Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 1: 15/07/24 – 19/07/24

Introducción

✓ En practicamente todos los aspectos de la vida siempre es necesario **tomar decisiones** que **envuelven muchos factores/variables**.

Introducción

- ✓ En practicamente todos los aspectos de la vida siempre es necesario **tomar decisiones** que **envuelven muchos factores/variables**.
- ✓ Agunos factores/variables suelen influir más en la toma decisiones. Así que, **es de suma importancia identificarlos**!

Introducción

- ✓ En practicamente todos los aspectos de la vida siempre es necesario **tomar decisiones** que **envuelven muchos factores/variables**.
- ✓ Agunos factores/variables suelen influir más en la toma decisiones. Así que, **es de suma importancia identificarlos**!
- ✓ Por ejemplo, para tratar de entender/explicar la realidad de algunos acontecimientos y fenómenos, los **expertos colectan** la información de **aquellas variables** que de acuerdo con su conocimiento **consideran intermitentes e importantes** en dicho fenómeno.

Introducción

- ✓ Hay fenomenos/acontecimientos que dependen de **muchas variables** y los especialistas en tales fenómenos suelen estar interesados en establecer **relaciones entre ellas**, por lo tanto, se debe realizar un análisis conjunto de las variables (análisis multivariado).
- ✓ Cuando no se perciben las relaciones existentes entre variables, **los efectos desconocidos entre variables dificultan la interpretación del fenómeno** en función de las variables consideradas.

Pontificia Universidad JAVERIANA Bogotá

Introducción

El **Análisis Multivariado** es una área de la Estadística que buscar estudiar y desarrollar métodos que permitan **describir y analizar datos multivariados.**

Introducción

El **Análisis Multivariado** es una área de la Estadística que buscar estudiar y desarrollar métodos que permitan **describir y analizar datos multivariados.**

¿Qué son Datos Multivariados?

Son datos en que **se observan** varias características de cada unidad de la muestra. Por ejemplo, un banco puede estar interesado en relacionar las características (variables) de sus clientes, tales como:

- ✓ saldo en la cuenta para una fecha determinada,
- ✓ uso de la tarjeta de crédito para una fecha determinada,
- ✓ si el cliente tiene o no algún tipo de crédito,

entre otras.

Introducción

En este caso:

- ✓ las unidades muestrales son los clientes
- ✓ las variables están **correlacionadas**, por ejemplo, "el saldo en la cuenta para una fecha determinada " puede estar relacionado con el "uso de la tarjeta de crédito para una fecha determinada".

Esa **correlación** entre las variables conlleva al estudio del **comportamiento conjunto** y **condicional** entre ellas.

Introducción

	Saldo en la Cta	Uso de la TC	Crédito
Cliente 1	x_{11}	x_{12}	x_{1p}
Cliente 2	x_{21}	x_{22}	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Cliente j	x_{j1}	x_{j2}	x_{jp}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Cliente n	x_{n1}	x_{n2}	x_{np}

Observación: Aquí se tienen $p = 3$ variables y n unidades muestrales

Introducción

En concreto:

- ✓ el **análisis multivariado** corresponde a una **variedad de métodos/técnicas** que **envuelven simultáneamente a todas (o partes) las variables** en la interpretación del conjunto de datos.
- ✓ Existen varios **métodos** de análisis multivariado, con **finalidades** muy **diferentes**. La **determinación** de cada método va **depender de los objetivos de la investigación** o del conocimiento que se pretenda generar. Por lo tanto, se debe tener precaución y elegir las más adecuadas para detectar los patrones esperados en sus datos. Además, se debe comprender las limitaciones de estos análisis.

Introducción

- ✓ En general, el análisis multivariado es un análisis de tipo **exploratorio** que es utilizado para **generar hipótesis**, y no para proporcionar confirmaciones sobre las mismas.
- ✓ Dentro de los **objetivos específicos** de los métodos multivariados está:
  - **Reducir los Datos o su Dimensionalidad**
    Tratan de representar a los datos de la forma más simple posible sin pérdida de la información.
  - **Agrupar y Ordenar**
    Tratan de crear grupos de objetos o de variables que sean "similares".

Introducción

- **Clasificar**
  Tratan de generar reglas para clasificar objetos dentro de grupos bien definidos.
- **Investigar Dependencia entre Variables**
  Generalmente las variables de interés están correlacionadas.
- **Predecir**
  Una vez se establezcan las relaciones entre las variables, se trata(quiere) predecir los valores de una o más variables sobre las base de las observaciones de las demás variables.
- **Construir Pruebas de Hipótesis**
  Tratan de validar supuestos o reforzar convicciones a priori.

Panorama de los Métodos Multivariados

Principales Técnicas Multivariadas:

- ✓ **Análisis de Componentes Principales**
  Es una técnica de reducción de datos, cuyo objetivo principal es construir combinaciones lineales[1] de las variables originales (componentes principales) que contengan gran parte de la variabilidad total original.
- ✓ **Análisis de Agrupamiento, Conglomerado o Cluster.**
  Es una técnica de reducción de datos, cuyo objetivo principal es la identificación de un número pequeño de grupos (cluster), donde las observaciones de cada grupo sean similares[2] y muy diferentes a las de los otros grupos.

[1] No Correlacionadas entre sí

[2] Con respecto alguna medida de distancia y con base a las variables

Panorama de los Métodos Multivariados

✓ **Análisis Discriminante**
Es una técnica análoga al análisis de regresión donde la variable dependiente (respuesta) es categórica (grupos predefinidos) y las variables independientes (regresoras) son continuas. La idea principal es encontrar relaciones (lineales o no lineales) entre las variables independientes que mejor discriminen a los grupos.

✓ **Análisis Factorial**
Es una técnica de reducción de datos, cuyo objetivo principal es describir a cada variable en términos de una combinación lineal de un número pequeño factores comunes no observables (describen gran parte de la variabilidad que comparten las variables) y un factor único para cada variable (variación exclusiva de cada variable).

Panorama de los Métodos Multivariados

- ✓ **Análisis de Correlación Canónica**
  Es una técnica de reducción de datos, cuyo objetivo principal es identificar y cuantificar la asociación entre dos conjuntos de variables. La idea básica consiste en realizar combinaciones lineales entre las variables de cada grupo (variables canónicas), de tal forma que se maximice la correlación entre estas dos combinaciones (correlación canónica).
- ✓ **Escalonamiento Multidimensional**
  Es una técnica que buscar detectar dimensiones significativas subyacentes a una distribución de datos que permitan explicar las similitudes, diferencias o regularidades observadas entre las mediciones del fenómeno estudiado.

- ✓ **Análisis de Correspondencia**
  Es una técnica descriptiva/exploratoria diseñada para el análisis de tablas de contingencia que contienen algún tipo de correspondencia entre sus filas y columnas.
- ✓ **Análisis Log-Lineal**
  Es una técnica que busca investigar la relación entre tres o más variables categóricas dadas en una tabla de contingencia. Su idea básica consite en expresar las probabilidades de las celdas de la tabla de contigencia en términos de efectos principales e interacción para las variables de dicha tabla.

Panorama de los Métodos Multivariados

- ✓ **Árboles de Clasificación/Decisión**
  Es una representación de un <u>conjunto de reglas</u> creadas para <u>tomar</u> cualquier <u>decisión</u>, en particular, <u>clasificar</u> un registro (para problemas de clasificación) o <u>estimar</u> un valor (para problemas de regresión). En cada pregunta del árbol se responde "SÍ" o "NO", las respuestas guiarán hasta la decisión final.
- ✓ **Regresión Lineal y NO Lineal**
  Es una técnica centrada en la relación (lineal o no lineal) entre una variable dependiente (respuesta) y un conjunto de variables regresoras/predictoras (independientes entre sí) que pueden ser discretas. El objetivo principal es medir el efecto que tiene cada una de las variables regresoras sobre la variable respuesta.
  Regresión Multiple, Análisis de Varianza, Logística, etc.

Panorama de los Métodos Multivariados

Es evidente que:

- ✓ algunos de las técnicas anteriores (**Regresión Lineal**, **Análisis Factorial**, **Análisis Discriminante**) se enfocan en atender problemas en que los que hay variables independientes (regresoras, predictoras, explicativas) y variables dependientes (respuesta). Este tipo de problemas se denominan problemas de **Aprendizaje Supervisado**.
- ✓ algunas de las técnicas anteriores (**Análisis de Componentes Principales**, **Análisis de Agrupamiento**, **Análisis de Correlaciones Canónicas**, **Análisis de Correspondencia**) centran su objetivo en estudiar la variabilidad de los datos de forma multivariada (no hay variable respuesta), principalmente con el interés en reducir su dimensionalidad. En este caso, los problemas son llamados **Aprendizaje NO Supervisado**.

Presentación y Visualización de los Datos

Presentación y Visualización de los Datos

Los datos y su organización:

- **Tipos de datos**: Los datos pueden ser recolectados pueden prevenir de:
  - ✓ **Experimentos**: a través de diseños experimentales (datos experimentales).
  - ✓ **Observaciones**: se recoge la observación existente (datos observacionales).
- **Presentación de los datos**: su objetivo es facilitar el análisis:
  - ✓ **Tablas**
  - ✓ **Arreglos matriciales**
  - ✓ **Medidas de resúmenes o descriptivas**.
  - ✓ **Gráficos**

Presentación de los Datos

Suponga que se observan $p \geq 1$ variables en n unidades muestrales, digamos que

x_{jk} : es la medida de la k-ésima variable en la j-ésima uni. muestral,

donde $j = 1, 2, \ldots, n$ y $k = 1, 2, \ldots, p$.

Estos datos pueden ser organizados en tablas:

	Var. 1	Var.2	$\cdots$	Var. k	$\cdots$	Var. p
Uni.Muestral 1	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Uni.Muestral 2	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Uni.Muestral j	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Uni.Muestral n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Presentación de los Datos

Note que:

- cada fila contiene la información de cada unidad muestral, así que **cada fila** es una **observación multivariada**
- cada columna contiene la información de cada variable.

Presentación de los Datos

Note que:

- cada fila contiene la información de cada unidad muestral, así que **cada fila** es una **observación multivariada**
- cada columna contiene la información de cada variable.

Recuerde que uno de los objetivos es comprender las **relaciones** entre **varias variables** (columnas).

Presentación de los Datos

Note que:

- cada fila contiene la información de cada unidad muestral, así que **cada fila** es una **observación multivariada**
- cada columna contiene la información de cada variable.

Recuerde que uno de los objetivos es comprender las **relaciones** entre **varias variables** (columnas). Para el tratamiento/manipulación de una gran cantidad de variables/datos y para "**relajar" la matemática envuelta en las técnicas estadísticas multivariadas, usaremos conceptos algebraicos**.

Presentación de los Datos

Podemos organizar estos datos en una matriz ($\mathbf{X}$) de n filas y p columnas:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

En las p columnas colocamos las informaciones por variables y en las filas las informaciones de las unidades muestrales.

Evidentemente, $\mathbf{X}$ es una matriz de dimensión $n \times p$, y a partir de ella, hacer un análisis conjunto (multivariado) de sus columnas (variables).

Presentación de los Datos

Si extraemos de **X**, la información de la $k-$ésima variable, ésta puede ser almacenada en un vector columna:

$$\mathbf{x}_k = \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{jk} \\ \vdots \\ x_{nk} \end{bmatrix}_{n \times 1}$$

y podemos analizar ella **individualmente (univariado)**. Por ejemplo, se pueden calcular las **estadísticas descriptivas** (medidas de localización, dispersión, asimetría y kurtosis).

Presentación de los Datos

Media Muestral

La media muestral (medida de **localización**) de la $k-$ésima variable, $k = 1, 2, \ldots, p$:

$$\overline{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$$

Varianza Muestral

La varianza muestral (medida de **dispersión**) de la $k-$ésima variable, $k = 1, 2, \ldots, p$:

$$s_k^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \overline{x}_k)^2$$

Presentación de los Datos

Observación: existe otra definición de la varianza muestral donde el ***denominador*** *es $n-1$ en lugar de n. Existen* ***razones teóricas*** *para hacerlo, especialmente cuando n es* ***pequeño****.*

Desviación Estandar Muestral

La desviación estandar muestral (medida de **dispersión** que posee las mismas unidades de medida de los datos) de la $k-$ésima variable:

$$s_k = \sqrt{s_k^2} \qquad k = 1, 2, \ldots, p$$

Presentación de los Datos

Coeficiente de Asimetría Muestral

El coeficiente de asimetría muestral es una medida que describe la **asimetría** de la distribución de los datos con respecto a la media muestral:

$$sk(x_k) = \frac{\sqrt{n}\sum_{j=1}^{n}(x_{jk} - \overline{x}_k)^3}{\left[\sum_{j=1}^{n}(x_{jk} - \overline{x}_k)^2\right]^{3/2}}$$

Observación:

- *Cuando los datos provienen de distribuciones* ***simétricas***

$$sk(x_k) \approx 0$$

- $sk(x_k) > 0$ *indica que la distribución es* ***asimétrica*** *positiva o a* ***derecha***.
- $sk(x_k) < 0$ *indica que la distribución es* ***asimétrica*** *negativa o a* ***izquierda***.

Presentación de los Datos

Coeficiente de Curtosis Muestral

El coeficiente de curtosis muestral es una medida que describe el **comportamiento en las colas** de la distribución de los datos:

$$k(x_k) = \frac{n \sum_{j=1}^{n} (x_{kj} - \overline{x}_k)^4}{\left[\sum_{j=1}^{n} (x_{kj} - \overline{x}_k)^2 \right]^2}$$

Presentación de los Datos

Observación:

- *Cuando los datos provienen de una distribución* ***normal***

$$k(x_k) \approx 3$$

- *Si* $k(x_k) > 3$ *se dice que la distribución es* ***leptocurtica***.
- *Si* $k(x_k) < 3$ *se dice que la distribución es* ***platicurtica***.

Presentación de los Datos

Observación:

- *Cuando los datos provienen de una distribución **normal***

$$k(x_k) \approx 3$$

- *Si $k(x_k) > 3$ se dice que la distribución es **leptocurtica**.*
- *Si $k(x_k) < 3$ se dice que la distribución es **platicurtica**.*

Esas son **medidas univariadas** (individuales). También, podemos analizar **cada par de variables (bivariado)**, **cada terna de variables (trivariado)**, etc.

Pontificia Universidad JAVERIANA Bogotá

Presentación de los Datos

Covarianza Muestral

La covarianza muestral (medida de **asociación lineal** entre **dos variables**) entre la $i-$ésima y la $k-$ésima variable, $i, k = 1, 2, \ldots, p$ $i \neq k$:

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k)$$

Presentación de los Datos

Observación:

- *Si* $s_{ik} = 0$, ***no hay asociación lineal*** *entre las dos variables.*
- *Si* $s_{ik} > 0$, *existe una* ***asociación lineal positiva*** *entre las dos variables.*
- *Si* $s_{ik} < 0$ *existe una* ***asociación lineal negativa*** *entre las dos variables.*
- $s_{ii} = s_i^2$ *es la* ***varianza muestral*** *de la* $i-$*ésima variable o equivalentemente la covarianza muestral de la* $i-$*ésima variable con ella misma*.

Presentación de los Datos

Observación:

- *Si* $s_{ik} = 0$, ***no hay asociación lineal*** *entre las dos variables.*
- *Si* $s_{ik} > 0$, *existe una* ***asociación lineal positiva*** *entre las dos variables.*
- *Si* $s_{ik} < 0$ *existe una* ***asociación lineal negativa*** *entre las dos variables.*
- $s_{ii} = s_i^2$ *es la* ***varianza muestral*** *de la* $i-$*ésima variable o equivalentemente la covarianza muestral de la* $i-$*ésima variable con ella misma.*

La **covarianza muestral** indica la relación lineal, sin embargo, **NO PROPORCIONA el grado de la fortaleza** de esa relación.

Presentación de los Datos

Correlación Muestral

La **correlación muestral** (medida de asociación lineal) entre la $i-$ésima y la $k-$ésima variable, $i, k = 1, 2, \ldots, p \; i \neq k$:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_i^2}\sqrt{s_k^2}} = \frac{\sum_{j=1}^{n}(x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \overline{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \overline{x}_k)^2}}$$

Presentación de los Datos

Observaciones:

- $|r_{ik}| \leq 1$
- $r_{ik} \approx 1$ *indica relación lineal* ***positiva fuerte***,
- $r_{ik} \approx -1$ *indica relación lineal* ***negativa fuerte***
- $r_{ik} \approx 0$ *indica que* ***NO*** *hay* ***asociación lineal***.

Presentación de los Datos

Las medias muestrales de cada una de las k variables, pueden ser agrupadas/escritas en un vector llamado **vector de medias**:

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_k \\ \vdots \\ \overline{x}_p \end{bmatrix}_{p \times 1}$$

Presentación de los Datos

Análogamente, las **varianzas muestrales** de cada una de las variables y las **covarianzas muestrales** entre las variables puede ser escrita en una matriz llamada **Matriz de Varianzas y Covarianzas Muestrales**:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{np} & s_{np} & \cdots & s_{pp} \end{bmatrix}_{p \times p}$$

Presentación de los Datos

Observe que:

- en la diagonal están las variazas,
- por fuera de la diagonal las covarianzas,
- la matriz **S** es simétrica.

También se puede obtener la **Matriz de Correlaciones Muestrales**:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{np} & r_{np} & \cdots & 1 \end{bmatrix}_{p \times p}$$

Presentación de los Datos

Example

En una libreria universitaria se colectaron informaciones sobre 4 registros de ventas de libros con el objetivo de investigar la naturaleza de las ventas de libros:

Monte total de cada venta	No. de libros vendidos
42	4
52	5
48	4
58	3

Obs: Example 1.1 from Johnson and Wicher (2014), Applied Multivariate Statistical Analysis, 6th Edition, pp. 6.

Pontificia Universidad JAVERIANA Bogotá

Presentación de los Datos

Evidentemente, se tienen $p = 2$ variables y $n = 4$ unidades muestrales.

La **matriz de datos** es

$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}_{4\times 2}$$

Las **medias muestrales**:

$$\overline{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{j1} = \frac{1}{4}(42 + 52 + 48 + 58) = 50$$

$$\overline{x}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{j2} = \frac{1}{4}(4 + 5 + 4 + 3) = 4$$

Presentación de los Datos

El **vector de medias muestral**:

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

Las **varianzas muestrales**:

$$s_1^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{j1} - \overline{x}_1)^2 = \frac{1}{4-1}\left[(42-50)^2 + \cdots + (58-50)^2\right] = 45.33$$

$$s_1^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{j2} - \overline{x}_2)^2 = \frac{1}{4-1}\left[(4-4)^2 + \cdots + (3-4)^2\right] = 0.67$$

Presentación de los Datos

La **covarianza muestral entre las dos variables**:

$$s_{12} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\overline{x}_i)(x_{jk}-\overline{x}_k) = \frac{1}{4}\left[(42-50)(4-4)+\cdots(58-50)(3-4) = -2\right]$$

Por lo tanto, la **matriz de varianzas y covarianzas muestrales**:

$$\mathbf{S} = \begin{bmatrix} 45.33 & -2 \\ -2 & 0.67 \end{bmatrix}$$

La **correlación muestral** entre las dos variables:

$$r_{12} = \frac{s_{12}}{\sqrt{s_1^2}\sqrt{s_2^2}} = \frac{-2}{\sqrt{45.33}\sqrt{0.67}} = -0.36$$

Presentación de los Datos

Por lo tanto, la **matriz de correlación muestral**:

$$\mathbf{R} = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

Usamos R para Calcular todas esas medidas.

```
# --- Ingresamos los Datos --- #
X<-matrix(c(42,52,48,58,4,5,4,3),ncol=2, nrow = 4)
X<-data.frame(X)
colnames(X)<-c("Ventas","No.Libros")
```

Presentación de los Datos

```
X
  Ventas No.Libros
1     42         4
2     52         5
3     48         4
4     58         3

# --- Vector de Medias Muestrales ---  #
xbarra<-apply(X,2,mean)
xbarra
     Ventas No.Libros
       50         4
```

Presentación de los Datos

De ahí,

$$\bar{\mathbf{x}} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

```
# --- Matriz de Varianzas y Covarianzas Muestral --- #
S<-cov(X)
S
             Ventas  No.Libros
Ventas     45.33333 -2.0000000
No.Libros  -2.00000  0.6666667
```

luego,

$$\mathbf{S} = \begin{bmatrix} 45.33 & -2.00 \\ -2.00 & 0.67 \end{bmatrix}$$

Pontificia Universidad JAVERIANA Bogotá

Presentación de los Datos

```
# --- Matriz de Correlación Muestral --- #
R<- cor(X)
             Ventas   No.Libros
Ventas      1.0000000 -0.3638034
No.Libros  -0.3638034  1.0000000
```

Así,

$$\mathbf{R} = \begin{bmatrix} 1.00 & -0.36 \\ -0.36 & 1.00 \end{bmatrix}$$

```
# --- Coeficientes de Asimetría y de Kurtosis --- #
require(moments)
```

Presentación de los Datos

```
# --- Coeficiente de Asimetría
skewness(X$Ventas)
skewness(X$No.Libros)

# --- Coeficiente de Kurtosis
kurtosis(X$Ventas)
kurtosis(X$No.Libros)
```

Luego,

$$sk(x_1) = 0 \qquad sk(x_2) = 0$$
$$k(x_1) = 1.78 \qquad k(x_2) = 2$$

Presentación de los Datos

También se pueden construir **gráficos** hasta para 3 variables (tridimensionales).

Presentación de los Datos

También se pueden construir **gráficos** hasta para 3 variables (tridimensionales).

Dentro de los **gráficos unidimensionales (variables individuales)** más recomendados están:

- ✓ **Diagrama de Puntos**.
- ✓ El Diagrama de Caja y Bigotes o **Boxplot**.
- ✓ El **Histograma**.

Todos proporcionan información sobre el **centro**, la **dispersión** y la **forma** de la distribución de donde provienen los datos.

Presentación de los Datos

Gráfico de Puntos

Este gráfico es recomendado para **variables cuantitativas discretas**, principalmente cuando el **conjunto de datos** es razonablemente **pequeño** o existen **pocos valores de datos distintos**.

Histogramas

Son recomendados para **muestras moderadas o grandes**.

Presentación de los Datos

Boxplots

- Es un gráfico basado en los **cuartiles**.
- Recomendado para **muestras moderadas o grandes**.
- Permite detectar **valores extremos**, **discrepantes** o **outliers** (en caso que estos existan)
- son útiles para la **comparación de varios conjuntos de datos**

Presentación de los Datos

Los **gráficos bidimensionales (pares de variables)** proporcionan información sobre la **orientación** de los datos en el plano cartesiano y la **asociación** que hay entre ellos. El **diagrama de dispersión o scatter plot** es el más ampliamente utilizado en la práctica. En el scatter plot, los puntos se representan en dos dimensiones (cada eje representa una variable).

Presentación de los Datos

Para el estudio de aspectos **tridimensionales** de los datos:

- **Matrices de Dispersión** o múltiples diagramas de dispersión: Se presentan conjuntamente, todos los diagramas de dispersión de los datos para cada par variables.
- **Representaciones Pictóricas**:
  Se emplean para su reconocer **observaciones similares**, por lo que las variables deben estar medidas en la **misma escala**.
  Principales: **Estrellas**, **Caras de Chernoff**.
- Diagramas de dispersión tridimensionales con rotación (**Spinning 3D Scatterplots**).

Presentación de los Datos

Example

Los datos `Iris` fueron originalmente presentados por Fisher (1936)[a], cuantifican la variación morfológica (**cuatro características**) de la flor del iris en sus **tres especies** (setosa, virginica, versicolor).

[a] Fisher, R. A. (1936), The use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7, 179-188.

```
# --- Cargamos los datos y los visualizamos --- #
 data("iris")  # Cargamos los datos
 View(iris)    # Visualizamos los datos
 X<-iris[,1:4] # Extraemos las variables cuantitativas
```

Presentación de los Datos

```
# --- Vector de Medias Muestrales --- #
Xbarra<-apply(X, 2,mean)
round(Xbarra,2)
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
        5.84         3.06         3.76         1.20
```

De ahí

$$\bar{x}_1 = 5.84 \qquad \bar{x}_2 = 3.06 \qquad \bar{x}_3 = 3.76 \qquad \bar{x}_4 = 1.20$$

por lo tanto,

$$\bar{\mathbf{x}} = \begin{bmatrix} 5.84 \\ 3.06 \\ 3.76 \\ 1.20 \end{bmatrix}$$

Presentación de los Datos

```
# --- Matriz de Varianzas y Covarianzas Muestral --- #
S<-cov(X)
> round(S,2)
              Sepal.   Sepal.   Petal.   Petal.
              Length    Width   Length   Width
Sepal.Length   0.69    -0.04    1.27     0.52
Sepal.Width   -0.04     0.19   -0.33    -0.12
Petal.Length   1.27    -0.33    3.12     1.30
Petal.Width    0.52    -0.12     1.30    0.58
```

De ahí,

$$\mathbf{S} = \begin{bmatrix} 0.69 & -0.04 & 1.27 & 0.52 \\ -0.04 & 0.19 & -0.33 & -0.12 \\ 1.27 & -0.33 & 3.12 & 1.30 \\ 0.52 & -0.12 & 1.30 & 0.58 \end{bmatrix}$$

Presentación de los Datos

De acuerdo con estos resultados, se puede decir que el ancho del sépalo tiene una relación lineal negativa con las demás variables. Por lo tanto, cuanto mayor (menor) sea el ancho del sépalo, menor (mayor) será la longitud del sépalo, la longitud y el ancho del pétalo.

```
# --- Matriz de Correlación Muestral --- #
R<-cor(X)
> round(R,2)
             Sepal.   Sepal.   Petal.   Petal.
             Length    Width   Length    Width
Sepal.Length 1.00     -0.12    0.87     0.82
Sepal.Width  -0.12    1.00    -0.43    -0.37
Petal.Length 0.87     -0.43    1.00    0.96
Petal.Width  0.82     -0.37    0.96    1.00
```

Presentación de los Datos

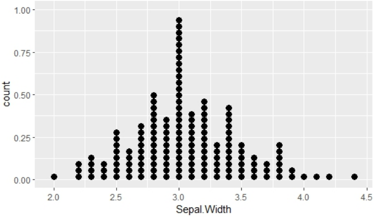
De ahí,

$$\mathbf{R} = \begin{bmatrix} 1.00 & -0.12 & 0.87 & 0.82 \\ -0.12 & 1.00 & -0.43 & -0.37 \\ 0.87 & -0.43 & 1.00 & 0.96 \\ 0.82 & -0.37 & 0.96 & 1.00 \end{bmatrix}$$

Observe que los coefecientes de correlación lineal entre el ancho del sépalo y el resto de las variables es negativo y débil. Mientras que el coefeciente de correlación lineal de la longitud del sépalo y la longitud y ancho del pétalo es positiva y fuerte. Así mismo, la longitud y el ancho del pétalo, presentan una correlación lineal positiva y fuerte (0.96).

Presentación de los Datos

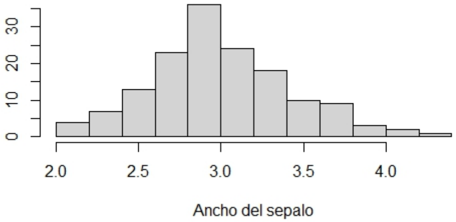
```
# --- Diagrama de Puntos --- #
require(ggplot2)
ggplot(X, aes(x = 'Sepal.Width')) +
geom_dotplot(binwidth = 0.05)
```



De acuerdo con el gráfico, la distribución del ancho del sépalo es levemente asimétrica a derecha y el pico aproximadamente 3.

Presentación de los Datos

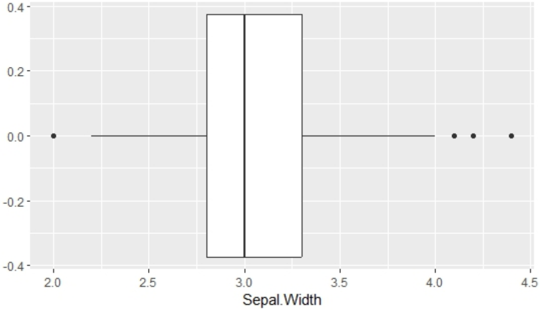
```
# --- Histograma --- #
hist(X$Sepal.Width, xlab = "Ancho del sepalo",
     ylab = "",  main="")
```



La distribución del ancho del sépalo parece ser asimétrica a derecha y su pico esta alrededor de 3.

Presentación de los Datos

```
# --- Boxplot --- #
 ggplot(X, aes(x=Sepal.Width))+
  geom_boxplot()
```



Presentación de los Datos

De acuerdo con el gráfico:

- El centro de la distribución es 3.
- La distribución del ancho del sépalo parece ser asimétrica a derecha (Q_2 está más cerca de Q_1)
- Se observan datos atípicos.

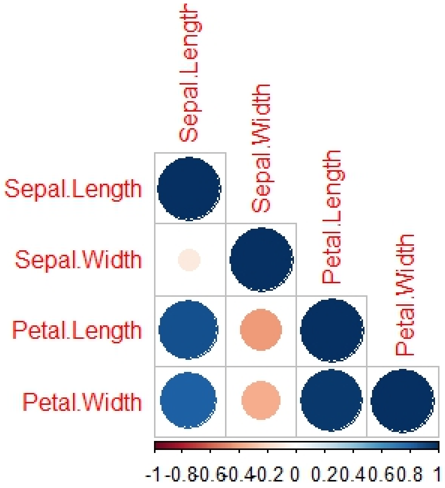
Presentación de los Datos

De acuerdo con el gráfico:

- El centro de la distribución es 3.
- La distribución del ancho del sépalo parece ser asimétrica a derecha (Q_2 está más cerca de Q_1)
- Se observan datos atípicos.

```
# --- Grafico de Correlaciones --- #
require(corrplot)
corrplot(R,type="lower")
```

Presentación de los Datos



Presentación de los Datos

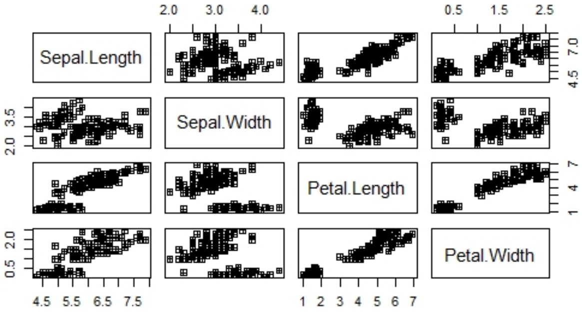
Note que:

- La correlación más fuerte (más azul) es entre la longitud y el ancho del pétalo. Seguida por la longitud del pétalo y el ancho del sépalo. Y, por la del ancho del pétalo y longitud del sépalo.
- Se observan correlaciones bajas y negativas (naranjas) del ancho del sépalo con ambas medidas del pétalo.
- La longitud y el ancho del sépalo presentaron las correlaciones (negativas) más bajas.

Observación: el tamaño de las bolas también indica la fuerza de la correlación. Entre más grandes, mayor es la correlación.

Presentación de los Datos

```
# --- Gráfico de Matrices de Dispersión --- #
pairs(X,       # Datos
      pch=12, # Simbolo
      )
```



Presentación de los Datos

En este gráfico se evidencia:

- algunos grupos, por ejemplo, entre la longitud del pétalo y el ancho del sépalo. Estos grupos se deben a que hay 3 grupos de especies.
- correlación lineal positiva entre la longitud y el ancho del pétalo.
- poca asociación entre la longitud y el ancho del sépalo.

Presentación de los Datos

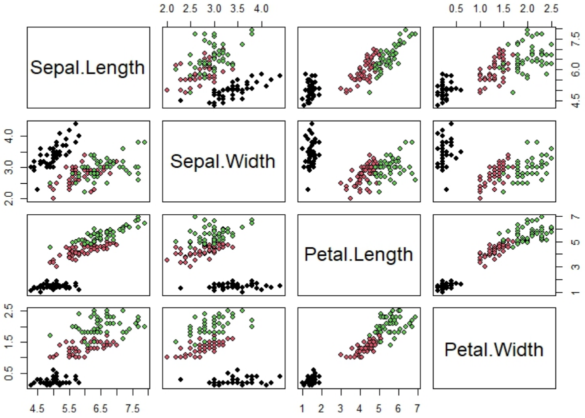
En este gráfico se evidencia:

- algunos grupos, por ejemplo, entre la longitud del pétalo y el ancho del sépalo. Estos grupos se deben a que hay 3 grupos de especies.
- correlación lineal positiva entre la longitud y el ancho del pétalo.
- poca asociación entre la longitud y el ancho del sépalo.

Podemos determinar los grupos (especies) por colores

```
pairs(X, main = "",
      pch = 21,
      bg = c(1, 2, 3)[unclass(iris$Species)])
```

Presentación de los Datos



Presentación de los Datos

*Observación: En la **diagonal principal del gráfico** pueden ser agregados algunos gráficos (**histogramas, boxplot**, etc) y medidas descriptivas (medias, varianzas, etc) de cada una de las variables.*

También, podemos **extraer** las estadísticas descriptivas para **cada una de las especies** usando el comando `subset`. Por ejemplo, para analizar la **especie Setosa**, extraemos los datos de las 4 variables de esta especie:

```
Setosa<-subset(iris, Species =='setosa' )
```

Presentación de los Datos

Observación: En la ***diagonal principal del gráfico*** *pueden ser agregados algunos gráficos (****histogramas, boxplot****, etc) y medidas descriptivas (medias, varianzas, etc) de cada una de las variables.*

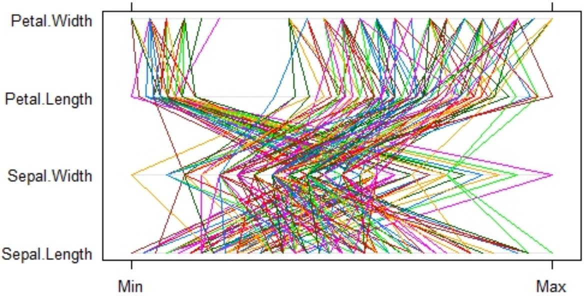
También, podemos **extraer** las estadísticas descriptivas para **cada una de las especies** usando el comando `subset`. Por ejemplo, para analizar la **especie Setosa**, extraemos los datos de las 4 variables de esta especie:

```
Setosa<-subset(iris, Species =='setosa' )
```

Gráficos de **curvas de crecimiento** (principalmente para ver el **signo de la asociación**)

Presentación de los Datos

```
# --- Curvas de Cremiento --- #
require(lattice)
parallelplot(X)
```



Presentación de los Datos

Observe que algunas plantas que tienen longitudes de pétalo pequeñas, el ancho del sépalo tiende a ser grande y para aquellas con anchuras grandes de sépalo, la longitud tiende a ser pequeñas. Así que en ambos casos pareciera existir una asociación lineal negativa.

Observación: Si todas las variables crecieran en el mismo sentido, tendriamos coordenadas paralelas.

Presentación de los Datos

Caras de Chernoff

- Se recomiendan para *n* pequeño
- Cada unidad muestral es representada por una cara, donde las características de ésta (ojos, boca, cejas, etc) describe a una variable.
- Es útil para agrupar unidades muestrales y detectar posibles outliers.

Presentación de los Datos

```
# --- Caras de Chernoff --- #
require(aplpack)
Setosa<-subset(iris,
               Species =='setosa' )
Plantas<-1:50
Setosa<-cbind(Plantas,Setosa)
faces(Setosa[1:25,1:4],
      face.type = 1,
      scale = TRUE,
      labels = Setosa$Plantas[1:25],
      plot.faces = TRUE,
      nrow.plot = 5,
      ncol.plot = 5)
```

Presentación de los Datos

Presentación de los Datos

Observación: Investigar cada una de las características de la cara (curvatura de la boca, el ángulo de la ceja, la nariz, etc).

Presentación de los Datos

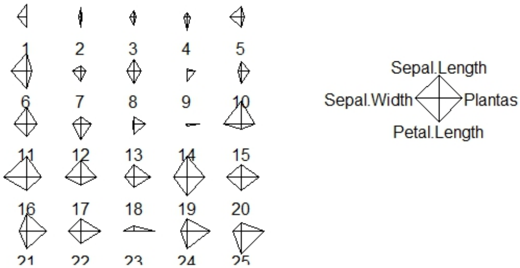
Observación: Investigar cada una de las características de la cara (curvatura de la boca, el ángulo de la ceja, la nariz, etc).

Gráfico de Estrellas o Radar

- Se recomienda para n y p pequeño.
- En el gráfico se muestra una estrella para cada observación.
- Cada estrella tendrá tantos rayos o ejes como variables haya.
- Las logitudes de los rayos son proporcionales a los valores de las variables.
- Es últil para agrupar unidades muestrales.

Presentación de los Datos

```
# --- Gráfico de Estrellas o Radar --- #
 stars(Setosa[1:25,1:4],
      key.loc=c(20,8))
```



Presentación de los Datos

Observación:

- *En la leyenda está la orientación de cada variable (la longitud de los rayos representan los valores de las variables)*
- *Leer Johnson and Wicher (2014), Applied Multivariate Statistical Analysis, 6th Edition, pp. 26-27.*

El gráfico de estrellas es usado por ejemplo para mostrar o **ilustrar** algunas las **características de los jugadores deportivos**.

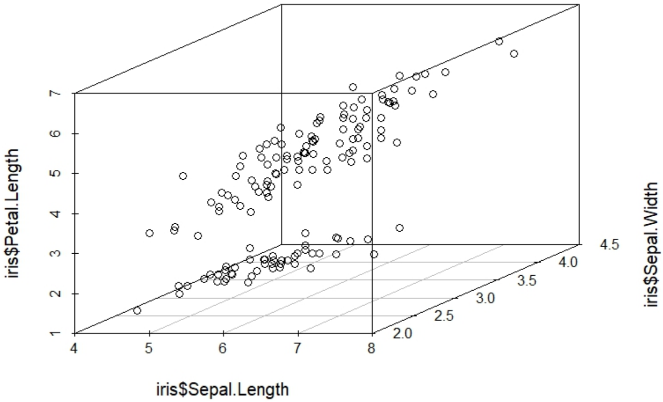
`https://www.pesmaster.com/pes-2021/`

Spinning 3D Scatterplots

Visualizan la relación entre tres o más variables.

```
# --- Scatterplot 3D --- #
require(scatterplot3d)
scatterplot3d(iris$Sepal.Length,
              iris$Sepal.Width,
              iris$Petal.Length)
# --- Con rotación
require(rgl)
plot3d(iris$Sepal.Length,
       iris$Sepal.Width,
       iris$Petal.Length,
       col="blue",size=3)
```

Presentación de los Datos



Vectores Aleatorios

Vimos que los datos muestrales pueden ser organizados/presentados en una matriz:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}_{n\times p}$$

$$= \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k & \cdots \mathbf{x}_p \end{bmatrix}$$

Vectores Aleatorios

donde las columnas (por ejemplo, la $k-$ésima)

$$\mathbf{x}_k = \begin{bmatrix} x_{1k} \\ \\ x_{2k} \\ \\ \vdots \\ \\ x_{nk} \end{bmatrix}$$

son **vectores** $(n \times 1)$ que contiene la información de las variables, $k = 1, 2, \ldots, p$.

Éstos son **datos muestrales** que provienen de alguna **población p variada**.

Vectores Aleatorios

Si denotamos por

$$\mathbf{X}_{p\times 1} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \\ \vdots \\ X_p \end{bmatrix}_{p\times 1}$$

al **vector (aleatorio)** que contiene las p **variables aleatorias**, entonces

✓ **<u>Cada</u> variable aleatoria** $X_i\,(i = 1, 2, \ldots, p)$ tiene su propia **distribución** de probabilidad (**marginal**) que permite estudiar su comportamiento, por ejemplo,

- su **Media Poblacional**:

$$\mu_i = \mathrm{E}[X_i]$$

- su **Varianza Poblacional**:

$$\sigma_i^2 = \mathrm{Var}[X_i] = \mathrm{E}\left[(X_i - \mu_i)^2\right]$$

✓ El comportamiento conjunto de **cada par de variables aleatorias**, digamos X_i e X_k $(i, k = 1, 2, \ldots, \overline{k,\ i} \neq k)$ está descrito por su **función de probabilidad conjunta**, la cual permite estudiar las **asociaciones lineales** entre ellas:

- **Covarianza Poblacional**:

$$\sigma_{ik} = \mathrm{Cov}[X_i, X_k] = \mathrm{E}\left[(X_i - \mu_i)(X_k - \mu_k)\right]$$

- **Correlación Poblacional**:

$$\rho_{ij} = \frac{\sigma_{ik}}{\sqrt{\sigma_i^2}\sqrt{\sigma_k^2}}$$

✓ El comportamiento conjunto de las p **variables aleatorias**, $X_1, X_2, \ldots, X_k, \ldots, X_p$ está descrito por la **función de densidad de probabilidad conjunta (fdpc)**:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_k, \ldots, x_p)$$

Si las X_i's son **mutuamente independientes**, entonces

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_k}(x_k) \cdots f_{X_p}(x_p)$$

y las **covarianzas son cero** (lo contrario no necesariamente es cierto).

Algunas características poblacionales (**parámetros**) de interés:

- ✓ **Vector de Medias Poblacional**:

$$\boldsymbol{\mu}_{p\times 1} = \mathrm{E}\left[\mathbf{X}\right] = \mathrm{E}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \\ \vdots \\ \mu_p \end{bmatrix}_{p\times 1}$$

donde $\mu_k = \mathrm{E}[X_k]$ es la **media poblacional (marginal)** de la $k-$ésima variable, $k = 1, 2, \ldots, p$.

✓ **Matriz de Varianzas y Covarianzas Poblacional**:

$$\boldsymbol{\Sigma}_{p\times p} = \mathrm{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}_{p\times p}$$

donde:

- $\sigma_{kk} = \sigma_k^2 = \mathrm{E}[(X_k - \mu_k)^2]$ es la **varianza poblacional (marginal)** de la $k-$ésima variable, $k = 1, 2, \ldots, p$.
- $\sigma_{ik} = \mathrm{E}[(X_i - \mu_i)(X_k - \mu_k)]$ es la **covarianza poblacional** entre la $i-$ésima y la $k-$ésima variable, $i \neq k = 1, 2, \ldots, p$.

- **Matriz de Correlación Poblacional**:

$$\mathbf{P}_{p\times p} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{bmatrix}_{p\times p}$$

donde

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

es la **correlación poblacional** entre la $i-$ésima y la $k-$ésima variable, $i \neq k = 1, 2, \ldots, p$.

Observación: Relación entre $\boldsymbol{\Sigma}$ *e* $\boldsymbol{P}$*:*

- $\boldsymbol{\Sigma} = \mathbf{V}^{1/2}\,\mathbf{P}\,\mathbf{V}^{1/2}$
- $\mathbf{P} = \left[\mathbf{V}^{1/2}\right]^{-1} \boldsymbol{\Sigma} \left[\mathbf{V}^{1/2}\right]^{-1}$

donde

$$\mathbf{V} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$