Unidad 6. Simulación Monte Carlo, Estimación puntual y Estimación por Intervalos de Confianza

1. Simulación Monte Carlo

Ya vimos que es posible conocer la ditribución exacta de algunos funcionales $T(X_1, \ldots, X_n)$ donde $(X_1, \ldots, X_n)$ es una muestra aleatoria de una variable aleatoria X con función de distribución F. También hemos calculado características numéricas de algunos funcionales como esperanza y varianza. Por ejemplo, si $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$, entonces su distribución es normal o gamma si la distribución de X es normal o exponencial, respectivamente. Análogamente, conocemos la distribución de $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i/(n-1)$, si X tiene distribución normal.

Pero estas situaciones son excepcionales y no siempre se puede hallar la forma explícita de la distribución de una función de la muestra aleatoria o de una característica numérica de esa función.

En general, la simulación Monte Carlo es un instrumento muy potente para el estudio por experimentación de propiedades de objetos aleatorios.

Consideremos una variable aleatoria Y con esperanza finita y sea G su función de distribución. Si $y_1, \ldots, y_R$ son R valores simulados de Y de forma independiente entonces podemos establecer las siguientes conclusiones:

- Por la ley de los grandes números,

$$\frac{1}{R} \sum_{i=1}^{R} y_i \approx \mathrm{E}(Y).$$

- Por el Teorema de Glivenko-Cantelli,

$$G_R^*(y) = \frac{1}{R}\#\{i : y_i \leq y\} \approx G(y),$$

provisto que R sea grande.

El desarrollo anterior se aplica al estudio del comportamiento de funciones de muestras aleatorias, como indicamos arriba. Si contamos con una función $Y = T(X_1, \ldots, X_n)$, la cual es una variable aleatoria, para hallar una aproximación a $G(y) = P(Y \leq y)$ procedemos de la siguiente forma:

- Generamos R muestras o réplicas independientes de $(X_1, \ldots, X_n)$:

$$\begin{array}{ll} \text{Muestra 1} & (X_1^{(1)}, \ldots, X_n^{(1)}) \\ \text{Muestra 2} & (X_1^{(2)}, \ldots, X_n^{(2)}) \\ \ldots & \\ \text{Muestra } R & (X_1^{(R)}, \ldots, X_n^{(R)}) \end{array}$$

  las que se denominan réplicas Monte Carlo.

- Para cada muestra i computamos $y_i = T(X_1^{(i)}, \ldots, X_n^{(i)})$, $i = 1, \ldots, R$.

Entonces $G_R^*(y)$ es una aproximación a $G(y)$ y, $\frac{1}{R}\sum_{i=1}^{R} T(X_1^{(i)}, \ldots, X_n^{(i)})$ aproxima a $\mathrm{E}(T(X_1, \ldots, X_n))$.

El anterior es un típico ejemplo de aplicación de la simulación Monte Carlo.

2. Estimación puntual

2.1. Introducción

En general, en un problema de investigación se plantea un diseño, concebido como planeamiento y desarrollo de una investigación, una instancia de captura (observación o producción) de las observaciones y un tercer momento denominado inferencia el cual es concebido- en términos esquemáticos- como el pasaje de la muestra a la población. La estadística descriptiva, como ya se estudió, juega un rol decisivo en la caracterización de las observaciones, en la determinación de su estructura.

En la instancia del diseño el muestreo tiene un rol decisivo. Aunque hay diversos tipos de muestreos en esta asignatura sólo utilizaremos el muestreo aleatorio simple, definido en la unidad anterior, en el que las unidades de análisis son recolectadas independientemente unas de otras y todas tienen la misma probabilidad de aparición.

Ya abordamos los conceptos de *población* y *muestra*, *parámetro*, *estadístico* y *estimador*[1]. Interesa estudiar propiedades de los estimadores y contar con métodos de construcción de los mismos.

Como parte de la estadística descriptiva estudiamos diferentes estadísticos como los siguientes:

- la media, que mide el valor promedio de una muestra,
- la mediana, midiendo el centro de la muestra ordenada,
- los cuantiles y cuartiles, mostrando dónde se ubican ciertas porciones de una muestra (ordenada) y
- la varianza, el desvío estándar y el rango intercuartílico, que miden la variabilidad, dispersión y amplitud de en una muestra.

[1] En esta unidad seguiremos principalmente el libro de Baron (2014)

Consideremos, como antes, una muestra aleatoria $(X_1, \ldots, X_n)$ con $X_1 \sim F$ siendo F una función de distribución. En general, en el contexto inferencial, el objetivo es "estimar" alguna característica desconocida de la distribución F como por ejemplo un parámetro θ, perteneciente a cierto espacio paramétrico Θ. A veces interesa también estimar funciones de los parámetros.

Ejemplo 1.

a) Si X tiene una distribución Gamma de parámetros α, λ su densidad es

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} \mathrm{e}^{-\lambda x}, \ \forall x > 0,$$

(λ, α) es el parámetro de la distribución y $\Theta = I\!R^+ \times I\!R^+$ es el espacio paramétrico. Notar que $\mathrm{E}(X) = \alpha/\lambda$ es una característica asociada a la distribución pero no es un parámetro.

b) Si $X \sim \mathrm{N}(\mu, \sigma^2)$ entonces (μ, σ^2) es el parámetro de la distribución y $\Theta = I\!R \times I\!R^+$. □

Observación 1. *En a) λ y α son funciones del parámetro (λ, α) al igual que en b) μ, σ^2 y σ son funciones de (μ, σ^2).*

Cada uno de los estadísticos listados arriba describen características de la muestra de n observaciones y también pueden utilizarse para dar estimaciones de características de F (población).

Definición 1. *Un estimador puntual de un parámetro θ es una función $\widehat{\theta}_n$ de la muestra aleatoria $(X_1, \ldots, X_n)$. Una estimación es el valor que asume $\widehat{\theta}_n$ para una realización de la muestra; i.e. $\theta_n(X_1(\omega), \ldots, X_n(\omega))$ para un elemento particular $\omega \in \Omega$.*

Notar que cualquier estadístico es también un estimador y la diferencia entre el significado de estadístico y el de estimador la da el contexto de utilización.

Introducimos a continuación dos conceptos muy importantes.

Definición 2. *El error cuadrático medio (ECM) de un estimador $\sigma(\widehat{\theta}_n)$ se define como*

$$\text{ECM}[\widehat{\theta}_n] = E[(\widehat{\theta}_n - \theta]^2.$$

Y, el error estándar del estimadorse define como su desvío estándar

$$\sigma(\widehat{\theta}_n) = \sqrt{\text{VAR}(\widehat{\theta}_n)}.$$

El error estándar de un estimador mide su variabilidad y por ende su precisión y confiabilidad. Este error nos informa cuánto los estimadores pueden variar entre diferentes muestras. Los estimadores con buenos desempeños son insesgados y con errores estándares bajos.

Ejemplo 2. *ECM y error estándar de la media muestral*

Si $\theta = \mu$ representa la esperanza de una variable aleatoria X con varianza finita $\text{VAR}(X) = \sigma^2$ entonces se prueba fácilmente que $\text{ECM}[\overline{X}_n] = \text{VAR}(\overline{X}_n) = \sigma^2/n$ y que $\sigma(\overline{X}_n) = \sigma/\sqrt{n}$.

Tener en cuenta que no estamos asumiendo ninguna distribución particular para la variable X.

□

Observar que el ECM y el error estándar de un estimador puede depender del parámetro a estimar o de otros parámetros desconocidos. En el ejemplo anterior está dependiendo de σ, con lo cual una estimación del error estándar se obtiene reemplazando ("plug-in") el parámetro por su valor estimado:

$$\widehat{\sigma}_{\overline{X}_n} = \widehat{\sigma}/\sqrt{n} = s_n/\sqrt{n},$$

donde s_n es el desvío estándar muestral.

2.2. Propiedades de los estimadores

Los estimadores al ser funciones de una muestra aleatoria, tienen una distribución en probabilidad. Para que un estimador tenga un desempeño adecuado su distribución debe tener propiedades como las de insesgamiento, normalidad asintótica y consistencia, definidas a continuación.

2.2.1. Insesgamiento

Dada una característica θ asociada a la distribución F, a un estimador lo denotaremos con $\widehat{\theta}_n$. Hay que notar que un estimador es un sucesión (de variables aleatorias) y a veces, para simplificar, omitimos la dependencia de n. Por un abuso de lenguaje diremos estimador en lugar de sucesión de estimadores.

Definición 3. *Un estimador $\widehat{\theta}_n$ se dice insesgado para un parámetro θ si, para todo n, se cumple que $\forall\theta \in \Theta:\ E(\widehat{\theta}_n) = \theta$. El sesgo de $\widehat{\theta}_n$, dado n, se define como $\text{SES}(\widehat{\theta}_n) = E(\widehat{\theta}_n - \theta)$.*

Notar que, dado un estimador, el ECM se descompone del siguiente modo:

$$\text{ECM}[\widehat{\theta}_n] = [\text{SES}(\widehat{\theta}_n)]^2 + \text{VAR}[\widehat{\theta}_n].$$

El siguiente ejemplo se trabaja en el práctico.

Ejemplo 3. *Estimadores insesgados de la media y la varianza de una variable aleatoria X*

Si X es una variable aleatoria con $\text{E}(X) = \mu$ y $\text{VAR}(X) = \sigma^2$ entonces la media muestral $\overline{X}_n$ y la varianza muestral s_n^2 son estimadores insesgados para μ y σ^2, respectivamente. □

Ejemplo 4. *Distribución Normal*

Si $X \sim \mathrm{N}(\mu, \sigma^2)$ entonces, como una consecuencia del ejemplo anterior, $\overline{X}_n$ es un estimador insesgado para μ y s_n^2 lo es para σ^2. Además la mediana muestral M_n es también un estimador insesgado para la "posición" μ.

□

Observación 2. *Insesgamiento y simulación Monte Carlo.*

Asumamos que $\widehat{\theta}_n$ es un estimador insesgado (para estimar θ). Generemos R muestras de tamaño n de X y para cada una de ellas computemos $\widehat{\theta}_n^{(j)}$, $j = 1, \ldots, R$. Por lo estudiado en la Sección 1, el insesgamiento implica que la media muestral de $\widehat{\theta}_n^{(j)}$, $j = 1, \ldots, R$ debe estar cercana a θ si R es grande, para cualquier tamaño de muestra n. □

En el contexto de esta observación introduzcamos el siguiente ejemplo.

Ejemplo 5. *Simulación Monte Carlo para el estimador mediana basado en la distribución gamma.*

Sea $X \sim \Gamma(3, 1/9)$, denotemos $\theta = \mathrm{E}(X) = \alpha/\lambda = 3/(\frac{1}{9}) = 27$ y consideremos la mediana muestral $\tilde{x}_n^{(j)}$ como un estimador de θ.

A los fines de evaluar el desempeño del estimador realicemos un experimento de simulación generando $R = 500$ muestras de tamaño $n = 50$ de $\tilde{x}_{50}$. Para $j = 1 \ldots, R$, denotemos con $\tilde{x}_{50}^{(j)}$ al estimador mediana muestral basado en la j-ésima muestra. A partir de las siguientes líneas[2]

```
Cómputo de R=500 medias y medianas basadas en muestras de tamaño n=50
R=500                           # Número de réplicas
n=50                           # Tamaño de la muestra
vecmedianas <- rep(0,R)         # Vector de ceros de longitud R
set.seed(1)                     # Semilla
for(i in 1:R){
vecmedianas[i] <- median( rgamma(n, shape=3, scale=9))# Cómputo de la mediana
}
mean(vecmedianas)
24.39684
```

obtenemos,

$$\frac{1}{500}\sum_{j=1}^{500} \tilde{x}_{50}^{(j)} = 24.39$$

[2]En R el parámetro de escala es $1/\lambda$.

lo que sugiere que la mediana muestral, a diferencia de la media muestral, no es un estimador insesgado para $\theta = \mathrm{E}(X) = 27$. □

2.2.2. Consistencia

Otra propiedad deseable para un estimador es la consistencia, definida a continuación.

Definición 4. *Un estimador $\widehat{\theta}_n$ se dice consistente para un parámetro θ si*

$$\forall \epsilon > 0 : P\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right) \to 0 \text{ cuando } n \to \infty.$$

Una notación equivalente es la siguiente

$$\widehat{\theta}_n \xrightarrow{p} \theta, \text{ cuando } n \to \infty$$

y se lee "que la sucesión $\widehat{\theta}_n$ converja en probabilidad a θ cuando n tiende a infinito".

Bajo ciertas condiciones la media muestral es un estimador consistente para $\theta = \mathrm{E}(X)$, tal como lo afirma el siguiente teorema, denominado Ley Débil de los Grandes Números.

Teorema 1. *Si $X_1, X_2, \ldots$ es una sucesión de variables aleatorias iid con $E(X_1) = \theta$ y $\mathrm{VAR}(X_1) = \sigma^2$ entonces $\overline{X}_n$ es consistente para θ.*

Demostración. Ver Apéndice.

En la Figura 1 se grafica, para la variable aleatoria Bernoulli de parámetro $p = 0.5$ que modela el lanzamiento de una moneda, $\overline{X}_n$ versus n a lo largo de 10^5 lanzamientos.

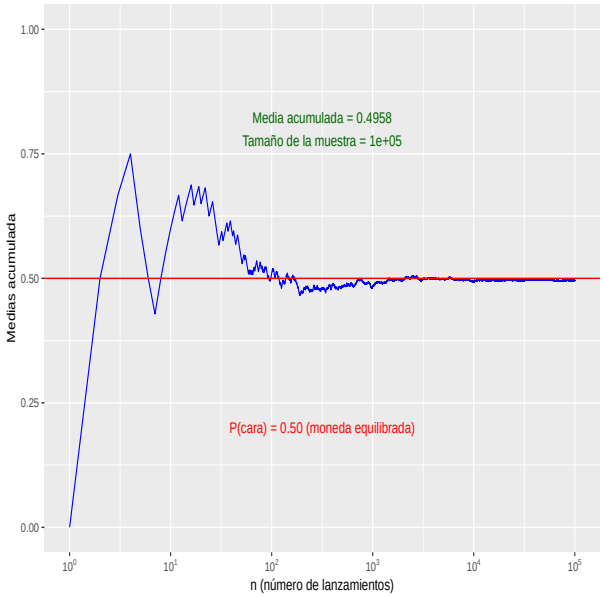


Figura 1: Medias muestrales (del número de caras) acumuladas en sucesivos lanzamientos.

Observación 3. *Sin la hipótesis de la finitud de la varianza de X el teorema anterior es falso.*

Ejemplo 6. *Estimadores de la escala de la distribución Normal*

Si $X \sim \mathrm{N}(\mu, \sigma^2)$, se puede probar que s_n- basado en una muestra aleatoria de X, $X_1, \ldots, X_n$- es un estimador consistente para estimar la escala σ.

Definamos ahora

$$Y_n = \text{mediana}\{|X_i - M_n| : 1 \leq i \leq n\}.$$

Es posible demostrar que

$$Y_n \xrightarrow{p} \sigma B, \text{ cuando } n \to \infty$$

donde B es la mediana de la función de distribución $G(u) = 2\Phi(u) - 1$. Notar que B es solución de

la ecuación $G(B) = 1/2$, es decir $B = \Phi^{-1}(0.75) = 0.6745$. De este modo,

$$\begin{aligned} \text{mad}_n &= 1.4826 Y_n \\ &= \frac{1}{B} Y_n \xrightarrow{p} \frac{1}{B}\sigma B = \sigma, \text{ cuando } n \to \infty, \end{aligned}$$

probando que mad_n es consistente para estimar σ. En "Notas de Estadística" de Boente y Yohahi, Cap. 7, pp. 18 y ss. se da una demostración rigurosa de cada paso.

De modo análogo se muestra que $\text{d}_{I,n}/1.349$, donde $\text{d}_{I,n}$ es la distancia intercuartil, es también consistente para estimar la escala σ.

□

2.2.3. Normalidad asintótica

Damos finalmente la última propiedad de los estimadores, que consideramos en este curso.

Definición 5. *Un estimador $\widehat{\theta}_n$ de θ se dice asintóticamente normal si*

$$\forall x \in I\!R : P\left(\sqrt{n}\left(\frac{\widehat{\theta}_n - \theta}{\sqrt{V(\theta)}}\right) \leq x\right) \longrightarrow \Phi(x), \text{ si } n \longrightarrow \infty$$

donde Φ es la función de distribución de la $N(0,1)$ y $V(\theta)$ es una constante denominada la varianza asintótica del estimador.

La definición de normalidad asintótica de una sucesión de estimadores afirma que si n es grande entonces

la función de distribución de $\sqrt{n}\left(\widehat{\theta}_n - \theta\right)$ es aproximadamente $\text{N}(0, V(\theta))$.

Ejemplo 7. *El Teorema Central del Límite sostiene que si* $(X_1, \ldots, X_n)$ *es una muestra aleatoria con* $E(X_1) = \theta$ *y* $\text{VAR}(X_1) = \sigma^2$ *finitas entonces* $\overline{X}_n$ *es asintóticamente normal con varianza asintótica* σ^2.

Mostrar que un estimador tiene distribución asintótica normal en general no es simple. A continuación enunciamos una proposición que establece, bajo ciertas condiciones, normalidad asintótica de la mediana muestral. La demostración está fuera del alcance de los contenidos del curso.

Proposición 1. *Sea* $X_1, X_2, \ldots$ *una sucesión de variables aleatorias iid con función de distribución común* F *con* θ *tal que* $F(\theta) = \frac{1}{2}$. *Asumamos que* $F'(\theta)$ *existe y es positiva. Entonces la mediana muestral* M_n *es asintóticamente normal con*

$$V(\theta) = \frac{1}{4\left[F'(\theta)\right]^2}.$$

Ejemplo 8. *Estudio de la distribución asintótica de la mediana muestral si* $X \sim$ *exponencial*(λ) .

En el panel de la izquierda de la Figura 8 se halla el gráfico de la densidad de $X \sim$ exponencial(3). Sea M_n la mediana muestral. De acuerdo a la proposicion anterior si $\theta = F^{-1}(1/2)$ entonces M_n posee, para n grande, una distribución aproximadamente

$$N(\theta, V(\theta)/n).$$

Observar que, si $x > 0$, $F(x) = 1 - \mathrm{e}^{-3x}$, $F'(x) = 3\mathrm{e}^{-3x}$ y por lo tanto

$$\theta = -\log(1/2)/3 \text{ y } V(\theta) = \frac{1}{4[(3/2)]^2} = 1/9 \approx 0.1111111.$$

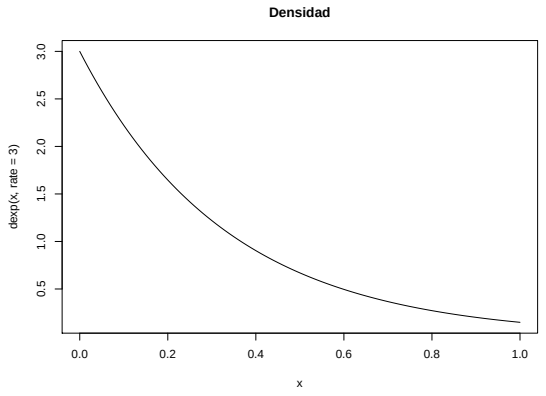
A continuación estimaremos θ y la varianza asintótica del estimador a través de un experimento de simulación Monte Carlo. Las siguientes líneas generan $N = 1000$ muestras de tamaño $n = 50$ de X y, el vector `vecmediana` contiene los 1000 valores de M_{50}.

```
set.seed(1234);
lambda<-3;
vecmediana<- c()
for(i in 1:1000) {
vecmediana[i] <-median(  rexp(n=50, rate=lambda)    );
}
```

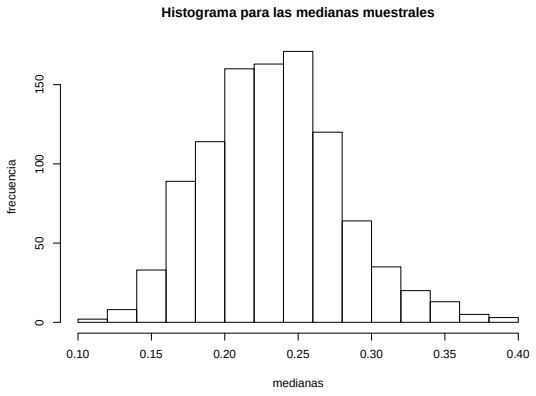
Calculemos en R la media y el desvío de las medianas muestrales

```
mean(vecmediana)
[1] 0.2336642
sd(vecmediana)^2
[1] 0.002140756
```

Estos valores obtenidos por simulación Monte Carlo aproximan muy bien a los valores exactos $\theta = -\log(1/2)/3$ y a $V(\theta)/50 = \dfrac{1}{9 \cdot 50} \approx 0.0022222$ (o sea $V(\theta) \approx 0.11111$). □



a) Densidad de una $X \sim$ exponencial(3)



b) Histograma para 1000 valores de M_{50}

Figura 2: Gráficos correspondientes al Ejemplo 8.

La varianza asintótica $V(\theta)$ del estimador no necesariamente coincide con la varianza del estimador $\text{VAR}\left(\widehat{\theta}_n\right)$. $V(\theta)$ es la varianza de la distribución límite de la sequencia $\left\{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)\right\}_{n\geq 1}$ y no depende de n mientras que la varianza de un estimador es la varianza de cada término $\widehat{\theta}_n$ de la sucesión de estimadores.

Métodos de generación de estimadores

Analizamos propiedades que se requieren para que un estimador tenga un despempeño adecuado. Ahora bien, ¿cómo se generan los estimadores? Hay diversos métodos de generación de familias de estimadores. Aqui veremos sólo un método y en clase se abordarán otros.

Definición 6. *Sea X una variable aleatoria con función de probabilidad de masa o de densidad $f(x;\theta)$, con $\theta \in \Theta$. Consideremos un valor observado $(x_1, \ldots, x_n)$ de la muestra. Definimos la verosimilitud de la muestra observada como la función $L : \Theta \to [0,1]$ definida por*

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Observar que $L(\theta)$ es igual a la función de probabilidad de masa conjunta o densidad conjunta del vector aleatorio $(X_1, \ldots, X_n)$ evaluada en la muestra observada. Cuando f es una función de probabilidad de masa, $L(\theta)$ puede interpretarse como un valor de probabilidad.

Definición 7. *En el contexto de la definición 6, el estimador de máxima verosimilitud (emv) $\widehat{\theta}_n$ se define como el valor de θ (si existe) que maximiza $L(\theta)$.*

$\widehat{\theta}_n$ es una función del valor observado $(x_1, \ldots, x_n)$ y es una variable aleatoria cuando se lo considera como función de la muestra aleatoria $(X_1, \ldots, X_n)$.

Notar que dado que la función logaritmo $\log(\cdot)$ es creciente entonces hallar el máximo de $L(\theta)$ es equivalente a encontrar el máximo de la función de log-verosimilitud definida como

$$l(\theta) = \log L(\theta).$$

Problema 1.

a) Si X tiene una distribución de Poisson de parámetro λ entonces el emv del parámetro es $\overline{x}_n$.

b) Si $X \sim$ exponencial(λ) entonces el emv de λ es

$$\widehat{\lambda}_n = \frac{1}{\overline{x}_n}.$$

c) Si $X \sim$ uniforme$[0, b]$ entonces el emv de b es

$$\widehat{b}_n = \text{máx}\{x_1, \ldots, x_n\}.$$

d) Si $X \sim N(\mu, \sigma^2)$ y $\boldsymbol{\theta} = (\mu, \sigma^2)$, el emv de $\boldsymbol{\theta}$ es

$$\widehat{\boldsymbol{\theta}} = (\overline{x}_n, \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2).$$

Solución.

a) Para demostrar esta afirmación consideremos una observación $(x_1, \ldots, x_n)$ de una muestra aleatoria de X.

$$\begin{aligned} l(\lambda) &= \log(\prod_{i=1}^{n} f(x_i; \lambda)) \\ &= \log\left[\prod_{i=1}^{n} \lambda^{x_i} \mathrm{e}^{-\lambda}/x_i!\right] \\ &= \log(\lambda)\sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n}\log(x_i!) \end{aligned}$$

Derivando e igualando a cero

$$\frac{d}{d\lambda} l(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$

hallamos que existe un único un punto crítico, $\sum_{i=1}^{n} x_i/n$, en el cual la derivada de segundo orden de $l(\lambda)$ es negativa. Además cuando $\lambda \to \infty$ o $\lambda \to 0$, $l(\lambda) \to -\infty$ y en consecuencia $\widehat{\lambda} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ es el emv de λ.

c) La función de verosimilitud es

$$\begin{aligned} L(b) &= \prod_{i=1}^{n} f(x_i, b) = \prod_{i=1}^{n} b^{-1} \mathrm{I}_{[0,b]}(x_i) \\ &= \begin{cases} b^{-n} & si \quad 0 \leq x_i \leq b, \, \forall i = 1, \ldots, n \\ \\ 0 & cc \end{cases} \\ &= \begin{cases} b^{-n} & si \quad b \geq x_{(n)} = \text{máx}\{x_1, \ldots, x_n\} \\ \\ 0 & cc \end{cases} \end{aligned}$$

y su gráfica se encuentra en la Figura 3. Claramente $\widehat{b} = \text{máx}\{x_1, \ldots, x_n\}$.

Las afirmación b) del problema se deja como ejercicio y la demostración de d) requiere de contenidos que están fuera del alcance de este curso. □

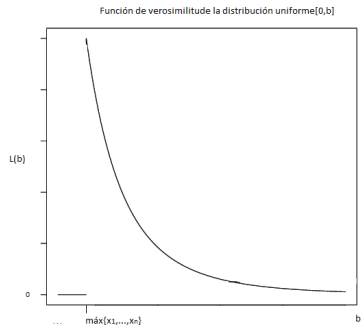


Figura 3: Gráfica de la función de verosimilitud $L(b)$ cuando $X \sim$ uniforme$[0, b]$

3. Estimación por intervalos de confianza

3.1. Introducción

Al informar un valor de estimador $\widehat{\theta}_n$ de un parámetro poblacional θ sabemos que es probable que

$$\widehat{\theta}_n \neq \theta$$

debido al error de muestreo.

Nos preguntamos entonces ¿hasta dónde podemos confiar en el valor estimado del parámetro? Es importante precisar mejor la pregunta y obtener una mejor respuesta. La siguiente definición permite hacerlo, proponiendo la construcción de un intervalo tal que contenga al parámetro con cierta probabilidad.

Definición 8. *Sea $\alpha \in (0,1)$. Un intervalo de confianza de nivel $(1-\alpha)\cdot 100\,\%$ para el parámetro θ es un intervalo $[a,b]$ tal que*

$$P\left(a \leq \theta \leq b\right) = 1-\alpha.$$

A $(1-\alpha)$ se le denomina nivel de confianza o probabilidad de cubrimiento.

Necesitamos efectuar varias observaciones. En primer lugar, dado que θ no es una variable aleatoria si los límites del intervalo $[a,b]$ no son aleatorios entonces $P\left(a \leq \theta \leq b\right) = 0$ ó 1 según $\theta \notin [a,b]$ o $\theta \in [a,b]$. Pero la definición impone que esta probabilidad sea $0 \neq 1-\alpha \neq 1$ ya que $\alpha \in (0,1)$. Con lo cual, para que el intervalo exista, a o b o ambos deben ser variables aleatorias.

En la Figura 4 se ilustra, parcialmente, el significado de la definición anterior. En la intersección de los ejes se dibuja el parámetro θ. Se seleccionan diez muestras independientes y del total de intervalos, uno por muestra, 8 intersecan el eje de ordenadas indicando que que contienen al parámetro y dos de ellos no contienen a θ. Para este experimento ¿cuál podría ser el valor del nivel de confianza?.

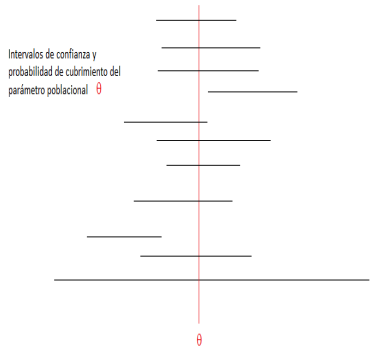


Figura 4: Representación de la probabilidad de cubrimiento

Nos preguntamos, ¿los dos intervalos que no contienen al parámetro fueron el resultado de datos erróneos? La respuesta es no, es consecuencia del carácter aleatorio del fenómeno.

Es importante tener en cuenta que una afirmación como la siguiente es **falsa** : "computé un intervalo de confianza del 90 %, hallé que el intervalo es el $[1.7, 6.8]$ y por la tanto la probabilidad de que este intervalo contenga al parámetro θ es de 0.9".

Y es falsa ya que, dado que el parámetro es una constante, la probabilidad de que $[1.7, 6.8]$ contenga a θ es 0 o 1 como se enfatizó arriba.

La afirmación **correcta** es "$[1.7, 6.8]$ contiene al parámetro θ con una *confianza* del 90 %" y la interpretamos del siguiente modo[3]: "si un día realizamos el experimento un número grande N de veces y por cada realización computamos el intervalo obtendríamos aproximadamente $0.9 \cdot N$ intervalos que contienen a θ y los restantes no. Nuestro intervalo particular $[1.7, 6.8]$ forma parte de los N intervalos, con lo cual tenemos una alta chance que forme parte del grupo de $0.9 \cdot N$ intervalos que contiene a θ".

3.2. Método para hallar intervalos de confianza

Iniciaremos con una exposición de un problema muy simplificado pero que tendrá varias aplicaciones de interés. Asumamos que contamos con un parámetro θ a ser estimado y un estimador $\widehat{\theta}_n$, que sea insesgado y que posea distribución normal.

Producimos la siguiente estandarización

$$\frac{\widehat{\theta}_n - \mathrm{E}(\widehat{\theta}_n)}{\sigma(\widehat{\theta}_n)} = \frac{\widehat{\theta}_n - \theta}{\sigma(\widehat{\theta}_n)}$$

[3]en clase debatimos sobre una discusión nteresante que da Wasserman, L. (2010). All of Statistics: A Concise Course in Statistical Inference. Springer.

donde $\mathrm{E}(\widehat{\theta}_n) = \theta$ por ser el estimador insesgado y $\sigma(\widehat{\theta}_n)$ es el error estándar del estimador. Luego

$$P\left(-z_{\alpha/2} \leq \frac{\widehat{\theta}_n - \theta}{\sigma(\widehat{\theta}_n)} \leq z_{\alpha/2}\right) = 1 - \alpha \tag{1}$$

donde

$$-z_{\alpha/2} = q_{\alpha/2} \text{ y } z_{\alpha/2} = q_{1-\alpha/2}$$

siendo $q_{\alpha/2}$ y $q_{1-\alpha/2}$ los cuantiles de la normal estándar como se indica en la Figura 5.

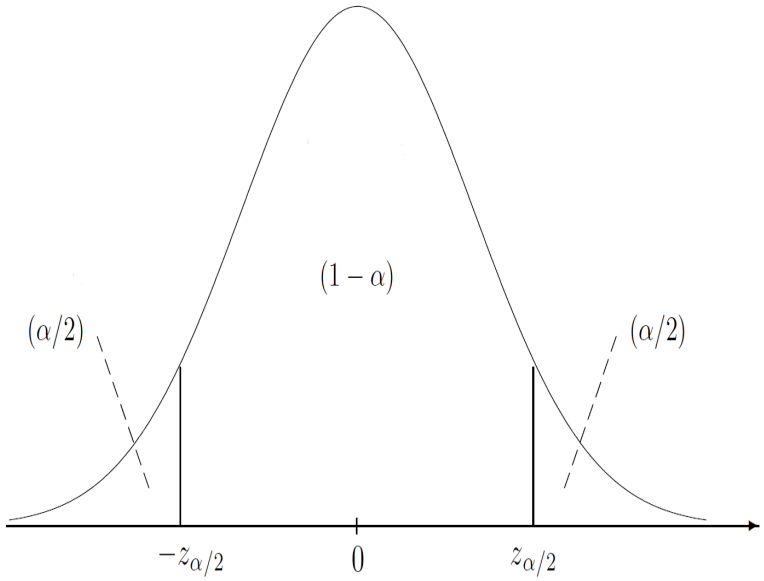


Figura 5: Representación de la ecuación (1) para un intervalo de confianza de nivel $1 - \alpha$.

Obtenemos así la siguiente expresión equivalente a (1)

$$P\left(\widehat{\theta}_n - z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n) \leq \theta \leq \widehat{\theta}_n + z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n)\right) = 1 - \alpha. \tag{2}$$

Arribamos entonces a la definición de intervalo de confianza dado a continuación.

Definición 9. *Sea $\widehat{\theta}_n$ un estimador insesgado de θ con distribución normal. Definimos el intervalo de confianza de nivel $1-\alpha$ como $[a,b]$ siendo*

$$a = \widehat{\theta}_n - z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n) \text{ y } b = \widehat{\theta}_n + z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n), \tag{3}$$

con $-z_{\alpha/2} = q_{\alpha/2}$ y $z_{\alpha/2} = q_{1-\alpha/2}$ los cuantiles de la normal estándar.

De aquí en más $z_{\alpha/2}$ denotará el cuantil que satisface

$$P(Z \leq z_{\alpha/2}) = 1 - \alpha/2 \tag{4}$$

donde $Z \sim \mathrm{N}(0,1)$.

Observación 4. *Cuando no conocemos la distribución exacta del estimador pero es asintóticamente normal entonces definimos el intervalo de confianza como en la definición anterior teniendo en cuanto que el nivel $1-\alpha$ se alcanza aproximadamente.*

Veamos ejemplos de construcción de intervalos de confianza en diferentes contextos y supuestos.

3.2.1. Intervalo de confianza para la media de una población con varianza conocida

Sea X una variable aleatoria con

$$\theta = \mu = \mathrm{E}(X),$$

asumimos que $\mathrm{VAR}(X) = \sigma^2$ es conocida y, queremos hallar un intervalo de confianza para θ, para $1-\alpha$ fijo.

Si utilizamos como estimador del parámetro a $\overline{X}_n$ tendremos entonces las siguientes dos posibilidades.

- Si la muestra aleatoria $(X_1, \ldots, X_n)$ proviene de una $X \sim \mathrm{N}(\theta, \sigma^2)$ se tiene que $\overline{X}_n \sim \mathrm{N}(\theta, \sigma^2/n)$ y así el intervalo de confianza de nivel $1-\alpha$ se construye como en (3) de la Definición 9.

- En cambio si la muestra aleatoria $(X_1, \ldots, X_n)$ no proviene de una variable aleatoria X normal, por el Teorema Central del Límite $\overline{X}_n$ posee, para n grande, una distribución "aproximadamente normal". Luego el intervalo de confianza es el dado por (3) de la Definición 9 y alcanza un nivel de confianza "aproximado" de $1-\alpha$.

Problema 2. *Construir un intervalo del* $95\,\%$ *de confianza para la media de una población basado en las mediciones*

$$2.5, 7.4, 8.0, 4.5, 7.4, 9.2$$

y asumiendo que el desvío poblacional es σ *=2.2.*

Solución. $n = 6$, $\overline{X}_n = 6.50$. Para alcanzar un nivel del

$$95\,\% \text{ ó } 1-\alpha = 0.95,$$

debe ser $\alpha = 0.05$, $\alpha/2 = 0.025$ y en consecuencia $-z_{0.025} = q_{0.025} = -1.96$ y $z_{0.025} = q_{0.975} = 1.96$. Luego, el intervalo de confianza es

$$\overline{X}_n \pm z_{0.025}\frac{\sigma}{\sqrt{n}} = 6.50 \pm (1.96)\frac{2.2}{\sqrt{6}} = 6.50 \pm 1.76 \text{ ó } [4.74, 8.26]$$

Concluimos entonces que

Con una confianza del 95 % el intervalo $[4.74, 8.26]$ contiene a la media poblacional.

□

Observación 5. *¿Qué ocurriría si no conocemos la distribución, queremos estimar la esperanza pero no conocemos la varianza? Observemos que el inconveniente es que el error estándar del estimador depende del parámetro σ^2. Una alternativa es que si n es grande podemos reemplazar $\sigma(\widehat{\theta}_n)$ por $\widehat{\sigma}(\widehat{\theta}_n)$ y utilizar el intervalo dado en* (3) *de la Definición 9 sabiendo que el nivel alcanzado es aproximado.*

3.2.2. Intervalo de confianza para la media de una población normal con varianza desconocida

Supongamos que queremos estimar $\theta = \mu = \mathrm{E}(X)$ siendo $X \sim \mathrm{N}(\theta, \sigma^2)$ pero desconocemos σ^2. Es posible mostrar que

$$\frac{\overline{X}_n - \theta}{s_n/\sqrt{n}} \sim t_{n-1}$$

donde t_{n-1} es una distribución t de Student con parámetro $n-1$ denominado los grados de libertad. Esta distribución es simétrica respecto de cero, con lo cual

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\overline{X}_n - \theta}{s_n/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1 - \alpha$$

donde $-t_{n-1,\alpha/2}$ es el cuantil $\alpha/2$ de la distribución t_{n-1} (es decir, si $T \sim t_{n-1}$ entonces $P(T \leq -t_{n-1,\alpha/2}) = \alpha/2$)

Problema 3. *En el contexto del Problema 2 supongamos que no se pueda asumir conocido el desvío de la población. En este caso ¿cuál es el el intervalo de confianza para la media poblacional?*

Solución. La forma del nuevo intervalo es

$$\overline{X}_n \pm t_{n-1,0.025}\frac{s_n}{\sqrt{n.}}$$

En R calculamos el desvío de la muestra s_n y el cuantil $t_{6-1,0.025}$

```
mediciones<- c(2.5, 7.4, 8, 4.5, 7.4, 9.2)

sd(mediciones)

[1] 2.496397

> qt(0.975, df=6-1)

[1] 2.570582
```

Luego, el intervalo de confianza es

$$6.50 \pm (2.57)\frac{2.5}{\sqrt{6}} = 6.50 \pm 2.62 \text{ ó } [3.88, 9.12].$$

Notar que la estimación reemplazando el desvío poblacional por el muestral se tradujo en una pérdida de precisión, es decir en un intervalo de mayor longitud. □

3.2.3. Selección del tamaño de la muestra

Observar que el intervalo de confianza (3)

$$[\widehat{\theta}_n - z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n), \widehat{\theta}_n + z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n)]$$

tiene

$$\text{centro} = \widehat{\theta}_n \text{ y longitud} = l = 2 \cdot z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n).$$

donde $z_{\alpha/2}$ satisface (4).

Cuanto más pequeña se la longitud del intervalo más precisa es la estimación. Una forma de aumentar la precisión es bajar el nivel de confianza pero esto carece de sentido ya que queremos estimar al parámetro con un alto nivel de confianza. Una forma es, mantener un nivel alto de confianza y variar el n. De este modo nos podemos preguntar ¿cuál grande debe ser n tal que el alcance la precisión deseada?

$$\text{¿Cuál debe ser } n \text{ tal que } z_{\alpha/2} \cdot \sigma(\widehat{\theta}_n) \leq \Delta?$$

donde Δ es un valor prefijado y se le llama *margen de error*.

Así por ejemplo si deseamos estimar a través de un intervalo de confianza de nivel $1-\alpha$ la media de una población con varianza conocida y, con un margen de error de a lo sumo Δ deberíamos elegir

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{\Delta}\right)^2.$$

3.2.4. Intervalo de confianza para la diferencia de medias de dos poblaciones normales

Supongamos que $(X_1, \ldots, X_n)$ es una muestra aleatoria de una $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$, $(Y_1, \ldots, Y_M)$ es una muestra aleatoria de una $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, ambas muestras son independientes y las varianzas se asumen conocidas. Nuestro problema es estimar, a través de un intervalo de confianza, la diferencia

de las medias poblacionales, es decir

$$\theta = \mu_X - \mu_Y$$

A continuación presentamos un ejemplo a ser resuelto en el Práctico 6.

Ejemplo 9. *Un asesor técnico de una productora de hardware quiere evaluar la efectividad de una actualización corriendo cierto proceso* 50 *veces antes de la actualización y* 50 *veces luego. Basados sobre estos datos, el tiempo promedio es de* 8.5 *minutos antes de la actualización mientras que posterior a la misma el tiempo medio es de* 7.2. *Históricamente, el desvío estándar ha sido de* 1.8 *minutos y presumiblemente no ha cambiado. El objetivo del asesor es construir un intervalo del* 90 % *de confianza mostrando cuánto se ha reducido el tiempo medio debido a la nueva actualización (se asume normalidad de las variables de las que provienen los datos e independencia de las muestras correspondientes a "antes" y "después").*

Para construir el intervalo realicemos las siguientes consideraciones:

- Proponemos como estimador de θ a $\widehat{\theta} = \overline{X}_n - \overline{Y}_m$.

- Observar

$$\overline{X}_n - \overline{Y}_m \sim \mathrm{N}(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$$

  y, por ende, el desvío estándar del estimador es

$$\sigma(\widehat{\theta}) = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

de donde, el intervalo de confianza de nivel $1-\alpha$ para $\theta = \mu_X - \mu_Y$ es

$$\overline{X}_n - \overline{Y}_m \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

3.2.5. Intervalo de confianza para proporciones

Problema 4. *Supongamos que una organización en defensa del consumidor quiere llevar a cabo una encuesta para estimar la proporción p de consumidores que manifiestan estar conformes con la compra de la última versión de un reproductor digital de audio. La organización tiene como una primera pregunta ¿cuán grande debe ser el tamaño de muestra de tal modo de estimar p con un margen de error del $2\,\%$ y con un $95\,\%$ de confianza?*

Para arribar a una respuesta a esta necesitamos dar un breve recorrido teórico.

Sean $X_1, \ldots, X_n$ variables aleatorias iid con distribución Bernoulli de parámetro p y asumamos que queremos estimar el parámetro. El emv de p, $\widehat{p}_n$, es la media muestral $\overline{X}_n$ de las observaciones, la cual al mismo tiempo es la proporción muestral (de éxitos) en las n observaciones. Así entonces,

- Proponemos como estimador insesgado de la proporción poblacional p a la proporción muestral $\widehat{p}_n$

- Por el Teorema Central del Límite, para n grande, $\widehat{p}_n$ tiene una distribución aproximadamente normal con
$$\mathrm{E}(\widehat{p}_n) = p \text{ y } \textsc{var}(\widehat{p}_n) = \frac{p(1-p)}{n}.$$

- El desvío estándar del estimador es $\sigma(\widehat{p}_n) = \sqrt{\frac{p(1-p)}{n}}$, con lo cual una estimación del mismo

es

$$\widehat{\sigma}(\widehat{p}_n) = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

En consecuencia, un intervalo de confianza con nivel aproximado $1-\alpha$ es

$$\widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}} \tag{5}$$

donde $z_{\alpha/2}$ satisface (4).

Es importante realizar aquí una observación. Por el teorema central del límite sabemos que $\dfrac{\widehat{p}_n - p}{\sigma(\widehat{p}_n)}$ tiene una distribución aproximadamente $N(0,1)$ si n es grande. Es posible mostrar que si reemplazamos $\sigma(\widehat{p}_n)$ por $\widehat{\sigma}(\widehat{p}_n)$ todavía seguimos teniendo la misma distribución límite y por lo tanto es lícito utilizar $z_{\alpha/2}$ en (5). Para una demostración detallada de este hecho ver Yohai (2006, p. 254).

Revisita al Problema 4 El objetivo es construir un intervalo de confianza de nivel $1-0.05 = 0.95$ con un margen de error de $\Delta = 0.02$, para estimar la proporción poblacional p de consumidores que manifiestan estar conformes . De este modo queremos hallar n tal que

$$z_{0,05/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \Delta = 0.02$$

donde $z_{0.025}$ viene determinado por el nivel de confianza 0.95. Luego, $P(Z \leq z_{\alpha/2}) = 0.975$ determina que $z_{0.025} = 1.96$.

¿cómo logramos hallar los valores de n que satisfagan la condición $1.96\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} \leq \Delta = 0.02$ si no podemos determinar de antemano el valor $\hat{p}$? La propuesta es acotar la función $f(p) = p(1-p)$ por su valor máximo en el intervalo $[0,1]$ el cual se alcanza[4] en $p = 1/2 = 0.5$. Así, si hallamos el

[4]con un análisis como el realizado en Cálculo para hallar el máximo de una función en un intervalo cerrado

mínimo valor de n tal que

$$1.96\sqrt{\frac{0.5(1-0.5)}{n}} \leq \Delta = 0.02, \tag{6}$$

como $1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 1.96\sqrt{\frac{0.5(1-0.5)}{n}}$ tendremos resuelto el problema. La desigualdad (6) es equivalente a

$$n \geq 0.25\left(\frac{1.96}{0.02}\right)^2 = 2401.$$

Esto es, necesitamos entrevistar al menos a 2401 personas para que la estimación tenga un margen de error del 2 % y una confianza del 95 %. □

4. Apéndice

Antes de demostrar el Teorema 1 enunciemos primero dos resultados con múltiples aplicaciones en probabilidad y en estadística.

Lema 1. *Desigualdades de Markov y de Chebyshev*

- *Desigualdad de Markov. Si X es una variable aleatoria no-negativa, entonces*

$$\forall a > 0 : P(X \geq a) \leq \frac{E(X)}{a}.$$

- *Desigualdades de Chebyshev. Si X una variable aleatoria con media μ y varianza σ^2 entonces*

$$\forall c > 0 : P\left(|X - \mu| \geq c\right) \leq \frac{\sigma^2}{c^2}.$$

Demostración. Para demostrar la afirmación del inciso a) definamos la variables aleatoria $Y_a = a\mathrm{I}_{[a,\infty)}$; i.e.

$$Y_a = \begin{cases} 0 & \text{si } X < a \\ a & \text{si } X \geq a. \end{cases}$$

Luego, las funciones (variables) Y_a y X satisfacen $Y_a \leq X$ y, en consecuencia $\mathrm{E}(Y_a) \leq \mathrm{E}(X)$. Al mismo tiempo $\mathrm{E}(Y_a) = aP(Y = a) = aP(X \geq a)$, de donde se deduce la tesis.

La tesis de b) se prueba aplicando la desigualdad de Markov a la variable aleatoria $(X - \mu)^2$ con $a = c^2$ □

Demostración del Teorema 1. Por la desigualdad de Chebyshev aplicada a $\overline{X}_n$, obtenemos:

$$\forall \epsilon > 0 : P\left(\left|\overline{X}_n - \mathrm{E}(\overline{X}_n)\right| \geq \epsilon\right) \leq \frac{\mathrm{VAR}(\overline{X}_n)}{\epsilon^2}.$$

Por otro lado

$$\mathrm{E}(\overline{X}_n) = \theta$$

y

$$\mathrm{VAR}(\overline{X}_n) = \mathrm{VAR}(\frac{1}{n}\sum X_i) = \frac{1}{n^2}\sum \mathrm{VAR}(X_i) = \frac{1}{n^2} n \mathrm{VAR}(X_1) = \sigma^2/n$$

Luego

$$0 \leq \lim_{n\to\infty} P\left(\left|\overline{X}_n - \theta\right| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{1}{n}\frac{\sigma^2}{\epsilon^2} = 0.$$

□