Distribución exacta de algunos estadísticos básicos

Introducción

Hemos indicado en el capítulo anterior que los estimadores de máxima verosimilitud tienen asintóticamente una distribución Normal. Sin embargo, parece evidente que sería mucho más interesante conocer la distribución exacta en el muestreo de los estadísticos y estimadores que vayamos a utilizar que la distribución que seguirían si la muestra fuese de tamaño infinito. La determinación de esta distribución exacta será especialmente importante en aquellas situaciones, habituales en la investigación psicológica, en las que debamos trabajar con muestras pequeñas, pues en estas condiciones la distribución real puede diferir sensiblemente de la distribución asintótica.

Más aún, el que existan discrepancias entre la distribución exacta y la distribución asintótica no puede ser determinado hasta tanto no se estudie la primera y la velocidad con la que converge a la segunda. De tal forma que el establecer si una muestra es pequeña o grande no es una cuestión que haga referencia a la magnitud absoluta del tamaño de la muestra sino que más bien, hace referencia a la distribución del estimador o del estadístico que se va a calcular a partir de ella. De forma que la muestra que puede considerarse grande para estimar la media de una población Normal, puede ser pequeña o incluso absolutamente insuficiente para estimar la varianza o el coeficiente de correlación.

Desde el punto de vista teórico, el problema está resuelto, especialmente en el caso de muestras aleatorias simples, recuérdese que la función de densidad de una muestra aleatoria simple puede calcularse fácilmente como el producto n veces de la función de densidad de la población. Conocida la función de densidad de la muestra, conocer la distribución de cualquier estadístico o estimador que es una función $g(x_1, x_2, \ldots x_n)$ de los valores de la muestra, requiere solamente un cambio de variable. En el peor de los casos, este cambio de variable puede resolverse por medio de aproximaciones numéricas. El problema surge cuando deseamos que la solución pueda expresarse por medio de funciones explícitas conocidas. Tal tipo de solución sólo ha podido ser alcanzada en un reducido número de casos, de forma que la única situación que ha podido ser investigada con resultados satisfactorios es el muestreo de distribuciones poblacionales Normales. Como ejemplo de este tipo de estudios enunciaremos el teorema de Fisher y las consecuencias que se derivan de él, resultados que emplearemos con profusión en temas sucesivos.

Teorema de Fisher

Sea una variable aleatoria X que sigue en la población una distribución Normal de parámetros μ y σ. De esa población obtenemos muestras aleatorias simples de tamaño n que genéricamente designamos $(x_1, x_2, \ldots x_n)$. Entonces se verifica que:

La media $\bar{x}$ y varianza S^2 muestrales son independientes

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}$$

la media muestral sigue una distribución Normal de media, la media de la población, y de desviación típica, la desviación típica de la población dividida por la raíz cuadrada del tamaño de la muestra y la suma de los cuadrados de las diferencias a la media dividida por la varianza de la población sigue una distribución Ji-cuadrado con n-1 grados de libertad. Es decir:

$$\bar{x} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \text{y} \qquad \frac{n \cdot S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$$

Razón de Student

Como consecuencia del teorema anterior es sencillo comprobar que la razón de Student:

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n-1}}}$$

sigue una distribución t de Student con n-1 grados de libertad. En efecto, en virtud de los resultados del teorema anterior y de la definición de la t de Student, tenemos que el siguiente cociente:

$$\frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{1}{n-1} \quad \frac{nS^2}{\sigma^2}}} \rightarrow t_{n-1}$$

sigue una distribución t de Student con n-1 grados de libertad, ya que en el numerador tenemos una variable que sigue una distribución Normal cero, uno y en el denominador la raíz cuadrada de una variable que sigue una distribución Ji-cuadrado dividida por sus grados de libertad y esa es precisamente la definición de la distribución t de Student. Simplificando la expresión anterior tenemos:

$$\frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{1}{n-1} \quad \frac{nS^2}{\sigma^2}}} = \frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{1}{\sqrt{n-1}} \cdot \frac{\sqrt{n} \cdot S}{\sigma}} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n-1}}}$$

Este resultado nos será de gran utilidad, porque nos proporciona un estadístico que depende de la media de la población, pero no de la varianza de la misma, del cual conocemos su distribución exacta que tampoco depende de la varianza de la población. Todo lo anterior nos permitirá hacer inferencias, tanto estimaciones como contrastes, de la media de la población sin necesidad de tener que estimar la varianza de la misma. Esto tiene la ventaja de que el tamaño de muestra necesario para inferir acerca de una media es sensiblemente menor que el que se necesita para conocer, con la misma precisión, el valor de la varianza.

De la misma, sino de mayor importancia es el resultado relativo a la diferencia entre los valores medios de dos muestras independientes. Sean dos variables aleatorias independientes:

$$X_1 \rightarrow N(\mu_1, \sigma) \quad \text{y} \quad X_2 \rightarrow N(\mu_2, \sigma)$$

de las cuales se extraen muestras aleatorias simples de tamaños n_1 y n_2. En virtud del teorema de Fisher se verificará que:

$$\bar{x}_1 \rightarrow N\left(\mu_1, \frac{\sigma}{\sqrt{n_1}}\right) \quad \text{y} \quad \bar{x}_2 \rightarrow N\left(\mu_2, \frac{\sigma}{\sqrt{n_2}}\right)$$

$$\frac{n_1 \cdot S_1^2}{\sigma^2} \rightarrow \chi^2_{n_1-1} \quad \text{y} \quad \frac{n_2 \cdot S_2^2}{\sigma^2} \rightarrow \chi^2_{n_2-1}$$

por las propiedades de la distribución Normal y de la distribución Ji-cuadrado que vimos en su día, se verifica que:

$$\bar{x}_1 - \bar{x}_2 \rightarrow N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}\right)$$

$$\frac{n_1 \cdot S_1^2}{\sigma^2} + \frac{n_2 \cdot S_2^2}{\sigma^2} \rightarrow \chi^2_{n_1+n_2-2}$$

En consecuencia, de acuerdo con la definición de la distribución t de Student el cociente:

$$\frac{\dfrac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}}}{\sqrt{\dfrac{1}{n_1 + n_2 - 2}\left(\dfrac{n_1 S_1^2}{\sigma^2} + \dfrac{n_2 S_2^2}{\sigma^2}\right)}} \rightarrow t_{n_1+n_2-2}$$

sigue una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad, simplificando se obtiene la siguiente expresión:

$$\frac{\dfrac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2}}}}{\sqrt{\dfrac{1}{n_1+n_2-2}\left(\dfrac{n_1S_1^2}{\sigma^2}+\dfrac{n_2S_2^2}{\sigma^2}\right)}} = \frac{\dfrac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}}{\dfrac{1}{\sigma}\sqrt{\dfrac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}}} = \frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{\sqrt{\dfrac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$$

que utilizaremos profusamente en los temas siguientes.

Una observación que debe tenerse muy en cuenta es que a las dos distribuciones poblacionales, aunque distintas en media, se les ha exigido que tengan la misma desviación típica σ, lo cual ha sido imprescindible en la deducción de la distribución del estadístico. Esta suposición deberá tenerse en cuenta en las aplicaciones que hagamos de este resultado, controlando su cumplimiento.