Computación Científica

Ingeniería de Sistemas y Computación

¿Qué es la estadística?

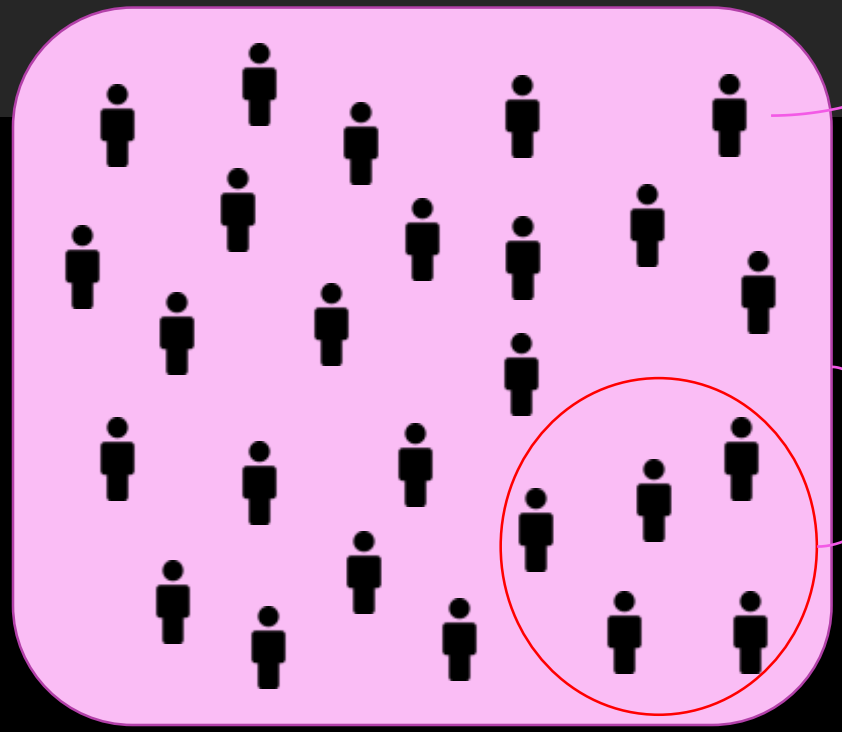
Definición

Es una disciplina científica que se ocupa de la obtención, orden y análisis de un conjunto de datos con el fin de obtener explicaciones y predicciones sobre fenómenos observados.

Tiene como objetivo mejorar la comprensión de los hechos a partir de la información disponible.

Tipos de estadística

- **Descriptiva:** Métodos de recolección, organización, resumen y presentación de un conjunto de datos. Describir las características fundamentales de los datos.
- **Inferencial:** Métodos utilizados para hacer predicciones, generalizaciones y obtener conclusiones a partir de los datos analizados.
- **- Paramétrica:** Asumen que los datos tienen una determinada distribución.
- **- No paramétrica:** No se puede determinar la existencia de una distribución.



Individuo: Elemento que aporta información sobre el fenómeno que se estudia.

Muestra: Subconjunto seleccionado de una población suficientemente representativo.

Población: Conjunto de todos los individuos que aportan información al fenómeno que se estudia.

Finitas

$$n = \frac{NZ^2pq}{d^2(N-1) + Z^2pq}$$

Infinitas

$$n = \frac{Z^2pq}{d^2}$$

p: proporción de la población de que aporta al fenómeno.
q: proporción de la población que no aporta al fenómeno.
Z: valor crítico dependiente del nivel de significancia.
- 99% -> 2.58
- 95% -> 1.96
- 90% -> 1.645

d: nivel de precisión absoluta.
- 90% -> 0.1
- 95% -> 0.05
- 99% -> 0.001

N: tamaño de la población.

Tipos de variables

Variables cuantitativas (numéricas):
Tienen valor numérico

Discretas:
Sólo toman valores enteros. Se asocian con conteos.

1, 2, 100, 10.000

Continuas:
Toman valores dentro del rango real. Se asocian con mediciones.

1.5, 3.1416, 2.18

Variables cualitativas (categóricas):
No tienen valor numérico

Nominales:
Expresa con su nombre una cualidad.

manzana, alto, amarillo

Ordinales:
Expresa un posible orden.

Primero, quinto, décimo

Medidas de Tendencia Central

Media

Se identifica como el punto de equilibrio de una muestra poblacional. Está dada por:

$$Media = \frac{\sum x_i}{N}$$

Mediana

Identifica el individuo que parte exactamente a la mitad el total de la población. Se define por.

$$Mediana = x_{\frac{N}{2}} \qquad Mediana = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$$

Moda

Es aquel dato que tiene una mayor frecuencia de generación.

$$Moda = mayor(F(xi))$$

Dan información del centro de los datos

Medidas de Dispersión o Variabilidad

Desviación Estándar

Es una medida de la dispersión de los datos de una muestra, expresando que tanta variabilidad tienen.

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

Varianza

Identifica que tanto varían las muestras de una población.

$$s = \sigma^2$$

Rango

Es el intervalo entre el mayor y el menor dato de una muestra

$$R = máx - min$$

Dan información sobre la disparidad de un conjunto de datos

Medidas de Posición

Cuartiles Son los valores que dividen la variable ordenada en 4 partes iguales

$$Q2 = Mediana$$

Deciles Son los valores que dividen la variable ordenada en 10 partes iguales

$$D5 = Mediana$$

Percentiles Son los valores que dividen la variable ordenada en 100 partes iguales

$$P50 = Mediana$$

Dan información sobre la cantidad de datos bajo una posición específica

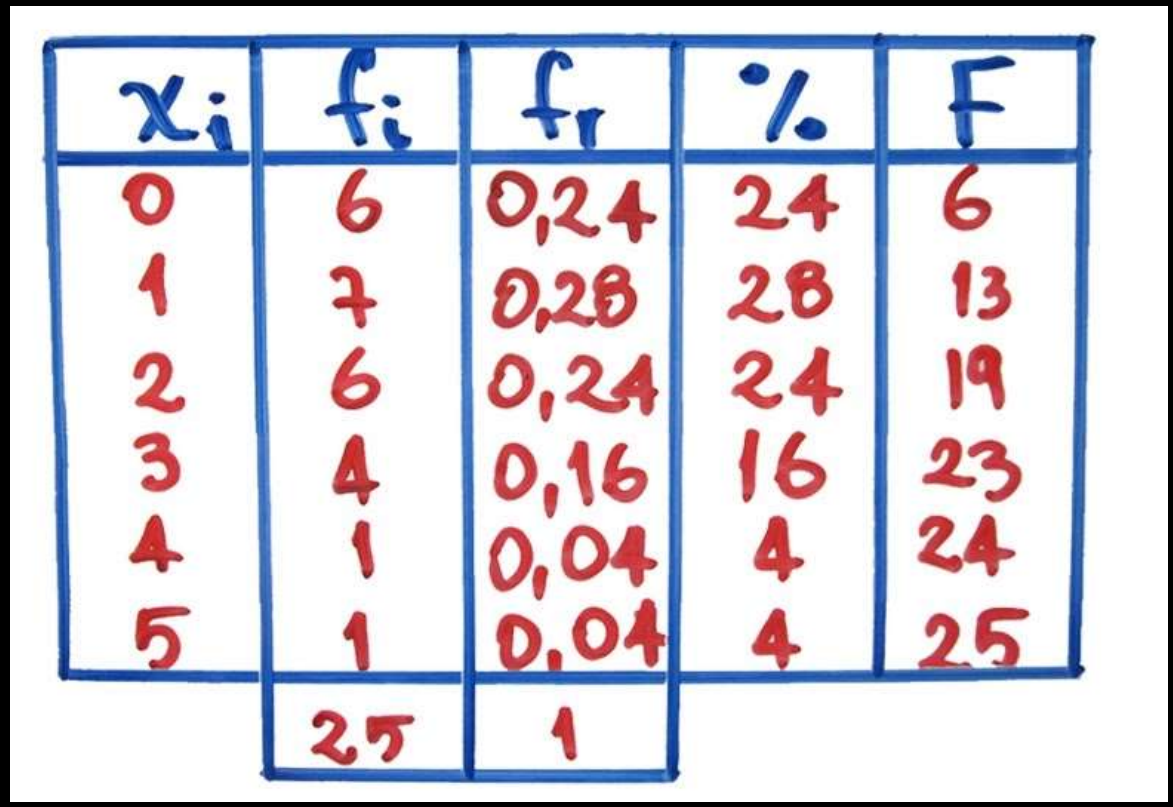
Frecuencia

Medición de la ocurrencia de un evento, esta frecuencia puede ser absoluta o relativa.

Frecuencia absoluta: Número total de veces que se repite una observación.

Frecuencia relativa: Es el valor relativo en porcentaje de la ocurrencia de una observación sobre el total de individuos.

Frecuencia acumulada: Cantidad acumulada de observaciones en una posición.



x_i	f_i	f_r	%	F
0	6	0,24	24	6
1	7	0,28	28	13
2	6	0,24	24	19
3	4	0,16	16	23
4	1	0,04	4	24
5	1	0,04	4	25
	25	1		

Intervalos de confianza

- Permite calcular un rango alrededor de una media muestral en el cual dicha medida existe con una probabilidad determinada.

$$IC = Media \pm Margen$$

Factores del intervalo de confianza

- Tamaño de la muestra: La cantidad de datos de cálculo de la variable muestral permite acercarse más o menos a la variable poblacional.
- Nivel de confianza: Porcentaje de casos en los que la estimación acierta.
- Margen de error: Probabilidad que el valor poblacional esté por fuera del intervalo.
- Medida de estimación: Fórmula pivote para estimación.

Cálculo de intervalo de confianza para la media

Estadístico pivote:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$\bar{X} = Media\ muestral$
$\mu = Media\ poblacional$
$\sigma = Desviación\ muestral$
$n = Tamaño\ de\ la\ muestra$
$\alpha = Nivel\ de\ confianza$
$Z\alpha_{/2} = probabilidad\ de\ un\ nivel\ de\ confianza$

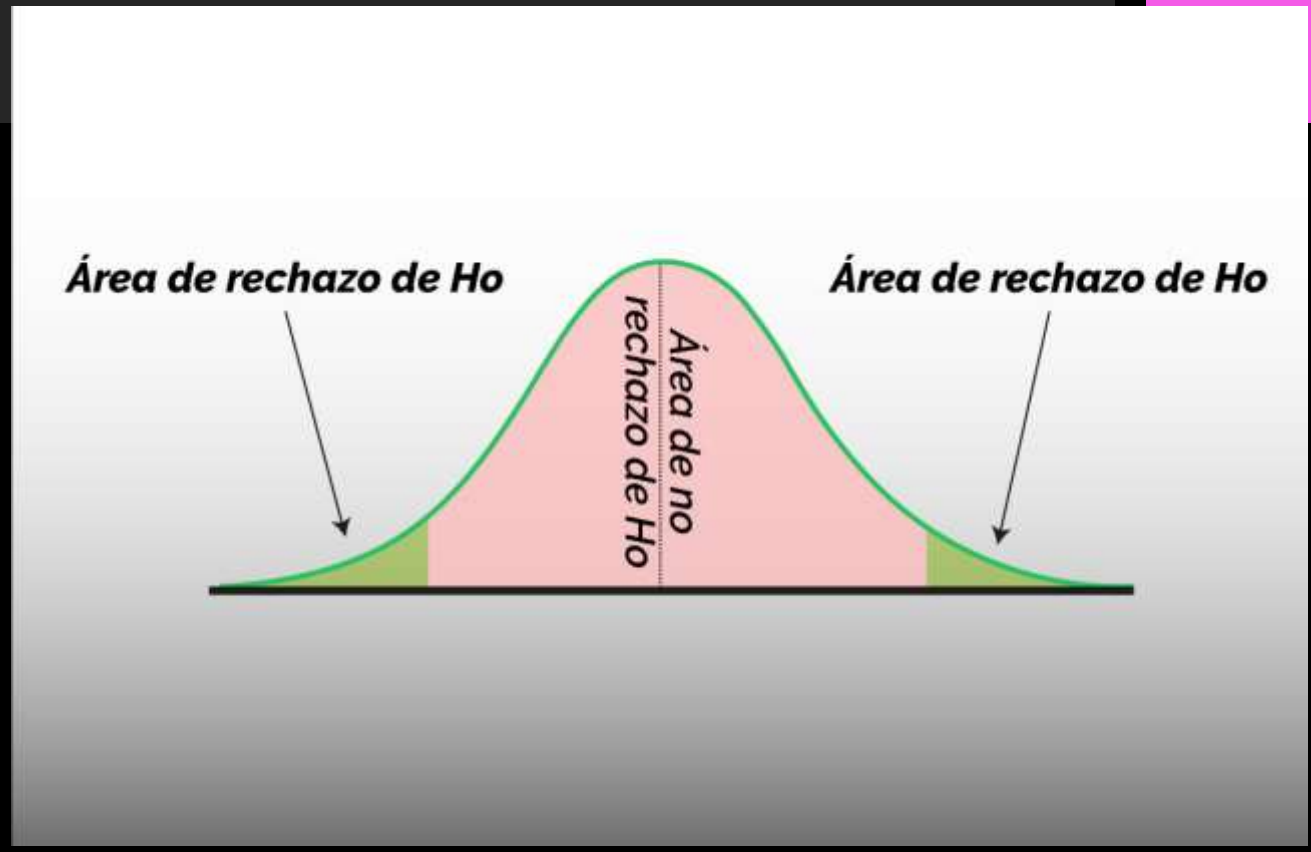
Intervalo resultante:

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} * Z\alpha_{/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} * Z\alpha_{/2}\right) = 1 - \alpha$$

Pruebas de hipótesis

- Procedimiento que busca tomar una decisión sobre el valor de verdad de una hipótesis estadística.

$$H0 = Hipótesis\ nula$$
$$H1 = Hipótesis\ alternativa$$



Análisis de correlación

- Trata de establecer la relación o dependencia entre dos variables que intervienen en una distribución bidimensional.

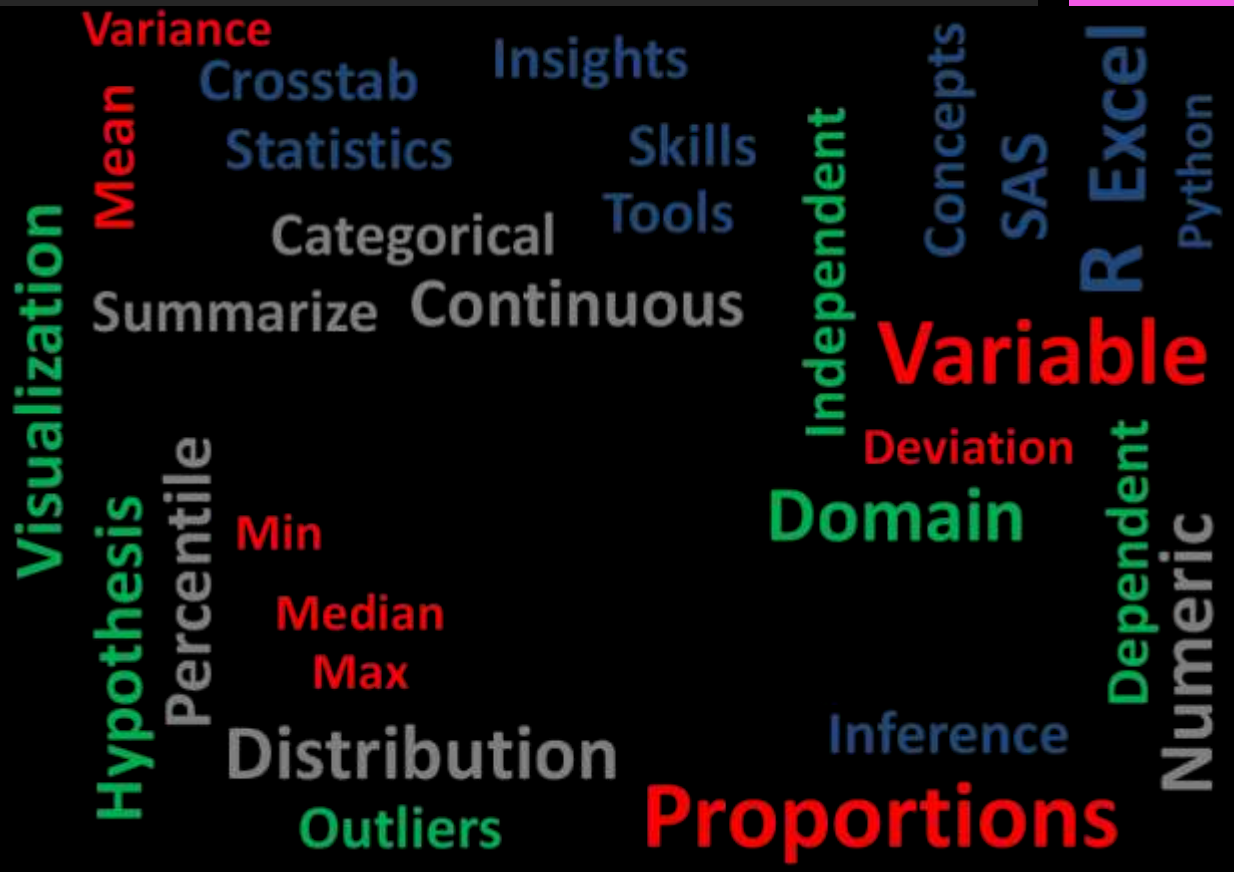
$$r = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Tipos de Correlación

- **Directa:** A medida que una variable crece la otra crece.
  - **Fuerte:** Correlación positiva por encima del 80%.
  - **Débil:** Correlación positiva por encima del 50% por debajo del 80%.
- **Inversa:** A medida que una variable crece la otra decrece.
  - **Fuerte:** Correlación negativa por encima del 80%.
  - **Débil:** Correlación negativa por encima del 50% por debajo del 80%.
- **Nula:** No se percibe cambio en una variable con respecto a la otra.

Análisis exploratorio de datos (EDA)

Es un acercamiento a analizar conjuntos de datos para resumir sus principales características, a menudo con métodos visuales. Se puede dar uso de un modelo estadístico, pero este proceso se usa esencialmente para visualizar qué nos pueden decir los datos más allá de un esquema formal de modelado.



Elementos del análisis exploratorio

- Fuentes de datos
- Limpieza de datos
- Análisis univariado
- Análisis bivariado
- Análisis mutivariado

Ingeniería de Características

Definición:

Es el proceso mediante el cual se seleccionan o transforman los datos, de forma que puedan ser entradas provechosas para un proceso de modelado. Tiene como objetivo logar el mejor desempeño de los modelos de datos.

X1	X2	X3	Y
a1	b1	c1	y1
a2	b2	c2	y2
a3	b3	c3	y3

Características

Importancia de la Ingeniería de Características

Mejores características implican mayor flexibilidad: Un modelo simple con buenas características puede generar un buen desempeño, son más fáciles de ejecutar y entender por el usuario.

Mejores características son modelos simples: Se pueden obtener modelos con parámetros no óptimos, pero con una identificación de los patrones subyacentes que ayudan a obtener un buen resultado.

Mejores características son mejores resultados: Obtener buenas características implica obtener los mejores resultados de modelado.

Técnicas comunes en Ingeniería de Características

Datos faltantes:

- **Eliminación de variables:** Se estima que una variable con una cantidad menor al 60% de los registros puede ser eliminada. Se recomienda si la variable no es relevante para el problema particular.
- **Imputación por media o mediana:** Mecanismo de completitud de variables numéricas. Se recomienda cuando los datos no cuentan con distribuciones sesgadas.
- **Imputación por regresión:** Mecanismo de completitud mediante modelos de regresión lineal de múltiples variables. Se recomienda con cualquier conjunto de datos faltantes.
- **Imputación por moda:** Mecanismo de completitud que usa la moda como mecanismo para variables numéricas y categóricas.

Técnicas comunes en Ingeniería de Características

Variables Continuas:

- **Normalización Min-Max:** Las características numéricas se ven escaladas en valores entre 0 y 1, mediante la siguiente fórmula:

$$x_{nueva} = \frac{x - min(x)}{max(x) - min(x)}$$

- **Estandarización:** Las características numéricas se convierten en variables cuya media es 0 y desviación estándar 1. Las lleva a una misma magnitud. Se usa la siguiente fórmula:

$$x_{nueva} = \frac{x - \mu}{\sigma}$$

Técnicas comunes en Ingeniería de Características

Variables categóricas:

- **Codificación ordinal:** Cada valor categórico se reemplaza con un valor numérico ordinal.

Fruta	Fruta - Codificada
Manzana	1
Pera	2
Papaya	3

- **Codificación matricial - dummy:** Cada valor categórico se reemplaza por un conjunto de 0 y 1, dependiendo de la cantidad de estados que tome la variable.

Fruta	Manzana	Pera	Papaya
Manzana	1	0	0
Pera	0	1	0
Papaya	0	0	1

Técnicas comunes en Ingeniería de Características

Importancia de las características:

- **Análisis univariado:** Mediante pruebas estadísticas se determina la relevancia de una variable sobre el problema tratado.
- **Puntuación:** Se establece un puntaje relacionado a su capacidad estadística para determinar si es relevante ante la salida esperada.
- **Correlación:** Mediante el índice de correlación de Pearson podemos identificar la fortaleza de la relación entre las variables.

Reducción de la dimensionalidad

Maldición de la dimensionalidad:

A medida que la cantidad de características alrededor de un problema aumenta significativamente, pero la cantidad de muestras no crece, la eficiencia del modelo cae exponencialmente.

X1	X2	X3	Y
a1	b1	c1	y1
a2	b2	c2	y2
a3	b3	c3	y3

Características

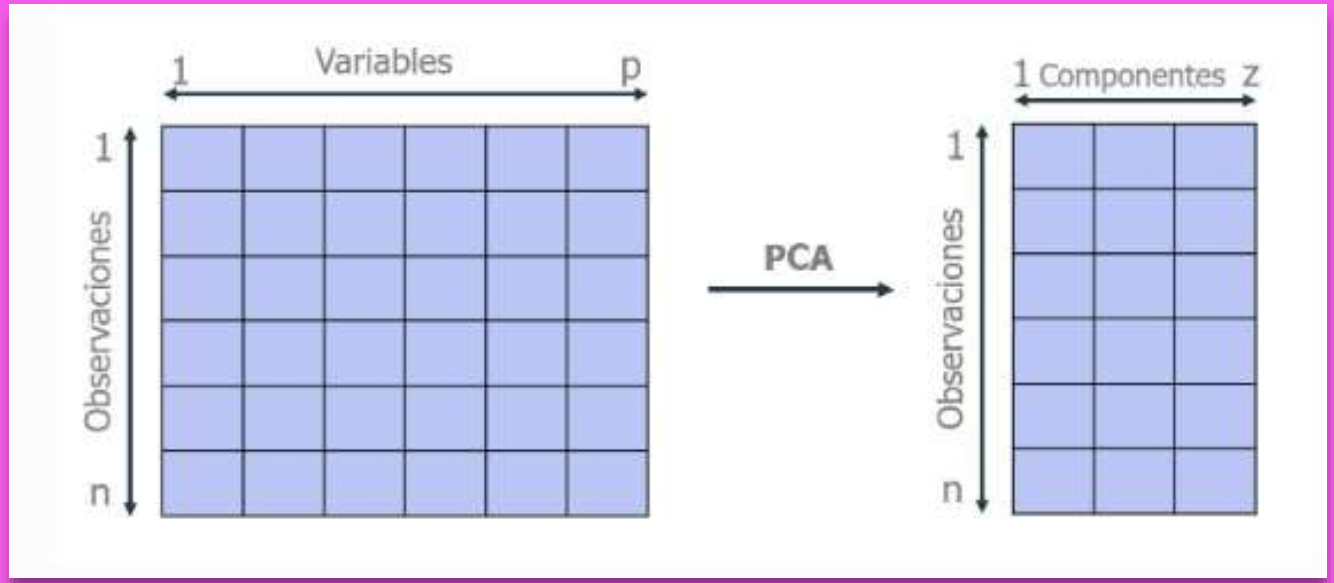
Análisis de Componentes Principales (PCA)

Definición:

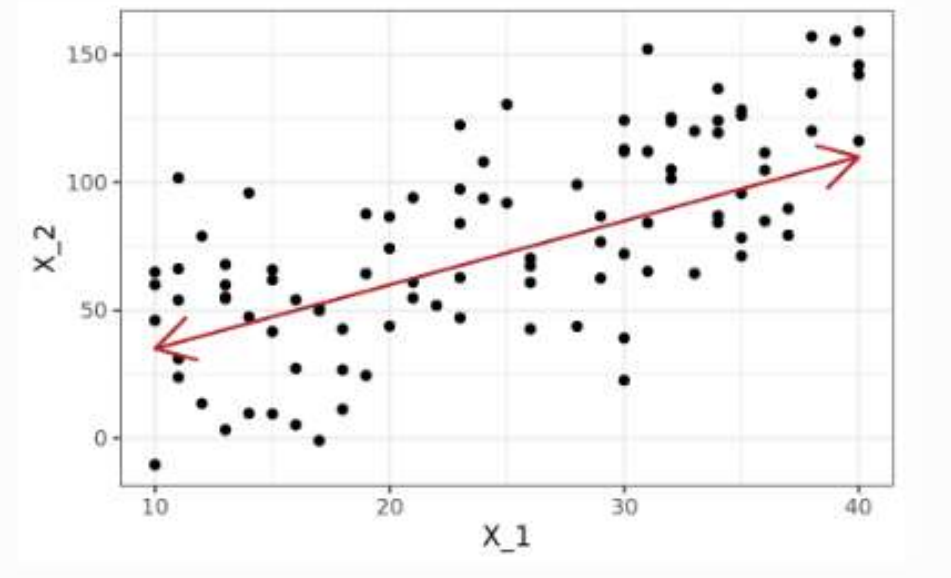
Procedimiento estadístico que transforma ortogonalmente las n dimensiones numéricas originales de un conjunto de datos en un nuevo conjunto de datos de m dimensiones llamados componentes principales.

Cada componente principal satisface la hipótesis de no correlación basado en que se tiene la mayor varianza posible manteniéndose ortogonal.

Nota: Para usar PCA los datos deben ser normalizados previamente. Aplicar PCA genera que el modelo de datos no sea explicable.

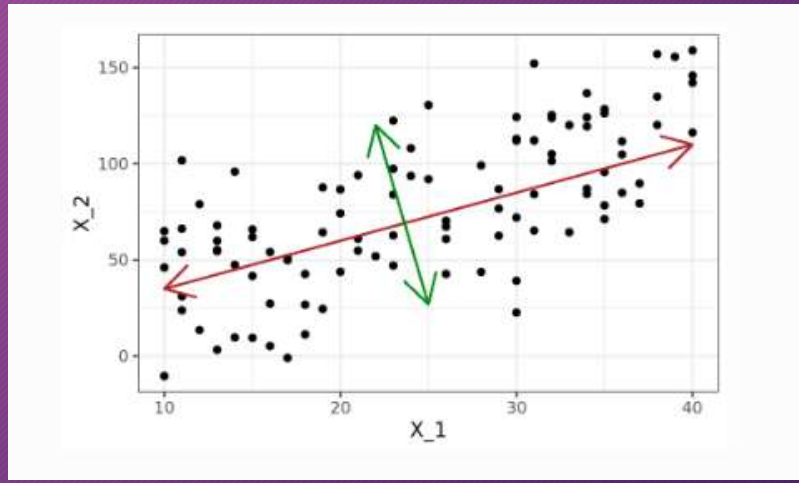


Análisis de Componentes Principales (PCA)



El vector que define la primer componente principal sigue la dirección de mayor varianza de las muestras. La proyección de cada observación sobre esa dirección equivale a la primer componente principal.

La segunda componente principal sigue la dirección en la que los datos muestran mayor varianza pero que no está relacionada con la anterior componente, es decir, es ortogonal.



Análisis de Componentes Principales (PCA)

Cálculo de las componentes principales:

Cada componente Z_i se obtiene de la combinación lineal de las variables originales, es decir cumplen con:

$$Z_1 = \varphi_{11}X_1 + \varphi_{12}X_2 + \cdots + \varphi_{1p}X_p \qquad \sum_{j=1}^{p} \varphi_{j1}^2 = 1$$

El problema se reduce a un ejercicio de optimización para encontrar los φ_{ij} que maximicen la varianza y que cumplan con la condición de ortogonalidad.

Dentro de PCA se mide la varianza explicada que indica la cantidad de información capaz de capturar por las componentes principales, esto se obtiene mediante:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p} \varphi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}$$

$$\frac{\text{Varianza explicada por la componente m}}{\text{Varianza propia del conjunto de datos}}$$

Análisis de Componentes Principales (PCA)

- **Elección de la mejor cantidad de componentes:**

Identificar el número de componentes mínimo donde el índice de varianza explicada deja de ser sustancial.