Indicadores Descriptivos: ¿Qué es la Media?

Introducción
Un Experimento Docente de Premonición
Cálculo vs Interpretación
La Visión Matemática de la Media
La correcta Definición de la Media
Inseparabilidad de la Media y la Varianza
Conclusiones

Por Prof. Carlos Araújo[1]

Departamento de Estadística, Facultad de Matemáticas, PUC

Las opiniones vertidas en este documento son de exclusiva responsabilidad del autor. El Departamento de Estadística de la PUC no necesariamente concuerda con dichas opiniones. Todo comentario u observación puede ser enviado a: araujo@mat.puc.cl

Introducción

Iniciamos ahora una sucesión de artículos referidos a qué enseñar en la respectiva área temática de los métodos estadísticos. En este primer artículo de la sucesión trataremos un caso especial porque, debido a una mala elección de transmisión del mensaje educativo, ha motivado injustas críticas a la Estadística

Un Experimento Docente de "Premonición"

Propongo al docente en Estadística que está leyendo este artículo que realice el siguiente experimento en una clase de revisión de indicadores de Estadística Descriptiva:

Comience por preguntar a los alumnos qué es la Moda. Los que lo recuerdan responderán que es el valor observado con mayor frecuencia.

Pregunte a continuación qué es la Mediana y algunos alumnos dirán que es un valor que "divide el 50% de las observaciones". Al solicitarles una mayor precisión podrán decir que es un valor tal que a lo más el 50% de las observaciones tienen valores menores o iguales que él y simultáneamente, a lo más, el 50% de las observaciones presentan valores mayores o iguales que este valor.

Antes de preguntar qué es la Media, anote en un papel estas dos frases: 1) "El promedio"; 2) "La sumas de los valores dividida por la cantidad de observaciones". Oculte este papel y pregunte a los alumnos sobre la definición de este indicador. Una vez oídas sus respuestas, Usted sorprenderá a los alumnos por su capacidad de premonición al mostrarles el papel con sus anotaciones porque casi seguramente todas las respuestas dadas por los alumnos coincidirán con una de las dos alternativas de respuesta anotadas.

Los que responden "el promedio" cambian una palabra por otra y es en consecuencia una tautología (habría que preguntar qué es un promedio). La segunda respuesta es la que más preocupa y a la cual le dedicaremos lo que sigue en este artículo.

Cálculo vs Interpretación

El resultado del experimento del punto anterior no es sorprendente por cuanto la mayoría de los textos (el 75% de los textos del SIBUC indicados en el Artículo 02) que tratan la Media

[1] ***Prof. Carlos A. Araújo Ayesta*** fue Profesor del Centro Interamericano de Enseñanza de Estadística - CIENES (1967-1997), Asistente General del Director del CIENES (1974-1994) y Secretario Técnico de la Conferencia Interamericana de Estadística (CIE) de la OEA. A partir de enero de 2005 es Profesor en la Pontificia Universidad Católica de Chile - PUC.

como indicador descriptivo de un conjunto finito de datos cuantitativos, la "definen" como la suma de los números de dicho conjunto dividida por la cardinalidad (número de unidades) del conjunto. Esta "definición" no indica qué *significa* una Media sino cómo se *calcula*.

Se observa además que ésta no es la situación en los demás indicadores clásicos (Moda, Percentil etc.) donde la definición se refiere al significado o interpretación del respectivo indicador (asunto recordado por los alumnos) para posteriormente referirse su cálculo (mediante fórmulas generalmente olvidadas por los alumnos).

Por otra parte, si la Estadística Descriptiva tiene como propósito resumir la información contenida en el conjunto de números para favorecer la realización de inferencias inductivas, (ver Artículo A03) resulta de fundamental importancia conocer la interpretación o significado de los indicadores de resumen y en especial el de la Media.

Con la "definición" presentada ¿cuál es la interpretación de la Media más plausible para un usuario? La respuesta está dada en los primeros cursos de aritmética: *la Media es lo que le corresponde a cada elemento del conjunto de números si a todos los elementos le correspondiera el mismo valor*.

Mediante esta interpretación se ridiculiza a la Estadística en general diciendo que es una disciplina poco seria ("la Estadística miente") por cuanto afirma, por ejemplo, que: "si una persona come dos pollos y la otra no come alimento alguno, en promedio cada una come un pollo"[2].

La Visión Matemática de la Media

Esta "definición" de la Media tan difundida y que tan pobre imagen ha dado a la Estadística, tiene su origen en la creencia de que "la Estadística es una rama de las Matemáticas". En efecto, la siguiente proposición matemática es válida:

Sea $X = \{x_i \mid x_i \in R,\ i = 1, 2, ..., n\}$ Entonces:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \Leftrightarrow \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \le \sum_{i=1}^{n}\left(x_i - k\right)^2 \quad \forall k \in R$$

Debido a la doble implicación, desde un punto de vista matemático es lo mismo definir la Media de un conjunto finito de números usando cualquiera de las expresiones que figuran a los costados de la doble implicación.

En consecuencia desde el punto de vista matemático se escoge la expresión de la izquierda para definir la Media porque es más fácil de decir en palabras y por otra parte es la forma como se calcula (prueba su existencia).

La Correcta Definición de la Media

Pero como ya se explicó esta definición de Media conduce a explicables equivocaciones en su *interpretación.* La correcta definición de esta indicador se origina en la parte derecha de la doble implicación observando primeramente que la Media no pretende describir alguna propiedad del conjunto de números sino que pretende ***representar*** al conjunto de números.

Para ello, se recurre a una temprana utilización del ***principio de mínimo cuadrados***, por lo que se procede de la siguiente manera:

[2] Hay muchos otros ejemplos de absurdas conclusiones. La cabeza en el horno y los pies en el congelador podría dar una temperatura promedio aceptable; o bien si la profundidad media de un río es 156 cm. Una persona que mide 170 cm. y no sabe nadar podría cruzarlo sin peligro. Se recuerda que esto sólo refleja lo mal que se enseña la Estadística

Sea $X=\{x_i \mid x_i \in R,\ i=1,2,...,n\}$ el conjunto de observaciones y sea $k \in R$. La diferencia $(x_i - k)$ mide cuán distante o disperso se encuentra el valor de la i-ésima observación respecto del número k. Si esta diferencia es positiva x_i se encuentra a la derecha de k y si es negativa se encuentra a la izquierda de k . Como no se desea considerar el signo (positivo o negativo) de esta "distancia" o "dispersión" usaremos como indicador de esta dispersión, el cuadrado de la diferencia[3], vale decir: $(x_i - k)^2$

Así este cuadrado de la diferencia se interpreta también como un indicador de cuán bien representa k al valor de la i-ésima observación. Si está cerca de cero diremos que k representa adecuadamente a x_i en otro caso diremos que no representa adecuadamente a este valor.

La suma de estos valores para todas las observaciones ($\sum_{i=1}^{n}(x_i - k)^2$) nos proporcionará entonces un indicador de la dispersión de **todas** las observaciones respecto del número k o dicho de otra forma, un indicador de la representatividad de k respecto del conjunto de observaciones.

Si este número es grande k representa mal al conjunto de datos, y si es pequeño se dirá que k es un buen representante del conjunto de datos u observaciones $X=\{x_1, x_2, ..., x_n\}$.

El principio de mínimo cuadrados establece elegir como mejor representante del conjunto X es el número k_0 que verifica

$$\sum_{i=1}^{n}(X_i - k_0)^2 \leq \sum_{i=1}^{n}(X_i - k)^2 \quad \forall k \in R$$

En otras palabras el principio de mínimo cuadrados establece como representante del conjunto X el número k que minimiza $\sum_{i=1}^{n}(x_i - k)^2$. Utilizando o bien el álgebra elemental o bien el cálculo diferencial, es fácil demostrar que este mínimo se obtiene para $k=\frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}$.

La Media es entonces el mejor representante mínimo cuadrático de un conjunto de números y consecuentemente sólo pretende representar dichos números y no explicar algún tipo de comportamiento de los mismos.

Se hace notar que usando las actuales facilidades computacionales (hoja de cálculo) es posible enseñar esta definición de la Media a alumnos del segundo ciclo de la enseñanza formal.

Inseparabilidad de Media y Varianza

Debido a que el mejor representante mínimo cuadrático no es necesariamente un buen representante, la correcta interpretación de la Media requiere cuán bien dicha Media representa al conjunto de datos numéricos, es decir, se requiere definir un indicador de la representatividad de la Media.

---

[3] Si en lugar del cuadrado de la diferencia se emplea el valor absoluto de la misma $|x_i - k|$, se puede demostrar que el mejor representante del conjunto de datos es la Mediana de dicho conjunto (Med_x). El equivalente a la Varianza sería la Mediana del conjunto $\{|x_i - k| \mid i = 1, 2, ..., n\}$

Para ello si $\left(x_i-\overline{x}\right)^2$ es un indicador de la representación de $\overline{x}$ respecto del punto x_i se usa nuevamente el principio de mínimo cuadrado en la búsqueda del mejor representante del conjunto $\left\{\left(x_i-\overline{x}\right)^2 \mid i=1,2,\ldots,n\right\}$ que es la Media de este conjunto de números y que se define como la Varianza del conjunto X original es decir: $V\left(X\right)=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$.

Se observa que en este contexto, resulta además incoherente definir la Cuasi-Varianza[4] dentro de Estadística Descriptiva.

A efectos de interpretar su valor, este indicador puede ser expresado en otras unidades de medida mediante la Desviación Estándar (DE) o, mejor aún, el Coeficiente de Variación (CV).

Precisamente usando la regla práctica de que si $CV\left(X\right)>0{,}1$ la Media no representa adecuadamente a los datos, es fácil mostrar que si una persona come dos pollos y la otra no come, el Coeficiente de Variación es 1.0, es decir 10 veces mayor que el máximo aceptable y por lo tanto, en esta situación, la Media es un mal representante de lo que come cada persona (y se recuerda que, bajo el principio de mínimo cuadrado, no hay mejor).

Conclusiones

Siendo el propósito de la Estadística Descriptiva resumir la información contenida en un conjunto de datos, para que el usuario de estos resúmenes pueda formulas inferencias inductivas con mayor comodidad, es imperativo que dicho usuario tenga una correcta interpretación de todos y cada una de los elementos de este resumen.

En el caso de la Media resulta altamente insatisfactorio definirla mediante su fórmula de cálculo puesto que la única interpretación posible lleva a conclusiones ridículas aunque divertidas.

Asimismo la correcta definición de la Media como el representante mínimo cuadrático del conjunto de datos, muestra que para su apropiada interpretación resulta inevitable el conocimiento de la Varianza del referido conjunta de datos.

[4] Se define la Cuasivarianza como $\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\frac{n-1}{n}V\left(X\right)$. Carece de significado en Estadística Descriptiva y sólo puede ser considerado en Inferencia Estadística.