Análisis estadístico de datos simulados

Índice

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 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.

Tabla 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.

2. Selección de una distribución de probabilidad

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.

Dentro de las distribuciones de probabilidad más utilizadas están las siguientes:

- Distribuciones continuas:

  a) Uniforme: Para cantidades que varían aleatoriamente entre valores a y b, y que no se conocen más datos.

  b) Exponencial: Tiempos entre llegadas de clientes a un sistema, y que ocurren a una tasa constante. Tiempos de falla de máquinas.

  c) Gamma, Weibull: Tiempo de servicio, tiempos de reparación.

  d) Normal: Errores. Aplicación del Teorema central del límite.

  e) Otras distribuciones: Ver (Law & Kelton, cap. 6)

- Distribuciones discretas:

  a) Bernoulli.

  b) Uniforme discreta.

  c) Geométrica: número de observaciones hasta detectar el primer error.

  d) Binomial negativa: número de observaciones hasta detectar el n-ésimo error.

  e) Poisson: Número de eventos en un intervalo de tiempo, si ocurren a tasa constante.

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.

Presentamos un resumen de las distribuciones más utilizadas. Un análisis más completo y detallado puede encontrarse en [1], capítulo 6.

2.1. Distribuciones continuas

En los siguientes casos se define la función de densidad sólo en el rango de la variable aleatoria.

- **Distribución uniforme.** $\mathcal{U}(a, b)$.

  Figura 1. Su función de densidad está dada por

  $$f(x) = \frac{1}{b-a}, \qquad a < x < b.$$

  Parámetros:

  - a: posición, $b - a$: escala.
  - Media: $\frac{a+b}{2}$.
  - Varianza: $\frac{(b-a)^2}{2}$.

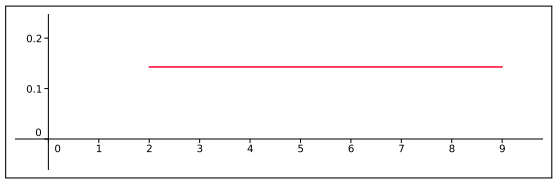

Figura 1: Distribución uniforme. Función de densidad.

- **Distribución Gamma:** $\Gamma(\alpha, \beta)$: Su función de densidad está dada por:

  $$f(x) = \frac{1}{\Gamma(\alpha)}\, \beta^{-\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, \qquad x > 0.$$

  En la Figura 2 se muestran los gráficos de $\Gamma(\alpha, 1)$, para $\alpha = 0.5$, 1, 2 y 3. Notar de la definición de f que $\alpha = 1$ corresponde a la distribución exponencial $\mathcal{E}(1)$.

  Parámetros:

  - α: forma, β: escala.
  - Media: $\alpha\beta$.
  - Varianza: $\alpha\beta^2$.

- **Distribución Weibull**(α, β). Su función de densidad está dada por:

  $$f(x) = \alpha\beta^{-\alpha} x^{\alpha-1}\, e^{-(x/\beta)^{\alpha}}, \qquad x > 0.$$

  En la Figura 3 se muestran gráficos para $\beta = 1$ y $\alpha = 0.5$, 1, 2 y 3.

  Parámetros:

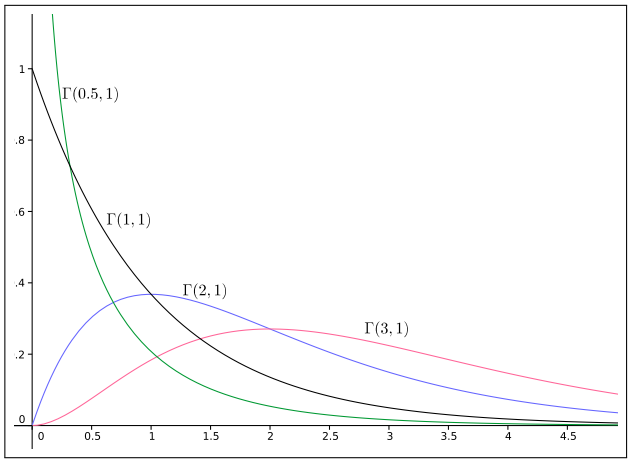

Figura 2: Distribuciones Gamma

- α: forma, β: escala.
- $$\text{Media: } \frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right) \qquad \text{Varianza: } \frac{\beta^2}{\alpha}\left[2\Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha}\left(\Gamma\left(\frac{1}{\alpha}\right)\right)^2\right].$$

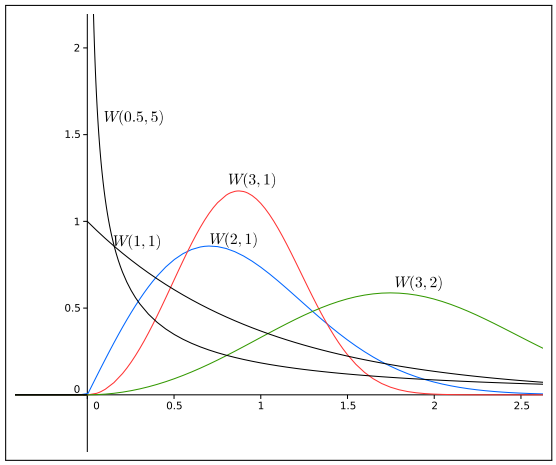

Figura 3: Distribución Weibull

- **Distribución Normal** $N(\mu, \sigma^2)$ Su función de densidad está dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-(x-\mu)^2/(2\sigma^2)), \qquad -\infty < x < \infty.$$

Parámetros:

- μ: posición, σ: escala.

- Media: μ.
- Varianza: σ^2.

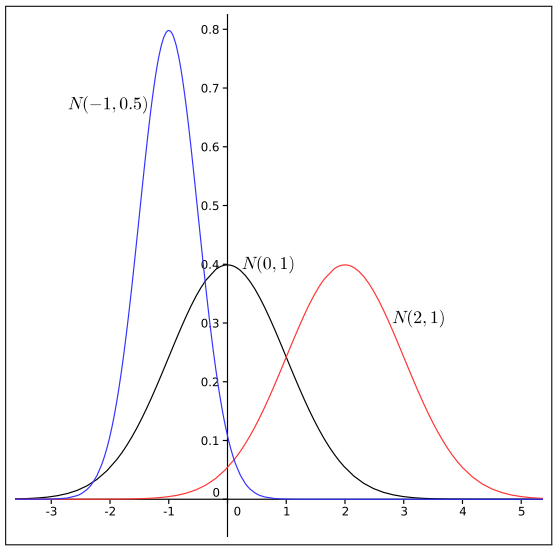

Figura 4: Distribución normal

- **Distribución Lognormal** $LN(\mu, \sigma^2)$ Su función de densidad está dada por:

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\log(x)-\mu)^2/(2\sigma^2)}, \qquad x > 0.$$

Parámetros:

- σ: forma, μ: escala.
- Media: $e^{\mu+\sigma^2/2}$.
- Varianza: $e^{2\mu+\sigma^`}(e^{\sigma^2} - 1)$.

2.2. Distribuciones discretas

Todas las siguientes distribuciones toman valores en un subconjunto de $\mathbb{Z}$.

- **Distribución uniforme** $U[a, b]$. Ocurrencia de un suceso aleatorio de un conjunto de eventos con igual probabilidad de éxito.

$$p(i) = \frac{1}{b - (a - 1)}, \qquad a \leq i \leq b.$$

Parámetros:

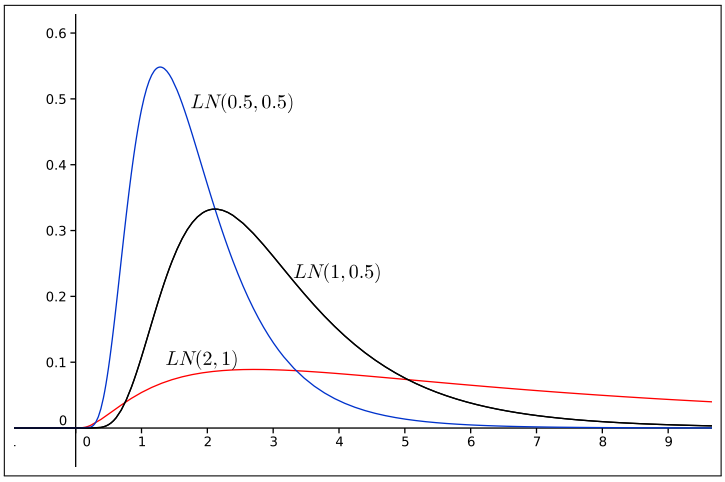

Figura 5: Distribución lognormal

- $a \leq b$. a: de posición. $b - a$: de escala.
- Media: $\frac{a+b}{2}$.
- Varianza: $\frac{(b-a+1)^2}{6}$.

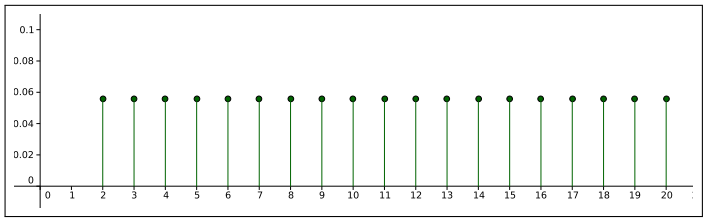

Figura 6: Distribución discreta. $U[2, 20]$

- **Distribución binomial** $B(n, p)$

  Número de éxitos en n ensayos independientes con probabilidad p de éxito.

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad 0 \leq i \leq n.$$

  Parámetros:

  - n, p.
  - Media: np.
  - Varianza: $np(1-p)$.

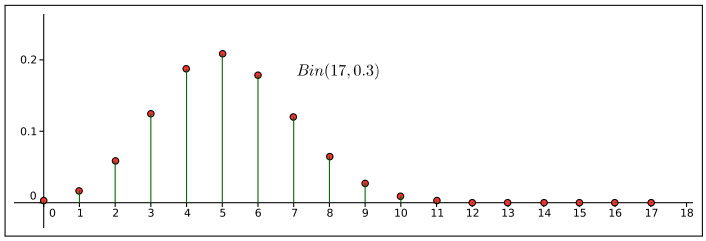

Figura 7: Distribución binomial. $Bin(17, 0.3)$

- **Distribución geométrica** $Geom(p)$

  Número de ensayos en una secuencia de eventos independientes con probabilidad p de éxito, hasta obtener el primer éxito.

$$p(i) = p\,(1-p)^i, \qquad i \geq 1.$$

  Parámetros:

  - p.
  - Media: $\frac{1}{p}$.
  - Varianza: $\frac{1-p}{p^2}$.

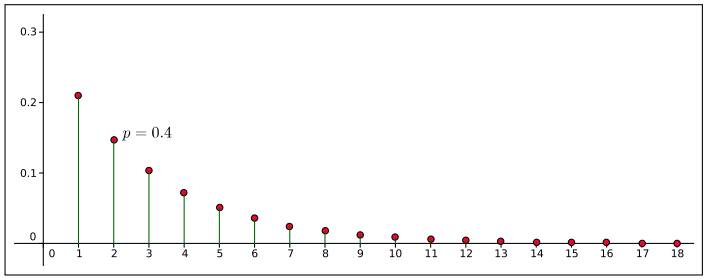

Figura 8: Distribución geométrica. $p = 0.4$

- **Distribución de Poisson** $\mathcal{P}(\lambda)$ Número de eventos que ocurren en un intervalo de tiempo cuando los eventos ocurren a una tasa constante.

$$p(i) = e^{-\lambda}\,\frac{\lambda^i}{i!}, \qquad i \geq 0.$$

  Parámetros:

  - $\lambda > 0$.
  - Media: λ.
  - Varianza: λ.

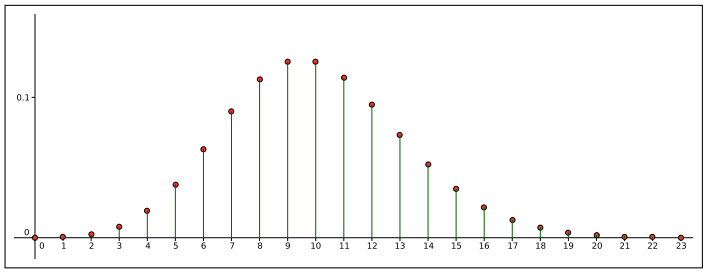

Figura 9: Distribución de Poisson. $\lambda = 15$

2.3. La distribución empírica

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.

Si la distribución es continua, la distribución empírica puede definirse a partir de las observaciones de la siguiente manera. Si se han observado datos

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

entonces en primer lugar se los ordena de forma creciente:

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

donde la notación $X_{(i)}$ es la observación que ocupa el i-ésimo lugar en el ordenamiento. Una posibilidad es tomar como distribución empírica a la distribución discreta que da a cada valor observado una probabilidad igual a la frecuencia observada. Por ejemplo, si los datos observados son $X_1 = 3.2$, $X_2 = 4.3$, $X_3 = -2.0$, $X_4 = 3.2$, $X_5 = 0$, entonces

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

y

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

La desventaja es que a partir de esta distribución sólo se simularán datos iguales a los observados. Entonces otra posibilidad es suavizar esta distribución empírica y darle forma de una curva lineal a trozos:

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

La Figura 10 ilustra ambas distribuciones empíricas para los datos del ejemplo. Así, será posible simular

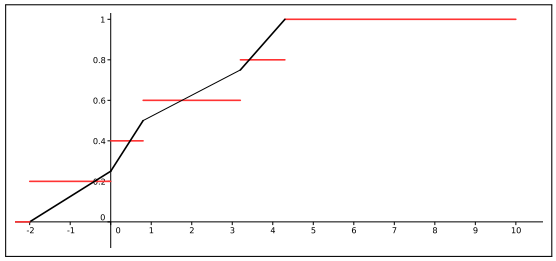

Figura 10: Distribuciones empiricas

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

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

y

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

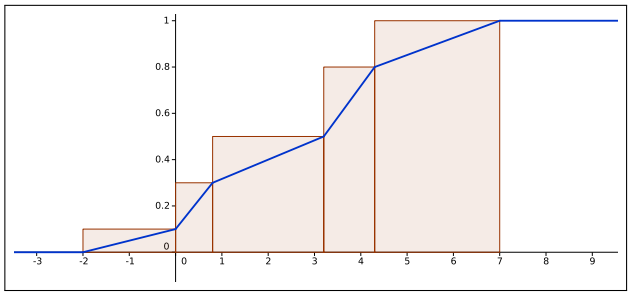

Figura 11: Distribución empírica a partir de datos agrupados

Si en cambio se asume que la distribución es discreta y se conocen los datos observados $X_1, X_2, \ldots, X_n$, la distribución empírica asigna una función de masa de probabilidad empírica a cada x dada por

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

Es decir, $p(x)$ es la frecuencia relativa observada de x.

3. Análisis de independencia de los datos

Muchos de los test estadísticos se basan en la independencia de las observaciones $X_1, X_2, \ldots, X_n$: estimaciones de máxima verosimilitud, test chi-cuadrado, Kolmogorov-Smirnov, y otros. Por lo tanto es importante analizar este aspecto de los datos observados.

Dos técnicas que sirven para analizar independencia, de un modo algo informal, son las siguientes:

- Gráficos de correlación: $\hat{\rho}_j$.
- Diagramas de dispersión (scattering): (X_i, X_{i+1}).

En los gráficos de **correlación**, se grafican los valores de las correlaciones muestrales $\hat{\rho}_j$, dadas por

$$\hat{\rho}_j = \frac{\hat{C}_j}{S^2(n)}, \qquad \hat{C}_j = \frac{\sum_{i=1}^{n-j}(X_i - \overline{X}(n))(X_{i+j} - \overline{X}(n)}{n-j}$$

donde $\overline{X}(n)$ y $S^2(n)$ denotan la media muestral y varianza muestral, cuya definición veremos más adelante.

Si los datos X_i son independientes, entonces los valores de $\hat{\rho}_j$ están próximos a cero. Por lo tanto, si los valores de $\hat{\rho}_j$ se alejan significativamente de 0 hay una fuerte evidencia de que los datos están correlacionados, es decir, no son independientes.

Los **diagramas de scattering** o **de dispersión** son el gráfico de los pares de puntos (X_i, X_{i+1}). Si los datos son independientes, estos pares de puntos deberían estar distribuidos aleatoriamente en algún

sector del plano. Por ejemplo, si son valores positivos, estarán distribuidos en el primer cuadrante. Por otra parte, si los pares de puntos se agrupan sobre una recta hay una fuerte evidencia que los datos están correlacionados.

4. 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 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

Tabla 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.

5. 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 5.1. 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, $X_1, X_2, \ldots, X_n$, los siguientes son estimadores:

$$\hat{\theta}_1 = \frac{X_1 + X_n}{2}, \qquad \hat{\theta}_2 = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{1}$$

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.

5.1. 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 (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 (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: utiliza toda la información obtenida de la muestra.

5.2. 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 $X_1, X_2, \ldots, X_n$ 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 $L(\theta)$:

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

El estimador de máxima verosimilitud tiene, en general, las siguientes propiedades:

a) Es único: $L(\hat{\theta}) > L(\theta)$ para cualquier otro valor de θ.

b) La distribución asintótica de $\hat{\theta}$ tiene media θ.

c) Es invariante: $\phi = h(\theta)$, entonces $\hat{\phi} = h(\hat{\theta})$.

d) Su distribución asintótica está normalmente distribuida.

e) Es fuertemente consistente: $\lim_{n\to\infty} \hat{\theta} = \theta$.

Ejemplo 5.1. 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 β, y este parámetro se estimará a partir de la muestra.

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

$$f_\beta(x) = \frac{1}{\beta} e^{-x/\beta}, \qquad x > 0,$$

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

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

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

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}(n).$$

Ejemplo 5.2. 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:

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

5.3. 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.

5.4. 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:

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

Notemos que si $E(X_i) = \theta$, $1 \leq i \leq n$, entonces

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

Por ello, la media muestral se utiliza como estimador de la media de una distribución.

Como estimador de la media de una distribución F, su error cuadrático medio está dado por su varianza. Esto es:

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

Así, cuanto mayor sea el tamaño de la muestra, menor será el Error Cuadrático Medio de la media muestral como estimador de la media de la población.

5.5. La varianza muestral

En el caso de la media muestral, el error cuadrático medio para la estimación de la media está acotado por $\frac{\sigma^2}{n}$. Entonces, si se quiere que este error 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 ecuación anterior da poca información y se hace necesario tener 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.

5.6. Simulaciones

Si en una 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

$$P\left(|\overline{X}(n) - \mu| > c\,\frac{\sigma}{\sqrt{n}}\right) \sim P(|Z| > c) = 2(1 - \Phi(c)).$$

Entonces, si queremos que la media muestral esté a una distancia menor que h de la media, con una probabilidad del 95 %, debemos considerar $c = 1.96$, y por lo tanto n debe satisfacer

$$1.96 \frac{\sigma}{\sqrt{n}} < h.$$

Así, h/c será una cota aceptable de la desviación del estimador ($\sigma/\sqrt{n}$).

Así, un procedimiento para determinar hasta qué valor de n deben generarse datos X_n es el siguiente:

SIMULACIÓN

1 Elegir d como valor aceptable de $\sigma/\sqrt{n}$.
2 Generar al menos 100 valores de X.
3 **while** $S(n)/\sqrt{n} > d$
4 $\quad n = n + 1$
5 $\quad$ Generar X
6 **return** $\overline{X}(n)$

En el algoritmo anterior, se deben calcular la media muestral al final del algoritmo y la varianza muestral en cada iteración. Por ello, es conveniente tener un método iterativo que permita calcular $\overline{X}(n)$ y $S^2(n)$ en cada paso, sin reutilizar todos los valores generados previamente.

Para el caso de la media muestral, la recursividad puede verse como sigue:

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

En el caso de la varianza muestral, resulta:

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

El segundo término del miembro derecho de la última igualdad puede reemplazarse usando las identidades para $\overline{X}(n)$ y $\overline{X}(n+1)$:

$$\begin{aligned}(n+1)\left(\overline{X}(n+1)-\overline{X}(n)\right) &= X_{n+1}-\overline{X}(n)\\ n\overline{X}(n+1)+\overline{X}(n+1)-n\overline{X}(n) &= X_{n+1}\\ n\left(\overline{X}(n+1)-\overline{X}(n)\right) &= X_{n+1}-\overline{X}(n+1)\end{aligned}$$

Luego tenemos:

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

Así, un algoritmo para la estimación de la media con una varianza del estimador menor a d será como el siguiente:

ESTIMACIÓN DE $\overline{X}(n)$

1 $Media = 0$
2 $Var = 0$
3 $j = 0$
4 **while** $\sqrt{\frac{Var}{j}} > d$ or $j < 100$
5 $\quad j = j + 1$
6 $\quad$ Generar X;
7 $\quad MediaAnt = Media$
8 $\quad Media = Media + \frac{1}{j}(X - Media)$
9 $\quad Var = (1 - \frac{1}{j-1}) \cdot Var + j \cdot (Media - MediaAnt)^2$
10 **return** $Media$

Ejemplo 5.3. El estimador $\overline{X}(n)$ puede utilizarse también para estimar la proporción de casos en una población. Por ejemplo, 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. Esto es:

$$X_i = \begin{cases}1 & \text{probabilidad } p\\ 0 & \text{probabilidad } 1-p.\end{cases}$$

El objetivo será entonces estimar la probabilidad p de éxito. Dado que $p = E[X_i]$, el estimador de p será la media muestral:

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

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. 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 es

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

Dado que la varianza y error cuadrático medio del estimador $\overline{X}(n)$ para la estimación de p es

$$ECM(\overline{X}(n); p) = E[(\overline{X}(n) - p)^2] = \text{Var}(\overline{X}(n)) = \frac{p(1-p)}{n},$$

y siendo p desconocido, se estima la varianza del estimador $\overline{X}(n)$ por:

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

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

ESTIMACIÓN DE p

1 $p = 0$
2 $j = 0$
3 **while** $\sqrt{\frac{p(1-p)}{j}} > d$ or $j < 100$
4 $\quad j = j + 1$
5 $\quad$ Generar X
6 $\quad p = p + \frac{1}{j}\,(X - p)$
7 **return** p

6. 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.

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

El estimador $\overline{X}(n)$ es un estimador puntual de la media poblacional. En particular, sabemos que si $E(X) = \theta$ y Var$(X) = \sigma^2$, entonces

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

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

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

La ecuación (2) 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

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

6.2. 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:

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

7. Técnica de Bootstrap para estimar el ECM

En términos generales, una técnica **bootstrap** es aquella que recupera una información a partir de los datos, sin asumir ninguna hipótesis sobre ellos. Un caso particular es el que veremos en esta sección.

Hemos visto que la media muestral $\overline{X}(n)$ es un estimador insesgado de la media, y por ende su error cuadrático medio es $S^2(n)/n$. Luego, independientemente de la distribución de los datos, existe una fórmula explícita para calcular el error cuadrático medio.

Ahora bien, supongamos el caso general en que se tiene un determinado estimador $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$, que se utiliza para estimar un determinado 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 estandar, entonces se tendrá:

$$ECM(\hat{\theta};\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\hat{\theta}(x_1,x_2,\ldots,x_n)-\theta)^2\frac{1}{(\sqrt{2\pi})^n}\,e^{-(x_1^2+\cdots+x_n^2)/(2\sigma^2)}\,dx_1dx_2\ldots dx_n.$$

En cambio, si la distribución F de los datos es desconocida no es posible determinar de manera exacta este valor, y en tal caso una posibilidad es utilizar la distribución empírica F_e de los datos en lugar de la distribución real F.

La distribución empírica F_e utilizada en este caso es aquella que asigna a cada elemento de la muestra una probabilidad igual a la frecuencia relativa con que aparece. Por ejemplo, si la muestra tiene tamaño 4 y los datos son $X_1 = 4.3$, $X_2 = X_3 = 1.8$ y $X_4 = -2.1$, entonces la distribución empírica está dada por:

$$F_e(x) = \begin{cases} 0 & x < -2.1 \\ 0.25 & -2.1 \leq x < 1.8 \\ 0.75 & 1.8 \leq x < 4.3 \\ 1 & x \geq 4.3 \end{cases}$$

o equivalentemente, asigna una probabilidad de masa dada por $p_e(4.3) = p_e(-2.1) = 0.25$ y $p_e(1.8) = 0.5$. Para n suficientemente grande, la distribución empírica F_e converge uniformemente a F, y por lo tanto los correspondientes parámetros (media, varianza, mediana, etc.) $\theta(F_e)$ convergen a $\theta(F)$. Luego, el error cuadrático medio $E_F[(\hat{\theta} - \theta)^2]$ puede aproximarse con el error cuadrático medio calculado con la distribución empírica:

$$E_{F_e}[(\hat{\theta} - \theta(F_e))^2].$$

Así, si la muestra es de tamaño n, y los datos obtenidos son $x_1, x_2, \ldots, x_n$, el error cuadrático medio se calculará de la siguiente manera:

a) Calcular $\theta(F_e)$. Por ejemplo, si θ_e es la varianza, entonces

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

b) Calcular el error cuadrático medio utilizando la distribución empírica:

$$\frac{1}{n^n} \sum_{i_1=1}^{n} \cdots \sum_{i_n=1}^{n} (\hat{\theta}(x_{i_1}, \ldots, x_{i_n}) - \theta(F_e))^2. \tag{3}$$

Notemos que la fórmula (3) es un promedio con n^n términos, donde se suma sobre todas las n-uplas posibles utilizando los datos observados. Para un n grande, este promedio puede estimarse por Monte Carlo seleccionando N términos de manera aleatoria.

Esto es, del conjunto $\{(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \mid x_{i_j} \in \{x_1, \ldots, x_n\},\ 1 \leq j \leq n\}$ de cardinal n^n, se toma una muestra de tamaño N:

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

y se estima el error cuadrático medio empírico con Monte Carlo:

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

7.1. Un ejemplo

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, e incluso puede desconocerse su distribución. 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 está dado por:

$$\theta = \lim_{M\to\infty} \frac{D_1 + D_2 + \cdots + D_M}{n_1 + n_2 + \cdots + D_M},$$

donde el límite se toma considerando muestras de tamaño M de pares (D_j, n_j). Si se conociera la distribución de las variables D y n, podría efectuarse una simulación de estas muestras o estimar el límite que define a θ. Notemos que por el Teorema Central del Límite, θ es el cociente entre los valores esperados de D y n:

$$\theta = \lim_{M\to\infty} \frac{\frac{D_1+D_2+\cdots+D_M}{M}}{\frac{n_1+n_2+\cdots+D_M}{M}} = \frac{\lim_{M\to\infty} \frac{D_1+D_2+\cdots+D_M}{M}}{\lim_{M\to\infty} \frac{n_1+n_2+\cdots+D_M}{M}} = \frac{E[D]}{E[n]}.$$

Así, un estimador natural de θ es el cociente de las medias muestrales:

$$\hat{\theta} = \frac{\overline{D}}{\overline{n}}.$$

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

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

Si en cambio se desconocen las distribuciones de las variables D y n, se puede aplicar la técnica bootstrap utilizando la distribución empírica de los datos obtenidos en una determinada 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 correspondiente a esta distribución está dado por:

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

El objetivo es ahora determinar cuál es el error cuadrático medio del estimador. Para esto, consideramos todas las M uplas posibles

$$(D_{i_1}, n_{i_1}),\ (D_{i_2}, n_{i_2}), \ldots, (D_{i_M}, n_{i_M}),$$

y calculamos:

$$\frac{1}{M^M} \sum_{i_1,\ldots,i_M} \left(\frac{D_{i_1} + D_{i_2} + \cdots + D_{i_M}}{n_{i_1} + n_{i_2} + \cdots + n_{i_M}} - \theta(F_e) \right)^2.$$

Si el tamaño de la muestra M es grande, el cálculo anterior puede implicar una suma de una gran cantidad de términos. Por ejemplo, si $M = 20$,

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

En tal caso, se puede aplicar Monte Carlo para estimar el promedio con un número menor de términos.

Referencias

[1] AVERILL M. LAW, W. DAVID KELTON, *Simulation Modeling and Analysis*, 3ra. edición. Edit. Mc. Graw Hill. 2000.

[2] SHELDON ROSS, *Simulation*, 3ra. edición, Edit. Academic Press. 2002.