Media, varianza y desviación estándar

Pedro Cervantes-Hernández *

La media ($\overline{X}$ en adelante) y la varianza (σ^2 en adelante) son estadísticos que se estiman a partir de una o varias muestras obtenidas de una población (Sharon 1999). La $\overline{X}$ es clasificada según Pérez (2002), como una medida de posición central y la σ^2 como una medida de dispersión.

La mayoría de las funciones estadísticas (univariadas, multivariadas y bayesianas) que se utilizan para describir y modelar datos, frecuentemente consideran dentro de su estructura matemática a $\overline{X}$ y σ^2, resaltando la importancia de éstas en el ámbito estadístico. Sin embargo, lo anterior, en ocasiones, no es bien reconocido y comprendido a plenitud, debido a una falta de claridad e interpretación que se tiene de sus conceptos. Una de las causas, que han propiciado lo anterior, se debe a que en la mayoría de los libros estadísticos dichos conceptos, al igual que la desviación estándar (σ en adelante), son abordados únicamente a nivel de función, sin considerar una explicación alterna que permita aclarar su importancia y aplicabilidad filosófica.

En este trabajo se utilizaron los conceptos de $\overline{X}$, σ^2 y σ descritos en Pérez (2002), los cuales fueron complementados, añadiendo una breve explicación concerniente a la importancia y aplicación de las funciones respectivas.

Para una población de tamaño N, $\overline{X}$ se define como la suma de todos los valores o datos (X_i) dividida por el número total de éstos ó N_i, la función que cuantifica a $\overline{X}$ es:

$$\overline{X} \quad \frac{\sum_{i\ 1}^{n} X_i}{N} \qquad (1)$$

Dentro de su estructura matemática, la ecuación 1 no proporciona información acerca del por qué $\overline{X}$ es una medida de posición central. Para abordar lo anterior, se utilizó la Figura 1, que muestra una dispersión espacial de datos hipotéticos y que en este caso, cada uno de éstos con respecto al total, serán descritos en términos de $\overline{X}$, σ^2 y σ. Para realizar lo antes mencionado, inicialmente se plantea la siguiente pregunta: ¿a partir de qué sitio en la Figura 1 se procederá con la descripción de lo datos?

La respuesta a la pregunta anterior, está en asociación a la manera de cómo en 1607 los ingleses, desde la diminuta aldea de Jamestown, Virginia, Estados Unidos (un sitio económico estratégico), comenzaron a explorar las tierras interiores de América del Norte, para detectar y seleccionar las más fértiles y prosperas, culminando en 1733, con el establecimiento de las 13 colonias a lo largo de la costa del Atlántico, desde New Hampshire hasta Georgia (Anónimo 2008).

La relación entre el ejemplo anterior y la Figura 1, se halla a que para el primero caso, fue necesario establecer un sitio estratégico mediante el cual se organizaron y ejecutaron las exploraciones a las tierras interiores de América del Norte. Este sitio estratégico se equiparó a establecer un punto de referencia dentro de la dispersión espacial, en el segundo caso, con base en el cual y de manera ordenada, se procederá a realizar la descripción de todos y cada uno de los datos con respecto al total. Este punto de referencia se estima con base en la función 1 y su posición dentro de la dispersión espacial (Fig. 1), está confinado al

* *Universidad del Mar, Instituto de Recursos, Ciudad Universitaria, campus Puerto Ángel, Apdo. Postal 47, Puerto Ángel, Oaxaca, 70902, México.*
Correo electrónico: pch@angel.umar.mx

sitio en donde se concentra la mayor cantidad de datos; por tanto, debido a las características antes señaladas, a $\overline{X}$ se le clasifica como una medida de posición central según Pérez (2002).

La descripción de los datos consiste en estimar el valor de la distancia que existe entre cada X_i con respecto a $\overline{X}$ (Fig. 2). La razón práctica de este cálculo, radica en conocer cuáles y cuántos de los X_i están cercanos y/o alejados de $\overline{X}$. Sin embargo, debido a que algunos de estos pueden estar mucho más cercanos y/o alejados que otros con respecto a $\overline{X}$. Se presenta un problema de posición, ocasionado por la relatividad de la distancia.

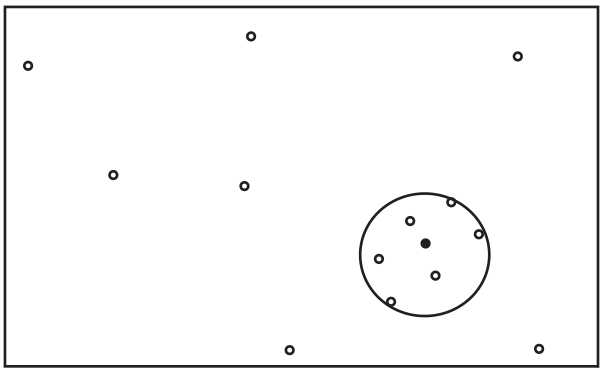

Figura 1. Dispersión espacial de datos hipotéticos. Círculos pequeños = datos ó X_i, círculo negro = $\overline{X}$, círculo mayor (zona con mayor cantidad de datos).

Para solucionar el problema anterior, es preferible considerar un promedio de dispersión de todos los X_i con respecto a $\overline{X}$; de manera que, a partir de éste, se cuantifique un porcentaje de datos cercanos y/o alejados de $\overline{X}$. A este promedio de dispersión se le conoce como σ^2 y de acuerdo con Meyer (1973), el proceso por el cual es estimado se le denomina "análisis de las desviaciones" o "análisis de dispersión".

Para una poblacion de tamaño N, σ^2 es una medida de dispersión de los valores o datos X_i, con respecto a $\overline{X}$, la ecuación que cuantifica a σ^2 es:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{N} \qquad (2)$$

La Figura 2 muestra la explicación alterna, que permite aclarar el concepto implícito en la ecuación 2. El valor particular de una distancia o desviación entre X_i con respecto a $\overline{X}$, se estima con base en el numerador $(X_i - \overline{X})$ (ecuación 2), en este caso, representado por una línea recta en la Figura 2. Meyer (1973) señaló que una de las propiedades de σ^2 es ser positiva, razón por la que el residuo anterior es elevado al cuadrado, esto es: $(X_i - \overline{X})^2$. Dado que se debe estimar el total de las distancias para obtener el promedio de dispersión, se aplica al numerador la sumatoria desde x= *i* a *n* y finalmente, éste se divide entre N datos (ecuación 2).

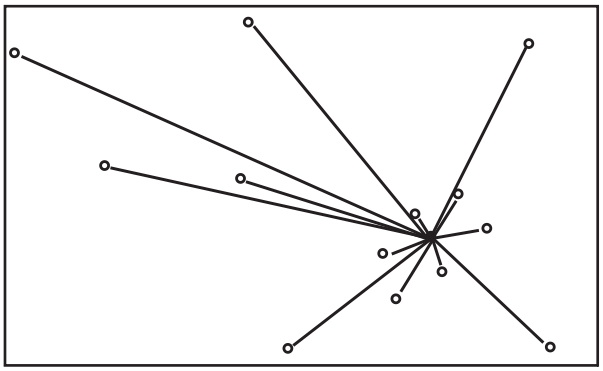

Figura 2. Dispersión espacial de datos hipotéticos. Líneas = distancias o desviaciones.

σ^2 en la ecuación 1 toma valores entre cero y uno, un valor cercano o igual a cero, indica que en promedio los datos se encuentran más cercanos a $\overline{X}$; mientras que, un valor cercano o igual a uno, indica que en promedio éstos se encuentran lejanos a $\overline{X}$. Sin embargo, en la práctica, estos valores no ocurren

comúnmente, predominando valores intermedios; razón por la cual, persiste la incertidumbre de cuáles y cuántos de estos datos están más cercanos y/o alejados de $\overline{X}$.

La solución al problema anterior, tiene su fundamento en el concepto de la desviación estándar (σ) o la raíz cuadrada de σ^2 según Mendenhall & Reinmuth (1981). La σ se utiliza para cuantificar un intervalo de confianza o límite de dispersión, dentro del cual los X_i incluidos se consideran cercanos a $\overline{X}$, mientras que fuera de éste se les considera alejados (Fig. 3). De acuerdo con esto, los límites de confianza se colocan sobre de σ^2, para describir lo antes señalado.

Los límites de confianza se cuantifican con

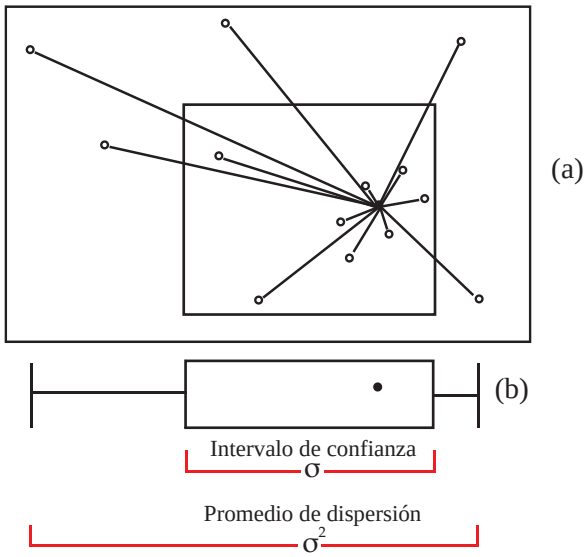

Figura 3. Dispersión espacial de datos hipotéticos con proyección a un diagrama estadístico de caja (a), señalando al intervalo de confianza (b) y el promedio de dispersión σ^2. Círculo negro =$\overline{X}$

base en el teorema de Tchebysheff en Mendenhall & Reinmuth (1981). Esto es: dado un número k mayor o igual a 1 y un conjunto de observaciones $X_1, X_2, \ldots X_n$, al menos $(1-1/k^2)$ de éstas caen dentro de k desviaciones estándar de la media.

La definición anterior hace referencia a que un intervalo de confianza se estima de la siguiente manera: $\overline{X} \pm (k \cdot \sigma)$, donde k es el número de veces que σ se aleja de $\overline{X}$. Para un valor de k = 1.3, el intervalo de confianza es: $\overline{X} \pm (1.3 \cdot \sigma)$ y de acuerdo con Tchebysheff, éste incluye el 41% de los datos; esto es: $(1-1/1.3^2)=0.408$.

Por acuerdo internacional, el intervalo de confianza se debe de cuantificar unificadamente o de manera estándar (de ahí el término desviación estándar), con un valor de k= 1.96, esto es: $\overline{X} \pm (1.9 \cdot \sigma)$, que genera un intervalo al 95%. Este acuerdo es considerado en todos los software de aplicación estadística, con opciones a modificar, según las necesidades que se requieran en el análisis y descripción de los datos.

Agradecimientos

Se agradecen los comentarios y sugerencias de Margarito Álvarez Rubio (ICMyL, UNAM).

Referencias

Anónimo. 2008. Trece Colonias. Consultado en junio de 2008: http://es.wikipedia.org/wiki/13_colonias

Mendenhall, W. & J.E. Reinmuth. 1981. Estadística para administración y economía. Grupo Editorial Iberoamericana, México, 707 pp.

Meyer, P.L. 1973. Probabilidad y aplicaciones estadísticas. 2a ed., Addison Wesley Iberoamericana, México, 480 pp.

Pérez, C. 2002. Estadística aplicada a través de Excel. Prentice Hall, Madrid, 616 pp.

Sharon, L. 1999. Muestreo, diseño y análisis. International Thomson Editores, México, 480 pp.