

# Estadística Aplicada III

## Introducción a los temas del curso

### Análisis exploratorio de datos

Jorge de la Vega Góngora

Departamento de Estadística,  
Instituto Tecnológico Autónomo de México

Semana 1: 15/17 de agosto de 2018

# Introducción al curso

# Temario I

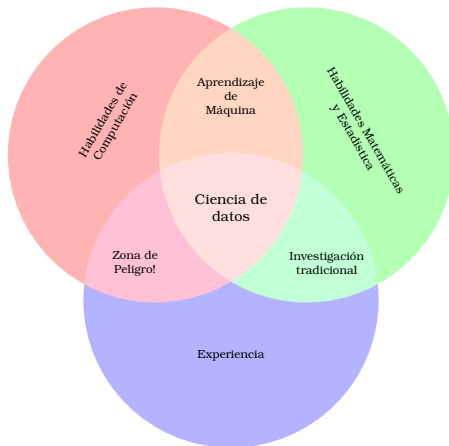
El principal tema de este curso es el análisis estadístico de vectores aleatorios, su manejo y organización y los modelos de análisis multivariado.

Revisar el **calendario**. El **Temario** resumido incluye:

- Análisis exploratorio de datos (EDA)
- Vectores y matrices aleatorios y la distribución normal multivariada
- Análisis de estructuras de covarianza (dependencia lineal de vectores aleatorios)
  - ▶ Análisis factorial
  - ▶ Componentes principales
  - ▶ Correlación canónica
- Técnicas y agrupamiento y clasificación:
  - ▶ Clasificación
  - ▶ Análisis discriminante
  - ▶ Escalamiento multidimensional
  - ▶ Análisis de correspondencias
- Análisis de datos categóricos
  - ▶ Tablas de contingencia
  - ▶ Modelos lineales generalizados
  - ▶ Regresión logística

Algunos temas que se consideran en esta materia se consideran como métodos básicos de la Ciencia de Datos.

## Temario II



## Diagrama de Venn de Conway

# Requerimientos I

- Este es un curso para aprender varios métodos y técnicas estadísticas para estudiar y analizar datos multivariados y con cierta estructura particular.
- Estas técnicas se basan en métodos analíticos, cuantitativos y geométricos en la mayor parte de los casos. Los requisitos incluyen:
  - ▶ Conocimiento de Estadística univariada
  - ▶ Álgebra lineal y cálculo matricial (determinantes, análisis espectral)
  - ▶ Geometría analítica
  - ▶ Conocimientos básicos de algún software estadístico (en la clase se utilizará principalmente R)
- Este **NO** es un curso de R. Lo utilizaremos como herramienta en los laboratorios y repasaremos algunas de sus estructuras y su uso, pero el curso es sobre métodos estadísticos, su fundamento y su aplicación. R es el medio, no el fin.

## Algunos modelos que vamos a estudiar en este curso

- **Creación de índices e indicadores:** Cómo combinar varias variables en una sola cantidad o número que mida adecuadamente una característica de interés. (Análisis de componentes principales).

## Algunos modelos que vamos a estudiar en este curso

- **Creación de índices e indicadores:** Cómo combinar varias variables en una sola cantidad o número que mida adecuadamente una característica de interés. (Análisis de componentes principales).
- **Análisis de factores:** describe las relaciones de covarianza entre varias variables en términos de otras cuantas variables aleatorias subyacentes llamadas *factores*, que no pueden ser observadas directamente. Por ejemplo, correlaciones de resultados en materias como inglés, música, matemáticas, sugieren un factor subyacente: *inteligencia*.

## Algunos modelos que vamos a estudiar en este curso

- **Creación de índices e indicadores:** Cómo combinar varias variables en una sola cantidad o número que mida adecuadamente una característica de interés. (Análisis de componentes principales).
- **Análisis de factores:** describe las relaciones de covarianza entre varias variables en términos de otras cuantas variables aleatorias subyacentes llamadas *factores*, que no pueden ser observadas directamente. Por ejemplo, correlaciones de resultados en materias como inglés, música, matemáticas, sugieren un factor subyacente: *inteligencia*.
- **Análisis de correlación canónica:** busca identificar y cuantificar las asociaciones entre 2 conjuntos de variables. E.g. la relación entre variables de política gubernamental con las variables de objetivos económicos.



## Algunos modelos que vamos a estudiar en este curso

- **Creación de índices e indicadores:** Cómo combinar varias variables en una sola cantidad o número que mida adecuadamente una característica de interés. (Análisis de componentes principales).
- **Análisis de factores:** describe las relaciones de covarianza entre varias variables en términos de otras cuantas variables aleatorias subyacentes llamadas *factores*, que no pueden ser observadas directamente. Por ejemplo, correlaciones de resultados en materias como inglés, música, matemáticas, sugieren un factor subyacente: *inteligencia*.
- **Análisis de correlación canónica:** busca identificar y cuantificar las asociaciones entre 2 conjuntos de variables. E.g. la relación entre variables de política gubernamental con las variables de objetivos económicos.
- **Discriminación:** cómo separar diferentes conjuntos de objetos u observaciones; con qué criterios se pueden separar. Por ejemplo, ¿Cómo separa un banco a los buenos pagadores de los que no lo son o de los que tienen riesgos mayores?

## Algunos modelos que vamos a estudiar en este curso

- **Creación de índices e indicadores:** Cómo combinar varias variables en una sola cantidad o número que mida adecuadamente una característica de interés. (Análisis de componentes principales).
- **Análisis de factores:** describe las relaciones de covarianza entre varias variables en términos de otras cuantas variables aleatorias subyacentes llamadas *factores*, que no pueden ser observadas directamente. Por ejemplo, correlaciones de resultados en materias como inglés, música, matemáticas, sugieren un factor subyacente: *inteligencia*.
- **Análisis de correlación canónica:** busca identificar y cuantificar las asociaciones entre 2 conjuntos de variables. E.g. la relación entre variables de política gubernamental con las variables de objetivos económicos.
- **Discriminación:** cómo separar diferentes conjuntos de objetos u observaciones; con qué criterios se pueden separar. Por ejemplo, ¿Cómo separa un banco a los buenos pagadores de los que no lo son o de los que tienen riesgos mayores?
- **Clasificación:** Cómo distribuir objetos o datos entre grupos previamente definidos.

## Algunos modelos que vamos a estudiar en este curso

- **Creación de índices e indicadores:** Cómo combinar varias variables en una sola cantidad o número que mida adecuadamente una característica de interés. (Análisis de componentes principales).
- **Análisis de factores:** describe las relaciones de covarianza entre varias variables en términos de otras cuantas variables aleatorias subyacentes llamadas *factores*, que no pueden ser observadas directamente. Por ejemplo, correlaciones de resultados en materias como inglés, música, matemáticas, sugieren un factor subyacente: *inteligencia*.
- **Análisis de correlación canónica:** busca identificar y cuantificar las asociaciones entre 2 conjuntos de variables. E.g. la relación entre variables de política gubernamental con las variables de objetivos económicos.
- **Discriminación:** cómo separar diferentes conjuntos de objetos u observaciones; con qué criterios se pueden separar. Por ejemplo, ¿Cómo separa un banco a los buenos pagadores de los que no lo son o de los que tienen riesgos mayores?
- **Clasificación:** Cómo distribuir objetos o datos entre grupos previamente definidos.
- **Ordenamiento:** Cómo ordenar objetos usando como criterios varios atributos que pueden ser de carácter objetivo y/o subjetivo, eg el “ranking” de universidades.

# Clasificación de métodos en DM

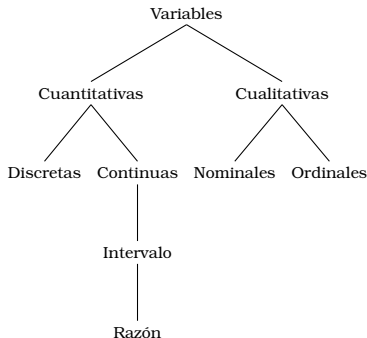
La clasificación típica desde el contexto de minería de datos de los modelos son:

- Aprendizaje supervisado
  - ▶ Regresión
  - ▶ Clasificación
  - ▶ Análisis Discriminante
  - ▶ Naive Bayes
- Aprendizaje no supervisado
  - ▶ Clustering
  - ▶ Reducción de dimensión
  - ▶ Reglas de asociación
  - ▶ Escalamiento dimensional
  - ▶ Componentes principales

## Organización de datos

# Datos multivariados I

- Un vector aleatorio usualmente corresponde a características observadas de una *unidad experimental*  $i$ :  $\mathbf{x}_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ , donde cada  $X_{ij}$  es una variable aleatoria asociada a una característica  $j$  de la unidad experimental  $i$ . Cada una de estas variables aleatorias puede ser *numérica* o *categorica*, de acuerdo a la siguiente clasificación:



## Datos multivariados II

- Una *variable categórica* con  $m$  distintos valores se puede representar como un vector  $f = (a_1, \dots, a_{m-1})$  de unos y ceros, llamado *factor*, donde cada componente corresponde a una de las posibles etiquetas de la variable (menos una) y toma el valor 1 para la  $a_i$  asociada a la etiqueta del valor de la variable. El caso en que todos los valores sean 0 corresponde a la categoría dejada fuera. Se necesitan  $m - 1$  variables dummies para una variable categórica con  $m$  valores.

### Ejemplo

Si  $X$  = religión con posibles valores en {católica, musulmán, judía, budista, agnóstico, ateo, otra}.

- Si se tiene un grupo de 10 personas con las siguientes religiones: { tt judía, musulmana, otra, budista, otra, agnóstica, católica, atea, atea, católica }, entonces la representación con factores queda de la siguiente manera:

Persona $i$	católica	musulmán	judía	budista	agnóstica	ateo
1	0	0	1	0	0	0
2	0	1	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	0	0
6	0	0	0	0	1	0
7	1	0	0	0	0	0
8	0	0	0	0	0	1
9	0	0	0	0	0	1
10	1	0	0	0	0	0

- Las variables categóricas son comunes en encuestas o en otras fuentes de datos.
- Usualmente los vectores de valores de las variables, incluyendo a los factores, para las unidades experimentales, se conjuntan en una **matriz de datos**  $X_{n \times p}$  con entradas  $X_{ij}$ , donde:
  - i. los renglones corresponden a casos, individuos, unidades experimentales
  - ii. columnas corresponden a atributos, variables o dimensiones.
  - iii. Los factores asociados a variables categóricas se pueden incorporar en las columnas de la matriz, pero deben estar vinculados a una sola variable.
- Las variables categóricas (nominales y ordinales) en un análisis multivariado deben convertirse a factores para poder trabajar con ellos.



## Datos multivariados IV

### Ejemplo

Estructuras de datos como la siguiente son comunes, este es un tipo de datos estructurado:

Autos	Calidad	km/hr	2puertas?	Años_garantía	Costo	Procedencia
Centra	Buena	20	1	2	120	USA
Vectra	Mala	18	1	3	93	USA
Megane	Excelente	24	0	2	150	Francia
Tsuru	Excelente	36	1	4	110	Japón
Lincoln	Mala	15	0	7	500	USA
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Calidad es una variable de tipo ordinal, 2Puertas? es binaria o categórica con dos niveles, Procedencia es una variable de tipo nominal.

# En relación al manejo de datos I

- En el análisis multivariado la recopilación de datos tiene dos problemas:
  - ▶ Usualmente se requiere un gran esfuerzo para recabar la información relevante y hacer la preparación de los datos.
  - ▶ En otros casos, el problema es el almacenamiento de datos y su extracción (data warehousing)
- Los datos multivariados tienden a ser masivos en la actualidad. En este contexto, la *gestión de datos* se vuelve un elemento relevante.
- En este curso no veremos muchos elementos de la gestión de datos, pero es importante tener en cuenta este componente en la experiencia laboral y contextos de explotación de datos.

Marco de gestión de datos de DAMA: Consultar DAMA-DMBOK2

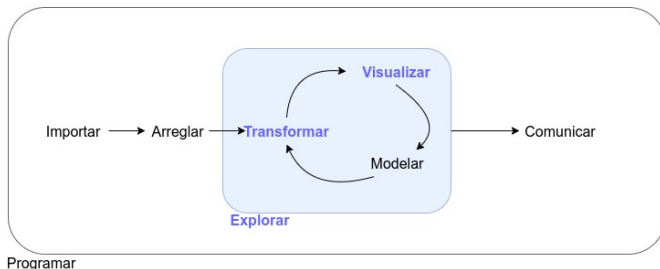
## En relación al manejo de datos II



Figure 1. The DAMA-DMBOK2 Guide Knowledge Area Wheel

# Tidyverse

- **Tidyverse** es un conjunto de herramientas (principalmente para R) desarrolladas por Hadley Wickham para aplicar la ciencia de datos.
- Ayudan principalmente a la manipulación de datos, su limpieza y su análisis exploratorio.
- Lo importante es que permite generar un esquema coherente y lógico para limpiar y ordenar datos, para facilitar su análisis.



Revisar los siguientes papers y libro en línea:

- **The Split-Apply-Combine Strategy for Data Analysis**
- **Tidy Data**
- **R for Data Science. Garret Grolemund, Hadley Wickham**
- **dplyr Tutorial**

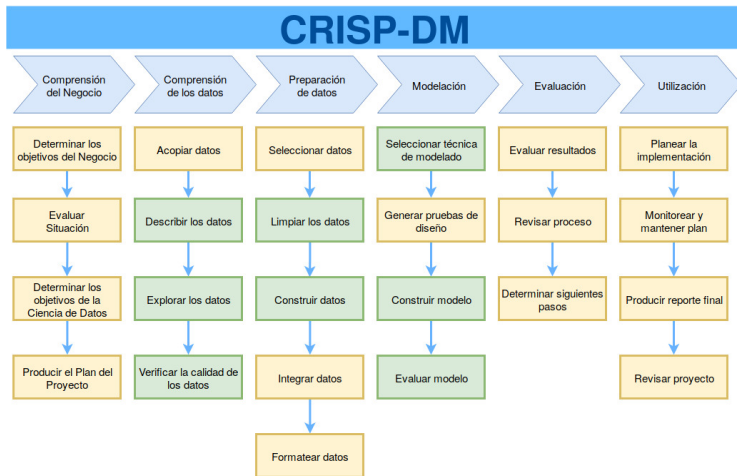
# Fuentes de datos I

- Las fuentes de datos actuales incluyen, pero no se limitan a fuentes como:
  - ▶ Redes sociales (twitter, facebook, Google)
  - ▶ Servicios en línea (Uber, Netflix, Amazon)
  - ▶ Registros de ventas en comercios (Wallmart, Costco)
  - ▶ **Kaggle**
  - ▶ **Minería de texto**
  - ▶ **Datos abiertos del Gobierno**, INEGI, Banco de México
  - ▶ **Datos abiertos NASA**
  - ▶ Consultas de Ley de Transparencia a las entidades Gubernamentales

## Proceso de análisis de información

# Proceso CRISP-DM I

El **Cross Industry Standard Process for Data Mining (CRISP-DM)**<sup>1</sup> es un buen referente como proceso para el análisis de datos.



<sup>1</sup>Fuente: Chapman, Clinton y Kerber, Khabaza, Reinartz, Shearer y Wirth, 1999 **CRISP-DM 1.0: Step-by-step Data Mining Guide**

## Análisis exploratorio de datos



# Análisis exploratorio de datos

- En lo que sigue veremos algunos ejemplos de datos, sus estadísticas descriptivas y de algunas herramientas para el análisis exploratorio.
- Una de las primeras cosas que se deben hacer con los datos es conocerlos y tratar de extraer la esencia de la información.
- EL análisis exploratorio de datos no es un proceso formal con reglas estrictas, es más bien la manera de conocer los datos y ganar información acerca de posibles estructuras que permitirán generar hipótesis y preguntas de investigación y que requerirán validarse con mayor análisis o a través de modelado.

## Pasos comunes del EDA

- Generar preguntas acerca de los datos (o antes de obtenerlos)
- Buscar respuestas ya sea por visualización, transformación y modelación de los datos
- Usar lo que se aprendió en el paso anterior para refinar las preguntas y/o generar nuevas preguntas.

# Análisis exploratorio de datos: gráficas I

- ¿Cómo se usan las gráficas en el análisis de datos?
  - ▶ Para comprender propiedades de los datos
  - ▶ Para encontrar patrones en los datos
  - ▶ Para sugerir estrategias de modelado
  - ▶ Para validar el análisis
  - ▶ Para comunicar resultados
- Las gráficas pueden ser de dos tipos: exploratorias o analíticas y de presentación o comunicación. Las características de las gráficas exploratorias incluyen:
  - ▶ Se hacen de manera rápida y/o interactiva
  - ▶ Se tiende a hacer muchas para comprender diversos aspectos de los datos.
  - ▶ El objetivo es lograr un entendimiento personal de la información: cuáles son los datos, cómo se ven, qué problemas pueden tener. Aquí conviene tener preguntas para guiar el análisis.
  - ▶ Diferir el uso de leyendas ni títulos (no son para presentación)
  - ▶ El color y el tamaño, así como otros atributos, se usan para incorporar información y no estética.

## Principios de gráficas analíticas (Roger Peng/Edward Tufte)

- ❶ Mostrar comparaciones entre grupos
  - ▶ Evidencia para una hipótesis es siempre relativa a otra hipótesis competitiva.
  - ▶ Siempre hay que preguntar ¿comparado a qué?
- ❷ Mostrar causalidad, mecanismo, explicación, estructura sistemática.  
¿Cuál es el marco causal para pensar acerca del problema?
- ❸ Mostrar datos multivariados: tratar de incorporar varias dimensiones al problema.
- ❹ Integrar múltiples modos de evidencia
- ❺ Describir y documentar la evidencia. Una gráfica debe decir una historia lo más completa posible y que sea creíble.
- ❻ El contenido es rey: ¿Cuál es la historia que se quiere contar? Si no hay historia, la gráfica no sirve.
  - ▶ Las presentaciones analíticas dependen de su calidad, relevancia e integridad del contenido.

Los datos se pueden resumir de múltiples maneras:

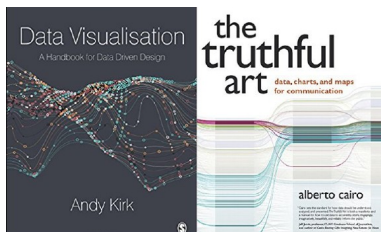
- En una dimensión
  - ▶ estadísticas sumarias univariadas: media, mediana, moda, desviación estándar, percentiles, varianza, rango, rango intercuartil, sesgo, kurtosis, coeficiente de variación, etc.
  - ▶ gráficas de caja o boxplots
  - ▶ histogramas
  - ▶ gráficas de densidad,
  - ▶ gráficas de barras (datos categóricos), gráficas de puntos, etc.
- En más dimensiones
  - ▶ Gráficas univariadas en retículas (lattice)
  - ▶ Gráficas o matrices de dispersión de puntos
  - ▶ Gráficas tridimensionales, etc.

# Recursos en temas de visualización y gráficas estadísticas I

- Curso gratuito de Alberto Cairo en [Journalismcourses.org](http://Journalismcourses.org)



- Libros de Alberto Cairo y Andy Kirk sobre datos y visualización:



- Libros sobre gráficas estadísticas

# Recursos en temas de visualización y gráficas estadísticas II



- Videos de Hans Rosing en TED
- Gapminder.org
- Portal BID: Números para el desarrollo

Consultar: [Guía gráfica del FT](#)

FT

# Ejemplos de análisis y datos I

## Ejemplo 1: Medidas físicas de estudiantes

Los datos corresponden a 8 medidas físicas de 27 estudiantes. Las variables son: sex sexo (0 para mujer), est estatura en cm, pes peso en kg, lpie, longitud del pie en cm, lbra longitud del brazo en cm, aes anchura de la espalda en cm, dcr diámetro del cráneo en cm, y lrt longitud entre la rodilla y el tobillo, en cm.<sup>a</sup> Los datos se organizan en una matriz  $\mathbf{X}_{27 \times 7}$ :

```
sex est pes lpie lbra aes dcr lrt
1  0 159 49  36  68 42.0  57 40
2  1 164 62  39  73 44.0  55 44
3  0 172 65  38  75 48.0  58 44
4  0 167 52  37  73 41.5  58 44
5  0 164 51  36  71 44.5  54 40
6  0 161 67  38  71 44.0  56 42
[1] 27 8
```

---

<sup>a</sup>los datos se pueden obtener del archivo: mefidis.dat, fuente original: Daniel Peña



## Algunas preguntas relevantes sobre los datos

Todo análisis comienza intentando responder algunas preguntas básicas que podamos responder obteniendo algunas estadísticas descriptivas de la información. Por ejemplo:

- ¿Hay diferencias significativas de peso entre mujeres y hombres?
- ¿Hay diferencias significativas de estatura entre mujeres y hombres?
- ¿Hay valores extremos en las variables de hombres y mujeres?

En las siguientes láminas se responderán a estas preguntas, utilizando algunas de las herramientas de EDA.

## Peso promedio I

Podemos simplemente analizar la distribución a través de las 5 estadísticas usuales:

```
summary(X)
```

sex	est	pes	lpie	lbra	aes	dcr
Min. :0.0000	Min. :152.0	Min. :43.00	Min. :34.00	Min. :66.00	Min. :36.00	Min. :54.00
1st Qu.:0.0000	1st Qu.:160.0	1st Qu.:52.00	1st Qu.:36.00	1st Qu.:69.50	1st Qu.:43.50	1st Qu.:56.00
Median :0.0000	Median :168.0	Median :65.00	Median :39.00	Median :73.00	Median :46.00	Median :57.00
Mean :0.4444	Mean :168.8	Mean :63.89	Mean :38.98	Mean :73.46	Mean :45.85	Mean :57.24
3rd Qu.:1.0000	3rd Qu.:177.0	3rd Qu.:73.50	3rd Qu.:41.00	3rd Qu.:76.50	3rd Qu.:48.00	3rd Qu.:58.50
Max. :1.0000	Max. :189.0	Max. :91.00	Max. :45.00	Max. :83.00	Max. :53.00	Max. :61.00

lrt
Min. :38.00
1st Qu.:41.00
Median :43.00
Mean :43.09
3rd Qu.:44.75
Max. :52.00

Al conjunto anterior también habría que agregar otras medidas relevantes para obtener información de la distribución univariada marginal:

- sesgo o asimetría:  $sk(x) = \frac{\sum (X_i - \bar{x})^3}{s^3}$  (0 para variables simétricas, en valor absoluto  $geq1$  para asimétricas).
- homogeneidad o curtosis:  $k = \frac{\sum (X_i - \bar{x})^4}{s^4}$  (con datos atípicos serán altos 7, u 8; si hay dos poblaciones mezcladas pero separadas,  $\approx 2$  y mientras más se separen las poblaciones tenderá a 1).

# Peso promedio II

- coeficiente de variación  $cv = \frac{s}{\bar{x}}$ . Mide variabilidad relativa

```
library(moments)
sapply(X, function(x) {c(var = var(x, na.rm=T),
  sesgo = skewness(x, na.rm = T),
  curtosis = kurtosis(x, na.rm = T),
  cv = sd(x, na.rm = T)/mean(x, na.rm = T))})
```

	sex	est	pes	lpie	lbra	aes	dcr	lrt
var	0.2564103	103.94871795	163.8717949	8.20156695	24.57549858	16.16951567	3.39173789	9.96225071
sesgo	0.2236068	0.16006185	0.1764735	0.28613403	0.39602759	-0.23513254	0.16821658	0.59632146
curtosis	1.0500000	1.95130781	2.2436490	2.08078793	2.27472053	2.84718675	2.17557298	3.64593359
cv	1.1393318	0.06040798	0.2003673	0.07346662	0.06748119	0.08769839	0.03217406	0.07324468

## Medidas multivariadas I

De manera multivariada podemos calcular el vector de medias y la matriz de varianzas y covarianzas con las siguientes fórmulas:

- Vector de medias:

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{X}' \mathbf{1}$$

donde  $\mathbf{1}$  es un vector de  $n$  unos.

- Matriz de varianzas y covarianzas:

$$\mathbf{S}_n = \frac{1}{n} \mathbf{X}' \mathbf{H} \mathbf{X}$$

y la versión insesgada:  $\mathbf{S} = \frac{n}{n-1} \mathbf{S}_n$ , donde la matriz  $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}'$  se conoce como la *matriz de centrado*.

- La matriz de centrado de  $n \times n$  tiene las siguientes propiedades:

- ▶  $\mathbf{H}$  es simétrica:  $\mathbf{H} = \mathbf{H}'$
- ▶  $\mathbf{H}$  es idempotente:  $\mathbf{H}^2 = \mathbf{H}$
- ▶  $\mathbf{H}$  tiene rango  $n - 1$  (ya que  $\mathbf{H} \mathbf{1} = \mathbf{0}$ )

- Matriz de correlaciones:

$$\mathbf{R} = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2}$$

donde  $\mathbf{D} = \text{diag}(s_i^2)$ .

# Medidas multivariadas II

Entonces, para los datos que tenemos,  $\bar{x}$ ,  $S$  y  $R$  son respectivamente:

```
X <- as.matrix(X)
n <- nrow(X)
uno <- rep(1,n)
xbar <- t(X) %*% uno / n
xbar
```

```
      [,1]
sex    0.4444444
est 168.7777778
pes  63.8888889
lpie 38.9814815
lbra 73.4629630
aes  45.8518519
dcr  57.2407407
lrt  43.0925926
```

```
#Equivalente a var(X)
H <- diag(n) - uno %*% t(uno)/n
S <- round(t(X) %*% H %*% X/(n-1),2)
S[1:7,1:3]
```

```
      sex    est    pes
sex 0.26   4.06   4.78
est 4.06 103.95 108.36
pes 4.78 108.36 163.87
lpie 1.24  27.09  31.15
lbra 1.98  45.86  52.05
aes  1.45  34.43  43.21
dcr  0.35  11.04  14.60
```

```
#Equivalente a cor(X)
D <- diag(apply(X,2,var))
R <- round(solve(D^0.5) %*% S %*%
           solve(D^0.5),3)
R[1:7,1:3]
```

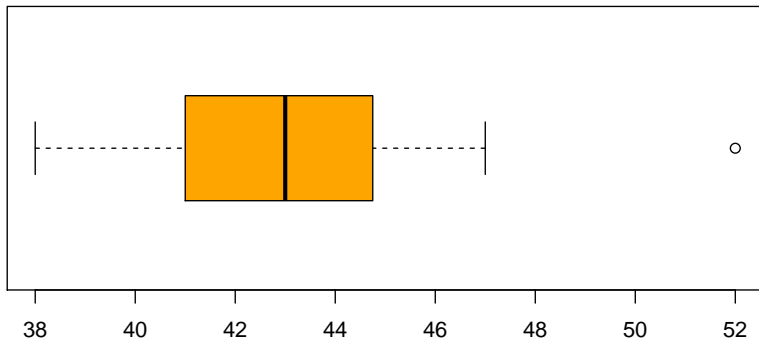
```
      [,1] [,2] [,3]
[1,] 1.014 0.786 0.737
[2,] 0.786 1.000 0.830
[3,] 0.737 0.830 1.000
[4,] 0.855 0.928 0.850
[5,] 0.789 0.907 0.820
[6,] 0.712 0.840 0.839
[7,] 0.375 0.588 0.619
```

Podemos obtener más información sobre el peso:

- Podemos obtener la distribución de la variable peso a través de un histograma
- distribución del peso por género

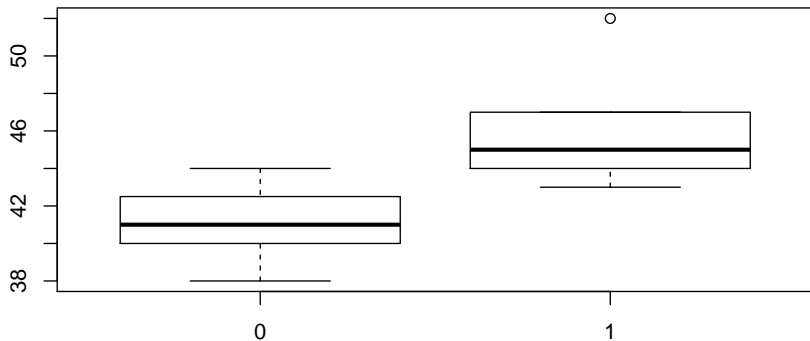
# Boxplot

```
X <- as.data.frame(X)
boxplot(X$lrt, horizontal = T, col = "orange")
```



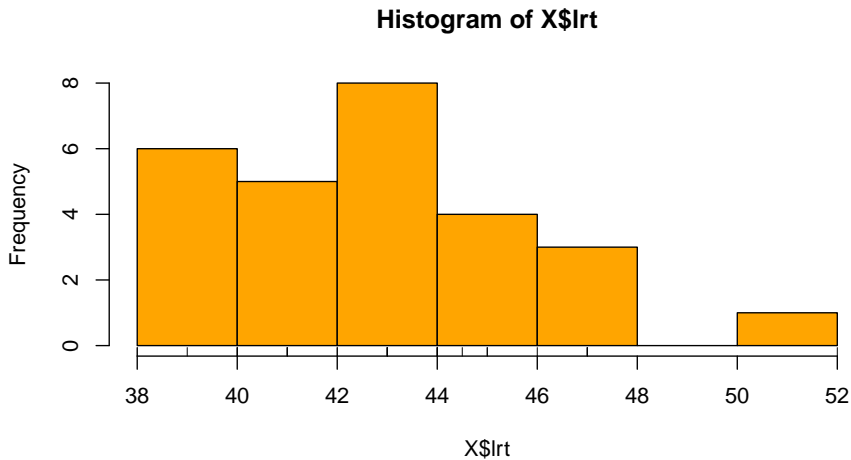
# Boxplot

```
with(X, boxplot(lrt ~ sex))
```



# Histograma

```
hist(X$lrt, col = "orange", breaks=5)  
rug(X$lrt)
```





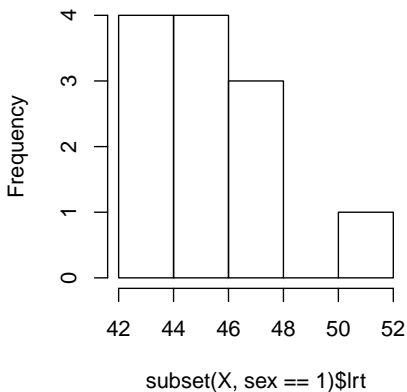
# Histogramas múltiples

```
par(mfrow=c(1,2))  
hist(subset(X, sex == 0)$lrt)  
hist(subset(X, sex == 1)$lrt)
```

Histogram of subset(X, sex == 0)\$lrt

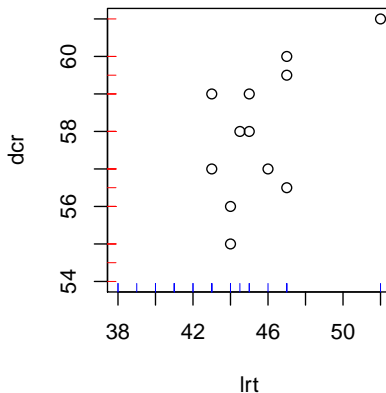


Histogram of subset(X, sex == 1)\$lrt



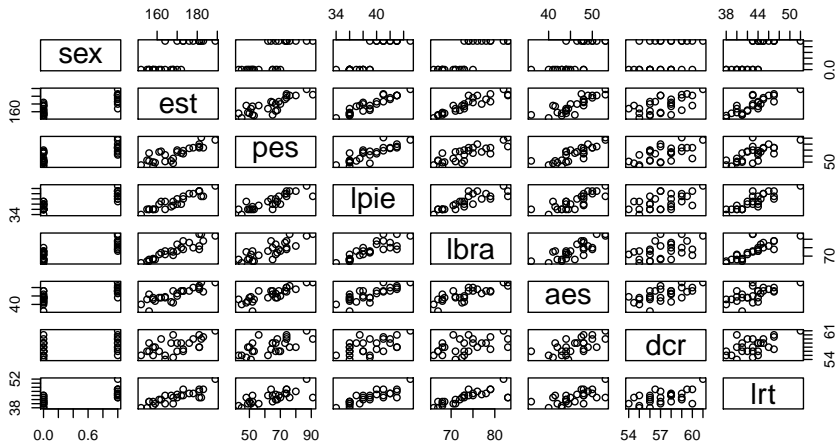
# Diagrama de dispersión de puntos (scatterplot)

```
par(pty="s")  
with(X, plot(lrt,dcr, col = sex))  
rug(X$lrt,side=1,col="blue")  
rug(X$dcr,side=2,col="red")
```



# Matriz de diagramas de dispersión de puntos (scatterplot matrices)

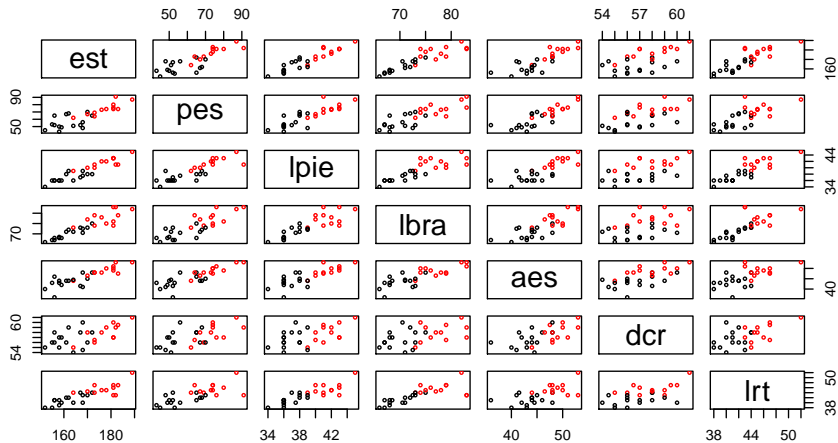
`pairs(X)`



Idealmente, la variable `sex` debería usarse para distinguir las observaciones de hombres y mujeres.

# Matriz de diagramas de dispersión de puntos (scatterplot matrices)

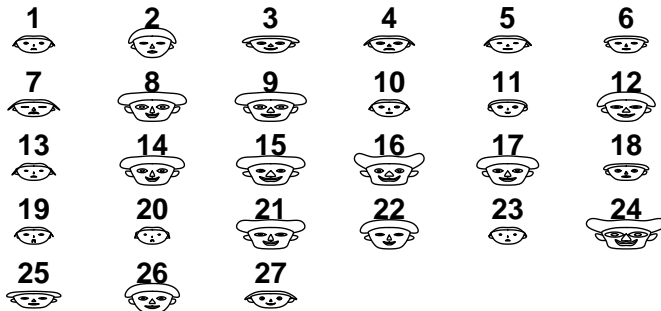
```
pairs(X[,-1],col=1 + X[,1],cex=0.5)
```



# Caras de Chernoff I

```
suppressMessages(library(aplpack))  
par(mar=c(1,1,1,1))  
faces(X, face.type=0)
```

## Caras de Chernoff II

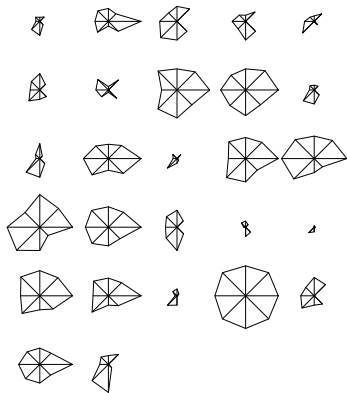


# Caras de Chernoff III

```
effect of variables:
modified item      Var
"height of face"  "sex"
"width of face"   "est"
"structure of face" "pes"
"height of mouth" "lpie"
"width of mouth"  "lbra"
"smiling"         "aes"
"height of eyes"  "dcr"
"width of eyes"   "lrt"
"height of hair"  "sex"
"width of hair"   "est"
"style of hair"   "pes"
"height of nose"  "lpie"
"width of nose"   "lbra"
"width of ear"    "aes"
"height of ear"   "dcr"
```

# Gráficas de estrellas

```
stars(X, key.loc=c(-5, 5))
```





# Otros Recursos

- Effective tables and graphs in official statistics
- Australian Bureau of Statistics, National Statistical Services. (2010) A guide to using evidence based policy. Canberra
- UNECE: Cómo hacer comprensibles los datos
- Exploratory Data Analysis, en *R for data Science*