

Capítulo 1

Fenómenos aleatorios. Espacios de probabilidad. Axiomas. Propiedades. Caso discreto. Caso continuo.

1.1. Introducción.

Capítulo 2

Convergencias de sucesiones de variables aleatorias.

Convergencia casi segura, convergencia en probabilidad, convergencia en media cuadrática, convergencia en ley. Relaciones entre ellas. Convergencia de sumas de variables aleatorias. Leyes débiles y fuertes de los grandes números. Aplicaciones a la inferencia estadística y al muestreo. Teorema Central del Límite.

2.1. Introducción.

Hemos visto que una variable aleatoria es una función medible definida de un espacio de probabilidad en \mathbb{R} .

Definimos como una sucesión de variables aleatorias a un conjunto $\{X_n\}_{n \in \mathbb{N}}$, de variables aleatorias definidas sobre el mismo espacio probabilístico. Es de interés desde el punto de vista de la probabilidad el comportamiento de dicha sucesión cuando el valor de n tiende a infinito. De esta forma, definiremos los diversos tipos de convergencia de variables aleatorias.

2.2. Convergencia de variables aleatorias.

2.2.1. Convergencia en probabilidad.

Definición 1. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico $(\Omega, \mathbb{A}, \mathbb{P})$, se dice que converge en probabilidad a otra variable aleatoria X definida sobre el mismo espacio y se denota por $X_n \xrightarrow{P} X$ si se cumple que:

$$P\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\} < \alpha \quad (2.1)$$

o, lo que es lo mismo, que para todo $\varepsilon > 0$:

$$\lim_{n \rightarrow \infty} P\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\} = 0 \quad (2.2)$$

Es decir, que para un n suficientemente grande, la diferencia entre X_n y la variable X es muy pequeña con probabilidad muy alta.

Como caso particular, diremos que $\{X_n\}_{n \in \mathbb{N}}$ converge a una constante a si:

$$\lim_{n \rightarrow \infty} P\{\omega \in \Omega : |X_n(\omega) - a| > \varepsilon\} = 0, \quad \forall \varepsilon > 0 \quad (2.3)$$

2.2.1.1. Propiedades.

1. $X_n \xrightarrow{P} X \Leftrightarrow X_n - X \xrightarrow{P} 0$.
2. Si $X_n \xrightarrow{P} X$ y $X_n \xrightarrow{P} Y$, entonces $P(X = Y) = 1$.

3. Si $X_n \xrightarrow{p} X$, entonces Si $X_n - X_m \xrightarrow{p} 0$ cuando $n, m \rightarrow \infty$.
4. Si $X_n \xrightarrow{p} X$ y $Y_n \xrightarrow{p} Y$ entonces $X_n \pm Y_n \xrightarrow{p} X \pm Y$.
5. Si $X_n \xrightarrow{p} X$ y k es una constante, entonces $kX_n \xrightarrow{p} kX$.
6. Si $X_n \xrightarrow{p} a$ y a es una constante, entonces $X_n^2 \xrightarrow{p} a^2$.
7. Si $X_n \xrightarrow{p} a$ y $Y_n \xrightarrow{p} b$, a, b constantes, entonces $X_n Y_n \xrightarrow{p} ab$.
8. Si $X_n \xrightarrow{p} 1$ y $X_n(\omega) \neq 0, \forall \omega \in \Omega$, entonces $X_n^{-1} \xrightarrow{p} 1$.
9. Si $X_n \xrightarrow{p} a$ y $Y_n \xrightarrow{p} b$, a, b constantes, $b \neq 0$, entonces $X_n Y_n^{-1} \xrightarrow{p} ab^{-1}$.
10. Si $X_n \xrightarrow{p} X$ y Y , es una variable aleatoria, entonces $Y X_n \xrightarrow{p} Y X$.
11. Si $X_n \xrightarrow{p} X$ y $Y_n \xrightarrow{p} Y$ entonces $X_n Y_n \xrightarrow{p} XY$.

Teorema 1. Si $X_n \xrightarrow{p} X$, y g es una función continua definida sobre \mathbb{R} , entonces $g(X_n) \xrightarrow{p} g(X)$

Corolario 1. Si $X_n \xrightarrow{p} c$, donde c es constante, y g es una función continua definida sobre \mathbb{R} , entonces $g(X_n) \xrightarrow{p} g(c)$

2.2.2. Convergencia casi segura.

Definición 2. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico $(\Omega, \mathbb{A}, \mathbb{P})$, se dice que converge casi seguramente a otra variable aleatoria X definida sobre el mismo espacio y se denota por $X_n \xrightarrow{c.s.} X$ si se cumple que:

$$P \left\{ \omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \right\} = 1 \quad (2.4)$$

Es decir, la probabilidad del conjunto del espacio probabilístico donde la sucesión numérica de variables converge puntualmente a la variable es igual a uno.

2.2.2.1. Propiedades.

1. Si $X_n \xrightarrow{c.s.} X$, entonces $X_n \xrightarrow{p} X$. En efecto, por la convergencia casi segura, podemos elegir un n_0 tal que ... La recíproca no es cierta.
2. Si $\{X_n\}_{n \in \mathbb{N}}$ es una sucesión estrictamente decreciente de variables aleatorias positivas, entonces $X_n \xrightarrow{p} 0$ implica que $X_n \xrightarrow{c.s.} 0$.

2.2.3. Convergencia en ley.

Definición 3. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico $(\Omega, \mathbb{A}, \mathbb{P})$, se dice que converge en ley o en distribución a otra variable aleatoria X definida sobre el mismo espacio y se denota por $X_n \xrightarrow{\ell} X$ si y solo si la correspondiente sucesión de funciones de distribución de las variables aleatorias $\{X_n\}_{n \in \mathbb{N}}$, denotada por $\{F_n\}_{n \in \mathbb{N}}$, converge a la función de distribución de la variable aleatoria X en todo punto de continuidad de esta función, es decir, si:

$$\lim_{n \rightarrow \infty} F_n(x) = F(x) \quad \forall x / F(x+0) = F(x-0) \quad (2.5)$$

2.2.3.1. Propiedades.

1.

Teorema 2. *La convergencia en probabilidad implica la convergencia en ley. Es decir,*

$$X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{\ell} X \quad (2.6)$$

2.

Teorema 3. *Sea una sucesión de variables aleatorias, con función de masa $p_n(k) = P\{X_n = k\}$ y sea la variable aleatoria X con función de masa $p(k) = P\{X = k\}$, entonces:*

$$p_n(k) \rightarrow p(k) \forall k \Leftrightarrow X_n \xrightarrow{\ell} X \quad (2.7)$$

3.

Teorema 4. *Sea una sucesión de variables aleatorias continuas, con función de densidad $f_n(x)$ y sea la variable aleatoria continua X con función de densidad $f(x)$, y se cumple que $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ para casi todo x , entonces $X_n \xrightarrow{\ell} X$.*

4. Si $X_n \xrightarrow{\ell} X$ y c es una constante, entonces $X_n + c \xrightarrow{\ell} X + c$ y $cX_n \xrightarrow{\ell} cX$ si $c \neq 0$.

5. Si k es una constante y $X_n \xrightarrow{\ell} k$, entonces $X_n \xrightarrow{p} k$.

2.2.4. Convergencia en media cuadrática.

Definición 4. *Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico $(\Omega, \mathbb{A}, \mathbb{P})$, y supongamos que se cumple $E[|X_n|^2] < \infty$, $\forall n \in \mathbb{N}$. Se dice que $\{X_n\}_{n \in \mathbb{N}}$ converge en media cuadrática hacia la variable aleatoria X definida sobre el mismo espacio y se denota por $X_n \xrightarrow{m.c.} X$ si y solo si:*

$$\lim_{n \rightarrow \infty} E[(X_n - X)^2] = 0 \quad (2.8)$$

2.2.4.1. Propiedades.

1. Si $X_n \xrightarrow{m.c.} X$, entonces $X_n \xrightarrow{p} X$.

2. Si $X_n \xrightarrow{m.c.} X$, entonces $E[X_n] \xrightarrow{n \rightarrow \infty} E[X]$ y $E[X_n^2] \xrightarrow{n \rightarrow \infty} E[X^2]$.

3. Si $X_n \xrightarrow{m.c.} X$, entonces $V[X_n] \xrightarrow{n \rightarrow \infty} V[X]$.

4. Sean $\{X_n\}_{n \in \mathbb{N}}$, $\{Y_m\}_{m \in \mathbb{N}}$ dos sucesiones de variables aleatorias tales que $X_n \xrightarrow{m.c.} X$ y $Y_m \xrightarrow{m.c.} Y$, entonces $E[X_n Y_n] \xrightarrow{m, n \rightarrow \infty} E[XY]$.

5. Sean $\{X_n\}_{n \in \mathbb{N}}$, $\{Y_m\}_{m \in \mathbb{N}}$ dos sucesiones de variables aleatorias tales que $X_n \xrightarrow{m.c.} X$ y $Y_m \xrightarrow{m.c.} Y$, entonces $Cov[X_n, Y_n] \xrightarrow{m, n \rightarrow \infty} Cov[X, Y]$.

2.2.5. Convergencia de las funciones características.

La función característica es una forma de caracterizar la distribución de probabilidad de una variable aleatoria. Parece lógico, por tanto, que tenga alguna relación con la convergencia en ley de las mismas.

En concreto, se puede demostrar que, sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico $(\Omega, \mathbb{A}, \mathbb{P})$, con funciones características $\varphi_n(t)$, y sea X una variable aleatoria definida sobre el mismo espacio y con función característica $\varphi(t)$, se cumple:

1. Si $X_n \xrightarrow{\ell} X$, $\Rightarrow \lim_{n \rightarrow \infty} \varphi_n(t) = \varphi(t)$, $\forall t \in \mathbb{R}$.

2. Si $\lim_{n \rightarrow \infty} \varphi_n(t) = \varphi(t)$, $\forall t \in \mathbb{R}$ y $\varphi(t)$ es continua en $t = 0$, entonces $X_n \xrightarrow{\ell} X$.

El primer apartado se demuestra mediante el teorema de Levy:

Teorema 5. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias tales que $X_n \xrightarrow{\ell} X$, siendo $F_n(x)$ y $F(x)$ las funciones de distribución de las X_n y X , respectivamente, y sea

$$\varphi_n(t) = \int_{\mathbb{R}} e^{itx} dF_n(x) \quad (2.9)$$

entonces,

$$\lim_{n \rightarrow \infty} \varphi_n(t) = \int_{\mathbb{R}} e^{itx} dF(x) \quad (2.10)$$

y ese límite se alcanza uniformemente en todo intervalo finito de t .

El segundo apartado es debido a Cramer.

2.3. Leyes de los grandes números.

La interpretación frecuentista de la probabilidad establece que, dado un experimento aleatorio, si consideramos un suceso A , la frecuencia relativa de aparición de dicho suceso tiende a estabilizarse hacia un valor definido a medida que aumentamos las repeticiones del experimento. A este valor se le llamaría probabilidad del suceso A .

Esta interpretación frecuentista de la probabilidad tiene una caracterización matemática dentro de la concepción axiomática de la probabilidad a través de las conocidas como **leyes de los grandes números**.

2.3.1. Ley débil de los grandes números.

Definición 5. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) tales que existe $E[X_i] = \alpha_i < \infty, \forall i \in \mathbb{N}$; sea

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Diremos que la sucesión $\{X_n\}_{n \in \mathbb{N}}$ obedece a la ley débil de los grandes números, y lo denotaremos por $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$ si y solo si la sucesión $\{\bar{X}_n - a_n\}_{n \in \mathbb{N}}$ converge en probabilidad hacia cero, siendo

$$a_n = \frac{1}{n} \sum_{i=1}^n E[X_i] = E[\bar{X}_n].$$

Es decir,

$$\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D} \Leftrightarrow \bar{X}_n - a_n \xrightarrow{p} 0.$$

En términos más generales, se dice que una sucesión $\{X_n\}_{n \in \mathbb{N}}$ obedece a la ley débil de los grandes números respecto de las constantes de normalización $B_n > 0$ si existe una sucesión de constantes $\{A_n\}_{n \in \mathbb{N}}$, llamadas de centralización tales que:

$$\frac{S_n - A_n}{B_n} \xrightarrow{p} 0$$

con $S_n = n\bar{X}_n = \sum_{i=1}^n X_i$. Llegamos a la definición inicial si $B_n = n$ y $A_n = \sum_{i=1}^n E[X_i]$.

Veremos ahora una serie de teoremas que establecen condiciones bajo las cuales una sucesión verifica la ley débil de los grandes números.

2.3.1.1. Teoremas.

Teorema 6. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) y tales que

- Las variables aleatorias de la sucesión están idénticamente distribuidas.
- Las X_n tienen media y varianza finitas, $\forall n \in \mathbb{N}$.

Entonces $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$.

Como las variables tienen la misma distribución, tendrán la misma media y varianza, por tanto, $\alpha_n = E[X_n] = \mu$, $V(X_n) = \sigma^2 < \infty$, $\forall n \in \mathbb{N}$. Por tanto, $a_n = E[\bar{X}_n] = \mu$, $V(\bar{X}_n) = \frac{\sigma^2}{n}$ por ser las variables independientes. Así pues,

$$\lim_{n \rightarrow \infty} P\{\omega \in \Omega / |\bar{X}_n - a_n| > \varepsilon\} = \lim_{n \rightarrow \infty} P\{\omega \in \Omega / |\bar{X}_n - \mu| > \varepsilon\} \leq \frac{E[(\bar{X}_n - \mu)^2]}{\varepsilon^2} = \lim_{n \rightarrow \infty} \frac{\sigma^2}{n\varepsilon^2} = 0$$

Donde hemos utilizado la acotación de Tchebychev, por tanto, $\bar{X}_n - a_n \xrightarrow{p} 0$ y $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$.

Teorema 7 (Tchebychev). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) y tales que toda variable X_n tiene varianza acotada, es decir, $V(X_n) = \sigma_n^2 < c$, $\forall n \in \mathbb{N}$ para alguna cota c , entonces $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$.

Como $a_n = E[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n E[X_i]$ y $V(\bar{X}_n) = \frac{\sum_{i=1}^n V(X_i)}{n^2} \leq \frac{c}{n}$, para todo $\varepsilon > 0$ tenemos que

$$\lim_{n \rightarrow \infty} P\{\omega \in \Omega / |\bar{X}_n - a_n| > \varepsilon\} \leq \frac{E[(\bar{X}_n - a_n)^2]}{\varepsilon^2} = \frac{V(\bar{X}_n)}{\varepsilon^2} \leq \lim_{n \rightarrow \infty} \frac{c}{n\varepsilon^2} = 0$$

Donde hemos utilizado la acotación de Tchebychev, por tanto, $\bar{X}_n - a_n \xrightarrow{p} 0$ y $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$.

Teorema 8 (Markov). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) y tales que $\lim_{n \rightarrow \infty} V(\bar{X}_n) = 0$, entonces $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$.

La demostración es inmediata utilizando la desigualdad de Tchebychev.

Teorema 9 (Khinchine). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) y tales que

- Las variables aleatorias de la sucesión están idénticamente distribuidas, con función de distribución F .
- Las X_n tienen media finita, $\forall n \in \mathbb{N}$.

Entonces $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$.

Teorema 10 (Teorema de Bernouilli). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) y tales que $X_n \sim B(1, p)$, entonces $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$.

Dado que las variables aleatorias son independientes e idénticamente distribuidas, con $E[X_n] = p$, $V(X_n) = p(1-p)$, ambas finitas, si aplicamos el primer teorema, vemos que la sucesión cumple la ley débil de los grandes números.

Además, $a_n = \frac{1}{n} \sum_{i=1}^n E[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$, y como $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D}$, $\bar{X}_n - a_n \xrightarrow{p} 0$, o sea, la frecuencia relativa de los éxitos en el experimento de Bernouilli converge en probabilidad a la probabilidad de éxito, que es equivalente a la interpretación frecuentista de la probabilidad.

2.3.2. Ley fuerte de los grandes números.

Definición 6. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) tales que existe $E[X_i] = \alpha_i < \infty, \forall i \in \mathbb{N}$; sea

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Diremos que la sucesión $\{X_n\}_{n \in \mathbb{N}}$ obedece a la ley fuerte de los grandes números, y lo denotaremos por $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{F}$ si y solo si la sucesión $\{\bar{X}_n - a_n\}_{n \in \mathbb{N}}$ converge hacia cero casi seguro, siendo

$$a_n = \frac{1}{n} \sum_{i=1}^n E[X_i] = E[\bar{X}_n].$$

Es decir,

$$\{X_n\}_{n \in \mathbb{N}} \in \mathcal{D} \Leftrightarrow \bar{X}_n - a_n \xrightarrow{p} 0.$$

En términos más generales, se dice que una sucesión $\{X_n\}_{n \in \mathbb{N}}$ obedece a la ley fuerte de los grandes números respecto de las constantes de normalización $B_n > 0$ si existe una sucesión de constantes $\{A_n\}_{n \in \mathbb{N}}$, llamadas de centralización tales que:

$$\frac{S_n - A_n}{B_n} \xrightarrow{c.s.} 0$$

con $S_n = n\bar{X}_n = \sum_{i=1}^n X_i$. Llegamos a la definición inicial si $B_n = n$ y $A_n = \sum_{i=1}^n E[X_i]$.

Veremos ahora una serie de teoremas que establecen condiciones bajo las cuales una sucesión verifica la ley fuerte de los grandes números.

2.3.2.1. Teoremas.

Desigualdad de Kolmogorov: Es una generalización de la desigualdad de Tchebychev.

Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) tales que existe $E[X_n] = \mu_n < \infty, V(X_n) = \sigma_n^2 < \infty, \forall i \in \mathbb{N}$. Entonces para todo H positivo se verifica:

$$P \left(\bigcup_{k=1}^n \{ \omega \in \Omega / |S_n - E[S_n]| \geq H V_n \} \right) \leq \frac{1}{H^2}$$

donde

$$S_n = \sum_{i=1}^n X_i \quad V_n^2 = V(S_n) = \sum_{i=1}^n \sigma_i^2$$

Teorema 11 (Kolmogorov). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) tales que existe $E[X_n] = \mu_n, V(X_n) = \sigma_n^2, \forall i \in \mathbb{N}$ y se cumple que $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$, entonces $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{F}$.

Como corolario, una sucesión con media y varianza acotada cumple la ley fuerte de los grandes números.

Teorema 12 (Borel-Cantelli). La frecuencia relativa de un suceso dicotómico obedece a la ley fuerte de los grandes números.

Teorema 13 (Khinchine). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con media finita μ . Entonces $\{X_n\}_{n \in \mathbb{N}} \in \mathcal{F}$.

2.4. Teorema central del límite.

2.4.1. Introducción.

2.4.2. Definición.

Definición 7. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el espacio probabilístico (Ω, \mathcal{A}, P) , con medias y varianza finitas. Diremos que la sucesión obedece al teorema central del límite, y lo denotaremos por

$\{X_n\}_{n \in \mathbb{N}} \in LN$ si y solo si la sucesión $\{S_n\}_{n \in \mathbb{N}}$ definida por $S_n = \sum_{i=1}^n X_i$ converge en ley hacia una distribución normal. Es decir:

$$\{X_n\}_{n \in \mathbb{N}} \in LN \Leftrightarrow \frac{S_n - E[S_n]}{D[S_n]} \xrightarrow{\ell} N(0, 1)$$

Teorema 14 (De Moivre). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas según una distribución $B(1, p)$, entonces $\{X_n\}_{n \in \mathbb{N}} \in LN$.

Teorema 15 (Levy-Lindeberg). Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con media y varianza finitas, entonces $\{X_n\}_{n \in \mathbb{N}} \in LN$.

Capítulo 3

Cadenas de Markov.

Distribución de la cadena. Cadenas homogéneas. Clasificación de los estados. Tipos de cadenas. Distribuciones estacionarias.

3.1. Introducción.

Las cadenas de Markov son un tipo especial de proceso estocástico. Un proceso estocástico es una sucesión de variables aleatorias que evolucionan en función de otra variable. En general, la variable en función de la que evolucionan suele ser el tiempo. Cada variable de un proceso estocástico tiene su propia función de distribución de probabilidad, y pueden estar correlacionadas entre sí o no estarlo.

En cuanto a los procesos estocásticos dependientes del tiempo, se pueden ver de dos maneras equivalentes: como un conjunto de realizaciones temporales dependientes de un índice, o como un conjunto de variables aleatorias dependientes del tiempo.

Formalmente definimos un proceso estocástico como un conjunto de variables aleatorias

Definición 8. Sea $(\Omega, \mathbb{A}, \mathbb{P})$ un espacio de probabilidad. Un proceso estocástico es una colección o familia de variables aleatorias $\{X(t)/t \in T\}$, dependientes de un parámetro t que en general se suele identificar con el tiempo. Al conjunto de valores del parámetro se le denomina espacio parametral y se denota por T , y al conjunto de valores posibles de las variables aleatorias se le llama espacio de estados y se denota por S .

A cada uno de los elementos del espacio de estados se le llama estado del proceso estocástico. Cada realización concreta del proceso se llamará trayectoria o función muestral.

Podemos considerar como casos particulares de esta estructura distintos conceptos asociados al cálculo de probabilidades. Así:

- Si $T = \{t_0\}$, se obtiene una variable aleatoria unidimensional, $X(t_0)$.
- Si $T = \{t_0, t_1, \dots, t_k\}$ tenemos una variable aleatoria k -dimensional, $(X(t_0), X(t_1), \dots, X(t_k))$.
- Si $T = \mathbb{N}$, tenemos una sucesión de variables aleatorias, $\{X(t_n)\}_{n \in \mathbb{N}}$.

Por otro lado, el parámetro asociado al proceso puede ser discreto, es decir, $T = \mathbb{N} \cup \{0\}$ o continuo, es decir, $T = [0, +\infty)$, o $T = \mathbb{R}$ si consideramos estados retrospectivos.

Los procesos estocásticos se pueden clasificar atendiendo a dos criterios:

- La cardinalidad del conjunto de estados: Si el espacio de estados es finito o infinito numerable (discreto) diremos que el proceso es una cadena. Si el espacio de estados es infinito no numerable (continuo) diremos que estamos ante un proceso.
- La cardinalidad del conjunto parametral, o conjunto de índices: Si el conjunto de índices es finito o infinito numerable (discreto) diremos que el proceso es en tiempo discreto. Si el conjunto de índices es infinito no numerable (continuo) diremos que estamos ante un proceso en tiempo continuo.

Por tanto, atendiendo a estas dos dimensiones, podemos clasificar los procesos estocásticos en:

Cadenas: Conjunto de estados discreto, tiempo discreto.

tiempo continuo: Conjunto de estados discreto, tiempo continuo.

tiempo discreto: Conjunto de estados continuo, tiempo discreto.

tiempo continuo: Conjunto de estados continuo, tiempo continuo.

En el caso de las cadenas, si S es un conjunto finito se habla de cadenas finitas.

Si fijamos la parte aleatoria de un proceso estocástico, obtendremos una función real de variable real, que sólo dependerá de T . Por ello, a las diferentes funciones que se generan en este caso se les llama trayectorias. Si T es un conjunto discreto, las trayectorias serán sucesiones reales, y si es continuo serán funciones reales.

Si fijamos la componente temporal o relativa al punto $t_0 \in T$ obtenemos una variable aleatoria unidimensional que refleja el comportamiento del proceso en el instante t_0 . La variable será discreta o continua según sea el espacio de estados.

Función de distribución del proceso estocástico: de primer y segundo orden.

Para un proceso estocástico se pueden definir las siguientes magnitudes:

- Función Media: $\mu(t) = E(X_t), t \in T$.
- Función Varianza: $\sigma^2(t) = E[(X_t - \mu(t))^2], t \in T$.
- Núcleo de Covarianza o Autocovarianza: $\gamma(r, s) = Cov(X_r, X_s) = E(X_r \cdot X_s) - \mu(r)\mu(s), r, s \in T$.
- Función de Autocorrelación: $\rho(r, s) = \frac{\gamma(r, s)}{\sigma(r)\sigma(s)}$.

Definición 9. Se dice que un proceso estocástico con un conjunto lineal de índices y con función de distribución $F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n)$ es estrictamente estacionario si y solo si se cumple:

$$F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) = F(x_1, x_2, \dots, x_n; t_1 + h, t_2 + h, \dots, t_n + h) \quad \forall t_1, t_2, \dots, t_n \in T, \forall h \in T, \forall n \in \mathbb{N}$$

Las condiciones para que un proceso sea estacionario son muy estrictas y difíciles de cumplir. Por ello se define la estacionariedad de orden r :

Definición 10. Se dice que un proceso estocástico con un conjunto lineal de índices y con función de distribución $F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n)$ es estacionario de orden r si y solo si se cumple:

$$F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) = F(x_1, x_2, \dots, x_n; t_1 + h, t_2 + h, \dots, t_n + h) \quad \forall t_1, t_2, \dots, t_n \in T, \forall h \in T, \forall n \in \mathbb{N} : n \leq r$$

Un proceso estrictamente estacionario será estacionario de cualquier orden, y un proceso estacionario de orden r será estacionario en los órdenes anteriores.

Teorema 16. Sea $\{X(t), t \in T\}$ un proceso estacionario de orden dos, entonces se cumple:

- Las variables aleatorias que lo integran están idénticamente distribuidas.
- Las funciones media y varianza del proceso son constantes.

Teorema 17. Sea $\{X(t), t \in T\}$ un proceso estacionario de orden dos, entonces las distribuciones conjuntas bidimensionales no dependen de los índices concretos, sino de la separación entre ellos.

Como consecuencia, las funciones de autocovarianza y autocorrelación solo dependerán de la distancia entre los índices.

Definición 11. Se dice que un proceso estocástico con un conjunto lineal de índices y con función de distribución $F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n)$ es débilmente estacionario si y solo si se cumple:

- $\mu(t) = \mu \quad \forall t \in T$.
- $\gamma(t, t + h) = C(h) \quad \forall t, h \in T$

Esta definición implica que la función de varianza es constante y la función de autocorrelación solo depende de la distancia entre los índices.

3.2. Cadenas de Markov.

Definición 12. Un proceso estocástico en tiempo discreto se dice que cumple la **propiedad de Markov** si para cualquier $n \geq 0$ y para cualesquiera estados $x_0, x_1, x_2, \dots, x_n, x_{n+1}$ se cumple que:

$$p(x_{n+1}|x_0, x_1, x_2, \dots, x_n) = p(x_{n+1}|x_n)$$

Es decir, la probabilidad de que el proceso esté en un estado en un momento t_i solo depende del estado en el que estuviera el proceso en el momento anterior.

Los procesos que cumplen la propiedad de Markov reúnen las características de ser aplicables a gran cantidad de fenómenos y de ser suficientemente sencillos como para ser analizados matemáticamente.

Definición 13. Una **cadena de Markov** es un proceso estocástico en tiempo discreto con un espacio de estados discreto que cumple la **propiedad de Markov**.

Una cadena de Markov cuyo conjunto de estados es finito se llama cadena finita.

3.2.1. Distribución de la cadena.

La distribución de probabilidad de una cadena de Markov nos dará la probabilidad de que la cadena ocupe cada uno de los estados del espacio de estados en cada instante de tiempo.

Así, la distribución de probabilidad de primer orden se la conoce como vector de estados en el instante t , es decir, la probabilidad de cada estado en el instante t . Se denota por V_t :

$$V_t = (p_i(t), i \in S), \forall t \in T \text{ con } p_i(t) = P(X_t = i).$$

Para la distribución de orden dos, es decir, la probabilidad de cada estado en el instante t condicionada por el estado de la cadena en el instante anterior, tendremos una matriz de estados:

$$P(X_m = i, X_n = j) = P(X_n = j|X_m = i)P(X_m = i) = P(X_n = j|X_m = i)P_i(m), \forall i, j \in S, \forall m, n \in T.$$

Y para la distribución de orden superior:

$$P(X_{t_1} = i_1, X_{t_2} = i_2, \dots, X_{t_n} = i_n) = P(X_{t_n} = i_n|X_{t_{n-1}} = i_{n-1}, \dots, X_{t_1} = i_1) \cdots P(X_{t_2} = i_2|X_{t_1} = i_1)P(X_{t_1} = i_1) = P(X_{t_n} = i_n|X_{t_{n-1}} = i_{n-1}) \cdots P(X_{t_2} = i_2|X_{t_1} = i_1)P(X_{t_1} = i_1), \forall i_1, \dots, i_n \in S, \forall t_1, \dots, t_n \in T.$$

A la probabilidad $P(X_n = j|X_m = i)$ se la denota por $p_{i,j}(m, n)$. Representa la probabilidad de la cadena de pasar del estado i al j entre los instantes m y n y se la conoce como probabilidad de transición.

A la matriz en la que se disponen las probabilidades de transición entre todos los estados del espacio de estados se le llama matriz de transición. A la matriz de transición entre dos momentos consecutivos se le llama matriz de transición en un paso o en una etapa. La matriz de transición entre los momentos m y n se denota por $P(m, n)$

- La matriz de transición es una matriz estocástica.
- El avance de la cadena entre dos estados se puede modelizar a partir del vector de estados y la matriz de transición: $V_t = V_s P(s, t)$, $\forall s, t \in T$, $s < t$.
- **Ecuación de Chapman-Kolmogorov:** $P(r, t) = P(r, s)P(s, t)$, $\forall r, s, t \in T$, $r < s < t$.
- $P(m, n) = P(m, m+1)P(m+1, m+2) \cdots P(n-1, n)$.
- $V_t = V_0 P(0, t)$.
- $p_i(n) = \sum_j p_{ij}(n, n-1)p_j(s)$.

3.2.2. Cadenas de Markov homogéneas.

Definición 14. Una **cadena de Markov** se dice que es **homogénea** si se cumple que $p_{i,j}(t, t+1) = p_{i,j}$, o también $P(t, t+1) = P$. Es decir, las probabilidades de transición en un paso entre dos estados no dependen del tiempo, son constantes.

Como consecuencia:

- $P(r+s) = P(r)P(s) \forall r, s \in T$, por la ecuación de Chapman-Kolmogorov.

- $P(h, t + h) = P(0, t) = P^t = P(t) \forall t, h \in T.$
- $V_t = V_0 P(t) = V_0 P^t \forall t \in T.$

A la hora de calcular P^t podemos hacer uso de la descomposición de Jordan, que facilitará el cálculo: $P = H J H^{-1}$, $P^t = H J^t H^{-1}$.

Una cadena homogénea queda especificada conociendo el vector de estados inicial y la matriz de transición.

3.2.3. Clasificación de los estados.

El hecho de que las cadenas de Markov homogéneas queden definidas por su matriz de transición entre estados aconseja un análisis de dicha matriz. Esto nos permitirá clasificar las relaciones entre los distintos estados de la cadena, así como definir una tipología dentro de los mismos.

Por tanto, definimos como tiempo de paso entre dos estados, i, j y representamos como N_{ij} a la variable aleatoria que representa el número de pasos que da la cadena para pasar del estado i al estado j . Representamos la probabilidad asociada al este suceso como $f_{ij}(n) = P(N_{ij} = n)$, que representa la probabilidad de que se necesiten n transiciones para pasar del estado i al estado j .

A la situación que representa N_{ii} se le llama tiempo de recurrencia y representa el número de transiciones para que la cadena regrese al estado i .

Si expresamos f_{ij} en función de las probabilidades de transición, tendremos que:

- $f_{ij}(1) = p_{ij}(1).$
- $p_{ij}(2) = f_{ij}(1)p_{jj}(1) + f_{ij}(2)$ y $f_{ij}(2) = p_{ij}(2) - f_{ij}(1)p_{jj}(1).$
- $p_{ij}(3) = f_{ij}(1)p_{jj}(2) + f_{ij}(2)p_{jj}(1) + f_{ij}(3)$ y $f_{ij}(3) = p_{ij}(3) - f_{ij}(1)p_{jj}(2) - f_{ij}(2)p_{jj}(1).$
- $p_{ij}(n) = f_{ij}(1)p_{jj}(n-1) + \dots + f_{ij}(n-1)p_{jj}(1) + f_{ij}(n)$ y $f_{ij}(n) = p_{ij}(n) - f_{ij}(1)p_{jj}(n-1) - \dots - f_{ij}(n-1)p_{jj}(1).$

Definimos $f_{ij} = \sum_n f_{ij}(n)$, que será la probabilidad de que la cadena pase en algún momento por el estado j habiendo partido del estado i . Por ser una probabilidad, $f_{ij} \leq 1$ y en particular, $f_{ii} \leq 1$. Por tanto, si $f_{ii} < 1$ existe una probabilidad no nula que una cadena que parta del estado i no regrese al mismo. A estos estados se les conoce como **estacionarios**.

Por el contrario, si $f_{ii} = 1$, es decir, si la cadena siempre regresa al estado i , diremos que es un estado **recurrente**. Una situación particular del estado recurrente es el caso en que $f_{ii}(1) = 1$, es decir, si la cadena accede al estado i ya no lo abandona. Estos estados se conocen como **absorbentes**.

Si calculamos $E[N_{ii}] = \sum_n n f_{ii}(n)$, que es el tiempo medio que tarda la cadena en regresar al estado i , tendremos dos posibilidades: $E[N_{ii}] < \infty$, estado **recurrente positivo** y $E[N_{ii}] = \infty$, estado **recurrente nulo**.

Otra forma de caracterizar los estados es calculando el número medio de veces que la cadena pasa por un estado partiendo de otro.:

Sea la variable aleatoria $Z_i(n)$:

$$Z_i(n) = \begin{cases} 1 & x_n = i \\ 0 & x_n \neq i \end{cases}$$

Tendremos que $P([Z_j(n) = 1/x_0 = i] = p_{ij}(n)$, y por tanto, $E[Z_j(n)/x_0 = i] = p_{ij}(n)$. Si definimos la variable aleatoria Z_i como el número de veces que la cadena pasa por el estado i , entonces $Z_i = \sum_n Z_i(n)$, y por tanto, $E[Z_j/x_0 = i] = E[\sum_n Z_j(n)/x_0 = i] = \sum_n p_{ij}(n)$, que es el número esperado de veces que la cadena pasa por el estado j partiendo del estado i .

Como además:

$$\sum_n p_{ij}(n) = \sum_n (f_{ij}(1)p_{jj}(n-1) + \dots + f_{ij}(n-1)p_{jj}(1) + f_{ij}(n)) = \left(\sum_n p_{jj}(n) \right) \left(\sum_n f_{ij}(n) \right) + \sum_n f_{ij}(n)$$

Y por tanto,

$$f_{ij} = \sum_n f_{ij}(n) = \frac{\sum_n p_{ij}(n)}{1 + \sum_n p_{jj}(n)}$$

Para la transición de un estado a sí mismo:

$$f_{ii} = \sum_n f_{ii}(n) = \frac{\sum_n p_{ii}(n)}{1 + \sum_n p_{ii}(n)} = \frac{1}{1 + \frac{1}{\sum_n p_{ii}(n)}}$$

Así, un estado es recurrente, esto es, $f_{ii} = 1$, si y solo si $\sum_n p_{ii}(n) = \infty$, es decir, la cadena vuelve a pasar de media por el estado un número infinito de veces.

Así, un estado es transitorio, esto es, $f_{ii} < 1$, si y solo si $\sum_n p_{ii}(n) < \infty$, es decir, la cadena vuelve a pasar de media por el estado un número finito de veces.

También podemos clasificar los estados según su relación entre ellos:

- Sean dos estados, i, j . Si se cumple que $f_{ij} > 0$, existe una probabilidad no nula de que partiendo del estado i la cadena acceda al estado j . Se dice entonces que el estado j es accesible desde el i , y se expresa $i \rightsquigarrow j$.
- Si se cumple que $i \rightsquigarrow j$ y además $j \rightsquigarrow i$, entonces se dice que los estados son comunicantes, y se representa así: $i \longleftrightarrow j$.

Si el estado j es recurrente y accesible desde el estado i , el número esperado de veces que la cadena pasa por el estado j es infinito:

$$f_{ij} \left(1 + \sum_n p_{jj}(n) \right) = \sum_n p_{ij}(n)$$

Y como por ser j recurrente $\sum_n p_{jj}(n) = \infty$, $\sum_n p_{ij}(n) = \infty$.

Si el estado i es recurrente y j es accesible desde el estado i , entonces es seguro que partiendo del estado j la cadena regresa al estado i .

Si j es accesible desde i , $f_{ij} > 0$, y la probabilidad de que saliendo del estado j no se acceda al i será $1 - f_{ji}$. Entonces, $f_{ij}(1 - f_{ji})$ será la probabilidad de que partiendo del estado i se llegue al j , y partiendo del j no se vuelva al i , que serían mayor que cero, contradiciendo la recurrencia del estado i , por tanto, tendrá que cumplirse $1 - f_{ji} = 0$ y $f_{ji} = 1$, es decir, partiendo del estado j se llega seguro al estado i . Por

Capítulo 4

Fundamentos de la Inferencia Estadística.

Concepto de muestra aleatoria. Distribución de la muestra. Estadísticos y su distribución en el muestreo. Función de distribución empírica y sus características. Teorema de Glivenko-Cantelli.

4.1. Introducción.

Uno de los objetivos de la estadística es obtener conclusiones acerca de una determinada población a partir de la observación de un subconjunto de miembros de la misma. A este proceso se le conoce con el nombre de **inferencia estadística**, y se basa en todas las conclusiones y teoremas que nos ofrece el cálculo de probabilidad. La inferencia se puede realizar de dos formas: prediciendo el valor de un parámetro que defina a la población en cuestión, en cuyo caso hablaremos de **estimación** del parámetro, o proponiendo una serie de hipótesis sobre la población y comprobando si se cumplen, caso en el que estaremos hablando de **contraste de hipótesis**.

4.2. Concepto de muestra aleatoria.

Para realizar cualquier inferencia acerca de una población necesitaremos obtener información sobre la misma. En ausencia de otros condicionantes, la situación óptima sería aquella en la que podemos estudiar a todos los miembros de la población, y así nos aseguraremos que las conclusiones que obtengamos serán completamente válidas. Si embargo, en muchas ocasiones esto no es posible, ya sea por no poder acceder a todos los individuos, por el coste en el que incurriríamos e incluso porque la observación de un individuo implique su destrucción. En estos casos hemos de conformarnos con observar un subconjunto de miembros de la población, y a partir de esa observación inferir las características poblacionales que nos interesen. A este subconjunto de la población que vamos a observar lo llamamos **muestra**. Por observar entendemos la medición de uno o varios parámetros asociados a los individuos seleccionados, a los que llamaremos **parámetros muestrales**.

Aparece ahora el dilema de cual será la mejor forma de seleccionar esos individuos de forma que las conclusiones que obtengamos sean lo más correctas posibles. Lo ideal sería que dicha muestra fuera una representación a escala de la población, es decir, sea representativa, pero dado que desconocemos las características de la misma, en general esto no es posible.

Si ese subconjunto lo seleccionamos de manera aleatoria, podemos decir que las desviaciones entre la muestra que hemos seleccionado y una muestra totalmente representativa se deben al azar, y por tanto podemos asignarlas una probabilidad. Así, nos permitirá asignar conocer la probabilidad de que los resultados de nuestra inferencia sean exactos. Si procedemos de esta manera, diremos que hemos seleccionado una **muestra aleatoria**.

Dado que el valor de nuestras observaciones no lo conocemos a priori, la muestra se compondrá de tantas variables aleatorias como elementos contenga nuestra muestra (variables aleatorias muestrales) y una vez tomada la información tendremos un conjunto de valores a los que llamaremos valores muestrales.

A esta forma de seleccionar la muestra se le llama muestreo probabilístico, y se puede realizar con reemplazamiento, en el que un mismo elemento puede estar presente más de una vez en la muestra, y sin reemplazamiento, en el que cada elemento de la población estará presente como mucho una vez en la muestra.

En la inferencia estudiaremos un tipo especial de muestras, las llamadas **muestras aleatorias simples**, que son muestras aleatorias con reemplazamiento en las que la probabilidad de un elemento de estar presente en la muestra

es la misma para toda la población. En este tipo de muestras las variables aleatorias muestrales son independientes e idénticamente distribuidas, y su distribución de probabilidad es la misma que la del total de la población.

Definición 15. Una *muestra aleatoria simple* de tamaño n está formada por n variables muestrales X_1, X_2, \dots, X_n independientes e idénticamente distribuidas, con la misma distribución de probabilidad que la característica poblacional, ξ , a investigar.

4.3. Distribución de la muestra.

La distribución de probabilidad de la muestra nos dará la probabilidad de obtención de cada muestra posible en una determinada población. En el caso de poblaciones cuya distribución de probabilidad sea discreta, la distribución de la muestra vendrá dada por todas las muestras posibles y sus respectivas probabilidades. En el caso continuo, consistirá en la función de densidad conjunta de las variables muestrales.

Sea una muestras, (x_1, \dots, x_n) . En el caso discreto tendremos:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2/X_1 = x_1) \cdots P(X_n = x_n/X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1})$$

Y, si hablamos de una muestra aleatoria simple, cada uno de los componentes de la muestra lo habremos extraído de forma independiente, por tanto:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n)$$

De forma análoga, para el caso continuo y en el supuesto de muestra aleatoria simple, podemos obtener que:

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n)$$

4.4. Estadísticos y su distribución en el muestreo.

Definimos como estadístico cualquier función de los parámetros muestrales siempre que no contenga ningún parámetro desconocido. Lo denotaremos en general por $T(\mathbf{X}) = T(x_1, \dots, x_n)$.

Dado que los elementos de la muestra son variables aleatorias, los estadísticos también serán variables aleatorias, con un campo de variación y una distribución de probabilidad propios y determinados por el campo de variación y distribución de la población. Su campo de variación constará de los valores que pueda tomar el estadístico para todos los elementos del espacio muestral, y su la probabilidad asociada a cada valor será igual a la suma de las probabilidades de todas las posibles muestras a partir de las cuales se obtenga ese valor.

A la distribución de probabilidad asociada a cada estadístico se la llama distribución de probabilidad del estadístico en el muestreo. Si conocemos esta distribución de probabilidad podremos realizar afirmaciones probabilísticas sobre nuestro estadístico.

Es muy importante tener en cuenta que los parámetros poblacionales son constantes, aunque en general desconozcamos su valor, mientras que los estadísticos muestrales son variables aleatorias.

A aquellos estadísticos cuyo valor nos sirve para estimar algún parámetro de la población se les llama **estimadores**. Es por eso que la distribución del estadístico en el muestreo es tan importante, ya que nos puede dar información acerca de cuan próximo es el estimador resultante al parámetro estimado.

4.5. Función de distribución empírica y sus características.

La función de distribución de probabilidad de una variable aleatoria se define como $F(x) = P(X \leq x)$. De manera análoga, definimos la función de distribución empírica de una muestra:

Definición 16. Sea una población con una función de distribución $F(x)$. Sea una muestra aleatoria simple de la población, (x_1, \dots, x_n) . Designamos por $N(x)$ el número de elementos de esa muestra cuyo valor en menor o igual que x . Entonces, definimos la función de distribución empírica de la muestra aleatoria, que denotaremos por $F_n(x)$ como:

$$F_n(x) = \frac{N(x)}{n}$$

La función de distribución empírica no tiene relación directa con la función de distribución de la población ni la función de distribución en el muestreo, sin embargo existe una relación indirecta ya que se ha obtenido aleatoriamente a partir de la población y parece lógico deducir que puede proporcionar una imagen aproximada de la distribución de probabilidad de la población de la que se extrajo la muestra.

Es por ello que es conveniente obtener sus momentos, llamados momentos muestrales. Dado que esta función de distribución le asigna una probabilidad de $\frac{1}{n}$ a cada elemento de la muestra, sus valores serán:

- Momentos de orden r respecto del origen:

$$a_r = \frac{\sum_{i=1}^n x_i^r}{n}$$

- Momentos de orden r respecto de la media:

$$m_r = \frac{\sum_{i=1}^n (x_i - \bar{x})^r}{n}$$

Que como podemos comprobar son estadísticos, y por tanto variables aleatorias.

Entre estos momentos, los más interesantes son la media muestral y la varianza muestral, a_1 y m_2 . Vamos a estudiar su esperanza y su varianza.

4.5.1. Esperanza y varianza de la media muestral.

4.5.1.1. Esperanza de la media muestral.

$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} n E(\xi) = E(\xi)$$

Ya que al ser una muestra aleatoria simple, los elementos de la muestra se distribuyen igual que la población y tendrán su misma esperanza. Por tanto, este resultado es válido para cualquier muestra aleatoria simple independientemente de la población de la que proceda.

4.5.1.2. Varianza de la media muestral.

$$V(\bar{x}) = V\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n x_i\right) = \frac{1}{n^2} n E(\xi^2) - \left(E(\xi)\right)^2 = \frac{1}{n} (E(\xi^2) - (E(\xi))^2) = \frac{\sigma^2}{n}$$

Y como las variables son independientes entre sí, por ser MAS:

$$\frac{1}{n^2} V\left(\sum_{i=1}^n x_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(x_i) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

Como podemos ver, siempre que la varianza de la población sea finita, la media de la muestra estará más concentrada en torno a la media poblacional cuanto mayor sea el tamaño de la muestra. En consecuencia, cuanto mayor sea el tamaño de la muestra más confianza tenemos en que la media muestral sea una buena estimación de la media poblacional.

Otro hecho de relevancia es que, como consecuencia del Teorema Central del Límite, independientemente del modelo de distribución de la variable poblacional, para tamaños muestrales elevados la media muestral aleatoria tiende a distribuirse como una normal con la esperanza y varianza anteriormente expuestas.

4.5.2. Esperanza y varianza de la varianza muestral.

4.5.2.1. Esperanza de la varianza muestral.

$$\begin{aligned}
 m_2 = S_x^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i - \mu + \mu - \bar{x})^2}{n} = \\
 &= \frac{\sum_{i=1}^n (x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2}{n} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} - (\bar{x} - \mu)^2 \\
 E(S_x^2) &= E\left(\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}\right) - E((\bar{x} - \mu)^2) = V(\xi) - \frac{v(\xi)}{n} = \frac{n-1}{n}\sigma^2
 \end{aligned}$$

4.5.2.2. Varianza de la varianza muestral.

Esta demostración es bastante engorrosa. El resultado final es:

$$V(S_x^2) = \frac{\mu_4 - \sigma^4}{n} - 2\frac{\mu_4 - 2\sigma^4}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}$$

4.6. Teorema de Glivenco-Cantelli.

Teorema 18. Sea (X_1, \dots, X_n) una muestra aleatoria simple obtenida de una población con función de distribución $F(x)$. Sea $F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i)$ su función de distribución empírica. entonces se cumple que

$$\sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| \xrightarrow{c.s.} 0$$

El Teorema nos dice que si definimos una banda de amplitud ε arbitrariamente estrecha entorno a la distribución de probabilidad teórica de la población en estudio, para un tamaño de muestra suficientemente grande podemos asegurar con probabilidad 1 que la función de distribución empírica estará contenida en esa banda.

Capítulo 5

Distribuciones en el muestreo asociadas con poblaciones normales.

Distribuciones de la media, varianza y diferencia de medias. Estadísticos ordenados. Distribución del mayor y menor valor. Distribución del recorrido.

5.1. Introducción.

La distribución normal está presente en muchos ámbitos de la ciencia, la economía y la ingeniería. Además, por el Teorema Central del Límite, sabemos que la suma de variables aleatorias independientes idénticamente distribuidas con media y varianza finitas converge en distribución a una normal. Por tanto, todos aquellos fenómenos que resulten de la adición de un gran número de efectos aleatorios que cumplan esta condición podrán describirse, al menos en primera aproximación, mediante una distribución normal.

5.2. Distribuciones en el muestreo asociadas con poblaciones normales.

Sabemos que la función característica de una suma de variables aleatorias independientes coincide con el producto de sus funciones características. También sabemos que $\varphi_{c\varepsilon}(t) = \varphi_{\varepsilon}(ct)$. Calculemos la función característica de una combinación lineal de variables normales, (X_1, \dots, X_n) :

$$\begin{aligned}\varphi_{a_1 X_1 + a_2 X_2 + \dots + a_n X_n}(t) &= \varphi_{X_1}(a_1 t) \varphi_{X_2}(a_2 t) \dots \varphi_{X_n}(a_n t) \\ \varphi_{a_1 X_1 + a_2 X_2 + \dots + a_n X_n}(t) &= e^{ia_1 t \mu - \frac{1}{2} \sigma^2 a_1^2 t^2} e^{ia_2 t \mu - \frac{1}{2} \sigma^2 a_2^2 t^2} \dots e^{ia_n t \mu - \frac{1}{2} \sigma^2 a_n^2 t^2} \\ \varphi_{a_1 X_1 + a_2 X_2 + \dots + a_n X_n}(t) &= e^{it(a_1 + a_2 + \dots + a_n)\mu - \frac{1}{2}(a_1^2 + a_2^2 + \dots + a_n^2)\sigma^2 t^2}\end{aligned}$$

Y por tanto una combinación lineal de variables aleatorias normales es otra variable aleatoria normal, así cuando tengamos estadísticos de poblaciones normales que sean combinación lineal de las observaciones, conoceremos su función de distribución muestral.

5.3. Distribuciones de la media, varianza y diferencia de medias.

5.3.1. Distribución de la media muestral de una población $N(\mu, \sigma)$.

5.3.1.1. Varianza Poblacional conocida.

Sabemos que $E(\bar{X}) = \mu$ y $Var(\bar{X}) = \frac{\sigma^2}{n}$ para cualquier población, sea o no normal.

Como la población es normal, $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ y por tanto:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Capítulo 6

Estimación puntual I.

Propiedades de los estimadores puntuales. Error cuadrático medio. Estimadores insesgados, consistentes y suficientes.

6.1. Introducción.

Una parte de la inferencia estadística consiste en obtener estimaciones acerca de los parámetros que definen la distribución de probabilidad de una población. Así, si tenemos una población normal, sin conocer los valores de μ y σ^2 no podremos calcular las probabilidades de los distintos sucesos, ni podremos realizar deducciones sobre la población.

La estimación de un parámetro consistirá en utilizar los datos muestrales en combinación con algún estadístico. Hay dos formas de llevar a cabo esta tarea: mediante la **estimación puntual**, en la que buscamos un estimador que en conjunción con los datos muestrales nos de una estimación univaluada del parámetro, y la **estimación por intervalos**, en la que definimos un intervalo dentro del cual, de forma probable, se encontrará el parámetro.

Formalmente, sea una variable aleatoria, φ , cuya función de distribución, $F(x; \theta)$ depende del parámetro θ definido en el espacio paramétrico Θ , la estimación puntual busca encontrar un estadístico que nos permita estimar a partir de una muestra aleatoria el valor de θ .

A este estadístico que va a utilizar para estimar θ lo llamamos **estimador**, y lo representamos por $\hat{\theta}$. Este estimador será una función de las variables aleatorias que forman la muestra, y debe quedar completamente definido una vez se produce la realización de la muestra.

Dado que para estimar un mismo parámetro podemos definir infinitud de estimadores, será necesario por un lado, establecer que propiedades es deseable que tenga un estimador para ser útil a nuestro propósito, y por otro descubrir que procedimientos nos permiten obtener estimadores que cumplan esas propiedades deseables.

Así, un estimador será una variable aleatoria, función de las variables aleatorias muestrales. Una estimación será una realización de esa variable aleatoria para una muestra determinada.

6.2. Propiedades de los estimadores.

Hemos visto que un estimador es un estadístico función de las variables aleatorias muestrales, y por tanto él mismo será una variable aleatoria con su función de distribución, su media y su varianza. De entre todos los estadísticos posibles, nos interesará utilizar como estimador aquel que nos produzca las mejores estimaciones del parámetro desconocido. Para ello definimos el **error cuadrático medio**, que utilizaremos como medida de la bondad del estimador.

Definición 17. Llamamos **error cuadrático medio** del estimador $\hat{\theta}$, y lo denotamos por $ECM(\hat{\theta})$ como el valor esperado del cuadrado de la diferencia entre el estimador $\hat{\theta}$ y el valor real de parámetro θ , es decir:

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

Si desarrollamos esta expresión:

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

$$ECM(\hat{\theta}) = V(\hat{\theta}) + \left(\theta - E\left[\hat{\theta}\right]\right)^2$$

Y vemos que el error cuadrático medio se tiene dos componentes:

- La varianza del estimador.
- El cuadrado de la diferencia entre el valor real del parámetro y la esperanza del estimador.

A la diferencia entre la esperanza del estimador y el valor real del parámetro la llamaremos **s sesgo del estimador**.

Parecería que lo que debemos buscar, por tanto, es un estimador que minimice el error cuadrático medio. Sin embargo, esto no es tan sencillo. Dejando aparte la dificultad de calcular el *ECM* de todos los estimadores posibles, normalmente este depende del valor del parámetro a estimar, y suele ocurrir que no exista ningún estimador que lo minimice para todos los posibles valores del parámetro. Por tanto, deberemos buscar otros criterios.

A partir del error cuadrático medio podemos deducir que propiedades es deseable que tenga un estimador. Así, vemos que para que el error cuadrático medio sea pequeño la varianza del estimador ha de ser pequeña, y su esperanza debe estar lo más cercana posible al valor real del parámetro, a ser posible debe coincidir con este. A la propiedad de que el estimador tenga varianza mínima se le conoce como **eficiencia** del estimador. Si la esperanza de un estimador coincide con el valor del parámetro que estima, se dice que es **insesgado**, en otro caso se dirá que el estimador es sesgado. Así, buscaremos estimadores insesgados cuya varianza sea lo más pequeña posible.

Por otro lado, dado que la estimación se obtiene a partir de una muestra, esta debe ser lo más representativa posible de la población en estudio. Esto también se puede alcanzar incrementando el tamaño de la muestra hasta el límite en que dicho tamaño coincide con el tamaño de la población, en cuyo caso el *ECM* será cero, ya que solo hay una estimación que coincide con el valor del parámetro. Por tanto, parece lógico exigir que cuanto mayor sea el tamaño de la muestra mayor probabilidad haya de que el estimador esté próximo al valor del parámetro. Esta propiedad se conoce como **consistencia**.

Además de estas propiedades, nos encontramos con otras tres, no inmediatas, pero importantes: **suficiencia**, **invarianza** y **robustez**.

La propiedad de suficiencia refleja el hecho de que al estimar nuestro parámetro estamos resumiendo la información contenida en la muestra en un único valor, con la esperanza de que este valor conserve toda la información contenida en la muestra. Esta situación no se da siempre, pero cuando se da, se dice que nuestro estimador es **suficiente**.

La propiedad de invarianza refleja la conveniencia de que obtenido una estimador de un parámetro, el estimador de una función del parámetro sea la función del estimador original.

En cuanto a la robustez, normalmente al estimar un parámetro debemos realizar una serie de hipótesis sobre la población en estudio. Un estimador es robusto si desviaciones de las hipótesis iniciales no afectan a la bondad del estimador, o lo hacen de forma débil.

6.3. Estimadores insesgados, consistentes y suficientes.

6.3.1. Estimadores insesgados.

La función de densidad en el muestreo de un estimador dependerá del parámetro o parámetros poblacionales, por lo que la esperanza matemática del estimador será:

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \hat{\theta}g(\hat{\theta}; \theta)d\hat{\theta}$$

Como el estimador es función de los elementos muestrales, y la densidad conjunta de una muestra aleatoria simple de tamaño n será $f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta)$, la esperanza del estimador se puede calcular también como

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{\theta}(x_1, \dots, x_n)f(x_1; \theta) \cdots f(x_n; \theta)dx_1 \cdots dx_n = \int_{\mathbf{X}} \hat{\theta}(\mathbf{X})L(\mathbf{X}; \theta)d\mathbf{X}$$

Así, en general podemos expresar la esperanza matemática del estimador como:

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

Ya hemos visto que $b(\theta)$ recibe el nombre de **sesgo** del estimador. Si es cero para todos los posibles valores de θ , diremos que el estimador es insesgado. Si es mayor que cero, diremos que el estimador tiene un sesgo positivo y por tanto sobreestima el parámetro. Si es menor que cero, diremos que el estimador tiene un sesgo negativo y por tanto subestima el parámetro.

Si un estimador es insesgado, no existirá error sistemático si lo utilizamos para estimar un parámetro. Un estimador es **asintóticamente insesgado** cuando $b(\hat{\theta} \rightarrow \theta) \rightarrow 0$ cuando $n \rightarrow \infty$. La insesgadez es una propiedad del estimador, no de una estimación concreta. Para verificar si un estimador es insesgado solo hay que calcular su esperanza matemática.

6.3.1.1. Propiedades de los estimadores insesgados.

1. Si dos estimadores $\hat{\theta}_1, \hat{\theta}_2$ de un mismo parámetro son insesgados, entonces para cualquier número c tal que $c \in (0, 1)$ el estimador definido por $\hat{\theta} = c\hat{\theta}_1 + (1 - c)\hat{\theta}_2$ es insesgado.

$$E(\hat{\theta}) = E[c\hat{\theta}_1 + (1 - c)\hat{\theta}_2] = cE(\hat{\theta}_1) + (1 - c)E(\hat{\theta}_2) = \theta$$

2. El momento muestral de orden r respecto al origen es un estimador insesgado del momento poblacional respecto al origen del mismo orden.

$$E(a_r) = E\left[\frac{1}{n} \sum_{i=1}^n x_i^r\right] = \frac{1}{n} \sum_{i=1}^n E(x_i^r) = \frac{1}{n} \sum_{i=1}^n \alpha_r = \alpha_r$$

3. El estimador $\hat{\mu} = \sum_{i=1}^n c_i x_i$ es un estimador insesgado de la media poblacional siempre que $\sum_{i=1}^n c_i = 1$.

$$E\left[\sum_{i=1}^n c_i x_i\right] = \sum_{i=1}^n c_i E(x_i) = \mu \sum_{i=1}^n c_i = \mu \Leftrightarrow \sum_{i=1}^n c_i = 1$$

6.3.2. Estimadores consistentes.

A medida que el tamaño de la muestra aumenta, tenemos cada vez más información acerca de la población. Por tanto, será deseable utilizar estimadores cuya bondad aumente a medida que aumenta el tamaño de la muestra. Bajo este concepto es bajo el que se sitúa la propiedad de la consistencia de un estimador. Para ello nos basaremos en los criterios de convergencia de variables aleatorias.

Definición 18. Diremos que una sucesión de estimadores de un parámetro θ , $\{\hat{\theta}_n\}$ es consistente si converge en probabilidad hacia el valor del parámetro θ , es decir, si:

$$\lim_{n \rightarrow \infty} P\left(|\hat{\theta}_n - \theta| \leq \varepsilon\right) = 1$$

O, de forma equivalente

$$\lim_{n \rightarrow \infty} P\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$$

Para todos los valores posibles de θ y para todo $\varepsilon > 0$.

Si definimos la sucesión de estimadores como el mismo estimador para tamaños cada vez mayores de muestra, la definición implica que para un estimador consistente al aumentar la muestra aumenta la probabilidad de que el valor de la estimación esté muy cercano al valor del parámetro a estimar. Es decir, la varianza del estimador disminuirá y su sesgo, si lo tiene, también será cada vez menor. Más formalmente, si consideramos la desigualdad de Tchebichev:

$$P\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \leq \frac{E(\hat{\theta}_n - \theta)^2}{\varepsilon^2}$$

Y como

$$E(\hat{\theta}_n - \theta)^2 = V(E(\hat{\theta}_n)) + b^2(E(\hat{\theta}_n))$$

Sustituyendo y tomando límites:

$$\begin{aligned} \lim_{n \rightarrow \infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) &\leq \lim_{n \rightarrow \infty} \frac{E(\hat{\theta}_n - \theta)^2}{\varepsilon^2} \\ \lim_{n \rightarrow \infty} E(\hat{\theta}_n - \theta)^2 &= \lim_{n \rightarrow \infty} V(\hat{\theta}_n) + \lim_{n \rightarrow \infty} b^2(\hat{\theta}_n) \end{aligned}$$

Y por tanto, para que se cumpla que $\lim_{n \rightarrow \infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0$ es suficiente que $\lim_{n \rightarrow \infty} V(\hat{\theta}_n) = 0$ y $\lim_{n \rightarrow \infty} b(\hat{\theta}_n) = 0$. Esta condición no es necesaria.

6.3.2.1. Propiedades de los estimadores consistentes.

1. Si $\hat{\theta}$ es un estimador consistente de θ , y sea g una función continua, entonces $g(\hat{\theta})$ es un estimador consistente de $G(\theta)$.
2. Los momentos muestrales con respecto al origen son estimadores consistentes de sus correspondientes momentos poblacionales.
3. Los momentos muestrales centrales son estimadores consistentes de sus correspondientes momentos poblacionales.

Estimador óptimo asintóticamente normal

6.3.3. Estimadores suficientes.

Hemos visto que los estimadores no son más que estadísticos que utilizamos para resumir la información que está presente en nuestra muestra aleatoria simple. Cabe preguntarse por tanto, si al efectuar ese resumen no estaremos perdiendo alguna parte de la información que contiene la muestra sobre el parámetro a estimar. Esto nos lleva a definir el concepto de suficiencia: intuitivamente, un estimador es suficiente si contiene toda la información acerca del parámetro a estimar que está presente en la muestra original. Claramente será deseable trabajar con estimadores suficientes.

Un estimador resume toda la información presente en la muestra acerca de un parámetro si una vez fijado el valor del estimador, la posible variabilidad de la muestra no está ligada al parámetro en cuestión, más formalmente, si la distribución de probabilidad de la muestra condicionada al valor del estimador no depende del parámetro a estimar. Formalmente:

Definición 19. Sea (X_1, \dots, X_n) una muestra aleatoria simple que proviene de una población cuya distribución de probabilidad depende de un parámetro θ desconocido. Diremos que el estadístico o estimador $T = T(X_1, \dots, X_n)$ es suficiente para el parámetro θ si la distribución condicionada de (X_1, \dots, X_n) dado el valor del estadístico $T = t$ no depende del valor del parámetro θ .

Esta definición nos proporciona una forma de comprobar si un estimador es suficiente, pero no nos permite encontrar uno. El teorema de factorización de Fischer-Neyman nos permite comprobar si un estadístico es suficiente de forma más sencilla, además de permitirnos encontrar un estimador suficiente.

Teorema 19. Teorema de factorización de Fischer-Neymann:

Sea (X_1, \dots, X_n) una muestra aleatoria simple que proviene de una población con función de distribución $F(x; \theta)$ y sea la función de cuantía de la muestra $P(x_1, \dots, x_n; \theta) = P_\theta(X_1 = x_1, \dots, X_n = x_n)$, o la función de densidad de la muestra $f(x_1, \dots, x_n; \theta)$, entonces el estadístico $T = T(X_1, \dots, X_n)$ es suficiente para el parámetro θ si y solo si se puede escribir:

$$P(x_1, \dots, x_n; \theta) = g(T(X_1, \dots, X_n); \theta) \cdot h(x_1, \dots, x_n)$$

o

$$f(x_1, \dots, x_n; \theta) = g(T(X_1, \dots, X_n); \theta) \cdot h(x_1, \dots, x_n)$$

donde g depende de θ y de la muestra a través del estadístico T y h solo depende de la muestra.

Teorema 20. Si el estadístico T_1 es suficiente y es función con inversa única del estadístico T_2 , $T_1 = f(T_2)$, entonces el estadístico T_2 también es suficiente.

Teorema 21. Si los estadísticos T_1 y T_2 son suficientes, están relacionados funcionalmente.

Cuando la población de estudio depende de dos parámetros, es interesante determinar dos estadísticos que sean conjuntamente suficientes para los dos parámetros, es decir, que entre ambos resuman la totalidad de información de la muestra para ambos parámetros. En este caso, el teorema se puede escribir así:

Teorema 22. Teorema de factorización de Fischer-Neymann:

Sea (X_1, \dots, X_n) una muestra aleatoria simple que proviene de una población con función de distribución $F(x; \theta_1, \theta_2)$ y sea la función de cuantía de la muestra $P(x_1, \dots, x_n; \theta_1, \theta_2) = P_{\theta_1, \theta_2}(X_1 = x_1, \dots, X_n = x_n)$, o la función de densidad de la muestra $f(x_1, \dots, x_n; \theta_1, \theta_2)$, entonces los estadísticos $T_1 = T_1(X_1, \dots, X_n)$ y $T_2 = T_2(X_1, \dots, X_n)$ son conjuntamente suficientes para los parámetros θ_1 y θ_2 si y solo si se puede escribir:

$$P(x_1, \dots, x_n; \theta_1, \theta_2) = g(T_1(X_1, \dots, X_n), T_2(X_1, \dots, X_n); \theta_1, \theta_2) \cdot h(x_1, \dots, x_n)$$

o

$$f(x_1, \dots, x_n; \theta_1, \theta_2) = g(T_1(X_1, \dots, X_n), T_2(X_1, \dots, X_n); \theta_1, \theta_2) \cdot h(x_1, \dots, x_n)$$

6.3.3.1. Estadístico minimal suficiente.

El concepto de suficiencia nos permite buscar un estadístico que contenga toda la información presente en la muestra acerca del parámetro a estimar. Ahora buscamos el **estadístico minimal suficiente**, entendiendo por éste un estadístico que resuma la información contenida en la muestra lo más posible, pero que siga siendo suficiente.

Definición 20. Diremos que un estadístico es **minimal suficiente** para un parámetro, θ , si es suficiente y cualquier reducción de la información definida por él ya no es suficiente.

Método de Lehmann y Scheffé para obtener un estadístico minimal suficiente: Si partimos de dos muestras aleatorias simples de igual tamaño, (X_1, \dots, X_n) e (Y_1, \dots, Y_n) , cuyas respectivas funciones de verosimilitud son:

$$L(x_1, \dots, x_n; \theta) = f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

$$L(y_1, \dots, y_n; \theta) = f(y_1, \dots, y_n; \theta) = \prod_{i=1}^n f(y_i; \theta)$$

Si podemos encontrar una función $g(x_1, \dots, x_n)$ tal que el cociente de las verosimilitudes no dependa de θ si y solo si $g(x_1, \dots, x_n) = g(y_1, \dots, y_n)$, entonces $g(x_1, \dots, x_n)$ es el estimador minimal suficiente de θ .

En el caso de tener k parámetros deberíamos encontrar k funciones para las que el cociente de verosimilitudes no dependa de los parámetros si y solo si $g_i(x_1, \dots, x_n) = g_i(y_1, \dots, y_n)$ para todo i .

6.3.3.2. Relación entre eficiencia y suficiencia.

Sabemos que si un estimador $\hat{\theta}$ es insesgado y su varianza alcanza la cota de Cramer-Rao, se verifica que

$$\frac{\partial \ln dF_n(x_1, \dots, x_n; \theta)}{\partial \theta} = A(\theta)(\hat{\theta} - \theta)$$

Si definimos:

$$\frac{\partial \ln g(\hat{\theta}, \theta)}{\partial \theta} = A(\theta)(\hat{\theta} - \theta)$$

Tendremos:

$$\frac{\partial \ln dF_n(x_1, \dots, x_n; \theta)}{\partial \theta} = \frac{\partial \ln g(\hat{\theta}, \theta)}{\partial \theta}$$

si integramos y expresamos la constante de integración como $\ln h(x_1, \dots, x_n)$, tenemos:

$$\ln dF_n(x_1, \dots, x_n; \theta) = \ln g(\hat{\theta}, \theta) + \ln h(x_1, \dots, x_n)$$

y por tanto:

$$dF_n(x_1, \dots, x_n; \theta) = g(\hat{\theta}, \theta)h(x_1, \dots, x_n)$$

que, aplicando el criterio de factorización de Fischer-Neymann, nos dice que el estimador es suficiente. Es decir, un estimador eficiente e insesgado es siempre suficiente.

6.3.3.3. Estimadores suficientes y estimadores UMVUE.

La suficiencia desempeña un papel importante en la obtención de estimadores insesgados uniformemente de mínima varianza.

Teorema 23. Teorema de Rao-Blackwell:

Sea una población con función de densidad representada por $f(x; \theta)$, sea $\hat{\theta}$ un estimador insesgado del parámetro θ y sea T un estadístico suficiente del mismo parámetro. Entonces, si definimos $g(T) = E[\hat{\theta}/T]$ se verifica:

- $g(T)$ es un estadístico, y es función del estadístico suficiente.
- $E[g(T)] = \theta$.
- $V(g(T)) \leq V(\hat{\theta})$.

Es decir, el estadístico $g(T)$ es función del estadístico suficiente, es un estimador insesgado de θ y su varianza es menos que la del estimador original.

Así, si tenemos un estimador insesgado y un estadístico suficiente, podemos usarlos para obtener un estimador insesgado de menor varianza.

Capítulo 7

Estimación puntual II.

Estimadores de mínima varianza. Estimadores eficientes. Estimadores robustos. Estimadores Bayesianos.

7.1. Introducción.

Una parte de la inferencia estadística consiste en obtener estimaciones acerca de los parámetros que definen la distribución de probabilidad de una población. Así, si tenemos una población normal, sin conocer los valores de μ y σ^2 no podremos calcular las probabilidades de los distintos sucesos, ni podremos realizar deducciones sobre la población.

La estimación de un parámetro consistirá en utilizar los datos muestrales en combinación con algún estadístico. Hay dos formas de llevar a cabo esta tarea: mediante la **estimación puntual**, en la que buscamos un estimador que en conjunción con los datos muestrales nos de una estimación univaluada del parámetro, y la **estimación por intervalos**, en la que definimos un intervalo dentro del cual, de forma probable, se encontrará el parámetro.

Formalmente, sea una variable aleatoria, φ , cuya función de distribución, $F(x; \theta)$ depende del parámetro θ definido en el espacio paramétrico Θ , la estimación puntual busca encontrar un estadístico que nos permita estimar a partir de una muestra aleatoria el valor de θ .

A este estadístico que va a utilizar para estimar θ lo llamamos **estimador**, y lo representamos por $\hat{\theta}$. Este estimador será una función de las variables aleatorias que forman la muestra, y debe quedar completamente definido una vez se produce la realización de la muestra.

Dado que para estimar un mismo parámetro podemos definir infinidad de estimadores, será necesario por un lado, establecer que propiedades es deseable que tenga un estimador para ser útil a nuestro propósito, y por otro descubrir que procedimientos nos permiten obtener estimadores que cumplan esas propiedades deseables.

Así, un estimador será una variable aleatoria, función de las variables aleatorias muestrales. Una estimación será una realización de esa variable aleatoria para una muestra determinada.

7.2. Estimadores de mínima varianza.

Sabemos que a la hora de utilizar un estimador puntual para estimar un parámetro asociado a una población que queremos estudiar, una buena medida de la bondad de ese estimador es su error cuadrático medio, es decir, $ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$, siendo mejor estimador aquel cuyo ECM sea menor. Como sabemos que el error cuadrático medio se puede descomponer como la suma de la varianza del estimador y el cuadrado de su sesgo, nos interesará buscar estimadores que, además de ser insesgados, tengan una varianza lo menor posible.

Sabemos que en general no es posible encontrar un estimador que minimice el error cuadrático medio para todos los valores posibles del parámetro. Sin embargo, si que se puede buscar aquel estimador que, siendo insesgado, minimice el error cuadrático medio. Esto equivale a encontrar el estimador insesgado que minimice la varianza. A este estimador se le llama **estimador insesgado de varianza mínima**.

Si además se verifica que la varianza es mínima para todos los posibles valores del parámetro a estudiar, entonces el estimador recibe el nombre de **estimador insesgado uniformemente de mínima varianza** (UMVUE por sus siglas en inglés.)

Definición 21. Diremos que un estimador insesgado para un parámetro θ , $\hat{\theta}_0$ es **insesgado y uniformemente de mínima varianza** si dado cualquier otro estimador insesgado de θ , $\hat{\theta}$ se verifica que $V(\hat{\theta}_0) \leq V(\hat{\theta})$ para todos los valores posibles de θ .

Para obtener el estimador insesgado y uniformemente de mínima varianza tendríamos que calcular las varianzas de todos los estimadores posibles de nuestro parámetro, lo cual es claramente inasumible. Para ayudarnos en esta tarea disponemos de la cota de Fretchet-Cramer-Rao, que nos da una cota inferior a la varianza del estimador.

Teorema 24. Sea (X_1, \dots, X_n) una muestra aleatoria simple proveniente de una población con una densidad de probabilidad $f(x; \theta)$. Designamos la función de densidad conjunta de la muestra por:

$$L(x_1, \dots, x_n; \theta) = dF_n(x_1, \dots, x_n; \theta) = f_n(x_1, \dots, x_n; \theta)$$

y sea $\hat{\theta}$ un estimador insesgado de θ , entonces si se verifican las condiciones de regularidad de Wolfowitz la varianza del estimador está acotada inferiormente según la siguiente desigualdad:

$$V(\hat{\theta}) \geq \frac{1}{E \left[\left(\frac{\partial \ln dF_n}{\partial \theta} \right)^2 \right]}$$

o bien, si la función de densidad de la población es $f(x; \theta)$

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

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

Las **condiciones de regularidad de Wolfowitz** son:

- El intervalo de variación de θ , D , es un intervalo abierto del eje real que nunca se reduce a un punto.
- El campo de variación de la variable aleatoria X que define la población no depende del parámetro θ .
- Para casi todo X y todo θ existe $\frac{\partial \ln dF_n}{\partial \theta}$.
- Se pueden diferenciar bajo el signo integral las expresiones $E[1]$ y $E[\hat{\theta}]$.
- Se verifica que:

$$E \left[\left(\frac{\partial \ln dF_n}{\partial \theta} \right)^2 \right] > 0, \forall \theta \in D$$

Fischer llamó a la expresión $E \left[\left(\frac{\partial \ln dF_n}{\partial \theta} \right)^2 \right]$ **cantidad de información de la muestra**, es decir, la cantidad de información que una muestra de tamaño n proporciona sobre el parámetro, medida en el sentido de la varianza de la variabilidad de la función de densidad respecto al parámetro.

Si el estimador hubiera sido insesgado, la cota sería la siguiente:

$$V(\hat{\theta}) \geq \frac{1 + \frac{\partial b(\hat{\theta})}{\partial \theta}}{E \left[\left(\frac{\partial \ln dF_n}{\partial \theta} \right)^2 \right]}$$

La cota FCR nos da un límite inferior para la varianza del estimador, pero esto no implica que la varianza de un estimador UMVUE sea igual a la de la cota. Es decir, puede haber un estimador UMVUE cuya varianza sea mayor que la cota de FCR.

7.3. Estimadores eficientes.

La insesgadez es una propiedad deseable en un estimador, pero por si sola no es suficiente para determinar si un estimador es útil, ya que solo requiere que el valor esperado del estimador sea igual al parámetro a estimar, y no requiere que haya valores del estimador estén próximos al mismo. Así, es deseable que los valores que tome el estimador para distintas muestras valores próximos unos de otros, de tal manera que su varianza sea pequeña. Así, ante dos estimadores insesgados, resultará más fiable aquel que presente una menor varianza. Definimos la eficiencia de un estimador comparando su varianza con la varianza de los demás estimadores insesgados. Así, el **estimador más eficiente** de un conjunto de estimadores insesgados será aquel que tenga una varianza menor.

Definición 22. Estimador eficiente:

*Diremos que un estimador $\hat{\theta}$ del parámetro poblacional θ es **eficiente** si es insesgado y además su varianza alcanza la cota de Fretcher-Cramer-Rao.*

Por tanto, un estimador eficiente será un estimador insesgado y uniformemente de mínima varianza cuya varianza coincide con la cota inferior de FCR. Estos estimadores son muy útiles en toda la inferencia, por lo que se intentarán obtener siempre que existan.

Definición 23. *Se define la **eficiencia de un estimador insesgado** $\hat{\theta}$ del parámetro poblacional θ como:*

$$\text{eff.}(\hat{\theta}) = \frac{\text{Cota F.C.R.}}{V(\hat{\theta})}$$

verificándose que $\text{eff.}(\hat{\theta}) \leq 1$

Así, dados dos estimadores del mismo parámetro, el más eficiente será el que tenga una eficiencia mayor. También podemos introducir el concepto de **eficiencia relativa**:

$$\text{eff. relativa}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)} = \frac{\text{eff.}(\hat{\theta}_1)}{\text{eff.}(\hat{\theta}_2)}$$

Si $\text{eff. relativa}(\hat{\theta}_1, \hat{\theta}_2) > 1$ $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$. Si $\text{eff. relativa}(\hat{\theta}_1, \hat{\theta}_2) < 1$ $\hat{\theta}_2$ es más eficiente que $\hat{\theta}_1$. Si $\text{eff. relativa}(\hat{\theta}_1, \hat{\theta}_2) = 1$ ambos estimadores son igual de eficientes.

Teorema 25. *Si un estimador $\hat{\theta}$ es insesgado, su varianza alcanza la cota de Fretcher-Cramer-Rao si se verifica:*

$$\frac{\partial \ln dF_n}{\partial \theta} = A(\theta)(\hat{\theta} - \theta)$$

siendo $A(\theta)$ una expresión que no depende de $\hat{\theta}$.

Teorema 26. *Si un estimador $\hat{\theta}$ es eficiente, entonces se verifica que:*

$$V(\hat{\theta}) = \frac{1}{A(\theta)}$$

Definición 24. *Diremos que un estimador $\hat{\theta}$ es asintóticamente eficiente si se cumple que:*

$$\lim_{n \rightarrow \infty} V(\hat{\theta}) = \text{Cota de Frechet-Cramer-Rao}$$

7.4. Estimadores robustos.

Un procedimiento estadístico es robusto si su comportamiento es relativamente insensible a desviaciones sobre las hipótesis iniciales en la que se haya basado su desarrollo. Dado que para encontrar estimadores normalmente necesitamos hacer hipótesis sobre la población en estudio, es conveniente contar con estimadores que sean lo más robustos posibles. Por ejemplo, es frecuente suponer la distribución de probabilidad de la variable aleatoria en estudio conocida, pero en general esto no es así, sino que se formula una hipótesis acerca de la misma. Si la distribución de probabilidad es distinta de la supuesta, y esta diferencia no es muy significativa y el procedimiento estadístico es insensible a estos cambios, se dice que el estimador es robusto.

Definición 25. *Diremos que un estimador es **robusto** cuando pequeños cambios en las hipótesis de partida del procedimiento de estimación no producen variaciones significativas en los resultados obtenidos.*

7.5. Estimadores Bayesianos.

Hasta ahora hemos estudiado la estimación puntual desde el punto de vista de la teoría del muestreo, que se basa en interpretar la probabilidad como una frecuencia relativa. Pasaremos ahora a estudiar el enfoque bayesiano de la inferencia estadística, en lo que se refiere a la estimación de parámetros.

En el enfoque bayesiano, un parámetro es visto como una variable aleatoria a la que se asigna una distribución de probabilidad a priori con base en el grado de creencia sobre la distribución del mismo, que se modifica con la información obtenida de la muestra, para obtener la distribución a posteriori. Con esta distribución a posteriori formularemos inferencias respecto al parámetro. Este enfoque resulta muy útil en aquellas situaciones en las que el parámetro a estimar no puede considerarse una cantidad fija, sino que puede variar dependiendo de las características del entorno.

Dado que consideramos el parámetro a estimar como una variable aleatoria, lo designamos por Θ , y por θ a la realización de dicha variable aleatoria. Suponemos que Θ es una variable aleatoria continua con una función de densidad incondicional a priori $f_{\Theta}(\theta)$, la cual refleja las creencias previas acerca de Θ . Si tomamos una muestra aleatoria simple de tamaño n (n variables aleatorias idénticamente distribuidas), X_1, \dots, X_n su función de densidad condicionada común será $f(x|\theta)$, y la función de densidad conjunta:

$$L(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta) \cdots f(x_n|\theta)$$

Como decimos que Θ es una variable aleatoria, el objetivo es estimar el valor particular θ para el cual la evidencia muestral que representa la densidad conjunta se encuentra condicionada. Por tanto, la función de densidad a posteriori de Θ será, aplicando el teorema de Bayes:

$$f(\theta|x_1, x_2, \dots, x_n) = \frac{L(x_1, x_2, \dots, x_n|\theta)f_{\Theta}(\theta)}{\int_{\Theta} L(x_1, x_2, \dots, x_n|\theta)f_{\Theta}(\theta)d\theta}$$

El denominador de esta expresión se denomina distribución predictiva, y representa la ponderación de todas las distribuciones posibles del parámetro, ponderados según la importancia que de a cada una la distribución a priori.

En la práctica, el cálculo se simplifica si observamos que el denominador no depende de θ , y actúa solo como constante normalizadora de la distribución a posteriori, para que su integral sea la unidad. Por tanto, vemos que la distribución a posteriori es proporcional a la distribución a priori multiplicada por la verosimilitud de la muestra. Por tanto, la distribución a posteriori combina la información previa de la que se dispone, representada por la distribución a priori, con la información aportada por la muestra. Si la distribución a priori es más o menos constante sobre el espacio paramétrico, la distribución a priori coincide con la verosimilitud, y se dice que la distribución a priori es no informativa.

Para tamaños muestrales grandes, se puede demostrar que en condiciones muy generales la distribución a posteriori está dominada por la verosimilitud, y adquiere una distribución aproximadamente normal con media y varianza coincidentes con la del estimador de máxima verosimilitud. En consecuencia, en estos casos el estimador bayesiano y el de máxima verosimilitud conducen a los mismos resultados.

Para obtener una estimación de θ necesitamos elegir una característica numérica de la distribución a posteriori que nos parezca representativa de la misma. Hay dos opciones:

- Elegir como estimación la moda de la distribución a posteriori, que es el valor más probable una vez observada la muestra. Esta situación tiene la misma justificación que la estimación de máxima verosimilitud en el contexto clásico.
- Elegir una función de pérdida que represente la consecuencia de haber escogido un valor de θ erróneo. Esta función debe ser una función no negativa de θ y su estimación, de manera que sea cero si coinciden.

Al depender también de θ , la función de pérdida también es una variable aleatoria. El estimador bayesiano del parámetro será aquél que minimice la esperanza de la función de pérdida.

Es obvio que para poder estimar el parámetro se debe especificar una función de pérdida. Esto es una tarea difícil, ya que las consecuencias no son siempre medibles. En muchos casos una función de pérdida razonable puede ser la forma cuadrática: $l(\theta, t) = (t - \theta)^2$. Para esta forma, se puede demostrar que el estimador de Bayes equivale a la distancia a posteriori de Θ .

Capítulo 8

Métodos de estimación.

Método de los momentos. Método de la mínima X2. Método de la mínima varianza. Método de los mínimos cuadrados. Métodos Bayesianos.

8.1. Introducción.

Dentro del proceso de inferencia estadística sobre una población, en el que queremos obtener estimadores para los parámetros que caracterizan esa población, sabemos que hay una serie de propiedades deseables en esos estimadores (insesgadez, consistencia, eficiencia). Otro problema es cómo obtener estimadores que presenten estas propiedades. Para ello veremos varios métodos de obtención de estimadores, y revisaremos que propiedades cumplen los estimadores obtenidos mediante esos métodos.

8.2. Método de los momentos.

Este método fue introducido por K. Pearson, y es el método general más antiguo. Consiste en igualar tantos momentos poblacionales como parámetros haya que estimar a sus correspondientes momentos muestrales, y resolver el sistema de ecuaciones así resultante para obtener los parámetros a estimar.

De manera formal, sea una población con una función de probabilidad $P(x; \theta_1, \theta_2, \dots, \theta_k)$ o bien una función de densidad $f(x; \theta_1, \theta_2, \dots, \theta_k)$, sea ésta discreta o continua, en las cuales aparecen k parámetros desconocidos que pretendemos estimar a partir de una muestra aleatoria simple de tamaño n , (X_1, X_2, \dots, X_n) . Designamos por $\alpha_1, \dots, \alpha_k$ los k primeros momentos con respecto al origen de la población, y por a_1, \dots, a_k los k primeros momentos muestrales respecto al origen. Igualando los momentos poblacionales a sus correspondientes momentos muestrales tenemos el sistema de ecuaciones:

$$\begin{aligned}\alpha_1 &= \int_{-\infty}^{\infty} x f(x; \theta_1, \theta_2, \dots, \theta_k) dx = \sum_{i=1}^n \frac{x_i}{n} = a_1 \\ \alpha_2 &= \int_{-\infty}^{\infty} x^2 f(x; \theta_1, \theta_2, \dots, \theta_k) dx = \sum_{i=1}^n \frac{x_i^2}{n} = a_2 \\ \alpha_k &= \int_{-\infty}^{\infty} x^k f(x; \theta_1, \theta_2, \dots, \theta_k) dx = \sum_{i=1}^n \frac{x_i^k}{n} = a_k\end{aligned}$$

Y resolviendo este sistema de ecuaciones para los parámetros a estimar, obtenemos los estimadores.

8.2.1. Propiedades de los estimadores.

- **Insesgadez:** Si los parámetros que vamos a estimar son momentos poblacionales, el estimador obtenido por el método de los momentos es insesgado. En este caso, $\hat{\alpha}_j = a_j = \frac{1}{n} \sum_{i=1}^n X_i^j$, y se puede demostrar fácilmente que $E(\hat{\alpha}_j) = \alpha_j$.

- **Consistencia:** Bajo condiciones bastante generales estos estimadores son consistentes. La demostración se basa en la consistencia de los momentos poblacionales como estimadores de los momentos muestrales.
- **Normalidad asintótica:** Si los parámetros desconocidos que pretendemos estimar son los momentos poblacionales, estos estimadores son asintóticamente normales. Se demuestra teniendo en cuenta que los momentos muestrales son variables aleatorias resultantes de la suma de n variables aleatorias IID y con la misma esperanza y varianza, por el Teorema Central del Límite, su distribución tenderá a una $N(\alpha_j, \sqrt{\frac{\alpha_{2j} - \alpha_j^2}{n}})$.

En resumen, estos estimadores son consistentes, pero en general insesgados y por tanto no eficientes. Es por esto que este método no se utiliza demasiado. Además, este método no utiliza la distribución de probabilidad de la población, solo utiliza los momentos, por lo que se pierde información.

8.3. Método de la mínima X2.

Es un método general para la obtención de estimadores puntuales que se aplica solo cuando hay una gran cantidad de datos, tanto en distribuciones discretas como en distribuciones continuas con datos agrupados.

Supongamos una población representada por la variable aleatoria X cuya función de probabilidad depende de k parámetros, $p(x; \theta_1, \dots, \theta_k)$. Suponemos que el campo de variación de la variable aleatoria lo dividimos en r subconjuntos excluyentes, S_1, \dots, S_r , a los que podremos asociar una probabilidad $p_i(\theta_1, \dots, \theta_k) = P(X \in S_i) > 0$, con $\sum_{i=1}^r p_i = 1$.

Tomamos una muestra aleatoria de tamaño n , y presentamos la muestra como una distribución de frecuencias según el número de observaciones que pertenecen a los r grupos que hemos definido, n_1, \dots, n_r con $\sum_{i=1}^r n_i = n$.

Por tanto, tenemos por un lado la probabilidad teórica que le corresponde a cada conjunto, y por otro las frecuencias relativas obtenidas a partir de la muestra aleatoria simple. Parece lógico tomar como estimadores de los parámetros para la muestra obtenida aquellos que minimicen la diferencia entre ambas distribuciones, y para ello minimizaremos los cuadrados de las diferencias, usando como medida de la discrepancia la expresión:

$$\sum_{i=1}^r c_i \left(\frac{n_i}{n} - p_i(\theta_1, \dots, \theta_k) \right)^2$$

Pearson demostró que si tomamos $c_i = \frac{n}{p_i(\theta_1, \dots, \theta_k)}$ obtenemos una medida de la desviación con propiedades relativamente fáciles, y de cierto interés para estudiar la desviación entre las distribuciones. Por tanto, tenemos que:

$$\chi^2 = \sum_{i=1}^r \frac{n}{p_i(\theta_1, \dots, \theta_k)} \left(\frac{n_i}{n} - p_i(\theta_1, \dots, \theta_k) \right)^2 = \sum_{i=1}^r \frac{(n_i - np_i(\theta_1, \dots, \theta_k))^2}{np_i(\theta_1, \dots, \theta_k)}$$

que sigue una distribución χ^2_{r-k-1} .

El método de la mínima χ^2 escoge los estimadores de manera que el valor de χ^2 sea mínimo. Así pues, se deriva respecto a los θ y se iguala a cero. Resolviendo el sistema resultante para los parámetros obtenemos sus estimadores.

Los estimadores de mínima χ^2 son asintóticamente equivalentes al estimador de máxima verosimilitud. Sin embargo, para n pequeños no se puede asegurar nada, pues el estimador de mínima χ^2 no tiene por qué ser función del estimador suficiente si existe.

En general son estimadores sesgados y no eficientes.

8.4. Método de la mínima varianza.

Es un método analítico, y consiste en hacer mínima la varianza del estimador. La técnica que se utiliza es encontrar ese mínimo condicionado por las restricciones que queramos imponer al estimador mediante multiplicadores de Lagrange. Se buscan estimadores lineales insesgados, es decir, estimadores insesgados que sean función lineal de las observaciones muestrales. Veamos dos aplicaciones:

8.4.1. Estimador de varianza mínima de la media poblacional.

Sea el estimador lineal $\hat{\mu} = a_1X_1 + \cdots + a_nX_n$, como ha de ser insesgado,

$$E[\hat{\mu}] = E[a_1X_1 + \cdots + a_nX_n] = a_1E[X_1] + \cdots + a_nE[X_n] = \mu \sum_{i=1}^n a_i = \mu$$

y por tanto, se tiene que cumplir que $\sum_{i=1}^n a_i = 1$ para que el estimador sea insesgado. La varianza del estimador será:

$$V(\hat{\mu}) = V(a_1X_1 + \cdots + a_nX_n) = a_1^2E(X_1) + \cdots + a_n^2V(X_n) = \sigma^2 \sum_{i=1}^n a_i^2$$

Y como la varianza ha de ser mínima, aplicamos el método de los multiplicadores de Lagrange:

$$\begin{aligned} \phi &= \sigma^2 \sum_{i=1}^n a_i^2 + \lambda \left(\sum_{i=1}^n a_i - 1 \right) \\ \frac{\partial \phi}{\partial a_i} &= 2a_i\sigma^2 + \lambda = 0 \\ \frac{\partial \phi}{\partial \lambda} &= \sum_{i=1}^n a_i - 1 = 0 \end{aligned}$$

Y resolviendo tenemos que $a_i = \frac{1}{n}$, $\lambda = -\frac{2\sigma^2}{n}$. Así, por tanto, $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ es el estimador lineal insesgado de varianza mínima para la media.

8.5. Método de los mínimos cuadrados.

En muchas ocasiones tenemos una variable aleatoria cuyo comportamiento se puede expresar mediante una función de un conjunto de variables aleatorias y no aleatorias que depende de una serie de parámetros. Así, nuestra muestra se compone de un conjunto de puntos en un espacio r -dimensional, y buscamos una función que pase lo más cerca posible de esos puntos. Para ello necesitamos estimar los parámetros.

Esta estimación se realiza minimizando la distancia entre el valor real de la variable aleatoria a estimar y el valor teórico que obtendríamos a partir de la función con nuestras estimaciones de los parámetros.

Así, si tenemos una variable aleatoria, y , tal que $y = g(\mathbf{X}; \theta_1, \dots, \theta_k)$. Para cada valor de la variable \mathbf{X} , x_i , tendremos un valor teórico de la variable y , Y_i , proporcionado por la función una vez ajustada. El error cometido al utilizar ese valor teórico en lugar del real será $e_i = y_i - Y_i = y_i - g(x_i; \theta)$.

La evaluación del error global cometido se realiza sumando los cuadrados de los errores:

$$\Phi = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - g(x_i; \theta)]^2$$

Y los estimadores de los parámetros serán aquellos que hacen mínima la suma de los cuadrados de los errores, es decir, la solución del sistema de ecuaciones representado por:

$$\frac{\partial \Phi}{\partial \theta_i} = -2 \sum_{j=1}^n [y_j - g(x_j; \theta)] \frac{\partial g(x_j; \theta)}{\partial \theta_i} = 0$$

8.6. Métodos Bayesianos.

Hasta ahora hemos estudiado la estimación puntual desde el punto de vista de la teoría del muestreo, que se basa en interpretar la probabilidad como una frecuencia relativa. Pasaremos ahora a estudiar el enfoque bayesiano de la inferencia estadística, en lo que se refiere a la estimación de parámetros.

En el enfoque bayesiano, un parámetro es visto como una variable aleatoria a la que se asigna una distribución de probabilidad a priori con base en el grado de creencia sobre la distribución del mismo, que se modifica con la

información obtenida de la muestra, para obtener la distribución a posteriori. Con esta distribución a posteriori formularemos inferencias respecto al parámetro. Este enfoque resulta muy útil en aquellas situaciones en las que el parámetro a estimar no puede considerarse una cantidad fija, sino que puede variar dependiendo de las características del entorno.

Dado que consideramos el parámetro a estimar como una variable aleatoria, lo designamos por Θ , y por θ a la realización de dicha variable aleatoria. Suponemos que Θ es una variable aleatoria continua con una función de densidad incondicional a priori $f_{\Theta}(\theta)$, la cual refleja las creencias previas acerca de Θ . Si tomamos una muestra aleatoria simple de tamaño n (n variables aleatorias idénticamente distribuidas), X_1, \dots, X_n su función de densidad condicionada común será $f(x|\theta)$, y la función de densidad conjunta:

$$L(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta) \cdots f(x_n|\theta)$$

Como decimos que Θ es una variable aleatoria, el objetivo es estimar el valor particular θ para el cual la evidencia muestral que representa la densidad conjunta se encuentra condicionada. Por tanto, la función de densidad a posteriori de Θ será, aplicando el teorema de Bayes:

$$f(\theta|x_1, x_2, \dots, x_n) = \frac{L(x_1, x_2, \dots, x_n|\theta)f_{\Theta}(\theta)}{\int_{\Theta} L(x_1, x_2, \dots, x_n|\theta)f_{\Theta}(\theta)d\theta}$$

El denominador de esta expresión se denomina distribución predictiva, y representa la ponderación de todas las distribuciones posibles del parámetro, ponderados según la importancia que de a cada una la distribución a priori.

En la práctica, el cálculo se simplifica si observamos que el denominador no depende de θ , y actúa solo como constante normalizadora de la distribución a posteriori, para que su integral sea la unidad. Por tanto, vemos que la distribución a posteriori es proporcional a la distribución a priori multiplicada por la verosimilitud de la muestra. Por tanto, la distribución a posteriori combina la información previa de la que se dispone, representada por la distribución a priori, con la información aportada por la muestra. Si la distribución a priori es más o menos constante sobre el espacio paramétrico, la distribución a priori coincide con la verosimilitud, y se dice que la distribución a priori es no informativa.

Para tamaños muestrales grandes, se puede demostrar que en condiciones muy generales la distribución a posteriori está dominada por la verosimilitud, y adquiere una distribución aproximadamente normal con media y varianza coincidentes con la del estimador de máxima verosimilitud. En consecuencia, en estos casos el estimador bayesiano y el de máxima verosimilitud conducen a los mismos resultados.

Para obtener una estimación de θ necesitamos elegir una característica numérica de la distribución a posteriori que nos parezca representativa de la misma. Hay dos opciones:

- Elegir como estimación la moda de la distribución a posteriori, que es el valor más probable una vez observada la muestra. Esta situación tiene la misma justificación que la estimación de máxima verosimilitud en el contexto clásico.
- Elegir una función de pérdida que represente la consecuencia de haber escogido un valor de θ erróneo. Esta función debe ser una función no negativa de θ y su estimación, de manera que sea cero si coinciden.

Al depender también de θ , la función de pérdida también es una variable aleatoria. El estimador bayesiano del parámetro será aquél que minimice la esperanza de la función de pérdida.

Es obvio que para poder estimar el parámetro se debe especificar una función de pérdida. Esto es una tarea difícil, ya que las consecuencias no son siempre medibles. En muchos casos una función de pérdida razonable puede ser la forma cuadrática: $l(\theta, t) = (t - \theta)^2$. Para esta forma, se puede demostrar que el estimador de Bayes equivale a la distancia a posteriori de Θ .

Capítulo 9

Método de estimación de máxima verosimilitud.

Propiedades. Distribución asintótica de los estimadores de máxima verosimilitud.

9.1. Introducción.

Dentro del proceso de inferencia estadística sobre una población, en el que queremos obtener estimadores para los parámetros que caracterizan esa población, sabemos que hay una serie de propiedades deseables en esos estimadores (insesgadez, consistencia, eficiencia). Otro problema es cómo obtener estimadores que presenten estas propiedades. Para ello veremos cómo obtener estimadores a partir del método de máxima verosimilitud y revisaremos que propiedades cumplen los estimadores obtenidos mediante dicho método.

9.2. Método de estimación de máxima verosimilitud.

El método de máxima verosimilitud se basa en el siguiente supuesto teórico: la obtención de una muestra a partir de una población no es más que realizar un experimento aleatorio y registrar el suceso que resulta del mismo. Pues bien, este método supone que el suceso que hemos obtenido será el más probable para la distribución de probabilidad asociada a la población, y a partir de esa suposición deducirá los parámetros a estimar. Pasemos a verlo de forma formal.

9.2.1. Función de verosimilitud de la muestra.

Definición 26. Definimos como **función de verosimilitud** de un conjunto de n variables aleatorias como la función de probabilidad o función de densidad conjunta de las n variables, y la denotamos por:

$$L(\mathbf{x}; \theta) = L(x_1, \dots, x_n; \theta) = f(x_1, \dots, x_n; \theta)$$

Para el caso de una muestra aleatoria simple, al ser variables aleatorias independientes idénticamente distribuidas, su función de verosimilitud será:

$$L(\mathbf{x}; \theta) = L(x_1, \dots, x_n; \theta) = f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

Por tanto, la función de verosimilitud es función de la muestra observada y depende del parámetro a estimar, θ .

El valor que toma la función de verosimilitud para una muestra concreta recibe el nombre de **elemento de verosimilitud** o **verosimilitud**, y solo depende del parámetro θ .

Si nuestra distribución de probabilidad es discreta, sustituiremos la función de densidad por la función de probabilidad.

9.2.2. Estimador de máxima verosimilitud.

Definición 27. El método de estimación de máxima verosimilitud consiste en elegir como estimador del parámetro desconocido a partir de una muestra aleatoria simple aquel valor que hace máxima la verosimilitud de la muestra, es decir, consiste en encontrar aquel valor $\hat{\theta}(X_1, \dots, X_n)$ para el que:

$$L(x_1, \dots, x_n; \hat{\theta}) = \max_{\theta \in \Omega} L(x_1, \dots, x_n; \theta)$$

A este estimador $\hat{\theta}(X_1, \dots, X_n)$ se le llama **estimador máximo-verosímil** o **estimador de máxima verosimilitud** del parámetro θ .

Así, el método elige el valor de θ para el que el valor de la verosimilitud de la muestra es máxima, y por tanto, elige el valor del parámetro de forma que la muestra que hemos obtenido sea la más probable. Otra forma de entenderlo es que elige el valor del parámetro más verosímil para la muestra considerada.

El hecho de que la función de verosimilitud sea el resultado de un producto de funciones complica en muchos casos la búsqueda del máximo. Es por eso que, dado que la función de densidad es siempre positiva, y por tanto maximizar $L(x_1, \dots, x_n; \theta)$ equivale a maximizar $\ln L(x_1, \dots, x_n; \theta)$, se calcula el estimador máximo verosímil a partir de la siguiente expresión:

$$\ln L(x_1, \dots, x_n; \hat{\theta}) = \max_{\theta \in \Omega} \ln L(x_1, \dots, x_n; \theta) = \max_{\theta \in \Omega} \sum_{i=1}^n \ln f(x_i; \theta)$$

Y obtendremos el estimador solucionando la siguiente ecuación:

$$\sum_{i=1}^n \frac{\partial \ln f(x_i; \theta)}{\partial \theta} = 0$$

El estimador así obtenido será función de las observaciones muestrales, y prescindiremos de aquellas soluciones que den lugar a que el estimador sea una constante.

Si la función de densidad o cuantía de la población depende de más de un parámetro, los estimadores vendrán dados por la solución del sistema de ecuaciones de verosimilitud.

Cualquier solución de las ecuaciones será un estimador de máxima verosimilitud. Si la solución es única, diremos que tenemos un estimador de máxima verosimilitud en sentido estricto. Si hay más de una solución, cada una de ellas será un estimador de máxima verosimilitud en sentido amplio.

9.3. Propiedades.