 = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{7.1}$$

Ahora bien, ¿qué relación hay entre el estimador y el parámetro a estimar? ¿Cuándo utilizar un estimador en particular para estimar un determinado parámetro? Esto tendrá que ver con las propiedades del estimador, ya sea en relación con el parámetro a estimar o en comparación con otros posibles estimadores.

7.3.1. Propiedades de un buen estimador

Un **buen estimador** debería cumplir con las siguientes propiedades.

- Insesgabilidad: se dice que el estimador es insesgado si $E[\hat{\theta}] = \theta$.

  Por ejemplo, si tomamos una muestra de tamaño n, los estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ de (7.1) son insesgados si lo que se quiere estimar es la media μ de la distribución, puesto que

  $$E(\hat{\theta}_1) = \frac{E(X_1) + E(X_n)}{2} = \mu, \qquad E(\hat{\theta}_2) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \mu.$$

- Consistencia: si al aumentar la muestra, el estimador se aproxima al parámetro.

  Notemos que el estimador $\hat{\theta}_1$ no es un estimador consistente ya que sólo utiliza dos elementos de la muestra, por lo cual no mejora la estimación incrementando el tamaño de esta muestra. En cambio por el Teorema Central del Límite, el estimador $\hat{\theta}_2$ tiende a la media de la distribución.

- Eficiencia: se calcula comparando su varianza con la de otro estimador. Cuanto menor es la varianza, se dice que el estimador es más eficiente.

  Por ejemplo, en (7.1), tenemos que para $i = 1, 2$,

  $$\text{Var}(\hat{\theta}_i) = E[(\hat{\theta}_i - E(\hat{\theta}_i))^2] = E[(\hat{\theta}_i - \mu)^2].$$

  Luego,

  $$Var(\hat{\theta}_1) = E((\hat{\theta}_1 - \mu)^2) = E\left[\left(\frac{X_1 - \mu}{2} + \frac{X_2 - \mu}{2}\right)^2\right] = \frac{\text{Var}(X_1) + \text{Var}(X_2)}{4} = \frac{1}{2}\,\text{Var}(X).$$

  En cambio, si se toman los n elementos de la muestra tenemos que:

  $$Var(\hat{\theta}_2) = \frac{1}{n}\,\text{Var}(X).$$

  Así, $\text{Var}(\hat{\theta}_2) < \text{Var}(\hat{\theta}_1)$ para $n > 2$, y por lo tanto $\hat{\theta}_2$ es más eficiente que $\hat{\theta}_1$.

- Suficiencia: significa que el estimador utiliza toda la información obtenida de la muestra.

7.3.2. Error cuadrático medio de un estimador

Si $\hat{\theta}$ es un estimador del parámetro θ de una distribución F, se define el **error cuadrático medio** (ECM) de $\hat{\theta}$ con respecto al parámetro θ como

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

Así, el ECM es una medida de dispersión del estimador con respecto al parámetro a estimar. Si el estimador es insesgado, es decir $E(\hat{\theta}) = \theta$, entonces el ECM coincide con la varianza del estimador. En general, se tiene:

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

El término $E(\hat{\theta} - \theta)$ se denomina **sesgo** del estimador. Así, el error cuadrático medio de un estimador es igual a su varianza más el sesgo al cuadrado. Si el estimador es insesgado, su ECM es igual a la varianza.

7.3.3. Estimadores de máxima verosimilitud

Hemos visto qué propiedades debería tener un buen estimador. Veremos ahora cómo podemos construir un estimador de un parámetro. Existen distintos métodos, y cada método hace alguna suposición sobre los datos que se obtienen en la muestra. Veremos el caso de los estimadores de máxima verosimilitud (maximum likelihood estimators (MLE)).

El estimador de máxima verosimilitud de un parámetro (o un conjunto de parámetros) θ, asume que la muestra obtenida tiene máxima probabilidad de ocurrir entre todas las muestras posibles de tamaño n, y que los datos X_1, X_2, ..., X_n son independientes.

Supongamos que se tiene la hipótesis de una distribución **discreta** para los datos observados, y se desconoce un parámetro θ. Sea $p_\theta(x)$ la probabilidad de masa para dicha distribución. Entonces, dado que se han observado datos $X_1, X_2, \ldots, X_n$, se define la función de máxima verosimilitud $L(\theta)$ como sigue:

$$L(\theta) = p_\theta(X_1) \cdot p_\theta(X_2) \cdots p_\theta(X_n).$$

Si la distribución supuesta es **continua**, y $f_\theta(x)$ es la densidad para dicha distribución, se define de manera análoga:

$$L(\theta) = f_\theta(X_1) \cdot f_\theta(X_2) \cdots f_\theta(X_n).$$

En cualquiera de los casos, el estimador de máxima verosimilitud es el valor $\hat{\theta}$ que maximiza $L(\theta)$:

$$L(\hat{\theta}) \geq L(\theta), \qquad \theta \text{ valor posible.}$$

Ejemplo 7.1. Supongamos que se ha tomado una muestra de tamaño n, y se tienen suficientes razones para suponer que tiene una distribución exponencial. Esta distribución depende de un parámetro λ, y este parámetro se estimará a partir de la muestra.

Dado que la función de densidad de una variable $X \sim \mathcal{E}(\lambda)$ es

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x > 0,$$

el estimador $\hat{\lambda}$ del parámetro λ será aquel que maximice la función $L(\lambda)$:

$$L(\lambda) \;=\; \left(\lambda e^{-\lambda X_1}\right)\left(\lambda e^{-\lambda X_2}\right)\cdots\left(\lambda e^{-\lambda X_n}\right) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} X_i\right)$$

El máximo de $L(\lambda)$ se alcanza donde su derivada es 0, y este valor corresponde a

$$\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1} = \frac{1}{\overline{X}(n)}.$$

Recordemos que si $X \sim \mathcal{E}(\lambda)$ entonces $E[X] = \frac{1}{\lambda}$. Luego el estimador de máxima verosimilitud para el valor esperado $\theta = 1/\lambda$ es en este caso:

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

Ejemplo 7.2. Consideremos ahora el estimador $\hat{p}$ de la probabilidad de éxito de una distribución geométrica $Geom(p)$. En este caso, el parámetro a estimar es $\theta = p$ $(0 < p < 1)$ y la probabilidad de masa está dada por

$$p_\theta(x) = \theta(1-\theta)^{x-1}, \qquad x = 1, 2, \ldots.$$

La función a maximizar es:

$$\begin{aligned} L(\theta) &= \theta(1-\theta)^{(X_1-1)}\,\theta(1-\theta)^{(X_2-1)}\,\ldots\theta(1-\theta)^{(X_n-1)} \\ &= \theta^n(1-\theta)^{\sum_{i=1}^n (X_i-1)} = \left(\frac{\theta}{1-\theta}\right)^n (1-\theta)^{\sum_{i=1}^n X_i} \end{aligned}$$

Derivando $L(\theta)$ con respecto a θ e igualando a 0 obtenemos la expresión $\hat{\theta} = \hat{p}$ en términos de la muestra de tamaño n que corresponde al estimador de máxima verosimilitud:

$$\hat{p} = \left(\frac{1}{n}\sum X_i\right)^{-1}$$

Nuevamente, dado que $\theta = 1/p$ es la esperanza de una variable geométrica con probabilidad de éxito p, tenemos que el estimador de la esperanza es

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

7.3.4. La media muestral

Dadas n observaciones: $X_1, X_2, \ldots, X_n$, con una misma distribución, la media muestral es el estimador definido por:

$$\overline{X}(n) = \frac{1}{n}\,(X_1 + X_2 + \cdots + X_n).$$

Hemos visto que en el caso de una variable exponencial y de una geométrica, la media muestral es el estimador de máxima verosimilitud del valor esperado. Ahora bien, este es **el** estimador utilizado para $E[X]$ cualquiera sea la distribución de X. En primer lugar, la media muestral es un **estimador insesgado** para $E[X]$. En efecto, si $E(X_i) = \theta$, $1 \leq i \leq n$, entonces

$$E[\overline{X}(n)] \;=\; E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \frac{n\theta}{n} = \theta.$$

En particular, la varianza de este estimador es igual a su error cuadrático medio:

$$\begin{aligned} ECM(\overline{X}(n), \theta) &= E[(\overline{X}(n) - \theta)^2] \\ &= \text{Var}(\overline{X}(n)) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

Así, cuanto mayor sea el tamaño de la muestra, menor será la varianza de la media muestral. Esto en principio permite aproximar el valor esperado con mayor certeza cuanto mayor sea el tamaño de la muestra.

7.3.5. La varianza muestral

En el caso de la media muestral, el error cuadrático medio para la estimación de la media es igual a su varianza: $\frac{\sigma^2}{n}$. Entonces, si se quiere que esta varianza sea menor que, por ejemplo, 0.001, la muestra deberá tener un tamaño n tal que

$$\frac{\sigma^2}{n} < 0.001.$$

Ahora bien, en general se desconoce el valor de σ por lo cual la fórmula anterior da poca información y se hace necesario tener a u vez un estimador para la varianza.

Se denomina **varianza muestral** para muestras de tamaño n al estimador $S^2(n)$ dado por:

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

Notemos que:

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

Además,

$$\begin{aligned} E[X_i^2] &= \text{Var}(X_i) + (E[X_i])^2 = \sigma^2 + \theta^2. \\ E[\overline{X}^2(n)] &= \frac{\sigma^2}{n} + \theta^2. \\ (n-1)E[S^2(n)] &= nE[X_1^2] - nE[\overline{X}^2(n)] = n(\sigma^2 + \theta^2) - n(\frac{\sigma^2}{n} + \theta^2) \\ E[S^2(n)] &= \sigma^2 \end{aligned}$$

Utilizaremos $S(n) = \sqrt{S^2(n)}$ como estimador de la desviación estándar.

7.4. Estimación con Simulaciones

La media muestral y la varianza muestral son estimadores que dependen del tamaño de la muestra, y hemos visto que cuanto mayor sea el tamaño de la muestra menor será la varianza del estimador media muestral.

En el caso particular en que se estén simulando datos con un programa determinado, el tamaño de la muestra se incrementa fácilmente aumentando el número de simulaciones. Entonces, por ejemplo, si un determinado programa simula la llegada de clientes a dos servidores en paralelo, y se desea estimar el promedio de clientes que son atendidos diariamente, cuanto mayor sea el número de simulaciones mejor será la estimación de este promedio. Ahora bien, ¿cuántas simulaciones es un "buen número" de simulaciones?

Además de un valor esperado, puede querer estimarse la probabilidad de atender a más de cinco clientes en una determinada hora, o que haya más de dos clientes sin atender a la hora del cierre. Estas probabilidades también pueden ser estimadas con el estimador de media muestral, pero aplicado a otra variable particular. Nuevamente, incrementar el número de simulaciones mejorará la estimación de esta probabilidad. Veamos con qué criterio podemos elegir un número aceptable de simulaciones.

7.4.1. Simulación de media muestral

Supongamos que con un programa de simulación es posible generar sucesivamente datos $\{X_i\}$, independientes, y se desea estimar la media μ de los datos, es decir $\mu = E[X_i]$. Entonces para un determinado n se podrá tomar el valor $\overline{X}(n)$ como una estimación de la media μ. Ahora bien, ¿cuán aproximado es este valor a la media que se desea estimar?

Sabemos por el Teorema Central del Límite, que la variable

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

tiene una distribución aproximadamente normal estándar, y por lo tanto

$$P\left(|\overline{X}(n) - \mu| < c\,\frac{\sigma}{\sqrt{n}}\right) \sim P(|Z| < c).$$

Entonces, el valor c tendrá que ver con el nivel de certeza de la aproximación a μ. Si queremos que la media muestral esté a una distancia menor que h de la media con una probabilidad del 95 %, debemos considerar $c = 1.96$. Por otra parte requerimos que el error cuadrático medio o varianza del estimador, $\sigma/\sqrt{n}$, no supere cierto valor d para tener una mayor precisión en la estimación. Así, en caso de conocerse el valor de σ el número n de simulaciones debe satisfacer

$$\frac{\sigma}{\sqrt{n}} < d. \tag{7.2}$$

Si en cambio el valor de σ es desconocido, entonces la desigualdad (7.2) se aplica con $S(n)$ en lugar de σ. Esto requiere que en cada simulación sea necesario calcular nuevamente la varianza muestral. Así, un procedimiento para determinar hasta qué valor de n deben generarse datos X_n es el siguiente:

```
def Media_Muestral_X(d):
    'Estimación de X(n) con ECM d'
    mediaX = simular X #X(1)
    Scuad, n = 0, 1 #Scuad = S^2(1)
    while n<=100 or sqrt(Scuad/n)>d:
        n += 1
        simular X
        actualizar mediaX #X(n)
        actualizar Scuad  #S^2(n)
    return mediaX
```

7.4.2. Fórmulas recursivas

En el algoritmo anterior, se debe calcular la media muestral y la varianza muestral en cada iteración. Por ello, es conveniente tener un método iterativo que permita calcular $\overline{X}(n)$ y $S(n)$ en cada paso, sin reutilizar todos los valores generados previamente. Para el caso de la media muestral, la recursividad puede verse como sigue:

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

$$\boxed{\overline{X}(n+1) = \overline{X}(n) + \frac{X_{n+1} - \overline{X}(n)}{n+1}.}$$

En el caso de la varianza muestral tenemos que las fórmulas para $S(n+1)$ y $S(n)$ están dadas por:

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

Para encontrar una relación entre ambas consideramos $nS^2(n+1)$ y sumamos y restamos $\overline{X}(n)$ en la expresión de la sumatoria:

$$\begin{aligned} nS^2(n+1) &= \sum_{i=1}^{n}(X_i - \overline{X}(n) + \overline{X}(n) - \overline{X}(n+1))^2 + (X_{n+1} - \overline{X}(n+1))^2 \\ &= \sum_{i=1}^{n}(X_i - \overline{X}(n))^2 + 2(\overline{X}(n) - \overline{X}(n+1))\underbrace{\sum_{i=1}^{n}(X_i - \overline{X}(n))}_{=0} + n(\overline{X}(n) - \overline{X}(n+1))^2 \\ &\quad + (X_{n+1} - \overline{X}(n+1))^2 \\ &= (n-1)S^2(n-1) + n(\overline{X}(n) - \overline{X}(n+1))^2 + (X_{n+1} - \overline{X}(n+1))^2. \end{aligned}$$

A su vez, de la fórmula de recurrencia para la media muestral tenemos que:

$$\begin{aligned} X_{n+1} - \overline{X}(n+1) &= X_{n+1} - \overline{X}(n) + \frac{X_{n+1} - \overline{X}(n)}{n+1} \\ &= \frac{n}{n+1}(X_{n+1} - \overline{X}(n)) = n \cdot (\overline{X}(n+1) - \overline{X}(n)). \end{aligned}$$

Volviendo al cálculo anterior tenemos que:

$$\begin{aligned} nS^2(n+1) &= (n-1)S^2(n) + n \cdot (\overline{X}(n) - \overline{X}(n+1))^2 + (n \cdot (\overline{X}(n) - \overline{X}(n+1))^2 \\ &= n(n+1) \cdot (\overline{X}(n) - \overline{X}(n+1))^2. \end{aligned}$$

Finalmente la fórmula de recurrencia para $S(n)$ está dada por:

$$\boxed{S^2(n+1) = \left(1 - \frac{1}{n}\right)\, S^2(n) + (n+1)\, \left(\overline{X}(n+1) - \overline{X}(n)\right)^2.}$$

Así, repasando las ideas, el estimador **media muestral** se calcula iteradamente para estimar el valor esperado de los valores simulados. Esta iteración prosigue hasta que el error cuadrático medio o varianza del estimador, σ^2/n, es menor a un valor d deseado. En caso de no conocerse σ se estima su valor con el estimador $S^2(n)$.

Así el algoritmo resulta:

```
def Media_Muestral_X(d):
    'Estimación del valor esperado con ECM<d'
    Media = simular X # X(1)
    Scuad, n = 0, 1 #Scuad = S^2(1)
    while n <= 100 or sqrt(Scuad/n) > d:
        n += 1
        simular X
        Media_Ant = Media
        Media = MediaAnt + (X - MediaAnt) / n
        Scuad = Scuad * (1 - 1 /(n-1)) + n*(Media - Media_Ant)**2
    return Media
```

7.4.3. Estimador de proporción

El estimador $\overline{X}(n)$ puede utilizarse también para estimar la proporción de casos en una población, o la probabilidad de ocurrencia de un evento. En este caso, en la i-ésima simulación se tendrá una variable de Bernoulli X_i con el valor 1 si ocurre el evento y 0 en caso contrario. Así:

$$E[X_i] = P(X_i = 1) = P(\text{ocurrencia del evento}).$$

Por lo tanto la media muestral es en este caso un estimador de esta proporción.

Ejemplo 7.3. En una simulación de llegada de clientes a un servidor que atiende entre las 8:00 y las 12:00, podría analizarse la proporción de días que quedan más de 2 clientes por atender a la hora del cierre. En ese caso, el día i (simulación i), se considera una variable aleatoria Bernoulli X_i que valdrá 1 si quedan más de dos clientes y 0 en caso contrario.

El objetivo será entonces estimar la probabilidad p de éxito, y el estimador de p será la media muestral:

$$\hat{p}(n) = \overline{X}(n),$$

donde n será el número total de simulaciones. Este número n dependerá de la *precisión* con que se quiera estimar p, en particular, del error cuadrático medio aceptable para el estimador: σ^2/n. Ahora bien, en el caso de una variable Bernoulli, la varianza σ^2 es $p(1-p)$, y por lo tanto un estimador para la varianza de la variable simulada es

$$\hat{\sigma^2} = \overline{X}(n)\,(1 - \overline{X}(n)),$$

y un estimador del error cuadrático medio del estimador, que coincide con su varianza es:

$$ECM(\hat{p}(n), p) = \text{Var}(\hat{p}(n)) = \frac{\overline{X}(n)(1-\overline{X}(n))}{n}.$$

Así, si $X_1, X_2, \ldots, X_n$, es una sucesión de v.a. independientes, Bernoulli, el algoritmo para la estimación de $p = E(X_i)$ es el siguiente:

```
def estimador_p(d):
    'Estimación de proporción con ECM<d'
    p = 0
    n = 0
    while n <= 100 or sqrt( p * (1-p) / n) > d:
        n += 1
        Simular X
        p = p + (X - p) / n
    return p
```

7.5. Estimador por intervalos

Al utilizar un estimador puntual para un parámetro, se elige un valor particular para el parámetro de acuerdo a la muestra obtenida. Así por ejemplo, si se está estimando la media de una distribución con una muestra de tamaño 100, y resulta $\overline{X}(100) = -2.5$, entonces se utilizará como parámetro exactamente ese valor.

Un **estimador por intervalo** de un parámetro es un intervalo para el que se predice que el parámetro está contenido en él. Es decir, en este caso se tiene un intervalo aleatorio con una cierta probabilidad de contener al parámetro buscado. La **confianza** que se da al intervalo es la probabilidad de que el intervalo contenga al parámetro.

7.5.1. Estimador por intervalo de $E(X)$

El estimador $\overline{X}(n)$ es un estimador puntual del valor esperado de X. Además sabemos que si $E(X) = \theta$ y Var$(X) = \sigma^2$, entonces la distribución de $\overline{X}(n)$ tiende a una normal estándar:

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

Recordemos que para $0 < \alpha < 1$, utilizamos la notación z_α para indicar el número real tal que $P(Z > z_\alpha) = \alpha$. Luego, dado que la normal estándar tiene una distribución simétrica con respecto a $x = 0$, para n suficientemente grande (> 100), tenemos que

$$P\left(-z_{\alpha/2} < \frac{\overline{X}(n) - \theta}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) = 1 - \alpha.$$

o equivalentemente

$$P\left(\overline{X}(n) - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{X}(n) + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{7.3}$$

La ecuación (7.3) determina un intervalo aleatorio que contiene al parámetro θ con una **confianza** de $1-\alpha$. Así por ejemplo, si se quiere un intervalo de confianza del $95\,\%$, entonces $\alpha/2 = 0.025$, y $z_{\alpha/2} = 1.96$, y para un $n > 100$ el intervalo será:

$$\left(\overline{X}(n) - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{X}(n) + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Análogamente, los intervalos de confianza del $99\,\%$ y del $90\,\%$ para un n determinado serán:

$$\left(\overline{X}(n) - 2.33\frac{\sigma}{\sqrt{n}},\ \overline{X}(n) + 2.33\frac{\sigma}{\sqrt{n}}\right) \qquad \text{y} \qquad \left(\overline{X}(n) - 1.64\frac{\sigma}{\sqrt{n}},\ \overline{X}(n) + 1.64\frac{\sigma}{\sqrt{n}}\right).$$

Si σ es desconocido, los intervalos anteriores se definen utilizando el estimador $\hat{\sigma} = \sqrt{S^2(n)}$. Notemos que si la muestra es de tamaño n, la longitud del intervalo de confianza del $100(1-\alpha)\,\%$ es

$$l = 2 \cdot z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad \text{o} \qquad l = 2 \cdot z_{\alpha/2}\frac{S(n)}{\sqrt{n}},$$

es decir que su longitud depende del valor de n, y más específicamente, es inversamente proporcional al valor de n. Así, en una simulación, si se quiere obtener un intervalo de confianza del $100(1-\alpha)\,\%$ y con una longitud menor a cierto número L, se continuarán generando valores hasta que

$$2 \cdot z_{\alpha/2}\frac{S(n)}{\sqrt{n}} < L.$$

```
def Media_Muestral_X(z_alfa_2, L): #z_alfa_2 = z_(alfa/2)
    'Confianza = (1 - alfa)%, amplitud del intervalo: L'
    d = L / (2 * z_alfa_2)
    Media = simular X
    Scuad, n = 0, 1
    while n <= 100 or sqrt(Scuad / n) > d:
        n += 1
        simular X
        Media_Ant = Media
        Media = MediaAnt + (X - MediaAnt) / n
        Scuad = Scuad * (1 - 1 /(n-1)) + n*(Media - Media_Ant)**2
    return Media
```

7.5.2. Estimador por intervalos de una proporción

En el caso de una variable Bernoulli, el estimador por intervalos del parámetro p también es el estimador por intervalos de la media poblacional. En este caso, el estimador para la varianza es $\overline{X}(n)(1-\overline{X}(n))$, y para n suficientemente grande se tiene que

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

Así, un intervalo de confianza del $100(1-\alpha)\,\%$ se obtiene a partir de la propiedad:

$$P\left(-z_{\alpha/2} < \sqrt{n}\,\frac{\overline{X}(n)-p}{\sqrt{\overline{X}(n)(1-\overline{X}(n))}} < z_{\alpha/2}\right) = 1-\alpha.$$

o equivalentemente, el intervalo de confianza es:

$$\left(\overline{X}(n) - z_{\alpha/2}\,\frac{\sqrt{\overline{X}(n)\,(1-\overline{X}(n))}}{\sqrt{n}},\ \overline{X}(n) + z_{\alpha/2}\,\frac{\sqrt{\overline{X}(n)\,(1-\overline{X}(n))}}{\sqrt{n}}\right)$$

```
def estimador_p(z_alfa_2,L):
    'Confianza: 100(1-alpha)%'
    'L: amplitud del intervalo'
    d = L / (2 * z_alfa_2)
    p = 0; n = 0
    while n <= 100 or sqrt(p * (1 - p) / n) > d:
        n += 1
        Simular X
        p = p + (X - p) / n
    return p
```

7.6. La técnica de Bootstrap

7.6.1. Muestras bootstrap

En términos generales, una técnica **bootstrap** es aquella que recupera una información a partir de los datos, sin asumir ninguna hipótesis sobre ellos.

Las técnicas bootstrap están asociadas al concepto de **muestra bootstrap**. Recordemos que si se han obtenido n observaciones de una variable X con distribución F: $X_1 = x_1$, $X_2 = x_2$, $\ldots$, $X_n = x_n$, entonces la distribución empírica F_e asigna a X la probabilidad:

$$p_i = P(X = x_i) = \frac{\#\{j \mid x_j = x_i\}}{n}.$$

Una **muestra bootstrap** es una muestra aleatoria de tamaño n tomada de X a partir de la distribución F_e. Dicho de otro modo, es una muestra aleatoria de tamaño n tomada con reposición del conjunto $\{x_1, x_2, \ldots, x_n\}$.

Ejemplo 7.4. Supongamos que se tienen las siguientes observaciones de una variable X:

$$x_1 = 1.4, \quad x_2 = 2.5, \quad x_3 = -0.5, \quad x_4 = 2.5.$$

La distribución empírica F_e asigna probabilidad

$$P_{F_e}(X = 2.5) = 0.5, \qquad P_{F_e}(X = 1.4) = P_{F_e}(X = -0.5) = 0.25.$$

Una muestra bootstrap es una muestra de tamaño $n = 4$ tomada de esta distribución empírica. Así por ejemplo,

$$(1.4, 1.4, 1.4, 1.4), \qquad (-0.5, 1.4, 2.5, 1.4), \qquad (2.5, -0.5, -0.5, 1.4),$$

son tres muestras bootstrap.

Si $\hat{\theta}$ es un estimador, entonces para cada muestra bootstrap también podemos evaluar $\hat{\theta}$ en esta muestra. Esto se denomina **replicación bootstrap** de $\hat{\theta}$. Por ejemplo, si $\hat{\theta} = \overline{X}(4)$ y la muestra bootstrap es $(-0.5, 1.4, 2.5, 1.4)$, entonces la replicación bootstrap es

$$\hat{\theta}(-0.5, 1.4, 2.5, 1.4) = \frac{-0.5 + 1.4 + 2.5 + 1.4}{4} = 1.2.$$

7.6.2. Estimación bootstrap

Supongamos ahora que se desea estimar determinado parámetro θ de la distribución F a partir de la muestra $X_1 = x_1$, $X_2 = x_2$, …, $X_n = x_n$. Para ello utilizamos un estimador $\hat{\theta}$, y tomamos como estimación de θ a $\hat{\theta}(x_1, \ldots, x_n)$. En algunos casos, como es el de la media muestral, sabemos que si la muestra es suficientemente grande la aproximación al parámetro a estimar $E_F[X]$ es buena, ya que la varianza del estimador es del orden de σ^2/n y además es un estimador insesgado. Pero esta suerte no ocurre con cualquier estimador, y entonces es necesario tener alguna medida de cuán bien este estimador aproxima al parámetro que se desea estimar. En particular, es importante conocer la varianza del estimador, y el error cuadrático medio del estimador con respecto al parámetro a estimar.

En casos que no se conozca la distribución F, el método bootstrap es una alternativa. Notemos que el error cuadrático medio y la varianza del estimador se definen ambos como un valor esperado:

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

La estimación bootstrap **estima** estos valores a partir de la distribución empírica de los datos y utilizando muestras bootstrap:

$$ECM(\hat{\theta}, \theta) \sim E_{F_e}[(\hat{\theta} - \theta_{F_e})^2], \qquad \text{Var}(\hat{\theta}) = E_{F_e}[(\hat{\theta} - E_{F_e}[\hat{\theta}])^2].$$

Del mismo modo se podría querer estimar una probabilidad, por ejemplo $P(a < \hat{\theta} < b)$. Dado que esta probabilidad puede verse como el valor esperado de una variable aleatoria Bernoulli:

$$X = \begin{cases} 1 & \hat{\theta}(X_1, \ldots, X_n) \in (a, b) \\ 0 & \text{en otro caso,} \end{cases}$$

entonces también puede realizarse una estimación bootstrap de la probabilidad:

$$P(a < \hat{\theta} < b) \sim \frac{1}{n^n} \#\{(x_1^*, x_2^*, \ldots, x_n^*) \mid a < \hat{\theta}(x_1^*, x_2^*, \ldots, x_n^*) < b\},$$

donde el supraíndice $*$ indica que es una muestra bootstrap de $\{x_1, \ldots, x_n\}$.

Estas estimaciones se denominan **estimaciones bootstrap ideales**, ya que toman todas las muestras bootstrap posibles. Sin embargo, en la práctica esto es computacionalmente muy costoso ya que requiere realizar n^n replicaciones bootstrap. Veamos un ejemplo para clarificar ideas.

7.6.3. Estimación bootstrap de una proporción

Ejemplo 7.5. Se han observado los siguientes datos:

$$x_1 = 1.4, \quad x_2 = 2.5, \quad x_3 = -0.5.$$

y se quiere analizar propiedades del estimador

$$\hat{\theta}(X_1, \ldots, X_n) = \frac{\overline{X}(n)}{S(n)}.$$

Por ejemplo, se quiere estimar la probabilidad

$$P(-0.2 \leq \hat{\theta} \leq 0.2). \tag{7.4}$$

La estimación bootstrap de esta probabilidad se determina usando la distribución empírica. Notemos que en nuestro caso $n = 3$, y que para este valor de n hay $3^3 = 27$ replicaciones bootstrap:

$$\hat{\theta}(a_1, a_2, a_3), \qquad a_i \in \{1.4,\ 2.5,\ -0.5\}.$$

La distribución empírica asigna una probabilidad de $1/3$ a cada valor x_i, y en consecuencia cada muestra (a_1, a_2, a_3) tiene probabilidad $1/27$. Consideramos aquí un caso donde los valores son todos distintos. Por ejemplo, una muestra bootstrap posible es $(a_1, a_2, a_3) = (2.5, -0.5, -0.5)$ y la replicación bootstrap del estimador es:

$$\hat{\theta}(a_1, a_2, a_3) = \frac{\overbrace{(2.5 - 0.5 - 0.5)/3}^{=0.5}}{\sqrt{(2.5 - 0.5)^2 + (-0.5 - 0.5)^2 + (-0.5 - 0.5)^2)/2}}.$$

Si la muestra tiene $a_1 = a_2 = a_3$, comprobamos si $-0.2\,S(n) \leq \overline{X}(n) \leq 0.2\,S(n)$, que para el caso particular de este ejemplo siempre será falso.

Entonces el valor buscado en (7.4) está dado por la probabilidad de que el estimador esté entre -0.2 y 0.2 bajo la distribución empírica. Esto es:

$$P_{F_e}(-0.2 \leq \hat{\theta} \leq 0.2) = \frac{1}{3^3}\#\{(a_1, a_2, a_3) \mid -0.2 \leq \hat{\theta}(a_1, a_2, a_3) \leq 0.2\},$$

que en este caso particular es $\frac{1}{9}$, o aproximadamente 0.11.

En el Ejemplo 7.5 el tamaño de la muestra es pequeño, por lo cual la estimación bootstrap de la probabilidad requiere sólo 27 cálculos. En tal caso se dice que es una **estimación bootstrap ideal**. Ahora bien, si el tamaño de la muestra n es mucho más grande, entonces la estimación bootstrap ideal requiere n^n cálculos lo cual puede ser computacionalmente muy costoso o hasta imposible.

En estos casos la técnica Boostrap consiste en seleccionar aleatoriamente N muestras bootstrap, con N menor a n^n, y estimar el valor esperado a calcular con el promedio para estas N muestras. Esto es una aplicación del método de Monte Carlo.

7.6.4. Estimación bootstrap del ECM

Veamos el caso particular de la estimación del error cuadrático medio de un estimador $\hat{\theta}$ con respecto a un parámetro θ. En el caso que se conozca la distribución F de los datos, se podrá calcular con mayor o menor complejidad el valor exacto del ECM:

$$ECM(\hat{\theta}, \theta) = E_F((\hat{\theta} - \theta)^2).$$

Aquí el subíndice F indica que el valor esperado se calcula en términos de esa distribución. Por ejemplo, si se asumen los datos con distribución normal estándar, entonces se tendrá:

$$ECM(\hat{\theta}; \theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\hat{\theta}(x_1, x_2, \ldots, x_n) - \theta)^2 \frac{1}{(\sqrt{2\pi})^n}\, e^{-(x_1^2+\cdots+x_n^2)/(2\sigma^2)}\, dx_1 dx_2 \ldots dx_n.$$

En cambio, si la distribución F de los datos es desconocida no es posible determinar de manera exacta este valor, y una alternativa es aproximarla con muestras bootstrap. Así, si la muestra es de tamaño n, y los datos obtenidos son x_1, x_2, …, x_n, el error cuadrático medio se calculará de la siguiente manera:

a) Se calcula el parámetro $\theta(F_e)$. Por ejemplo, si θ_e es la varianza, entonces

$$\theta(F_e) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \qquad \text{con} \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

b) Se consideran N muestras bootstrap y se calculan las respectivas replicaciones bootstrap del estimador. Esto es, del conjunto de muestras bootstrap:

$$\{(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \mid x_{i_j} \in \{x_1, \ldots, x_n\},\ 1 \leq j \leq n\}$$

de cardinal n^n se toman N elementos aleatoriamente:

$$(x_{i_1}^{(1)}, x_{i_2}^{(1)}, \ldots, x_{i_n}^{(1)}),\ (x_{i_1}^{(2)}, x_{i_2}^{(2)}, \ldots, x_{i_n}^{(2)}), \qquad (x_{i_1}^{(N)}, x_{i_2}^{(N)}, \ldots, x_{i_n}^{(N)}),$$

y en cada una de estas muestras se replica el estimador. Así por ejemplo, si $\hat{\theta}$ es la varianza muestral, entonces en la muestra bootstrap $(x_{i_1}, \ldots, x_{i_n})$ se evalúa:

$$\hat{\theta} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{i_j} - \overline{y})^2, \qquad \text{con } \overline{y} = \frac{1}{n}\sum_{j=1}^{n} x_{i_j}.$$

c) Por último, el error cuadrático medio para la distribución empírica es un valor esperado de las diferencias al cuadrado entre las realizaciones bootstrap y el parámetro a estimar. Dado que no se han tomado todas las muestras bootstrap, este valor esperado se aproxima por Monte Carlo como un promedio de los N valores obtenidos:

$$ECM(\hat{\theta}, \theta) \sim \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}(x_{i_1}^{(j)}, x_{i_2}^{(j)}, \ldots, x_{i_n}^{(j)}) - \theta(F_e)\right)^2.$$

Ejemplo 7.6. En el Ejemplo 7.5, la estimación bootstrap **ideal** del error cuadrático medio de la varianza muestral con respecto a la varianza se calcula tomando las 27 muestras bootstrap, y los parámetros utilizados se calculan con la distribución empírica. Así, como la media empírica es

$$\mu_{F_e} = \frac{1.4 + 2.5 - 0.5}{3} = \frac{3.4}{3},$$

entonces la varianza empírica es

$$\sigma_{F_e}^2 = \frac{(1.4 - \frac{3.4}{3})^2 + (2.5 - \frac{3.4}{3})^2 + (-0.5 - \frac{3.4}{3})^2}{3} = \frac{47.46}{27} \simeq 1.53556.$$

Luego la estimación bootstrap del error cuadrático medio del estimador con respecto a la varianza está dado por:

$$ECM(S^2(3), \sigma_{F_e}^2) = \frac{1}{27}\cdot\sum_{a=(a_1,a_2,a_3)}\left(\frac{1}{2}\sum_{i=1}^{3}(a_i - \overline{a})^2 - \underbrace{1.53556}_{\sigma_{F_e}^2}\right)^2, \qquad \overline{a} = \frac{a_1 + a_2 + a_3}{3}$$

Ejemplo 7.7. Supongamos que se tienen datos de un sistema durante M días. Para cada día, se conoce el número de clientes que han ingresado al sistema, que llamamos $n_1, n_2, \ldots, n_M$. A su vez, para el día j, se conoce el tiempo que cada cliente pasó en el sistema:

$$T_{1,j}, T_{2,j}, \ldots, T_{n_j,j}, \qquad 1 \leq j \leq M.$$

Una pregunta posible, es determinar cuál es el tiempo promedio que un cliente pasa en el sistema.

Notemos que en un día en particular, los tiempos de permanencia de los clientes pueden no ser variables aleatorias independientes. Sin embargo, podemos asumir que los tiempos totales de permanencia en días distintos sí son variables que provienen de una misma distribución, e independientes entre sí. Denotamos con D_j la suma de los tiempos de permanencia de todos los clientes en el día j:

$$D_j = T_{1,j} + T_{2,j} + \ldots, T_{n_j,j}, \qquad 1 \leq j \leq n.$$

El parámetro que se desea estimar es el promedio de permanencia de un cliente en el sistema, que por el Teorema Central del Límite puede determinarse como:

$$\theta = \lim_{K\to\infty} \frac{D_1 + D_2 + \cdots + D_K}{n_1 + n_2 + \cdots + n_K},$$

en caso que fuera posible tomar muestras de cualquier tamaño K. Así, un estimador natural de θ es el cociente de las medias muestrales para D y n, ya que:

$$\hat{\theta} = \frac{D_1 + D_2 + \cdots + D_M}{n_1 + n_2 + \cdots + n_M} = \frac{\frac{D_1+D_2+\cdots+D_M}{M}}{\frac{n_1+n_2+\cdots+n_M}{M}} = \frac{\overline{D}}{\overline{n}}.$$

Para determinar el error cuadrático medio del estimador, se debería conocer la distribución F de la variable (vector) aleatorio (D, n), y en base a esta información determinar:

$$ECM(\hat{\theta},\, \theta) = E_F\left[\left(\hat{\theta}(D,n) - \frac{E[D]}{E[n]}\right)^2\right].$$

Pero si no se conoce esta distribución, se puede aplicar la técnica bootstrap utilizando la distribución empírica de los datos obtenidos en la muestra. Así, asignamos la probabilidad:

$$P_{F_e}(D = D_j, n = n_j) = \frac{1}{M}, \qquad 1 \leq j \leq M,$$

y el parámetro θ a estimar en la distribución empírica está dado por:

$$\theta(F_e) = \frac{E_{F_e}(D)}{E_{F_e}(D)} = \frac{D_1 + D_2 + \cdots + D_M}{n_1 + n_2 + \cdots + n_M}.$$

Consideramos ahora B muestras bootstrap:

$$b^{(j)} = \big((D_{i_1}, n_{i_1}),\ (D_{i_2}, n_{i_2}), \ldots, (D_{i_M}, n_{i_M})\big)^{(j)}, \qquad 1 \leq j \leq B,$$

y calculamos las correspondientes realizaciones bootstrap:

$$\hat{\theta}(b^{(j)}) = \frac{D_{i_1} + D_{i_2} + \cdots + D_{i_M}}{n_{i_1} + n_{i_2} + \cdots + n_{i_M}}, \qquad 1 \leq j \leq B.$$

Por último, la estimación bootstrap del error cuadrático medio está dada por:

$$\frac{1}{B} \sum_{j=1}^{B} \left(\hat{\theta}(b^{(j)}) - \theta(F_e) \right)^2.$$

Notemos que si el tamaño de la muestra M es grande y se considerara una estimación boostrap ideal, el cálculo anterior puede implicar una suma de una gran cantidad de términos. Por ejemplo, si $M = 20$,

$$20^{20} = 104\,857\,600\,000\,000\,000\,000\,000.$$

7.6.5. Estimación bootstrap de Var$(\hat{\theta})$

En el caso de la estimación boostrap de la varianza de un estimador $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$, utilizamos el estimador varianza muestral. Es decir, si $\hat{\theta}$ es un estimador queremos calcular el valor:

$$E\left(\left(\hat{\theta} - E(\hat{\theta}) \right)^2 \right),$$

y que la técnica bootstrap aproxima con la distribución empírica. Esto es:

$$E_{F_e}\left(\left(\hat{\theta} - E_{F_e}(\hat{\theta}) \right)^2 \right) \sim E\left(\left(\hat{\theta} - E(\hat{\theta}) \right)^2 \right).$$

En la distribución empírica, $\hat{\theta}$ toma valores en n^n muestras posibles, por lo cual la varianza empírica del estimador puede no ser fácilmente calculable y en ese caso se estima a su vez con la varianza muestral a partir de una muestra de tamaño N de valores del estimador:

$$\frac{1}{N-1} \sum_{i=1}^{N} (\hat{\theta}_i - \overline{\hat{\theta}})^2, \qquad \overline{\hat{\theta}} = \frac{1}{N} \sum_{i=1}^{N} \theta_i.$$

Supongamos entonces que se tiene una muestra de tamaño n:

$$x_{i_1}, x_{i_2}, \ldots, x_{i_n}.$$

Para la estimación bootstrap se consideran N muestras bootstrap:

$$b_1 = (x_{i_1}^{(1)}, x_{i_2}^{(1)}, \ldots, x_{i_n}^{(1)}),\ b_2 = (x_{i_1}^{(2)}, x_{i_2}^{(2)}, \ldots, x_{i_n}^{(2)}), \qquad , b_N = (x_{i_1}^{(N)}, x_{i_2}^{(N)}, \ldots, x_{i_n}^{(N)}),$$

y se calculan las N replicaciones bootstrap correspondientes:

$$\hat{\theta}(b_1), \quad \hat{\theta}(b_2), \quad \ldots \quad \hat{\theta}(b_N).$$

Hasta aquí es el equivalente de haber tomado una muestra de tamaño N de la variable aleatoria $\hat{\theta}$. Ahora se evalúa la media muestral en la muestra obtenida:

$$\hat{\theta}_m = \frac{1}{N} \sum_{j=1}^{N} \hat{\theta}(b_j),$$

y la estimación bootstrap de la varianza del estimador será,

$$\hat{\text{Var}}_{F_e}(\hat{\theta}) = \frac{1}{N-1} \sum_{j=1}^{N} (\hat{\theta}(b_j) - \hat{\theta}_m)^2.$$