Punto de vista geométrico de algunos conceptos de estadística descriptiva

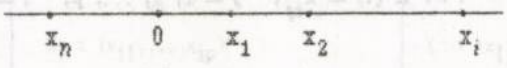
Hernán García

Para describir un conjunto $\mathcal{I}$ de n individuos u objetos, se considera la variable o característica constitutiva x.

x es una aplicación de $\mathcal{I}$ en $\mathbb{R}$ de tal manera que a cada individuo i le asocia el valor de la caracerística en él

$$x : \mathcal{I} \to \mathbb{R}$$
$$i \mapsto x(i) = x_i$$

Sea $E = \{x_1, x_2, \ldots, x_n\}$ el conjunto de observaciones. Este conjunto puede ser representado como puntos sobre la recta real o como un punto en $\mathbb{R}^n$.



Para resumir o condensar un conjunto de observaciones $E = \{x_1, x_2, \ldots, x_n\}$, se usan dos estadísticos: uno, una medida de tendencia central y otro, una medida de variación o dispersión.

Las medidas de tendencia central son valores que caracterizan la ubicación central de las observaciones. Algunas medidas de tendencia central son:

Promedio: $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$

Mediana: M(x) es un valor tal que, ordenadas en magnitud las observaciones, el 50% es menor que ella y el 50% mayor.

Promedio de valores extremos: $ME(x) = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$.

Las medidas de dispersión indican el grado de variabilidad entre las observaciones. Entre las principales medidas de dispersión se tiene:

Desviación Estandar: $S_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$

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

Desviación Mediana: $DM(x) = \sum_{i=1}^{n} \frac{|x_i - M(x)|}{n}$

Rango: $R(x) = \max\{x_i\} - \min\{x_i\}$.

Si a cada individuo $i \in \mathcal{I}$ se le asocian p variables o características cuantitativas, su representación se la hace mediante n puntos en $\mathbb{R}^p$

$$X^j : \mathcal{I} \to \mathbb{R}$$
$$i \mapsto X^j(i) = x_{ij}; \quad j = 1, 2, \ldots, p; \quad i = 1, 2, \ldots, n$$

El conjunto de observaciones $E = \left\{(x_{i1}, x_{i2}, \ldots, x_{ip}) : i=1,2, \ldots,n; \ j=1,2, \ldots,p\right\}$ se lo representa mediante una matriz:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

donde las n filas representan los n individuos y las p columnas las p variables.

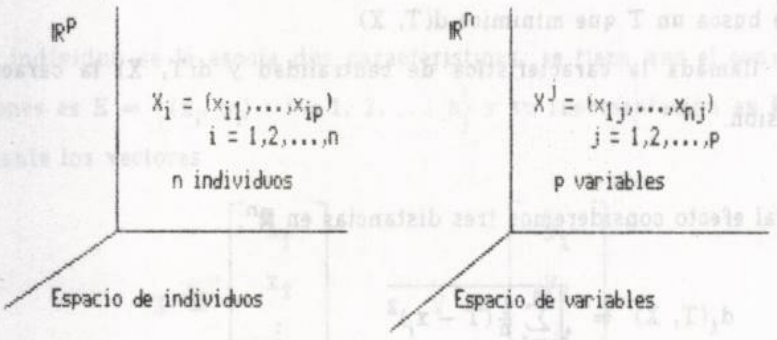
El vector: $X^j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$ representa las mediciones de la j-ésima variable

correspondiente a las n observaciones,

y el vector: $X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$ representa las $p$mediciones del individuo i,

o sea que x_{ij} es el valor de la caracteristica j correspondiente al individuo i

Desde un punto de vista geométrico es posible conceptualizar la matriz de datos multivariados de dos maneras: como un conjunto de n individuos en un espacio definido por las p variables, o como un conjunto de p variables definidas en un espacio de n dimensiones.



En el primer caso se comparan individuos considerados en función de sus características. Si por el contrario se comparan columnas, se obtendrá información acerca de la relación entre características consideradas en función de los individuos que se estudian.

Características de valor central y dispersión asociadas a la elección de un distancia en $\mathbb{R}^n$

En el espacio de variables $\mathbb{R}^n$, condensar un conjunto $E = \{x_1, x_2, \ldots, x_n\}$ al que solo se le considera una característica

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

es encontrar un valor único t que resuma la información de $X \in \mathbb{R}^n$. Se escoge un t de tal manera que

$$T = \begin{bmatrix} t \\ t \\ \vdots \\ t \end{bmatrix} \text{ sea el más próximo a } X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Para encontrar este valor de t se hace lo siguiente:

1) Se elige una distancia d en $\mathbb{R}^n$

2) Se busca un T que minimice d(T, X)

t será llamada la característica de centralidad y d(T, X) la característica de dispersión.

Para tal efecto consideremos tres distancias en $\mathbb{R}^n$:

$$d_1(T, X) = \sqrt{\sum_{i=1}^{n} \frac{1}{n}(T - x_i)^2}$$

$$d_2(T, X) = \max\left\{ |t - x_i| : i = 1, 2, \ldots, n \right\}$$

$$d_3(T, X) = \sum_{i=1}^{n} \frac{1}{n} |t - x_i|$$

Si $\mathbb{R}^n$ está provisto de la distancia d_1 se tiene que $d_1(T, X)$ es minimo para

$t = \sum_{i=1}^{n} \frac{1}{n} x_i = \bar{x}$, entonces $d_1(T, X) = \sqrt{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})^2} = S_x$.

Por lo tanto el promedio $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ y la desviación estandar $S_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$ son las medidas de centralidad y dispersión asociadas a la distancia d_1.

Si $\mathbb{R}^n$ está provisto de la distancia d_2, $d_2(T, X)$ es mínimo para $t = ME(x) = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$, entonces $d_2(T, X) = \max\left\{|x_i - ME(x)| : i=1,2, \ldots,n\right\} = \frac{R}{2}$ o sea que el promedio de valores extremos $ME(x) = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$ y el rango $R(x) = \max\{x_i\} - \min\{x_i\}$ son las medidas de centralidad y dispersión asociadas a d_2.

De la misma forma, si $\mathbb{R}^n$ está provisto de la distancia d_3, la mediana y la desviación mediana $DM(x) = \sum_{i=1}^{n} \frac{|x_i - M(x)|}{n}$ son las medidas de centralidad y dispersión asociadas a la distancia d_3.

Interpretación geométrica

Si a cada individuo se le asocia dos caracteristicas, se tiene que el conjunto de observaciones es $E = \left\{(x_i, y_i) : i = 1, 2, \ldots, n\right\}$ y su representación en $\mathbb{R}^n$ se la hace mediante los vectores

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Los valores de la media de las variables X e Y están dados por

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

Si $\vec{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ se tiene que

$$\bar{x} = \frac{1}{n}\langle \vec{1}_n, X \rangle = \frac{1}{n}[1, 1, \ldots, 1] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\bar{y} = \frac{1}{n}\langle \vec{1}_n, Y \rangle = \sum_{i=1}^{n} \frac{y_i}{n}$$

Los valores de las características centradas son

$$X - \bar{x}\,\vec{1}_n = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \bar{x} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$$

$$Y - \bar{y}\,\vec{1}_n = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$

La desviación estandar de la variable X viene dada por

$$S_x = \frac{1}{\sqrt{n}} \left\| X - \bar{x}\,\vec{1}_n \right\| = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

y la varianza

$$S_x^2 = \frac{1}{n} \left\| X - \bar{x}\,\vec{1}_n \right\|^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

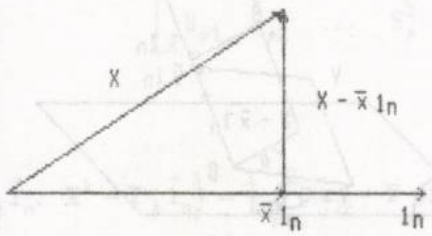
Para la variable Y:

$$S_y = \frac{1}{\sqrt{n}} \left\| Y - \bar{y}\, \vec{1}_n \right\| = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n}}$$

$$S_y^2 = \frac{1}{n} \left\| Y - \bar{y}\, \vec{1}_n \right\|^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n}$$

La proyección de la variable X sobre el vector unitario $\frac{1}{\sqrt{n}}\vec{1}_n$ es

$$X'\left(\frac{1}{\sqrt{n}}\vec{1}_n\right)\frac{1}{\sqrt{n}}\vec{1}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}\vec{1}_n = \bar{x}\,\vec{1}_n$$



o sea que la proyección de la variable X sobre la recta generada por el vector $\vec{1}_n$ es el vector $\bar{x}\,\vec{1}_n$. El vector $\bar{x}\,\vec{1}_n$ tiene longitud $\sqrt{n}\,|\bar{x}|$. Por lo tanto, la media $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ es la longitud de la proyección de la variable X sobre $\vec{1}_n$ y la desviación estandar S_x es la distancia de X a $\vec{1}_n$ multiplicada por $\frac{1}{\sqrt{n}}$.

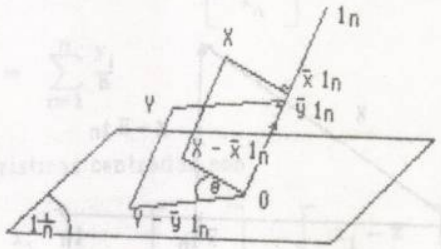
La dependencia entre las variables X e Y se la expresa mediante los conceptos de covarianza y el coeficiente simple de correlación lineal.

Desafortunadamente, es difícil utilizar la covarianza como una medida absoluta de la dependencia porque su valor depende de la escala de medición y por consiguiente es difícil determinar si una covarianza en particular es grande a simple vista. Se puede eliminar este problema al estandarizar su valor, utilizando el coeficiente simple de correlación lineal.

$$\operatorname{cov}(X, Y) = S_{xy} = \frac{1}{n}\langle X - \bar{x}\,\vec{1}_n,\ Y - \bar{y}\,\vec{1}_n\rangle = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\gamma(X, Y) = \gamma_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\langle X - \bar{x}\,\vec{1}_n,\ Y - \bar{y}\,\vec{1}_n\rangle}{\left\|X - \bar{x}\,\vec{1}_n\right\|\left\|Y - \bar{y}\,\vec{1}_n\right\|} = \cos\theta_{xy}$$

La interpretación geométrica es la siguiente: Sea $\vec{1}_n^{\perp}$ el complemento ortogonal de $\vec{1}_n$, entonces $\mathbb{R}^n = \vec{1}_n \oplus \vec{1}_n^{\perp}$.

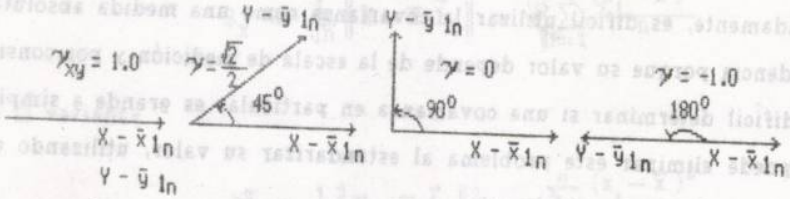


Las variables centradas $(X - \bar{x}\,\vec{1}_n)$ y $(Y - \bar{y}\,\vec{1}_n)$ son las proyecciones de las variables X e Y sobre $\vec{1}_n^{\perp}$; $n S_{xy}$ viene expresada por el producto interno de los vectores $(X - \bar{x}\,\vec{1}_n)$ y $(Y - \bar{y}\,\vec{1}_n)$.

El coeficiente de correlación lineal entre las variablex X e Y, es el coseno del ángulo entre $(X - \bar{x}\,\vec{1}_n)$ y $(Y - \bar{y}\,\vec{1}_n)$. Como $\gamma(X, Y) = \cos\theta_{xy}$, se tiene que

1) $-1 \le \gamma_{xy} \le 1$

2) $\gamma_{xy} = 0 \Leftrightarrow (X - \bar{x}\,\vec{1}_n) \perp (Y - \bar{y}\,\vec{1}_n)$



Representación del coeficiente de correlación en términos de los ángulos entre pares de variables centradas

Varianza generalizada

Cuando se analiza una variable, la varianza es usada para describir el grado de variación entre las medidas de esta variable. Cuando p variables son observadas en cada individuo, la variabilidad es descrita por la matriz de varianzas y covarianzas.

$$S = \begin{bmatrix} S_1^2 & S_{12} & \cdots & S_{1p} \\ S_{21} & S_2^2 & \cdots & S_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_p^2 \end{bmatrix}$$

donde,

$$S_{jk} = \frac{1}{n}\left\langle X^j - \bar{X}_j \vec{1}_n,\ X^k - \bar{X}_k \vec{1}_n \right\rangle = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{X}_j)(x_{ik} - \bar{X}_k);\quad j,k = 1,2,\ldots,p$$

$$S_{jj} = S_j^2,\ \text{varianza de la variable } X^j$$

La matriz de varianzas y covarianzas que contiene p varianzas y $\frac{1}{2}p(1-p)$ covarianzas, es una manera de expresar la dispersión de los datos alrededor de la media. Sin embargo, a veces es necesario disponer de un escalar que sintetice esta dispersión. Puede encontrarse un número que exprese la variabilidad multivariada a partir de la información contenida en la misma matriz S. Dada la matriz S, se denomina varianza generalizada al determinante de dicha matriz

$$V = |S|$$

y variación total a la traza de la matriz S:

$$\mathrm{Tr}(S) = \sum_{j=1}^{p} S_{jj} = S_1^2 + S_2^2 + \ldots + S_p^2$$

Tanto la varianza generalizada como la variación total serán mayores cuanto mayor sea la dispersión de los datos alrededor de la media.

Dadas las variables X^1, X^2, ..., X^p; el vector de medias $\bar{X}$ viene expresado mediante

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \quad \text{donde} \quad \bar{x}_j = \sum_{i=1}^{n} \frac{x_{ij}}{n}$$

Para las dos variables

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

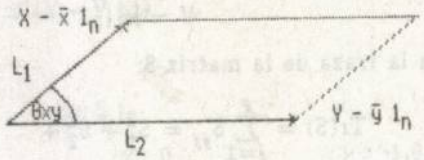
la varianza generalizada viene dada por

$$|S| = \begin{vmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_y \end{vmatrix} = \begin{vmatrix} S_{xx} & \sqrt{S_{xx}}\sqrt{S_{yy}}\,r_{xy} \\ \sqrt{S_{xx}}\sqrt{S_{yy}}\,r_{xy} & S_{yy} \end{vmatrix}$$

$$= S_{xx}S_{yy} - S_{xx}S_{yy}r_{xy}^2 = S_{xx}S_{yy}(1 - r_{xy}^2) = S_{xx}S_{yy}(1 - \cos^2\theta_{xy})$$

$$= S_{xx}S_{yy}\operatorname{sen}^2\theta_{xy} = \frac{1}{n^2}\left\|X - \bar{x}\vec{1}_n\right\|^2\left\|Y - \bar{y}\vec{1}_n\right\|^2\operatorname{sen}^2\theta_{xy}$$

$$= \frac{1}{n^2}\left(\left\|X - \bar{x}\vec{1}_n\right\|\left\|Y - \bar{y}\vec{1}_n\right\|\operatorname{sen}\theta_{xy}\right)^2$$

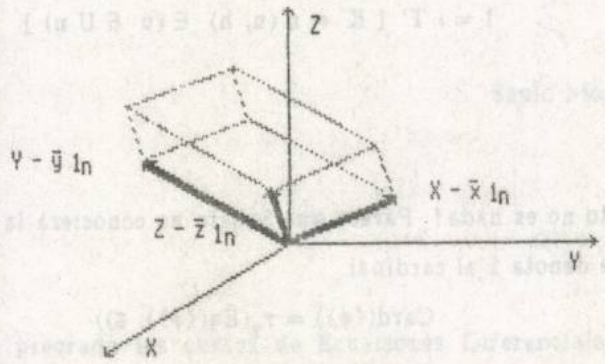


Area = $L_1L_2\operatorname{sen}\theta_{xy}$

Lo que implica que la varianza generalizada es proporcional al cuadrado del área del paralelogramo formado por las características centradas $X - \bar{x}\vec{1}_n$ y $Y - \bar{y}\vec{1}_n$.

Para p variables la varianza generalizada es proporcional al cuadrado del volumen generado por las características centradas $X^1 - \bar{x}_1 \vec{1}_n, \ldots, X^p - \bar{x}_p \vec{1}_n$.

Para un tamaño de muestra fijo, el volumen o $|S|$ crece cuando la longitud de cualquiera de las $X^i - \bar{x}_i \vec{1}_n$ o $\sqrt{S_{ii}}$ crece.



Varianza generalizada para p = 3

NOTA. La varianza generalizada $|S|$ es cero cuando los vectores $X^1 - \bar{x}_1 \vec{1}_n, \ldots, X^p - \bar{x}_p \vec{1}_n$ son linealmente dependientes.

BIBLIOGRAFIA

[1] CAILLEZ F., PAGES P. Introduction a L'Analise des Donnees. SMASH, Paris, 1976.

[2] JOHNSON R, WICHERN D. Applied Multivariate Statistical Analysis. Prentice – Hall, New Jersey, 1982.

[3] SEBER G. Multivariate Observations. John Wiley and Sons. New York, 1984.

UNIVERSIDAD DE NARIÑO
DEPTO. DE MATEMATICAS Y ESTADISTICA
PASTO

LOGISTICA

Algunos no creen en el porvenir de la ignorancia. He aquí la definición del número 1 que Burali-Forti da en su trabajo "Una questione sui numeri transfiniti":

$$1 = \iota\, T' \, [\, K \circ \eta\, (u,\, h) \in (u \in U\, n)\,]$$

E. Sábato

Uno y el universo

Nota: ¡Y esto no es nada! Parece que Sábato no conociera la de Bourbaki.

"Se denota 1 al cardinal

$$\mathrm{Card}(\{\phi\}) = \tau_Z(\mathrm{Eq}(\{\phi\}),\, Z)) \qquad (*)$$

(*) Es claro que no debe confundirse el *término* matemático *designado* (Chap.1, §1, No. 1) por el símbolo $\ll 1 \gg$ con la palabra $\ll 1 \gg$ del lenguaje ordinario. El término designado por $\ll 1 \gg$ es igual, en virtud de la definición dada arriba, al término designado por el símbolo

$$\begin{aligned}
&\tau_Z((\exists u)(\exists U)(u = (U, \{\phi\}, Z) \text{ et } U \subset \{\phi\} \times Z \text{ et}\\
&\quad (\forall x)((x \in \{\phi\}) \Rightarrow (\exists y)((x, y) \in U)) \text{ et}\\
&\qquad (\forall x)(\forall y)(\forall y')(((x, y) \in U \text{ et}\\
&\qquad\quad (x, y') \in U) \Rightarrow (y = y')) \text{ et}\\
&\qquad\quad (\forall y)((y \in Z) \Rightarrow (\exists x)((x, y) \in U)))).
\end{aligned}$$

Una estimación rápida muestra que el término así definido es una fórmula que consta de muchas decenas de miles de signos (cada uno de los cuales es uno de los signos τ, $\square$, $\vee$, $\neg$, $=$, $\in$, $\supset$)".

N. Bourbaki

Theórie des Ensembles

(Libro I, Cap.III, pág. 55)

Tomado de "LECTURAS MATEMATICAS", volumen II, No. 3.