CENTRADOS EN TI

Estadística descriptiva bivariada

Luis Andres Campos Maldonado

Maestría en analítica de datos

Introducción

La descripción de información para una sola variable, bien sea cualitativa o cuantitativa, proporciona información importante solo de aquella variable. Información adicional cuando se involucran más de una variable, resulta ser de gran utilidad en la descripción de un conjunto de datos. Tanto la información numérica como la gráfica, toma un mayor valor cuando se relacionan 2 o más variables. La creación de perfiles o segmentos nos puede dar una idea de que perfiles tienen ciertas características que pueden ser llamativas al ser contrastadas con otras. En caso de tener una variable objetivo de estudio, la información debe ser priorizada teniendo esta variable como referencia.

Tablas cruzadas

Supongamos que sobre la misma unidad experimental se han obtenido 2 características que son de tipo categórico, es este caso, un cruce de la información permitirá encontrar las cantidades absolutas y relativas de los diferentes perfiles. Consideremos que tenemos las características de: SEXO y ESTADO CIVIL (Soltero, Casado y Viudo), en este caso se obtienen 6 perfiles:

1. Mujeres solteras.
2. Mujeres casadas.
3. Mujeres viudas.
4. Hombres solteros.
5. Hombres casados.
6. Hombres viudos.

Al ser variables de tipo categórico, las medidas de estos perfiles serán de conteo.

Nota

Observe que además de lo anterior, y considerando una variable de tipo numérico, digamos la edad, podremos comparar perfiles vía esa variable cuantitativa, por ejemplo, comparar la media en la edad de la mujeres y hombres solteros.

Nota

Observe que además de lo anterior, y considerando una variable de tipo numérico, digamos la edad, podremos comparar perfiles vía esa variable cuantitativa, por ejemplo, comparar la media en la edad de la mujeres y hombres solteros.

Ejemplo: Tabla titanic3

SEXO / CLASE	HOMBRE	MUJER	% TOTAL
1	13.67	11.0	24.68
2	13.06	8.1	21.16
3	37.66	16.5	54.16
% TOTAL	64.40	35.6	100.00

Cuando los datos bivariados provienen de variables cuantitativas resulta de interés estudiar la relación que guarda una con la otra. La relación puede ser de diferente naturaleza: lineal, cuadrática, etc.

Covarianza

La covarianza es el valor que refleja en que medida dos variables aleatorias varían de forma conjunta respecto a sus medias. Adicionalmente, esta permite identificar la relación que existe entre las variables en este sentido, si una variable aumenta la otra también (igual si disminuye) en este caso se dice que hay una relación positiva. Y es negativa en el caso en que una aumenta y la otra tiende a disminuir.

Fórmula Covarianza

La covarianza esta dada por la siguiente fórmula

$$Muestra : Cov(X, Y) = S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

$$Poblacion : COV(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

donde $\overline{x}$ y $\overline{y}$ son las medias de las variables X y Y respectivamente y n es el total de observaciones.

Propiedades de la covarianza

1. $Cov(X, c) = 0$, donde c es una constante.
2. $Cov(X, X) = Var(X)$, la covarianza de una variable con sí misma es la varianza de la variable.
3. $Cov(X, Y) = C(Y, X)$, la varianza es simétrica, es decir, no importa el orden en que se coloquen la variables.
4. $Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X, Y)$.

Comentario

- La covarianza mide la fuerza de la relación lineal entre dos variables.
- Una covarianza alta no implica efecto causal.

Interpretación de la covarianza

1. $Cov(X, Y) > 0$; las variables X y Y tienden a moverse en la misma dirección.
2. $Cov(X, Y) < 0$; las variables X y Y tienden a moverse en direcciones opuestas.
3. $Cov(X, Y) = 0$: las variables X y Y no están relacionadas linealmente.

La covarianza presenta un inconveniente a tener en cuenta: Se expresa en términos de unidades no interpretables que provienen de las variables involucradas.
Para corregir esto, vamos a normalizar la expresión y obtendremos un nuevo coeficiente:

Coeficiente de correlación de Pearson

$$muestra : r_{xy} = \frac{S_{xy}}{S_x S_y} \qquad poblacion : \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

lo que es equivalente a

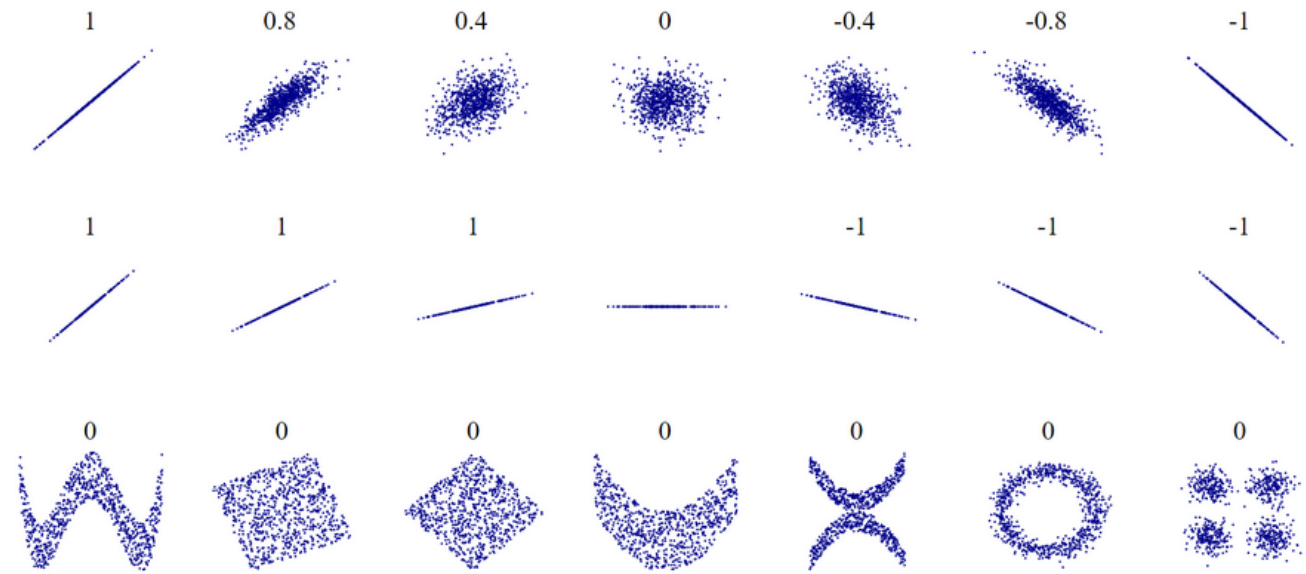
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{1/2}\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{1/2}}$$

Comentarios correlación

1. $-1 \leq r_{xy} \leq 1$
2. Si $r_{xy} = 1$, hay una relación positiva perfecta, es decir, indica una dependencia total entre las variables de manera directa. Si una de ellas aumenta, la otra también lo hace en proporción constante. Es decir, la relación es lineal.
3. Si $0 < r_{x,y} < 1$, existe una correlación positiva.
4. Si $r_{x,y} = 0$, no existe relación lineal. Pero pueden haber otro tipo de relaciones.
5. Si $-1 < r_{x,y} < 0$, existe una correlación negativa
6. Si $r_{x,y} = -1$, existe una correlación negativa perfecta, es decir, indica una dependencia total de manera inversa. Si una variable disminuye la otra aumenta en proporción constante.

Scatter plot

El gráfico en el cual se visualizan las variables se denomina diagrama de dispersión. Este gráfico muestra las parejas (x_i, y_i), donde x_i y y_i representan las mediciones del i-ésimo individuo en cada una de las variables.



https://upload.wikimedia.org/wikipedia/commons/d/d4/Correlation_examples2.svg