Sobre Parámetros Estadísticos desde un punto de vista Geométrico

Eugenio Saavedra Gallardo

1. Introducción

El propósito de este artículo es mostrar los conceptos de media y varianza, de un conjunto de datos, como el vértice de una parábola. Para ello, se introduce la noción de "mejor representante" a través de la distancia euclideana. Usamos una planilla de cálculo para ilustrar las ideas a través de un ejemplo.

Este método es una buena aplicación de la función cuadrática y puede ser usado como una aplicación de esta.

2. Revisión de conceptos básicos

Primeramente, recordemos que para cualquier conjunto numérico de datos $x_1, \ldots, x_n$, la media y la varianza de estos datos es definida por

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Además, desarrollando cuadrado de binomio, se obtiene que la varianza puede ser expresada como

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\,\bar{x} + \bar{x}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} 2x_i\,\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\,\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\,n\,\bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\,\bar{x} + \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \qquad (1)
\end{aligned}$$

También, si $(a_1, \ldots, a_n)$ y $(b_1, \ldots, b_n)$ son n-úplas de números reales, entonces la distancia entre $(a_1, \ldots, a_n)$ y $(b_1, \ldots, b_n)$, que anotamos $d((a_1, \ldots, a_n), (b_1, \ldots, b_n))$ es definida por

$$d((a_1, \ldots, a_n), (b_1, \ldots, b_n)) = \left(\sum_{i=1}^{n} (a_i - b_i)^2 \right)^{\frac{1}{2}}.$$

Así por ejemplo, si $n = 2$ y $A = (a_1, a_2)$, $B = (b_1, a_2)$, $C = (b_1, b_2)$ entonces el triángulo ABC es rectángulo, como el que muestra la figura siguiente

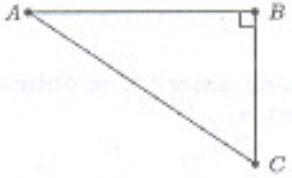


Figura 1

por lo que Teorema de Pitágoras implica que

$$\begin{aligned} AC &= \sqrt{AB^2 + BC^2} \\ &= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}. \end{aligned}$$

O sea,

$$AC = d((a_1, a_2), (b_1, b_2)).$$

3. Problema

Nuestro problema consiste en encontrar un valor, digamos c, que "represente" al conjunto de datos numéricos $x_1, \ldots, x_n$. La condición para encontrar al valor c es que sea un "buen representante".

¿Qué queremos decir con "buen representante"?.

Que se equivoque lo menos posible

¿Qué significa "equivocarse lo menos posible"?.

Significa que si t es otro representante cualquiera del conjunto de datos $x_1, \ldots, x_n$, entonces la distancia entre las n-úplas $(x_1, \ldots, x_n)$ y $(c, \ldots, c)$ debe ser menor que la distancia entre las n-úplas $(x_1, \ldots, x_n)$ y $(t, \ldots, t)$.

En consecuencia, nuestro problema consiste en encontrar el valor de c de modo que minimize la expresión $d((x_1, \ldots, x_n), (t, \ldots, t))$.

4. Solución

Si definimos la función $f(t)$ en la forma

$$f(t) = d^2((x_1, \ldots, x_n), (t, \ldots, t)),$$

entonces $d((x_1, \ldots, x_n), (t, \ldots, t)) = (f(t))^{\frac{1}{2}}$, por lo que minimizar $d((x_1, \ldots, x_n), (t, \ldots, t))$, equivale a minimizar la función $f(t)$.

Observemos que, de la definición de distancia entre dos n-uplas,

$$\begin{aligned} f(t) = & \quad d^2((x_1, \ldots, x_n), (t, \ldots, t)) \\ = & \quad \sum_{i=1}^{n} (x_i - t)^2. \end{aligned}$$

Desarrollando el cuadrado de binomio de la expresión anterior, se obtiene que la función cuadrática $f(t)$ puede escribirse en la forma,

$$\begin{aligned} f(t) = & \quad \sum_{i=1}^{n} (x_i^2 - 2x_i\, t + t^2) \\ = & \quad \sum_{i=1}^{n} x_i^2 \; - \; \sum_{i=1}^{n} 2t\, x_i \; + \; \sum_{i=1}^{n} t^2 \\ = & \quad \sum_{i=1}^{n} x_i^2 \; - \; 2 \left(\sum_{i=1}^{n} x_i \right) \, t \; + \; nt^2 \\ = & \quad \sum_{i=1}^{n} x_i^2 \; - \; 2n \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \, t \; + \; nt^2 \\ = & \quad \sum_{i=1}^{n} x_i^2 \; - \; 2(n\, \bar{x})\, t + nt^2. \end{aligned}$$

Pero, de (1), se deduce que

$$\sum_{i=1}^{n} x_1^2 = n\, \sigma^2 + n\, \bar{x}^2,$$

de donde

$$f(t) = n\, \sigma^2 + n\, \bar{x}^2 - 2(n\, \bar{x})\, t + n\, t^2. \tag{2}$$

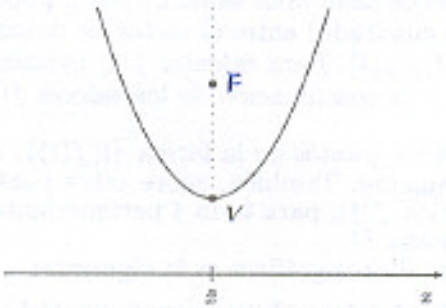
Finalmente, factorizando en esta última expresión resulta

$$f(t) = n\,[(t-\bar{x})^2 + \sigma^2]. \tag{3}$$

El gráfico de la función cuadrática $f(t)$, es la parábola de vértice $V = (\bar{x}, n\,\sigma^2)$ y foco $F = \left(\bar{x}, n\,\sigma^2 + \frac{1}{4n}\right)$.

Notar que tanto la ordenada del vértice como del foco son números positivos y además, $n\sigma^2 < n\sigma^2 + \frac{1}{4n}$, por lo que la parábola se abre hacia arriba. Así, el gráfico de $f(t)$ es como el de la figura siguiente

Figura 2



Observando la figura anterior vemos que el valor mínimo de la función $f(t)$ se alcanza en el vértice de la parábola, es decir, en el valor $c = \bar{x}$. Además, el valor mínimo de $f(t)$ es $n\,\sigma^2$.

En consecuencia, si t es real cualquiera,

$$d((x_1, x_2, \ldots, x_n), (\bar{x}, \bar{x}, \ldots, \bar{x})) \leq d((x_1, x_2, \ldots, x_n), (t, t, \ldots, t))$$

y

$$d((x_1, x_2, \ldots, x_n), (\bar{x}, \bar{x}, \ldots, \bar{x})) = \sqrt{n}\,\sigma.$$

Este último valor lo podemos llamar "error de representación".

En resumen, para un conjunto de datos $x_1, x_2, \ldots, x_n$, su media resulta ser el "mejor representante" de estos y su desviación estándar (salvo el factor $\sqrt{n}$) corresponde al "error de representarlos" por medio de $\bar{x}$.

5. Usando una hoja Excel en un ejemplo

Considere n datos. Puede escoger datos sacados, por ejemplo, de internet. Abriendo una hoja excel, en la columna B, ponemos los n datos $x_1, \ldots, x_n$. A continuación, para estos datos calculamos la media $\bar{x}$ y la varianza σ^2.

En la columna C se copian los n datos de la columna B y luego se ordenan en forma ascendente. Así, esta columna permite tener una mejor visualización de los datos (en particular, el rango donde estos varían).

En la columna D colocamos distintos valores de t, que pudiesen representar a los datos. El número de valores de t debe ser mayor que 1 y se sugiere menor que n.

El valor numérico escogido para cada t puede ser cualquiera (además de $\bar{x}$), pero es razonable pensar que un representante de los datos debería ser un valor que fluctúe entre el mínimo y el máximo de estos. Por ejemplo, entre los valores de t que podrían escogerse serían: el mínimo, el máximo, la mediana, la moda, algunos percentiles, etc.

En la columna E, y al lado de cada valor asumido por t, ponemos $f(t)$, es decir, ponemos la distancia (al cuadrado) entre el vector de datos $(x_1, x_2, \ldots, x_n)$ y el vector representante $(t, t, \ldots, t)$. Para calcular $f(t)$ usamos (3).

En la misma columna E, y a continuación de los valores $f(t)$, colocamos el valor mínimo de estos.

Para finalizar, graficamos los puntos de la forma $(t, f(t))$, obtenidos en las columnas D y E, respectivamente. También, sobre estos puntos, trazamos el gráfico de la función cuadrática $f(t)$, para todo t perteneciente al rango de los valores ingresados en la columna D.

La secuencia para obtener dichos gráficos es la siguiente:

- En las columnas D y E, seleccionar los valores de t y $f(t)$, respectivamente (incluya los encabezados t y $f(t)$). Ahora ir al menú asistente de gráficos y escoger XY (dispersión).
- Elegir comparar pares de valores - siguiente.

  Ahora en colocar gráfico, escoger la opción como objeto en: - finalizar.
- Finalmente, cliquear sobre un punto del gráfico (con botón derecho). Escoger la opción agregar línea de tendencia, ahora marcar polinomial (orden 2).

Referencias

[1] Allende, R. y otros (1996) *Comentarios sobre Propuesta Curricular para Estadística y Probabilidad en la Educación Media*, Revista de la Sociedad Chilena de Estadística, vol. 12, 57-72.

[2] Ministerio de Educación (2000) *Programa de Estudio Cuarto año Medio*, Formación General, Inscripción N^o 122.854, Santiago, Chile.

[3] Saavedra, E. (2003) *Cálculo de Probabilidades*, Editorial Universidad de Santiago, Santiago, Chile.

[4] Saavedra, E. (2005) *Contenidos Básicos de Estadística y Probabilidades*, Editorial Universidad de Santiago, Santiago, Chile. Por aparecer

Departamento de Matemáticas, Universidad de Santiago de Chile
Casilla 307, Correo 2, Santiago, Chile - email: keno@usach.cl