1. Introducción a probabilidad y estadística

1.1. Esperanza

Definición 1.1. (variable discreta) Sea X una variable discreta con función de probabilidad $P_X(x)$. La esperanza de X, denotada por $E[X]$, se define por

$$E[X] = \sum_{i=1}^{n} x_i \cdot P(X = x_i) \tag{1}$$

Definición 1.2. (variable absolutamente continua) Sea X una variable absolutamente continua con función de densidad $f_X(x)$. La esperanza de X, denotada por $E[X]$, se define por

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) \cdot dx \tag{2}$$

1.2. Varianza

La varianza da una medida de cuánto varían los valores de una muestra obtenida al realizar experimentos aleatorios.

Los x_i son los valores obtenidos. Mientras que $\bar{X}$ es el promedio de los resultados del experimento. N es la cantidad de resultados obtenidos. *Video Youtube min 5:40* **[lista_intro_data_science]**

$$Var(X) = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \bar{X})^2 \tag{3}$$

Cuandio la varianza es baja, los valores de la muestra se agrupn cerca de su valor esperado. Cuando la varianza es alta, los valores de la muestra se dispersan más.

Definición 1.3. (Varianza). Sea X una variable aleatoria con esperanza finita. La varianza de X se define por

$$Var(X) = E[(X - E[X])^2] \tag{4}$$

Definición 1.4. (Desviación estándar). La desviación estándar de X se define por

$$\sigma_X = \sqrt{Var(X)} \tag{5}$$

1.3. Covarianza

Definición 1.5. (Covarianza) La covarianza es una medida de cómo varían conjuntamente dos variables aleatorias.

Sean X e Y dos variables aleatorias de varianzas finitas definidas sobre el mismo espacio de probabilidad (Ω, A, P). La covarianza de X e Y se define por

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned} \tag{6}$$

Interpretación de la covarianza:

- Si $S_{xy} > 0$ hay dependencia directa (positiva), es decir, a grandes valores de X corresponden grandes valores de Y.
- Si $S_{xy} < 0$ hay dependencia inversa (negativa), es decir, a grandes valores de X corresponden pequeños valores de Y.
- Si $S_{xy} = 0$ no hay dependencia lineal entre X e Y.

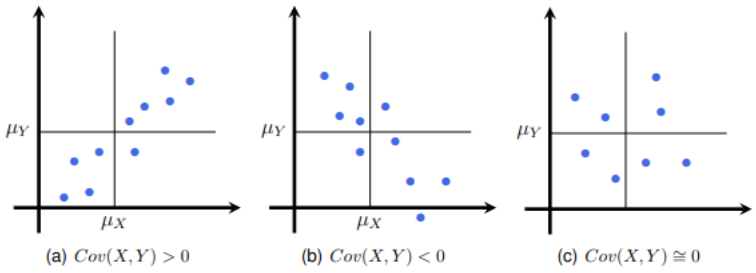

Figura 1: Covarianza

Definición 1.6. (Covarianza muestral)

$$S_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{7}$$

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \tag{8}$$

1.4. Correlación

Es una medida de relación lineal entre dos variables cualitativas continuas. Con esta medida, se logra determinar si las variables varían conjuntamente.

Es una medida normalizada, su valor va de -1 a 1. El caso en el que la correlación es 0, indica que no existe relación lineal entre las variables.

En caso de que se 1, se trata de correlación perfecta en sentido positivo. En caso de que sea -1, se trata de correlación perfecta en sentido negativo.

El sentido positivo, indica que varían en el mismo sentido. El sentido negativo, indica que varían en sentidos opuestos.

Definición 1.7. (Coeficiente de correlación)

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{9}$$

Definición 1.8. (Coeficiente de Correlación de Pearson)

$$\begin{aligned}\rho_{xy} &= \frac{S_{xy}}{S_x \cdot S_y} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}\end{aligned} \tag{10}$$

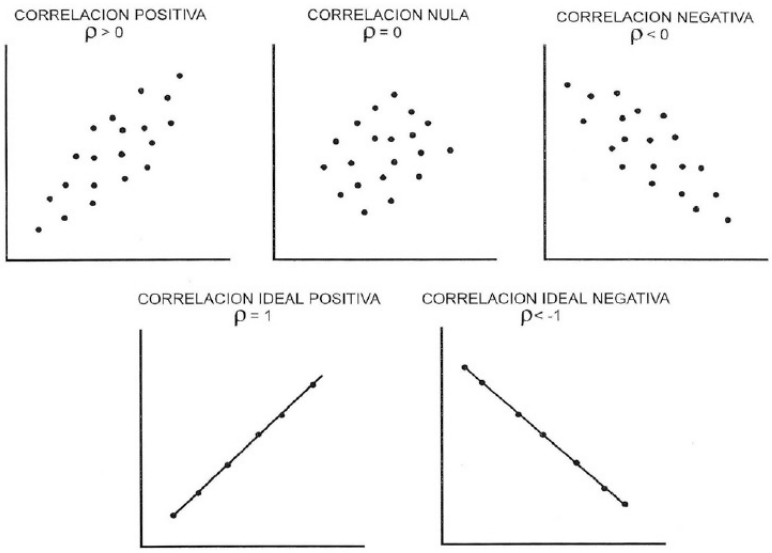

Figura 2: Coeficiente de Correlación de Pearson

Que las variables esten corelacionadas no implica que haya una relación de causalidad entre las mismas. Por ejemplo, si se toma la temperatura y la cantidad de helados vendidos, ambas variables están correlacionadas, pero no hay una relación de causalidad.

1.5. Ejemplos

Ejemplo 1.1. Ejemplo de como calcular datos variables aleatorias. Video Youtube [**ejemplo_variables_aleatorias**]

- **Esperanza**

$$\begin{aligned}E[X] &= \frac{1}{4} \cdot (2+3+5+6) = 4 \\ E[Y] &= \frac{1}{4} \cdot (1+2+2+3) = 2\end{aligned} \tag{11}$$

X	Y
2	1
3	2
5	2
6	3

Cuadro 1: Datos

- **Varianza**

$$\begin{aligned} Var(X) &= \frac{1}{4} \cdot [(2-4)^2 + (3-4)^2 + (5-4)^2 + (6-4)^2] = 2{,}5 \\ Var(Y) &= \frac{1}{4} \cdot ((1-2)^2 + (2-2)^2 + (2-2)^2 + (3-2)^2) = 0{,}5 \end{aligned} \tag{12}$$

- **Covarianza**

$$\begin{aligned} Cov(X,Y) &= \frac{1}{4} \cdot [(2-4)(1-2) + (3-4)(2-2) \\ &\quad + (5-4)(2-2) + (6-4)(3-2)] \\ &= \frac{1}{4} \cdot (2+0+0+2) \\ &= 1 \end{aligned} \tag{13}$$

- **Correlación**

$$\begin{aligned} \rho_{xy} &= \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \\ &= \frac{1}{\sqrt{2{,}5} \cdot \sqrt{0{,}5}} \\ &= \sqrt{\frac{4}{5}} = 0{,}8944 \end{aligned} \tag{14}$$