Boletín de Matemáticas
Nueva Serie, Volumen VI No. 1 (1999), pp. 43–51

NOTAS DE CLASE

PROMEDIO ARITMÉTICO Y VARIANZA EN GRUPOS FINITOS DE DATOS NUMÉRICOS

JORGE E. ORTIZ PINILLA (*)

Resúmen. Se recogen propiedades conocidas de la media y la varianza para mostrar que detrás de sus definiciones está el criterio de los mínimos cuadrados que rige buena parte de la estadística clásica. Se propone como alternativa pedagógica la presentación del promedio como el valor más cercano de los datos y la varianza como una transformación de los cuadrados de las distancias existentes entre ellos.

1. Introducción

Usualmente en los textos de estadística (Mendenhall, Sincich, 1997; Johnson, 1997; Canavos, 1988) la media y la varianza se presentan como definiciones que, eventualmente, se asocian con los conceptos de centro de gravedad y momento de inercia traídos de la física . Desde otra perspectiva complementaria se puede introducir tempranamente los modelos lineales a través del concepto de valores referenciales en un grupo de datos con la aplicación del criterio de mínimos cuadrados.

La identificación de grupos es una de las actividades más comunes del hombre. Cualquier forma de denominación corresponde implícita o explícitamente a un manejo más o menos generalizado de alguna clasificación de los objetos observados, sean estos concretos o abstractos.

(*) Texto recibido 10/22/99, revisado 5/30/00. Jorge E. Ortiz Pinilla, Departamento de Matemáticas, Universidad Nacional de Colombia;

Formalmente clasificar significa identificar o establecer una partición de un conjunto. Esto es, determinar varios subconjuntos disyuntos de manera que su unión cubra el conjunto completo. A los elementos de cada grupo se les da un nombre específico. Los seres humanos se clasifican, por ejemplo, en hombres y mujeres de acuerdo con la variable llamada sexo con la cual se establece la partición. Un conjunto de conejos puede descomponerse en varios grupos para ser asignados a tratamientos diferentes con algún objetivo específico. El grupo recibe algún nombre acorde con el tratamiento asignado y este procedimiento establece la partición del conjunto total que resulta en la clasificación de los conejos. En las variables que han servido como base para la clasificación, cada grupo tiene entonces, características propias que lo identifican como tal y permiten distinguirlo de otros.

Además de las características utilizadas para identificar la partición establecida, interesa observar el comportamiento de otras variables en los grupos. Cabe preguntarse si ellas revelan también características propias de cada grupo, que lo diferencian de los demás o si, por el contrario, esas diferencias son imperceptibles y su comportamiento es homogéneo en el conjunto total. La base de estos análisis se encuentra en el cálculo de expresiones que resumen de alguna manera el comportamiento de los datos del grupo.

En esta lectura se propone presentar la media y la varianza en conjuntos finitos, no como definiciones cuya justificación es posterior, sino como conceptos derivados de la aplicación del criterio de mínimos cuadrados para la búsqueda de valores referenciales que identifiquen el comportamiento global de variables numéricas en dichos conjuntos.

2. Valores referenciales de los grupos de datos numéricos

Cuando $x_1, x_2, \ldots, x_n$ son datos de una variable numérica observada en los elementos de un grupo, implícitamente se acepta la existencia de algo común en ellos, que evidencia la conformación del grupo. En caso contrario, difícilmente podría decirse que pertenecen al mismo grupo. La parte común es lo que se llamará un **valor referencial** del grupo y que lo especifica de alguna forma, al menos parcialmente.

Cada uno de los datos del grupo[1] puede descomponerse en dos partes:

$$
\begin{aligned}
& x_i = g + e_i, \qquad \text{en donde} \\
(1) \qquad & g \quad \text{es la parte común con el grupo y} \\
& e_i \quad \text{es la parte no común con el grupo y} \\
& \qquad \text{que es propia del dato individual}
\end{aligned}
$$

[1]los datos del grupo se entenderán como los datos de la variable numérica observada en los elementos del grupo

Esta última se puede interpretar como un **desvío individual** del dato con relación al valor referencial del grupo y corresponde exactamente a la diferencia entre el dato particular y el valor referencial.

Los desvíos individuales tienen una doble importancia desde el punto de vista práctico: (1) permiten examinar la homogeneidad o la heterogeneidad del grupo y (2) determinan en alguna forma los datos que son **típicos** dentro del grupo y los que son **atípicos** para el grupo. Generalmente se llaman atípicos a los datos que presentan grandes diferencias con el valor referencial.

Una vez definido, el valor referencial g no varía de individuo a individuo dentro del mismo grupo. Sin embargo, su definición es arbitraria y puede cambiar. Su selección determina los desvíos individuales y éstos son afectados de manera inmediata con los cambios en el valor referencial. Es posible que el mismo conjunto de valores atípicos también cambie.

A manera de ilustración se puede considerar el grupo de datos: $(3, 4, 4, 5)$ y suponer que se ha tomado arbitrariamente 5 como valor referencial. Los desvíos individuales son $-2, -1, -1, 0$ y el dato con mayor riesgo de ser calificado como atípico es el 3 por su mayor desvío. Si ahora se toma también arbitrariamente 20 como valor referencial, los desvíos individuales son muy grandes $(-17, -16, -16, -15)$ y todos los datos pueden ser calificados como atípicos. Por último, si se toma 4 como el valor referencial, los desvíos individuales son todos pequeños $(-1, 0, 0, 1)$ y lo más probable es que ahora ningún dato se tome como atípico. Esta última situación favorece la selección de 4 como el valor referencial adecuado para el conjunto de datos.

Parece claro que si el valor referencial representa la parte común de los datos del grupo, entonces su selección no debe precisamente generar condiciones para que los datos sean calificados como atípicos sino en los casos en que verdaderamente lo sean. De esta manera, el valor referencial debe estar lo más cerca posible de todos los datos del grupo. Por esta razón, los criterios para definirlo consisten en buscar el valor que minimice alguna expresión global de las diferencias con los datos del grupo. La más utilizada es la suma de los cuadrados de los desvíos individuales con relación al valor referencial. Entonces: g es el valor referencial del grupo de datos $x_1, x_2, \ldots, x_n$ si cumple la si-guiente condición:

$$\sum_{i=1}^{n}(x_i - g)^2 \leq \sum_{i=1}^{n}(x_i - y)^2, \qquad \forall y. \tag{2}$$

Sin mucha dificultad se puede demostrar que el valor buscado g es igual a:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{3}$$

conocido como **el promedio de los datos del grupo**:

En efecto, si los $x_i, i = 1, \ldots, n$ son los valores observados ya fijos, entonces $\sum_{i=1}^{n}(x_i - y)^2$ es función de y para la cual se busca el valor mínimo. Derivando,

entonces con respecto a y e igualando la derivada a cero se obtiene (Kreysig, 1974)

$$\frac{d}{dy}\sum_{i=1}^{n}(x_i - y)^2 = -2\sum_{i=1}^{n}(x_i - y) = 0,$$

entonces

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y = ny,$$

y por lo tanto, el valor mínimo se obtiene cuando

$$y = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

como se dijo antes. La segunda derivada es 2 lo que muestra que se trata de un valor mínimo. Se dice que **el promedio es el valor referencial del grupo, según el criterio de los mínimos cuadrados**.

Existe otra manera de ver el resultado anterior, que aporta elementos de utilidad:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - y)^2 &= \sum_{i=1}^{n}([x_i - \bar{x}] + [\bar{x} - y])^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - y)^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - y) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - y)^2 + 2(\bar{x} - y)\sum_{i=1}^{n}(x_i - \bar{x}).\end{aligned}$$

La suma que aparece en el último término es:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - n\bar{x} \\ &= \sum_{i=1}^{n} x_i - n(\frac{1}{n}\sum_{i=1}^{n} x_i) \\ &= 0.\end{aligned} \tag{4}$$

Por lo tanto, la siguiente igualdad es válida para cualquier conjunto de valores $y, x_1, x_2, \ldots, x_n$:

$$\sum_{i=1}^{n}(x_i - y)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - y)^2 \tag{5}$$

Se observa que si $y = \bar{x}$, entonces, $\sum_{i=1}^{n}(x_i - y)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ pero si $y \neq \bar{x}$, la cantidad positiva $n(\bar{x} - y)^2$ debe agregarse a $\sum_{i=1}^{n}(x_i - \bar{x})^2$ para que pueda igualar a $\sum_{i=1}^{n}(x_i - y)^2$.

Los resultados (4) y (5) destacan dos propiedades importantes de los promedios como valores referenciales de grupos:

1. La suma de todas las diferencias entre los valores individuales y el promedio del grupo es igual a cero (ecuación (4)).
2. Si se toma un valor referencial diferente del promedio, entre mayor sea la diferencia, mayor será la suma de los cuadrados de los desvíos individuales con relación a ese valor referencial (ecuación (5)).

Existen otras expresiones que sirven como base para definir el valor referencial de un grupo. Por ejemplo, en lugar de tomar la suma de los cuadrados de los desvíos, como se hizo en (2), se puede considerar la suma de sus valores absolutos, dando origen a otro valor referencial conocido como la mediana del grupo. En esta lectura, se trata solamente el promedio como valor referencial.

El modelo (1) se puede escribir ahora,

$$x_i = \bar{x} + e_i,$$

y el desvío individual es

$$e_i = x_i - \bar{x}.$$

En (4) se demostró que la suma de los desvíos individuales con respecto al promedio es igual a cero: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Un caso particular se presenta si todos los datos son iguales a su valor referencial (su promedio). Si esto ocurre, entonces no sólamente la suma de los desvíos es cero sino que también cada uno de ellos es también igual a cero. En el grupo no hay diferencias y todos los datos individuales se comportan exactamente igual.

La situación anterior es muy particular. En general en los grupos existen diferencias que, según se ha dicho antes, sirven como elemento de diagnóstico individual para cada dato, permitiendo así decidir si se lo considera cercano del valor referencial (dato típico) o si se encuentra demasiado lejos de él (valor atípico). Las técnicas orientadas a la elaboración de diagnósticos no necesitan presentación. Prácticamente todas las disciplinas realizan, de una forma u otra, este tipo de actividad.

3. Desvíos individuales y heterogeneidad del grupo

Como elemento de diagnóstico colectivo, los desvíos permiten analizar la homogeneidad o la heterogeneidad del grupo. Si se puede decir que los desvíos son globalmente demasiado grandes, el grupo se considera heterogéneo.

Lo más natural es tomar la suma de los desvíos como el indicador global de heterogeneidad. Sin embargo, ya se sabe por (4) que esa suma es igual a cero para cualquier grupo sin importar cómo sean los desvíos individuales, grandes o pequeños. Es la razón por la que anteriormente se debió utilizar la suma de los cuadrados de los desvíos u otra expresión adecuada.

Recordando que al tomar el promedio como el valor referencial, se está tomando el valor "más cercano" de todos los puntos en el sentido de los mínimos

cuadrados, la suma de los cuadrados de los desvíos se convierte en un indicador de la heterogeneidad del grupo. Si se obtiene un valor demasiado grande, el grupo es heterogéneo ; en caso contrario, el grupo es homogéneo.

El valor del desvío elevado al cuadrado, $(x_i - \bar{x})^2$, se denomina la variación del individuo i dentro del grupo. La suma de las variaciones individuales se llama la **variación interna** del grupo.

(6) Variación interna del grupo: $\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Para un conjunto de datos fijo, el promedio es un valor que no forzosamente coincide con alguno de los datos del grupo. Esto hace que en cierto modo sea considerado como un valor intruso que llega al grupo como referencia para establecer cuáles datos son atípicos y cuáles no. Intuitivamente, para examinar si un dato x_j es típico o no, se debería tomar como punto de partida la suma de sus diferencias al cuadrado con los datos del grupo:

$$\sum_{i=1}^{n}(x_i - x_j)^2$$

En este caso, la suma de las anteriores diferencias para todos los datos x_j, $j = 1, \ldots, n$ sería el indicador de la homogeneidad o heterogeneidad del grupo:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

La igualdad (5) es válida para cualquier valor y. En particular para x_j se tiene:

$$\sum_{i=1}^{n}(x_i - x_j)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - x_j)^2.$$

Acumulando estas sumas para todos los x_j se obtiene:

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - x_j)^2 &= \sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{n} n(\bar{x} - x_j)^2 \\ &= n\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\sum_{j=1}^{n}(x_j - \bar{x})^2 \\ &= 2n\sum_{i=1}^{n}(x_i - \bar{x})^2\end{aligned} \tag{7}$$

Se encuentra entonces que la suma total de las diferencias al cuadrado entre los datos de un grupo es precisamente un múltiplo de la variación interna del

mismo. El promedio es un dato referencial que permite simplificar los cálculos de la suma de los cuadrados de las diferencias individuales del grupo (entendiendo por diferencias individuales, las diferencias existentes entre los datos de los individuos). Tal vez sea más interpretable la suma de los cuadrados de las diferencias entre datos de individuos distintos y sin repetir comparaciones. Es decir, la siguiente suma de $n(n-1)/2$ términos:

$$\sum_{i=1}^{n}\sum_{j<i}^{n}(x_i - x_j)^2$$

No es difícil demostrar la siguiente igualdad

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = 2\sum_{i=1}^{n}\sum_{j<i}^{n}(x_i - x_j)^2. \tag{8}$$

De (7) y (8) se deduce que:

$$n\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}\sum_{j<i}^{n}(x_i - x_j)^2, \tag{9}$$

y como en esta última suma hay $n(n-1)/2$ términos, entonces el promedio:

$$\begin{aligned}\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i}^{n}(x_i - x_j)^2 &= \frac{2}{n(n-1)}n\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= 2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \end{aligned} \tag{10}$$

De este resultado se extrae una interpretación para la cantidad

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{11}$$

como el semipromedio de los cuadrados de las diferencias entre los datos de los individuos del grupo. Aparentemente debería ser más conocido el promedio que el semipromedio. Sin embargo la historia ha dado a conocer la expresión (11) como la **varianza muestral de los datos del grupo**. Nosotros la llamaremos también **la varianza interna del grupo** para mantener coherencia con la definición dada anteriormente en (6) para la variación interna.

Por otra parte si, en lugar de tomar solamente las diferencias entre datos de individuos distintos, se toman todas las n^2 diferencias que aparecen en la doble suma de (7), entonces el promedio será igual a:

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - x_j)^2 = \frac{1}{n^2}2n\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= 2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

por lo que la cantidad

$$s^2_{n,x} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{12}$$

corresponde al semipromedio de la suma de los cuadrados de las diferencias que se pueden formar con los datos de los n individuos del grupo, incluyendo las diferencias entre los datos de los individuos consigo mismos. Ellas son sistemáticamente iguales a cero y su inclusión ocasiona una disminución artificial que hace subvalorar la apreciación de la heterogeneidad de la variable en el grupo. Como puede verse entonces, esta expresión presenta dificultades para la interpretación y por lo tanto es menos recomendable, al menos desde esta perspectiva que la dada en la expresión (11).

4. Conclusiones

1. El promedio es el valor globalmente más cercano de los datos individuales de un grupo en el sentido de los mínimos cuadrados.
2. El promedio es la parte constante g de los datos x_i de un grupo, representados con el modelo

   $$x_i = g + e_i,$$

   en donde las partes no constantes e_i se han minimizado globalmente de acuerdo con el criterio de los mínimos cuadrados.
3. El promedio de los cuadrados de las diferencias entre todos los valores individuales de la variable numérica en estudio en un grupo está dado por:

   $$2s^2_x = \frac{2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

4. La varianza muestral es la mitad del promedio anterior.
5. El promedio de los cuadrados de las diferencias de los datos individuales con el promedio de los datos del grupo,

   $$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

incorpora los cuadrados de las diferencias de los datos consigo mismos, lo cual dificulta una interpretación para esta expresión, distinta de la de su definición.

6. El concepto de varianza en grupos finitos de datos numéricos puede definirse independientemente del valor referencial utilizado. Puede no utilizarse el promedio y de todas maneras la siguiente expresión define de manera única la varianza.

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j<i} (x_i - x_j)^2.$$

Referencias

[1] Canavos, George C., *Probabilidad y estadística: Aplicaciones y métodos*, McGraw Hill. Madrid, (1997).

[2] Johnson, R., *Probabilidad y estadística para ingenieros de Miller y Freund*, Prentice Hall. México (1997).

[3] Kreysig, E., *Introducción a la estadística matemática: Principios y métodos*, Limusa, México (1974).

[4] Mendenhall, W., Sincich, T., *Probabilidad y estadística para ingeniería y ciencias*, Prentice Hall Hispanoamericana, México (1997).

[5] Wonnacott, T. H., Wonnacott, R. J. , *Introducción a la estadística*, Limusa, México (1997).