Muestras

jueves, 16 de agosto de 2018 13:47

Introducción

Consideremos una situación experimental donde la variable de interés X, tiene función de densidad o de probabilidad $f(x)$

Definición 1 Un conjunto $X_1, X_2, \ldots, X_n$ de variables aleatorias que son mutuamente independientes y que tienen la misma función de densidad o de probabilidad $f(x)$ se llama muestra aleatoria de tamaño n de la población $f(x)$

Muchas veces decimos que $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes idénticamente distribuidas y lo abreviaremos como iid.

A partir de lo anterior la función de probabilidad o densidad conjunta de $X_1, \ldots, X_n$ está dada por

$$f(x_1, \ldots, x_n) = \prod f(x_i)$$

$$f(x_1, \ldots, x_n) = \prod_{i=1} f(x_i)$$

Si se tiene una familia paramétrica de densidades o funciones de probabilidad obtenemos

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Ejemplo: Supongamos que $X_1, \ldots, X_n$ es una muestra aleatoria de una población con distribución exponencial de parámetro β. o sea $f(x) = \frac{1}{\beta} e^{-x/\beta}$, $x \geq 0$

La densidad conjunta entonces es:

$$f(x_1, \ldots, x_n \mid \beta) = \prod_{i=1}^{n} f(x_i \mid \beta) = \frac{1}{\beta^n} e^{-\frac{(x_1 + \ldots + x_n)}{\beta}}$$

Podemos calcular la probabilidad que todas las X_i sean mayores a 2 mediante

$$P(X_1 > 2, X_2 > 2, \ldots, X_n > 2) = \int_2^{+\infty} \cdots \int_2^{+\infty} \frac{1}{\beta^n} e^{-\frac{\sum x_i}{\beta}} dx_1 \ldots dx_n$$

$$= \left(\int_2^{+\infty} \ldots e^{-x_i/\beta} \ldots \right) = \left(e^{-2/\beta} \right)^n = e^{-2n/\beta}$$

$$= \prod_{i=1}^{n} \left(\int_{2}^{+\infty} \frac{1}{\beta} e^{-x_i/\beta} \, dx_i \right) = \left(e^{-2/\beta}\right)^n = e^{-2n/\beta}$$

El caso mostrado en la definición 1 suele llamarse muestra de una población infinita.

¿Qué sucede cuando se tiene una población finita?

Supongamos que tenemos una población finita $\{x_1, x_2, \ldots, x_N\}$. Se toma una muestra $X_1, X_2, \ldots, X_n$ de esa población.

Consideremos dos formas de tomar esa muestra

a) con reemplazamiento b) sin reemplazamiento.

En el caso a) la probabilidad de elegir cada valor de la población en cualquier etapa del muestreo es $1/N$. Es claro que se cumplen las condiciones de la definición 1.

En el segundo caso si $X_1 = x_1$ entonces la probabilidad de $X_2 = x_2$ es 0 para $x_2 = x_1$,

probabilidad de $X_2 = x_2$ es 0 para $x_2 = x_1$ y
vale $1/N-1$ para $x_2 \neq x_1$
O sea no se satisface la condición de la definición 1

Las variables aleatorias $X_1, \ldots, X_n$ no son mutuamente independientes pues $P(X_2 = y / X_1 = y) = 0$ y $P(X_2 = y / X_1 = x) = 1/N-1$

Es interesante ver que $X_1, \ldots, X_n$ son idénticamente distribuidas, o sea la distribución (marginal) de X_i es la misma $\forall i = 1, \ldots, n$.

Por ejemplo

$$P(X_2 = x) = \sum_{i=1}^{N} P(X_2 = x \mid X_1 = x_i) P(X_1 = x_i)$$

$$= (N-1) \frac{1}{N-1} \frac{1}{N} = 1/N$$

A veces a este tipo de muestra se le llama muestra aleatoria simple.

Es claro que no se cumple la definición 1 pero si N es grande se está muy cercano a la independencia

Definición 2 Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n de una población y sea $T(x_1,\dots,x_n)$ una función a valores reales o vectoriales cuyo dominio incluye el espacio muestral de $(X_1,\dots,X_n)$. Entonces a $Y=T(X_1,\dots,X_n)$ se le llama estadístico. La función de distribución del estadístico Y es llamada distribución muestral de Y

Definición 3: Se define la media muestral de $X_1,\dots,X_n$ como

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{X_1+\dots+X_n}{n}$$

Definición 4: Se define la varianza muestral mediante

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

La desviación estandar muestral se define como $S=\sqrt{S^2}$

Teorema 5: Consideremos $x_1,\dots,x_n$ números cualesquiera y sea $\bar{x}=(x_1+\dots+x_n)/n$. Entonces

a) $\displaystyle\min_a\sum_{i=1}^{n}(x_i-a)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2$

b) $\displaystyle(n-1)s^2=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$

Se deja como ejercicio para el práctico

Teorema 6: Sea $X_1, ---, X_n$ una muestra aleatoria de una población con media μ y varianza $\sigma^2 < \infty$. Entonces

a) $E(\bar{X}) = \mu$

b) $Var(\bar{X}) = \sigma^2/n$

c) $E(S^2) = \sigma^2$

Dem: a) $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\, E(X_1) = \mu$

b) $Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\, n\, Var(X_1) = \frac{\sigma^2}{n}$

(en este caso se aplicó fuertemente la independencia de las variables X_i)

c) $E(S^2) = E\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\right) = \frac{1}{n-1}\left(n\, E(X_1^2) - n\, E(\bar{X}^2)\right)$

Sabemos que $\sigma^2 = E(X_1^2) - E^2(X_1) = E(X_1^2) - \mu^2$

luego $E(X_1^2) = \sigma^2 + \mu^2$

obs

Además $E(\bar{X}^2) = E\left(\left(\sum_{i=1}^{n} X_i / n\right)^2\right) =$

$V(X) = E(X^2) - E^2(X)$

luego $E(\bar{X}) = \mu$

$V(X) = E(X^2) - E^2(X)$

Además $E(\bar{X}^2) = E\left(\left(\sum_{i=1}^{n} X_i/n\right)^2\right) =$

$$= E\left(\frac{1}{n^2}\left(\sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j\right)\right) = \frac{1}{n^2}\, n\, E(X_1^2) + \frac{1}{n^2}\sum_{i\neq j} E(X_i)E(X_j)$$

$$= \frac{1}{n}(\sigma^2+\mu^2) + \frac{1}{n^2}\sum_{i\neq j}\mu^2 = \frac{1}{n}(\sigma^2+\mu^2) + \frac{n(n-1)}{n^2}\mu^2$$

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

Reemplazando en $E(S^2)$ obtenemos

$$E(S^2) = \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) =$$

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

Definición 7: Convergencia en Probabilidad.

Una sucesión de variables aleatorias $X_1, X_2, \ldots$ converge en probabilidad a una variable aleatoria X si para todo $\varepsilon>0$

$\lim_{n\to+\infty} P(|X_n - X| \geq \varepsilon) = 0$ o lo que es equivalente si $\lim_{n \to +\infty} P(|X_n - X| < \varepsilon) = 1$

Teorema 8: (Ley débil de los grandes números)
Sean $X_1, \ldots$ variables aleatorias iid con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2 < \infty$. Definamos $\bar{X}_n = (1/n) \sum_{i=1}^{n} X_i$ entonces para todo $\varepsilon > 0$

$$\lim_{n \to +\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$$

o sea $\bar{X}_n$ converge en probabilidad a μ

Dem: Usaremos la desigualdad de Chebyshev
Para todo $\varepsilon > 0$

$$P(|\bar{X}_n - \mu| \geq \varepsilon) = P((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \leq \frac{E(\bar{X}_n - \mu)^2}{\varepsilon^2} =$$

$$= \frac{Var \bar{X}}{\varepsilon^2} = \frac{\sigma^2}{n \varepsilon^2} \xrightarrow[n \to +\infty]{} 0$$ de donde se obtiene la tesis.

Definición 9: (Convergencia casi segura) Una sucesión de variables aleatorias $X_1, X_2, \ldots$ converge casi seguramente a una variable aleatoria X si para todo $\varepsilon > 0$

$$P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1$$

o lo que es lo mismo $P\left(\lim_{n\to+\infty} X_n = X\right) = 1$

$$P\left(\lim_{n\to+\infty} X_n \neq X\right) = 0$$

En el práctico se explorará la diferencias entre las definiciones 7 y 9. Si adelantamos que la convergencia casi segura implica la convergencia en probabilidad

Teorema 10: Ley Fuerte de los grandes números

Sean $X_1, X_2, \ldots$ una sucesión de variables aleatorias iid con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2 < \infty$ y definamos $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Entonces $\bar{X}_n$ converge casi seguramente a μ

Definición 11: Una sucesión de variables aleatorias $X_1, \ldots$ converge en distribución a una variable aleatoria X si:

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x)$$

para todo punto de continuidad de F_X.

$$\lim_{n} P(X_n \leq x) = P(X \leq x)$$

Teorema 12 Si $X_1, X_2, \ldots$ converge en probabilidad a X entonces también converge en distribución a X

Por último damos también una versión del teorema central del límite

Teorema 13: Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias con $E(X_i) = \mu$ y $0 < \operatorname{Var} X_i = \sigma^2 < \infty$

Sea $G_n(x)$ la función de distribución de $\sqrt{n}(\bar{X}_n - \mu)/\sigma$

Entonces para todo x; $-\infty < x < \infty$

$$\lim_{n \to +\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dy$$

o sea $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converge en distribución a la normal típica

Teorema 14 (Slutsky)

Si $X_n \to X$ en distribución y $Y_n \to a$ (a cte) en probabilidad, entonces

a) $Y_n X_n \longrightarrow aX$ en distribución

b) $X_n + Y_n \longrightarrow X + a$ en distribución

Ejemplo Supongamos que se cumple el teorema central del límite y

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

Supongamos que σ es desconocido y que estamos en las condiciones para que $\sigma/S_n \longrightarrow 1$ en probabilidad (ver Práctico 1). Entonces

$$\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \longrightarrow \mathcal{N}(0,1)$$