Estadística Española • Volumen 59, número 192 / 2017, pp. 37-44

Estimación de la desviación estándar

Mariano Ruiz Espejo(*)
Universidad Católica San Antonio de Murcia

Resumen

En el presente artículo estudiamos las propiedades del estimador "cuasidesviación estándar muestral" como estimador de la "desviación estándar poblacional" cuando el diseño muestral es el muestreo aleatorio simple con reemplazamiento de tamaño fijo, así como cuando este tamaño muestral tiende a infinito.

Palabras clave: cuasidesviación estándar muestral, desviación estándar poblacional, muestreo aleatorio simple con reemplazamiento.

Clasificación AMS: 62D05, 62E20, 62Pxx.

Estimation of the standard deviation

Abstract

In the present article we study the properties of the estimator "sample standard quasideviation" as estimator of the "population standard deviation" when the sampling design is simple random sampling with replacement of fixed size, as well as when this sample size tends to infinite.

Keywords: sample standard quasideviation, population standard deviation, simple random sampling with replacement.

AMS classification: 62D05, 62E20, 62Pxx.

1. Introducción

La "desviación estándar" se define como la raíz cuadrada de la varianza de una población o de una variable aleatoria que la representa. Tiene una gran importancia en la inferencia clásica, sobre todo en relación con el estudio de la distribución normal como uno de los parámetros que determinan la distribución además de la media poblacional, pero su interés es más reducido en la inferencia tradicional en poblaciones finitas (un estudio de ambos tipos de inferencia ha sido hecho por Ruiz Espejo, 2014). La "desviación estándar

* Reconozco los comentarios de los profesores Juan José Egozcue Rubí (Universidad Politécnica de Cataluña, Barcelona) en una cuestión planteada por éste en el portal ResearchGate.net y de Jorge Ortiz Pinilla (Universidad de Santo Tomás, Bogotá).

poblacional", también llamada "desviación típica poblacional", se denota usualmente como σ y su expresión, en el caso de una población finita de tamaño N, es la siguiente

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Donde y_i es el valor de la variable de interés definida en la población finita, y el subíndice i corresponde a la unidad i-ésima de la población finita o universo de unidades identificadas

$$U = \{1, 2, \dots, i, \dots, N\}$$

Y la media poblacional la hemos denotado por

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Estas definiciones tienen su extensión directa en poblaciones infinitas, así como los estimadores que a continuación presentamos.

Partimos del estimador insesgado (caso particular de la cuasicovarianza muestral, Ruiz Espejo, 1997) y óptimo para distribución libre de la varianza poblacional σ^2, en el sentido de mínima varianza, como demuestra en Ruiz Espejo *et al.* (2013), que es el estimador cuasivarianza muestral s^2 en el muestreo aleatorio simple (con reemplazamiento). La optimalidad se debe también a un resultado expuesto en Zacks (1971, p. 150). Usando la terminología del libro de Ruiz Espejo (2013), este estimador insesgado de σ^2 está determinado por la expresión

$$s^2 = \frac{1}{n-1}\sum_{i\in\boldsymbol{s}}(y_i - \bar{y}_{\boldsymbol{s}})^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_{i_j} - \bar{y}_{\boldsymbol{s}}\right)^2$$

Donde ahora n es el tamaño muestral fijo o número de unidades (repetidas o no) que aparecen en la muestra ordenada $\boldsymbol{s}$ que es la secuencia muestral de unidades ordenadas por orden de obtención por diseño de muestreo aleatorio simple con reemplazamiento, y hemos denotado por la media muestral a

$$\bar{y}_{\boldsymbol{s}} = \frac{1}{n}\sum_{i\in\boldsymbol{s}} y_i = \frac{1}{n}\sum_{j=1}^{n} y_{i_j}$$

La notación $i \in \boldsymbol{s}$ indica que se suma para cada unidad i que aparece en cada una de las n extracciones en la muestra ordenada o secuencia $\boldsymbol{s} = (i_1, i_2, \dots, i_j, \dots, i_n) \in U^n$, donde el subíndice j indica la j-ésima extracción ($j = 1, 2, \dots, n$), e i_j la unidad de U seleccionada en la j-ésima extracción de la muestra.

El problema, al que tratamos de dar una respuesta, es acerca de la estimación de la función paramétrica σ. Según Ruiz Espejo (1987) no existe un estimador UMECM, es decir, "uniformemente de mínimo error cuadrático medio" en el caso de *población finita fijada*, para el diseño de muestreo aleatorio simple con reemplazamiento de tamaño fijo n. Por esta población finita fijada nos referimos al modelo general en el que cada unidad de la población finita tiene definida un único valor de la variable de interés y éste es observable y fijo para cada unidad dentro del conjunto de números reales, en el muestreo aleatorio simple con reemplazamiento. Es decir, no existe estimador UMECM para todo parámetro N-dimensional $(y_1, y_2, \dots, y_N) \in \mathbb{R}^N$, siendo $\mathbb{R}$ la recta real.

Este tipo de población finita fijada es la que tiene mayor interés práctico, ya que seleccionar una muestra aleatoria simple de una población infinita supuestamente real no es posible ya que no podemos identificar todas las unidades de dicha población para después "acceder a" y "observar" la información de la variable de interés en las unidades seleccionadas. Es decir, es prácticamente imposible llevar a efecto un marco desde el que seleccionar la muestra para después poder observarla, o, lo que es lo mismo, no podemos identificar las unidades y, por tanto, no podemos acceder a las unidades de la población infinita si ésta existiera en la realidad, no solo en el terreno de las ideas o hipótesis.

Lo deseable sería incluso que la función paramétrica desviación estándar poblacional σ tuviera un estimador insesgado, pero este supuesto deseable estimador no ha podido ser encontrado entre los estimadores que se han ido proponiendo en la literatura en el contexto de poblaciones finitas. Sin embargo, vamos a estudiar algunas propiedades del estimador "cuasidesviación estándar muestral" de σ, que definimos por

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i \in \boldsymbol{s}} (y_i - \bar{y}_{\boldsymbol{s}})^2}$$

Indicamos que en el caso de una población normal, Bolch (1968) describe un estimador insesgado de la desviación estándar σ. Pero, como hemos explicado, este resultado tiene interés teórico más que práctico, pues la selección de la muestra de una población normal tiene que ser artificial pero no puede ser obtenida de la realidad con rigor probabilístico en el diseño muestral ya que hay infinitas unidades no identificadas y así éstas no pueden ser seleccionadas para su observación con un diseño muestral de unidades todas ellas accesibles.

Otro estimador de la desviación estándar poblacional en el muestreo aleatorio simple sin reemplazamiento de tamaño efectivo n, con $2 \leq n \leq N$, basado en el estimador insesgado de la varianza poblacional es el que toma la expresión (Ruiz Espejo, 1995)

$$v = \frac{N-1}{N(n-1)}\sum_{i \in \boldsymbol{s}} (y_i - \bar{y}_{\boldsymbol{s}})^2$$

Donde ahora $\boldsymbol{s}$ representa a la muestra subconjunto de U de cardinal n. El estimador que se propone para la desviación estándar poblacional es precisamente $\sqrt{v}$, que subestima en promedio a la desviación estándar poblacional excepto en casos muy particulares de modo

similar a como veremos en el caso de muestreo aleatorio simple con reemplazamiento. Es decir, excepto cuando la variable de interés es constante en todas las unidades de la población finita, o bien cuando el tamaño muestral sea $n = N$ o el diseño muestral se trate de un censo de la población finita. En este caso, el estimador $\sqrt{v}$ es consistente para estimar la desviación estándar poblacional, y coincide con ésta en un censo, es decir $\sqrt{v} = \sigma$. En los demás casos no contemplados a estos dichos, este estimador $\sqrt{v}$ tiene un sesgo negativo para estimar σ.

2. Sesgo negativo de la cuasidesviación estándar muestral

Propiedad 1. El estimador "cuasidesviación estándar muestral" s, en el muestreo aleatorio simple con reemplazamiento de tamaño fijo n, tiene sesgo negativo o cero para estimar la "desviación estándar poblacional" σ. El valor del sesgo cero se obtiene en el caso trivial particular en que la variable de interés sea constante y fija en todas las unidades de la población finita, y en ese caso $s = \sigma = 0$ en todas las muestras aleatorias simples con reemplazamiento, concluyendo que su sesgo sería entonces $B(s;\ \sigma) = B(0;0) = E(0) - 0 = 0$.

Demostración. El estimador cuasidesviación estándar muestral s subestima en promedio a la función paramétrica desviación estándar poblacional σ. Para ello basta observar que la función raíz cuadrada es cóncava en la semirrecta real positiva o cero, por lo que aplicando el teorema de Jensen a la variable aleatoria positiva s tenemos que su esperanza matemática resulta ser

$$E(s) = E\left(\sqrt{s^2}\right) \leq \sqrt{E(s^2)} = \sqrt{\sigma^2} = \sigma$$

Dándose el signo de igualdad solo en el caso de que la variable aleatoria s tome un solo valor en todas las muestras, pero esto solo sería posible en el caso de que la variable de interés fuera constante en todas las unidades de la población finita. Cosa que podríamos considerar como excepción no usual en la práctica, por lo que podemos concluir que si la variable de interés fuese realmente variable y no tomase un único valor constante en todas las unidades de la población, entonces se dará el resultado siguiente

$$E(s) < \sigma$$

Lo que significa que el sesgo de s para estimar σ es negativo, salvo casos triviales de variable de interés constante en todas las unidades. Es decir, denotando el sesgo por B,

$$B(s;\ \sigma) = E(s) - \sigma < 0. \blacksquare$$

3. Convergencia en probabilidad

Si bien el estimador "cuasidesviación estándar muestral" s subestima en promedio a la "desviación estándar poblacional" σ, asintóticamente es buen estimador porque converge en probabilidad a dicha función paramétrica.

Para demostrarlo basta aplicar el teorema siguiente.

Teorema. Si una sucesión de variables aleatorias X_n converge en probabilidad a cierta constante c cuando n tiende a infinito, y f es una función derivable definida en el recorrido de la sucesión de las variables aleatorias, y esta derivada está acotada en un entorno de c, entonces la sucesión de variables aleatorias $f(X_n)$ converge en probabilidad a $f(c)$ cuando n tiende a infinito.

Demostración. Para verlo hacemos uso de que dada una constante $k > 0$, la probabilidad siguiente

$$p\{|f(X_n) - f(c)| < k\} = p\left\{|X_n - c|\frac{|f(X_n) - f(c)|}{|X_n - c|} < k\right\}$$

$$\geq p\{|X_n - c|\delta < k\} = p\left\{|X_n - c| < \frac{k}{\delta}\right\} \to 1$$

Donde δ es la cota superior positiva del valor absoluto de la derivada de la función f en un entorno de c. Como esta última probabilidad converge a uno cuando n tiende a infinito por definición de la "convergencia en probabilidad" de X_n a c, concluimos que la primera probabilidad también converge a uno cuando n tiende a infinito, o lo que se deduce lógicamente, que la sucesión de variables aleatorias $f(X_n)$ converge en probabilidad a $f(c)$. ■

Como consecuencia, tenemos la propiedad siguiente.

Propiedad 2. En poblaciones finitas fijadas y en poblaciones infinitas con cuarto momento central poblacional finito, la "cuasidesviación estándar muestral" s, en el muestreo aleatorio simple con reemplazamiento de tamaño fijo n, converge en probabilidad a la "desviación estándar poblacional" σ (si $0 < \sigma < \infty$), cuando n tiende a infinito.

Demostración. Nuestro caso es una aplicación directa del *Teorema* visto, concretamente la sucesión de variables aleatorias es $X_n = s^2$, que depende del tamaño muestral n, converge en probabilidad a $c = \sigma^2$ cuando n tiende a infinito. Indicamos que s^2 toma valores positivos o cero, y que para esos valores usamos la función derivable $f(x) = \sqrt{x}$ que tiene derivada positiva, decreciente y acotada en un entorno de σ^2, con $0 < \sigma^2 < \infty$. Luego s converge en probabilidad a σ.

Que s^2 converge en probabilidad a σ^2 puede verse a partir de la desigualdad de Chebychev y de que la varianza de s^2 es una función de n que converge a cero cuando n tiende a infinito, concretamente

$$V(s^2) = \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)} = o(n^{-1}) \to 0$$

cuando $n \to \infty$. Siendo μ_4 el cuarto momento central poblacional

$$\mu_m = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^m$$

para $m = 4$, y $\sigma^4 = \mu_2^2$ es el cuadrado de la varianza poblacional; ambas funciones paramétricas están definidas y son constantes positivas finitas para una variable de interés determinada, en el caso de una población finita. Hemos denotado por $o(n^{-1})$ a un infinitésimo de orden n^{-1}, es decir, como consecuencia converge a cero cuando n tiende a infinito. El resultado sigue siendo válido para cualquier población infinita cuyo cuarto momento central poblacional exista y sea finito.

En concreto, el resultado se puede formalizar de este modo; por el teorema de Chebychev, para toda constante $k > 0$

$$p\{|s^2 - \sigma^2| < k\} \geq 1 - \frac{V(s^2)}{k^2}$$

Esta cota inferior que aparece en el segundo miembro, de la probabilidad que aparece en el primer miembro, tiende a uno cuando n tiende a infinito, o lo que implica, que la cuasivarianza muestral s^2 converge en probabilidad a la varianza poblacional σ^2.

Una mera comprobación nos permite ver que si el momento central poblacional de orden cuatro existe y es finito, también existe y es finito el momento central poblacional de orden dos. Esto se debe a que si X es una variable aleatoria y existe su momento respecto al origen de orden cuatro

$$E(X^4) = [E(X^2)]^2 + V(X^2) \geq [E(X^2)]^2$$

Si denotamos ahora el caso particular $X = (y - \bar{y})$, al tomar esperanzas matemáticas tenemos el resultado buscado, es decir si μ_4 existe y es finito

$$\infty > \mu_4 \geq \sigma^4 \geq 0$$

Por lo que $0 \leq \sigma^2 \leq \sqrt{\mu_4} < \infty$, es decir, la varianza poblacional σ^2 existe y es finita. Este último resultado es especialmente útil con poblaciones infinitas, ya que para poblaciones finitas todos los momentos existen y son finitos. ■

Finalmente, veamos una propiedad también interesante para el estimador propuesto.

Propiedad 3. El estimador "cuasidesviación estándar muestral" s es asintóticamente insesgado para estimar la "desviación estándar poblacional" σ, bajo muestreo aleatorio simple con reemplazamiento. Es decir, el límite cuando el tamaño muestral n tiende a infinito, de la esperanza matemática $E(s)$, es igual a σ. Es decir,

$$\lim_{n\to\infty} E(s) = \sigma.$$

Demostración. El estimador s es asintóticamente insesgado para estimar σ, pues de lo contrario, por la *Propiedad 1*, se verificaría que el límite de la esperanza matemática $E(s)$,

cuando n tiende a infinito, sería menor que σ, y esto nos llevaría a la conclusión de que s no converge en probabilidad a σ. Esto último nos lleva a contradecir la *Propiedad 2*, lo que es un absurdo que viene de suponer que el límite de $E(s)$, cuando n tiende a infinito, es menor que σ. Luego concluimos el resultado enunciado. ■

4. Conclusiones

En conclusión, hemos visto que el estimador "cuasidesviación estándar muestral" subestima en promedio a la "desviación estándar poblacional", es decir tiene un sesgo negativo en el muestreo aleatorio simple (con reemplazamiento), salvo en casos particulares de escasa importancia. Pero tiene la buena propiedad asintótica de que dicho estimador converge en probabilidad a tal función paramétrica, cuando el tamaño muestral fijo tiende a infinito, en la generalidad de las poblaciones finitas fijadas excluyendo el caso trivial de que la variable de interés sea una constante fija en todas las unidades de la población finita. Además, dicho estimador s es asintóticamente insesgado para estimar la desviación estándar poblacional.

Sin embargo, una condición suficiente para que este resultado "sobre el papel" sea cierto en poblaciones infinitas, éstas tienen que verificar que el momento central poblacional de orden cuatro tiene que existir y ser finito, algo garantizado de modo natural en poblaciones finitas. Y digo "sobre el papel" porque su interés práctico del muestreo de poblaciones infinitas se reduce a selecciones de muestras artificiales, pues no sería posible seleccionar y observar una muestra aleatoria simple de una población infinita si no tenemos identificadas todas las unidades de la población, pero al ser ésta infinita resulta imposible tener un marco a priori de todas las unidades de la población infinita para que sean todas accesibles, y una vez seleccionada la muestra, observar los datos de cada unidad de la muestra obtenida.

Por todo ello, el estimador "cuasidesviación estándar muestral" es sesgado, de sesgo negativo (o cero, en casos particulares de poca importancia) para estimar la desviación estándar poblacional σ, pero este sesgo se hace asintóticamente en el tamaño muestral cada vez más reducido y tiende a ser aproximadamente nulo, es decir

$$\lim_{n\to\infty} B(s;\, \sigma) = 0$$

Este resultado asintótico queda lejos de poder ser aplicado en la práctica, ya que su interés real se centra en el valor del tamaño muestral muy grande, es decir, en un entorno de infinito. Lo que es en la práctica difícil de verificar ya que con el "tamaño de la muestra" aumenta el "coste de observación", lo que supondría una utilidad costosísima del resultado y, por tanto, lejos de ser aprovechable salvo que la convergencia sea rápida y el coste por observación disminuyese mucho conforme aumentase el tamaño muestral.

En cuanto a su error cuadrático medio, no alcanza un mínimo uniformemente en el modelo objetivo de *población finita fijada* (Cassel *et al.*, 1977; Ruiz Espejo, 1987, 2013), pero tiene la propiedad de que su cuadrado s^2 es un estimador insesgado óptimo, en el

sentido de mínima varianza, para estimar la varianza poblacional σ^2 para el planteamiento teórico de *distribución libre* (Zacks, 1971; Ruiz Espejo *et al.*, 2013, 2016).

Referencias

BOLCH, BEN W. (1968). «More on unbiased estimation of the standard deviation». *The American Statistician* 22 (3), 27.

CASSEL, CLAES-MAGNUS; SÄRNDAL, CARL-ERIK; & WRETMAN, JAN HAKAN (1977). «Foundations of Inference in Survey Sampling». Nueva York, NY: Wiley.

RUIZ ESPEJO, MARIANO (1987). «Sobre estimadores UMV y UMECM en poblaciones finitas». *Estadística Española* 29 (115), 105-111.

RUIZ ESPEJO, MARIANO (1995). «Una relación entre cuasicovarianzas muestral y poblacional». *Revista de la Academia de Ciencias Exactas, Físico-Químicas y Naturales de Zaragoza (2)* 50, 51-53.

RUIZ ESPEJO, MARIANO (1997). «Sobre la cuasicovarianza muestral en el muestreo aleatorio simple». *Revista de la Academia de Ciencias Exactas, Físico-Químicas y Naturales de Zaragoza (2)* 52, 55-57.

RUIZ ESPEJO, MARIANO (2013). «Exactitud de la Inferencia en Poblaciones Finitas». Madrid: Bubok.

RUIZ ESPEJO, MARIANO (2014). «Fundamentos de la Inferencia Estadística Objetiva» (3ª edición). Raleigh, NC: Lulu Press.

RUIZ ESPEJO, MARIANO (2015). «Sobre estimación insesgada óptima del cuarto momento central poblacional». *Estadística Española* 57 (188), 287-290.

RUIZ ESPEJO, MARIANO; DELGADO PINEDA, MIGUEL; & NADARAJAH, SARALEES (2013; 2016). «Optimal unbiased estimation of some population central moments». *Metron* 71, 39-62; 74, 139.

ZACKS, S. (1971). «The Theory of Statistical Inference». Nueva York, NY: Wiley.