Capítulo 4. Teoría de muestras grandes

Índice

La teoría de muestras grandes proporciona resultados que se cumplen cuando el tamaño muestral tiende a infinito, por lo que a este campo se le denomina también estadística asintótica. En las aplicaciones reales las muestras necesariamente tienen un tamaño finito, por lo que estos resultados son aproximaciones que funcionan bien en muestras grandes.

Los resultados de este capítulo se agrupan en dos categorías, la *ley de los grandes números*, y el *teorem del límite central*. Ambas tienen diversas variantes en función de las condiciones en que se aplican y la generalidad de los resultados que proporcionan, por lo que en ocasiones aparece escrito en plural, leyes de los grandes números y teoremas del límite central). A continuación veremos las versiones más sencillas de ambos resultados.

La ley de los grandes números trata sobre la convergencia de la media muestral hacia la media poblacional cuando el tamaño de la muestra aumenta. El teorema del límite central tiene que ver con la distribución de una suma de variables aleatorias, que se aproxima a una distribución normal cuando el número de variables que intervienen en la suma es elevado. Cómo la media muestral es una suma de variables aleatorias, una por cada elemento de la muestra, dividida por el número de datos, el teorema del límite central implica que la distribución muestral de la media es normal en muestras grandes, y la ley de los grandes números significa que la distribución muestral de la media está centrada en la media poblacional.

Estos dos resultados, aparentemente sencillos, constituyen la base de la gran mayoría de los procedimientos de inferencia empleados en estadística aplicada. Procedimientos tales como los contrastes de una y dos medias, los contrastes sobre proporciones, la prueba chi-cuadrado, etc. tienen su base en estos métodos. Las propiedades asintóticas de los estimadores máximo-verosímiles se siguen de estos teoremas, gracias a los cuales podemos obtener intervalos de confianza basándonos en una aproximación normal.

Una razón por la que resultan tan útiles es que son procedimientos *libres de distribución*. La validez de ambos teoremas no depende de cual sea la distribución de los datos en la población de partida. Es indiferente que dicha distribución sea uniforme, Poisson, exponencial o cualquier otra, la media muestral converge a la media poblacional y la distribución de una suma de variables converge a una distribución normal. Esto permite utilizar la normal, u otras distribuciones basadas en ella como t o chi-cuadrado, para realizar contrastes o construir intervalos de confianza sobre medias sin necesidad de saber cual es la verdadera distribución de la variable.

En primer lugar veremos una introducción a lo que significa el límite $n \to \infty$ en teoría de la probabilidad y después veremos los resultados fundamentales del capítulo.

1. Ley de los grandes números

La media muestral es la suma de las observaciones de la muestra dividida por el tamaño muestral:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

En esta definición, cada uno de los elementos $X_1, \ldots, X_n$ es una variable aleatoria y, bajo las condiciones del muestreo aleatorio simple, todas ellas tienen el mismo valor esperado $E(X_i) = \mu$, siendo μ la media poblacional. Resulta intuitivo suponer que la media muestral, $\overline{X}$, será similar a la media poblacional, μ, y que cuanto mayor sea n más razonable es suponer que $\overline{X}$ estará próximo a μ. Esto es justamente lo que dice la ley de los grandes números, que expresa que $\overline{X}$ converge en probabilidad a μ.

> *Ley de los grandes números.* Sea $X_1, X_2, \ldots$ una secuencia de variables aleatorias idénticamente distribuidas y con valor esperado finito $E(X_i) = \mu$. Entonces, para cada $\epsilon > 0$
>
> $$P(|\overline{X} - \mu| \geq \epsilon) \to 0 \quad \text{cuando} \quad n \to \infty.$$

Este resultado también se escribe $\overline{X} \xrightarrow{p} \mu$, donde $\xrightarrow{p}$ quiere decir *convergencia en probabilidad*. En la formulación de este teorema, ϵ es la diferencia entre $\overline{X}$ y μ, y esta diferencia podemos hacerla tan pequeña como queramos aumentando el tamaño muestral. En definitiva, la probabilidad de encontrar valores de $\overline{X} - \mu$ mayores que ϵ tiende a cero al aumentar n sea cual sea el valor de ϵ.

La ley de los grandes números también tiene aplicación a variables dicotómicas. En este caso nos dice que la proporción muestral tiende a la probabilidad de éxito. Intuitivamente es un resultado obvio, si tenemos una moneda imparcial, cabe esperar que cuantos más lanzamientos realicemos más próxima estará la proporción de caras a 0,5. En el lenguaje de la teoría de la probabilidad esto se expresa del siguiente modo. Supongamos que X es una variable de Bernoulli (π), entonces $E(X) = \pi$ y la proporción muestral es

$$P = \frac{\sum_i X_i}{n}.$$

Como P es la media muestral, $P \xrightarrow{p} \pi$.

Ejemplo 1. La varianza muestral es la media de las puntuaciones diferenciales elevadas al cuadrado:

$$S_n^2 = \frac{1}{n} \sum_i (X_i - \overline{X})^2.$$

Podemos comprobar que la varianza muestral es asintóticamente insesgada mediante sucesivas aplicaciones de la ley de los grandes números. En primer lugar, sabemos que $\overline{X} \xrightarrow{p} \mu$, en consecuencia $\sum_i (X_i - \overline{X})^2/n \xrightarrow{p} \sum_i (X_i - \mu)^2/n$, por lo que asintóticamente tenemos una media de n términos, $(X_i - \mu)^2$, cuyo valor esperado es $E((X - \mu)^2) = \sigma^2$. Aplicando nuevamente la ley de los grandes números tenemos que la media de las variables $(X_i - \mu)^2$ converge en probabilidad a su valor esperado, por lo que $S_n^2 \xrightarrow{p} \sigma^2$.

Ejemplo 2. Según vimos al estudiar las distribuciones, si $X \sim Poisson(\lambda)$ entonces $E(X) = \lambda$. Supongamos ahora que tenemos una m.a.s. procedente de una distribución de Poisson y queremos estimar λ. Entonces el estimador natural es la media muestral, dado que gracias al resultado $\overline{X} \xrightarrow{p} \lambda$ sabemos que la media muestral es un estimador asintóticamente insesgado. Esto no resuelve el problema de si existen otros estimadores más eficientes (con menor varianza) o de cual sea la precisión del estimador, para ello habría que realizar un análisis más detallado y estudiar otros métodos de estimación, pero al menos proporciona una primera respuesta al problema de estimar el parámetro desconocido.

Ejemplo 3. En un estudio sobre tiempos de reacción hemos encontrado que un sujeto tarda los siguientes segundos en realizar cuatro tareas $\boldsymbol{x} = (4, 6, 1, 9)'$. Queremos estimar la velocidad de ejecución. Para ello, asumimos que $X \sim$ exponencial(ω) y como $E(X) = 1/\omega$ entonces $\overline{X} \xrightarrow{p} 1/\omega$. Aplicado a nuestros datos, $\overline{X} = 5$ y la velocidad estimada es $\hat{\omega} = 1/5 = 0,2$.

2. Teorema del límite central

El teorema del límite central es uno de los resultados más importantes de la teoría de la probabilidad y constituye la base de innumerables procedimientos estadísticos. Tanto los contrastes de hipótesis sobre medias hasta los de bondad de ajuste basados en chi-cuadrado, pasando por la obtención de estimadores por intervalos, tienen su base en este teorema.

Expresado en palabras, el teorema del límite central dice que la distribución de la suma de varias variables aleatorias se aproxima a una distribución normal a medida que el número de variables aumenta. Al igual que la ley de los grandes números, el teorema del límite central tiene distintas versiones en función de las características de las variables sumadas, aunque una de las más sencillas es la siguiente.

> *Teorema del límite central.* Sea X_1, X_2, ... una secuencia de variables aleatorias cada una de ellas con valor esperado $E(X_i) = \mu$ y varianza $Var(X_i) = \sigma^2$. Entonces, la distribución del estadístico
>
> $$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$
>
> tiende a una distribución normal estándar cuando $n \to \infty$.

Según se ha formulado, el teorema del límite central se refiere a la media de n variables. No obstante, como la media no es más que la suma de variables dividida por n, el teorema podría formularse de igual modo haciendo referencia a la suma y no a la media. En concreto, el estadístico Z puede escribirse del siguiente modo cuando el interés está en la suma de variables

$$Z = \frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}}.$$

El teorema del límite central no garantiza que la distribución muestral de la media es normal. Simplemente nos permite confiar en que en *muestras grandes* será aproximadamente normal. No existe un criterio exacto que diga cuando una muestra es grande y podemos confiar en que la aproximación normal es adecuada, pero dicha aproximación no debería utilizarse con muestras de menos de 20 observaciones.

Ejemplo 5. Un grupo de 36 personas ha realizado una prueba consistente en leer en voz alta un determinado texto. La variable X_i indica el número de errores cometidos por el sujeto i. Como la longitud del texto es elevada y basándonos en nuestra experiencia previa, asumimos que $X_i \sim \text{Poisson}(\lambda = 4)$. Supongamos que queremos conocer la probabilidad de que el número medio de errores cometidos sea menor o igual a cinco. Gracias al teorema del límite central sabemos que la variable

$$Z = \frac{\sqrt{n}(\overline{X} - \lambda)}{\sqrt{\lambda}}$$

sigue aproximadamente una distribución normal estándar en muestras grandes. Por tanto

$$Z = \frac{\sqrt{36}(5 - 4)}{\sqrt{4}} = 3.$$

Buscando en la tabla de la normal encontramos $P(\overline{X} \leq 5) = P(Z \leq 3) \approx 0,99865$.

Ejemplo 5. Aproximación normal a la binomial. Lanzamos 100 veces una moneda imparcial y queremos saber cual es la probabilidad de encontrar más de cincuenta caras. Como el resultado de cada lanzamiento es $X_i \sim$ Bernoulli $(\pi = 0,5)$ tenemos que $E(X_i) = \pi = 0,5$ y $Var(X_i) = \pi(1 - \pi) = 0,25$. Aplicando el teorema del límite central tenemos que la variable

$$Z = \frac{\sum_i X_i - n\pi}{\sqrt{n\pi(1 - \pi)}}$$

es aproximadamente normal(0, 1). Entonces

$$Z = \frac{51 - 100(0,5)}{\sqrt{100(0,25)}} = 0,2.$$

Por tanto $P(X \geq 51) \approx P(Z \geq 0,2) \approx 0,42$.

3. Ejercicios

Ejercicio 1. Sea U una variable distribuida según la uniforme en el intervalo (0, 1). Obtenga la probabilidad de que la media de una muestra de tamaño 27 sea superior a 0,6.

Ejercicio 2. Sea U una variable distribuida según la uniforme en el intervalo (0, 1). ¿Entre qué valores se encuentra la media de una muestra de tamaño 27 con una probabilidad de 0,95?

Ejercicio 3. Sea $X \sim$ Poisson(8) y tomamos una muestra de tamaño 64. ¿Cual es la probabilidad de encontrar una media mayor o igual a 9?

Ejercicio 4. ¿Cual es la probabilidad de que la suma de 25 variables independientes distribuidas según Poisson(16) sea inferior a 380?

Ejercicio 5. Sea $X \sim$ Poisson(10) y tomamos una muestra de tamaño 36. ¿Entre qué valores se encuentra la media muestral con una probabilidad de 0,99?

Ejercicio 6. Sea $X \sim$ exponencial $(\omega = 0,2)$. ¿Entre qué valores se encuentra la media muestral con probabilidad 0,95 si $n = 25$?

Ejercicio 7. Sea $X \sim$ chi-cuadrado con 8 grados de libertad (gl). Además sabemos que en una distribución chi-cuadrado $E(X) = gl$ y $Var(X) = 2gl$. Calcule la probabilidad de que la media de X sea inferior a 10 en una muestra de tamaño 25.