



# DIPLOMADO EN MINERÍA DE DATOS

## Análisis Exploratorio de Datos

### Estadística Descriptiva

#### Introducción

En la actualidad estamos expuestos a una gran cantidad de información sobre diversos tópicos de carácter económico, político, social, psicológico, biológico, físico, deportivo, etc., que se nos presentan generalmente a través de los medios de comunicación, ya sean audio-visuales o escritos. La forma en que se nos muestra esta información es mediante algún valor que la sintetiza o compendia, una tabla de resultados numéricos o una gráfica, con la finalidad de ilustrarnos sobre algún fenómeno, hecho u acontecimiento de interés. Todos estos valores y formas de mostrarla, constituyen breves resúmenes de la información generada y contenida en estos eventos.

La *estadística descriptiva* es una rama de la estadística cuyo objetivo es recolectar, procesar, analizar y representar un conjunto de datos, con el fin de describir apropiadamente sus características. Para realizar este *análisis exploratorio* de datos, la estadística descriptiva recurre a tablas numéricas, gráficas y medidas de resumen. Este tipo de elementos descriptivos de información nos circundan todos los días, por lo que podemos afirmar que tenemos una gran exposición a los métodos propios de la estadística descriptiva o que, dicho de otro modo, la estadística descriptiva es un acompañante cotidiano que nos proporciona información de manera breve y concisa para enterarnos de qué está pasando a nuestro alrededor.

Es conveniente destacar el carácter exploratorio de los métodos de la estadística descriptiva, ya que existe una tendencia errónea a generalizar sus resultados a *toda la población*; tarea,

esta última, de la que se ocupa la llamada *estadística inferencial*.

## Conceptos básicos de la estadística

Antes de iniciar con la presentación de las distintas técnicas descriptivas para analizar datos, es conveniente definir algunos elementos básicos que forman el cuerpo conceptual de la estadística. Dado el nivel de inserción que actualmente tiene la estadística en la sociedad, estos conceptos son de conocimiento general, así que la intención es únicamente proporcionar una definición escueta de los mismos.

### Población

El concepto de *población* es fundamental en estadística ya que, generalmente, todos los esfuerzos estadísticos están dirigidos a lograr un conocimiento lo más preciso posible sobre ella. Como debemos suponer, existe una gran cantidad de definiciones de población, nosotros adoptaremos la siguiente.

**Población:** Conjunto finito o infinito de personas u objetos que consta de todas las observaciones posibles de un fenómeno determinado, que tiene una característica común. También conocida como población estadística o universo.

De la definición es claro que la población está constituida por el total de elementos de un conjunto, que comparten alguna característica particular. Por ejemplo, podemos referirnos a la población de mexicanos, que está compuesta por todos los individuos de nacionalidad mexicana que existen en el mundo. Otro ejemplo de una población, pueden ser los estudiantes de la Facultad de Ciencias de la UNAM, integrada por todos los estudiantes matriculados en esta escuela. La población no sólo se refiere a un conjunto de individuos, por ejemplo, en una institución de crédito, una población puede estar formada por el total de créditos expedidos. En un proceso de producción, la población usualmente está determinada por el total de productos que genera algún proceso.

### Muestra

Este es, tal vez, uno de los conceptos más importantes en estadística. Mencionamos que por lo regular estamos interesados en el conocimiento preciso de una población; no obstante,

sabemos que obtener información sobre toda la población puede ser un proceso costoso y demorado. Para obtener información de manera rápida y expedita de la población, se recurre generalmente a la selección de una *muestra* de la misma. Sabemos que esta selección de los individuos u objetos que constituyen la muestra, se debe realizar mediante un procedimiento que, de alguna forma, garantice que es *representativa*<sup>1</sup> de la población, situación que se obtiene a través de un proceso de selección aleatoria. Por lo tanto, debemos diferenciar lo que es una muestra de una muestra aleatoria.

**Muestra:** Es un subconjunto de individuos u objetos de una población.

**Muestra Aleatoria:** Es una muestra en la que todos los individuos u objetos de la población, tienen la misma probabilidad de ser elegidos para formar parte de ella. Una muestra aleatoria se conoce también como una muestra probabilística.

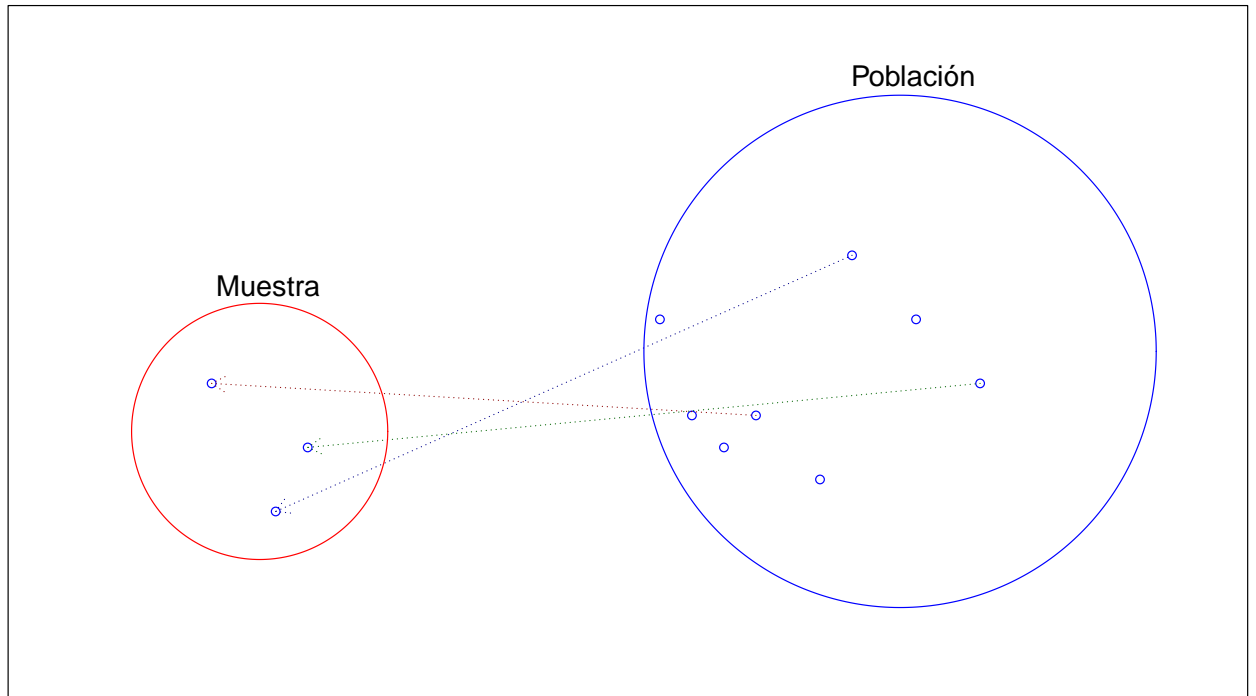
Las muestras aleatorias son preferidas en los procesos estadísticos porque la selección de ellas se hace de manera objetiva.

El conjunto de métodos o técnicas estadísticas para realizar la selección aleatoria de los individuos en la muestra, recibe el nombre de *técnicas de muestreo* y constituye la parte nodal del área estadística conocida como *muestreo*.

---

<sup>1</sup>De forma vaga, se dice que una muestra es *representativa* de la población, si en ella se pueden observar todas las características de esta última

## MUESTREO



## Censo

Una muestra *bien seleccionada* será suficiente, en general, para obtener una buena aproximación de las características de la población bajo estudio; pero, el hecho de que sólo se tenga un subconjunto de la población implica estar sujetos a cometer el llamado *error de muestreo*, que es el error que se comete debido al hecho de que se sacan conclusiones sobre características de la población total, a partir de la observación de sólo una parte de ella (la muestra). La única manera de garantizar que no está presente el error de muestreo en nuestras conclusiones, sería realizando un *censo* de la población.

Un **Censo** es encuestar, entrevistar o enumerar a *todos* los individuos o componentes de la población. Este proceso es, en general, demasiado tardado y costoso; basta recordar, por ejemplo, que el censo de población y vivienda en México se realiza cada diez años debido, entre otros factores, a su alto costo de operación y ejecución.

## Variables, datos y escalas de medición

Hemos mencionado reiteradamente que la materia prima de la estadística es la información sobre algún fenómeno aleatorio de interés. En una investigación, esta información se obtiene a través de variables asociadas a los individuos de la población sujetos a esta investigación, por lo tanto, es necesario definir algunas características de los elementos que constituyen esta información.

**Variable(estadística)**: Es una característica o atributo, observable o medible, que toma diferentes valores en los individuos de la población.

La información que tenemos sobre los individuos está constituida por sus variables recabadas en la investigación. Por ejemplo, una población sobre seres humanos, nuestras variables de interés pueden ser: altura, peso, edad, sexo, grupo sanguíneo, etc., en la evaluación de la calidad de un restaurante, las variables pueden ser: tiempo de atención, sabor de la comida, limpieza del lugar, amabilidad del personal, etc..

Al resultados de la observación de dicha variable para cada componente se le denomina *dato*, que es un valor particular de la variable.

En nuestros ejemplos anteriores, algunos posibles datos de las variables son:

- Altura:  $\{1.78, 1.65, 1.71, \dots\}$
- Peso:  $\{83, 57, 68, \dots\}$
- Edad:  $\{22, 49, 17, \dots\}$
- Sexo:  $\{Femenino, Femenino, Masculino, \dots\}$
- Grupo sanguíneo  $\{O, A, AB, \dots\}$
- Tiempo de atención:  $\{18.5, 23, 12.8, \dots\}$
- Sabor de la comida:  $\{Muy\ buena, Regular, Buena, \dots\}$

- Limpieza del lugar:  $\{Regular, Muy\ limpio, Limpio, \dots\}$
- Amabilidad del personal  $\{Nada\ amable, Amable, Muy\ amable, \dots\}$

En la investigación y para fines estadísticos, por lo general, las variables en estudio se pueden identificar de dos formas principalmente:

## Por tipo de dato de la variable

**Variables cualitativas:** Son variables cuyos datos son no numéricos que expresan distintas cualidades o características de un sujeto. Como por ejemplo, la variable *sexo* mencionada anteriormente.

**Variables cuantitativas:** Son aquellas características factibles de medición, existe un instrumento o una forma establecida para registrar la información, también son conocidas como variables numéricas. Se clasifican en discreta y continua.

- **Variables cuantitativas discretas:** son aquellas que sólo pueden tomar un conjunto finito o numerable de valores (generalmente valores enteros). En nuestros ejemplos, la *edad* medida en años cumplidos y *peso* en kilogramos, son variables cuantitativas discretas.
- **Variables cuantitativas continuas:** Este tipo de variables son las que pueden tomar *todos los valores en un intervalo definido*. La *estatura* y el *tiempo de atención* pueden considerarse variables cuantitativas continuas en nuestros ejemplos.

## Por escala de medición de la variable

Las variables se pueden clasificar también de acuerdo a su nivel o escala de medición, esta escala puede ser nominal, ordinal, de intervalo o de razón. Es importante mencionar que las nominales u ordinales se le atribuyen a variables de tipo cualitativas y las de intervalo o de razón a variables de tipo cuantitativo.

**Escala nominal:** Los valores de la variable sólo identifican alguna característica de la población. En nuestros ejemplos, variables cualitativas nominales son:

- *sexo*, con categorías: *Femenino*, *Masculino*.
- *Grupo sanguíneo* con categorías: *tipo O*, *A*, *AB*,...
- *Sabor de la comida*: cuyas categorías son: *Muy buena*, *Regular*, *Buena*,...
- *Limpieza del lugar*: que tiene categorías: *Regular*, *Muy limpio*, *Limpio*,...
- *Amabilidad del personal*: formada por las categorías: *Nada amable*, *Amable*, *Muy amable*,...

**Escala ordinal**: Son variables nominales en las que sus categorías representan cierto orden. En los ejemplos, estas variables corresponden a

- *Sabor de la comida*: cuyas categorías ordenadas son: *Muy buena*, *Buena*, *Regular*, *Mala*
- *Limpieza del lugar*: que tiene categorías ordenadas: *Muy limpio*, *Limpio*, *Regular*, *Sucio*, *Muy sucio*
- *Amabilidad del personal*: con orden de categorías, dado por: *Muy Amable*, *Amable*, *Poco amable*, *Nada amable*

Para estas variables, las “etiquetas” que definen las categorías no son intercambiables, puesto que, de serlo, romperían con el orden de la variable. No siempre es claro si la “distancia” entre categorías adyacentes es la misma. Por lo general, sólo indican que una categoría es mejor o peor que la otra, pero no en qué magnitud lo es.

Por el número de categorías que las componen, las variables nominales se suelen clasificar también como

- \* **Dicotómicas o binarias** que tienen únicamente dos categorías o que toman sólo dos valores posibles. En nuestros ejemplos, *sexo* es una variable dicotómica con categorías *Femenino* y *Masculino*. Un tipo de variable dicotómica muy importante en estadística,

es aquella cuyas categorías denotan presencia o ausencia de una característica, conocida como *variable indicadora* o *variable dummy*. Generalmente, la presencia de la característica se denota como *uno* y su ausencia como *cero*.

\* **Politómicas** que tienen más de dos categorías. En nuestros ejemplos, el grupo sanguíneo es una variable politómica con categorías *O*, *A*, *AB*, etc.

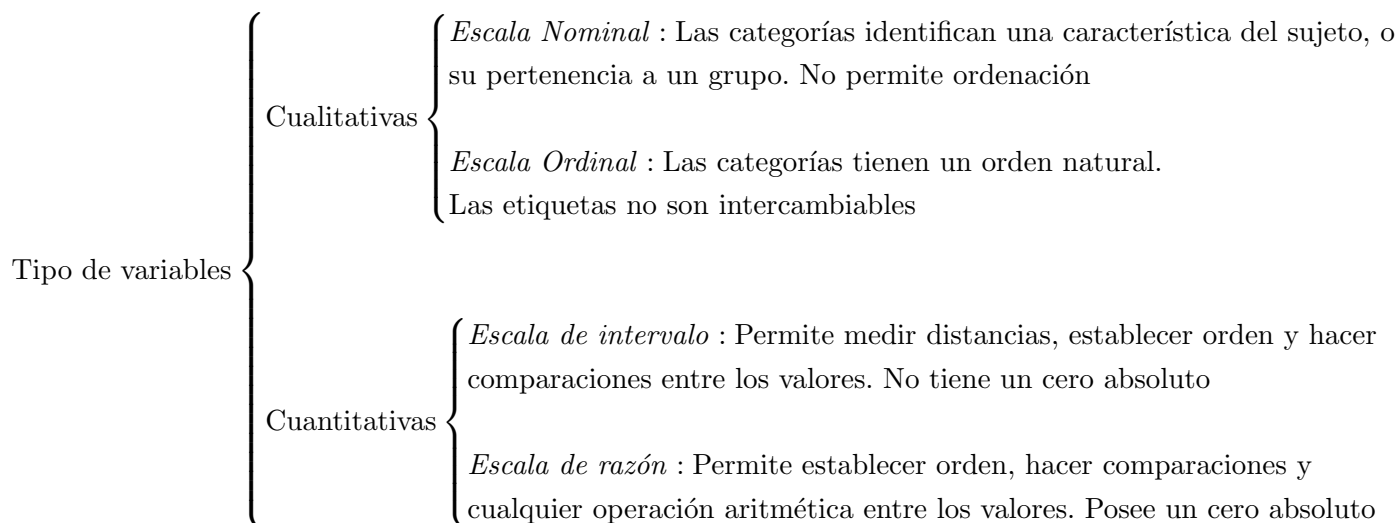
**Escala de intervalo:** Esta escala representa magnitudes, con la propiedad de igualdad de la distancia entre puntos de escala de la misma amplitud. Por ejemplo, si una persona mide 1.65 m, entonces tiene 5 cm más que otra que mide 1.70 m. Estos 5 cm representan la misma diferencia entre una persona que mide 1.82 m y otra que mide 1.77 m. Aquí puede establecerse orden entre sus valores, hacerse comparaciones de igualdad, y medir la distancia existente entre cada valor de la escala. El valor cero de la escala no es absoluto, sino un cero arbitrario que no refleja ausencia de la magnitud medida, es un valor más en la escala, por lo que las operaciones aritméticas de multiplicación y división *no son apropiadas*. El ejemplo más común de una variable medida en esta escala, es la temperatura. Por ejemplo, en la escala de grados centígrados puede decirse que la distancia entre 21° y 26° C es la misma que la existente entre 27° y 32° C, pero no puede afirmarse que una temperatura de 30° C equivale al doble de 15° C en cuanto a intensidad de calor se refiere, debido a la ausencia de cero absoluto.

**Escala de razón:** Tiene las mismas características que la escala de intervalo pero, además, posee el cero absoluto; es decir, el valor cero no es arbitrario, pues representa la ausencia de la magnitud que se está midiendo. Con esta escala se puede realizar cualquier operación de ordenamiento, de comparación y aritmética. A iguales diferencias entre los números asignados corresponden iguales diferencias en el grado de atributo presente en el objeto de estudio. En nuestros ejemplos, estas variables son *altura*, *peso* y *edad*.

La relevancia que tiene reconocer la escala de medición de nuestras variables, se debe a que no todos los procedimientos estadísticos son apropiados para cualquier tipo de variable. En general, el uso de las técnicas estadísticas dependen de la escala de medición de las variables involucradas. Es decir, el tipo de medidas de resumen, tablas, gráficas o modelos para presentar y analizar un conjunto de datos, depende de la clase de variables que lo constituyan.



En resumen, los tipos de variables se pueden agrupar de la siguiente forma:



## Análisis exploratorio de datos

Una vez que se han definido los objetivos de una investigación, y que se ha recabado la información relevante para cubrirlos, el investigador o analista está listo para iniciar la primera fase de análisis estadístico, que consiste en realizar un resumen de las características más importantes de los datos, a través de tablas, gráficas y valores que la describan y resuman, es decir, está preparado para hacer *estadística descriptiva*.

### Tablas y gráficos de frecuencias

Una primera forma de resumir nuestros datos recolectados, es por medio de tablas o distribuciones de frecuencias, para lo cual es necesario conocer los distintos tipos de frecuencias de datos.

Supongamos que quiere analizar la variable  $\mathbf{X}$  que tiene  $n$  observaciones,  $x_1, x_2, \dots, x_n$ , en las cuales pueden haber datos repetidos, entonces, un primer objetivo consiste en determinar el número de observaciones que se repite el valor  $x_i$ . Este número se conoce en estadística como la *frecuencia*, o más precisamente, la *frecuencia absoluta*, denotada por  $n_i$ ,  $i = 1, 2, \dots, k$ ,  $k \leq n$ . Además de las frecuencias absolutas, es común presentar las *frecuencias relativas*. La frecuencia relativa se define como la frecuencia absoluta dividida por el total de observaciones, es decir

$$f_i = \frac{n_i}{n}, \quad i = 1, 2, \dots, k$$

Un ejemplo podría ser con las calificaciones de 24 estudiantes que presentaron un examen de matemáticas, cuyos valores obtenidos en la prueba son:

5, 7, 8, 7, 9, 8, 6, 10, 9, 7, 7, 8, 8, 9, 8, 7, 8, 7, 9, 6, 8, 7, 8, 8

Las frecuencia absolutas del conjunto de datos son,  $n_5 = 1$ ,  $n_6 = 2$ ,  $n_7 = 7$ ,  $n_8 = 9$ ,  $n_9 = 4$  y  $n_{10} = 1$ ; las frecuencias relativas son  $f_5 = 0.0416$ ,  $f_6 = 0.0833$ ,  $f_7 = 0.2916$ ,  $f_8 = 0.375$ ,  $f_9 = 0.1666$  y  $f_{10} = 0.0416$

Se define como la *distribución de frecuencias* al registro de todas las frecuencias asociadas a las posibles categorías o valores de la variable. Ya que disponemos de dos tipos de frecuencias, la absoluta y la relativa, entonces tenemos también una distribución de frecuencias

absolutas y una de frecuencias relativas.

Si se trabaja con una variable de tipo ordinal (cuyas categorías se pueden ordenar de menor a mayor) también es posible calcular las frecuencias acumuladas.

Las frecuencias acumuladas permiten conocer rápidamente el número de observaciones que están por debajo de un determinado valor o categoría. Nuevamente, debemos distinguir entre *frecuencias acumuladas absolutas* y *frecuencias acumuladas relativas*. La frecuencia absoluta acumulada se define como:

$$N_i = \sum_{j=1}^i n_j \quad i = 1, 2, \dots, k$$

mientras que la frecuencia relativa acumulada se define por

$$F_i = \frac{\sum_{j=1}^i n_j}{n} = \frac{N_i}{n} \quad i = 1, 2, \dots, k$$

conviene remarcar que si no se tiene una variable ordinal o no se puede ordenar la variable en cuestión, no es posible referirse a valores acumulados y, por tanto, tampoco a distribuciones acumuladas.

Con los datos de calificaciones anteriores tenemos que, la distribución de frecuencia absoluta es  $N_5 = 1$ ,  $N_6 = 3$ ,  $N_7 = 10$ ,  $N_8 = 19$ ,  $N_9 = 23$  y  $N_{10} = 24$ ; las frecuencias relativas son  $F_5 = 0.0416$ ,  $F_6 = 0.125$ ,  $F_7 = 0.4166$ ,  $F_8 = 0.7916$ ,  $F_9 = 0.9583$  y  $F_{10} = 1$

## Tablas de frecuencias para datos no agrupados

Una vez calculadas las frecuencias, es común desplegarlas en la llamada *tabla de frecuencias*. El formato usual de estas tablas es como se muestra en el cuadro 1. En ella aparecen el total de observaciones condensadas en las  $k$  distintas categorías o valores de la variable, las frecuencias absolutas, las frecuencias relativas, las frecuencias absolutas acumuladas y las frecuencias relativas acumuladas.

**Cuadro 1**  
**Tabla de frecuencias**

Categorías	Frecuencias Absolutas	Frecuencias Relativas	Frecuencias Abs. Acumuladas	Frecuencias Rel. Acumuladas
$X_i$	$n_i$	$f_i$	$N_i$	$F_i$
$X_1$	$n_1$	$\frac{n_1}{n}$	$n_1$	$\frac{n_1}{n}$
$X_2$	$n_2$	$\frac{n_2}{n}$	$n_1 + n_2$	$\frac{n_1+n_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_k$	$n_k$	$\frac{n_k}{n}$	$n_1 + n_2 + \cdots + n_k$	$\frac{n_1+n_2+\cdots+n_k}{n}$

La tabla de frecuencias para nuestro ejemplo es

Calificación	Frecuencias Absolutas	Frecuencias Absolutas	Frecuencias Abs. Acumuladas	Frecuencias Rel. Acumuladas
5	1	0.0416	1	0.0416
6	2	0.0833	3	0.125
7	7	0.2916	10	0.416
8	9	0.375	19	0.7916
9	4	0.166	23	0.9583
10	1	0.041	24	1

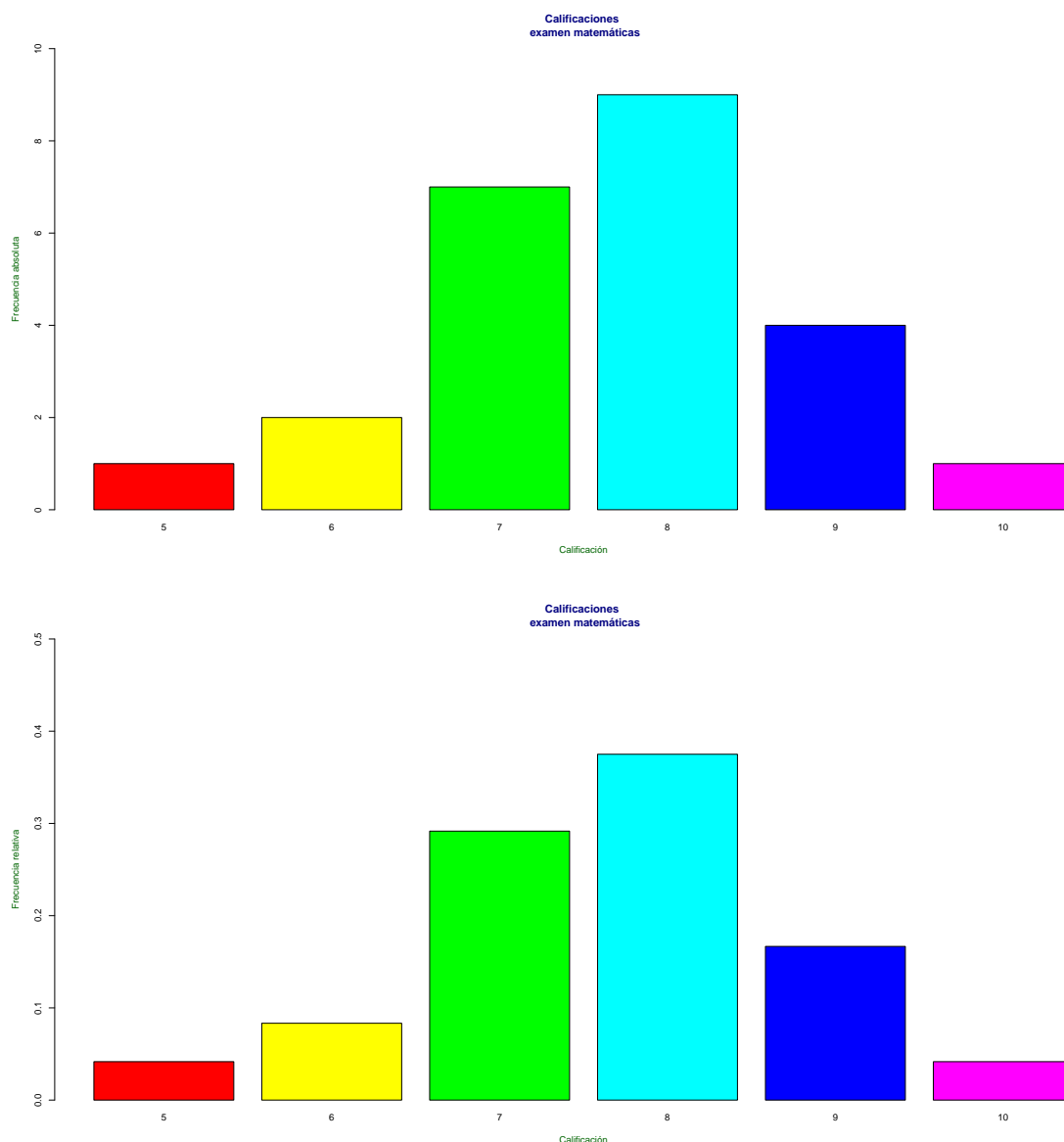
Para poder tener un uso “correcto” de este resumen de los datos, es necesario que la variable no tome un número grande de valores (k no muy grande), la posibilidad de tener una variable que tome pocos valores, por lo regular no suele suceder para casos en el que ésta es cuantitativa, teniendo entonces que la representación es más utilizada para variables cualitativas.

Asimismo, una de las formas más populares y comunes de presentar la información contenida en un conjunto de datos, es a través de representaciones gráficas. Las gráficas estadísticas tienen la finalidad de presentar los aspectos relevantes de los datos de una manera sencilla y clara, lo que permite que baste una simple mirada para observar muchas de sus características. En esencia, las gráficas responden al dicho popular de que “una imagen dice más que mil palabras” , que podemos adaptar a la estadística como “una gráfica estadística dice más que mil medidas y tablas de resumen” . Es importante remarcar que en muchos cursos de estadística básica, no se le da la debida importancia a la presentación y uso de gráficas, y sólo se muestran como un complemento de los métodos descriptivos numéricos.

## Diagrama de barras

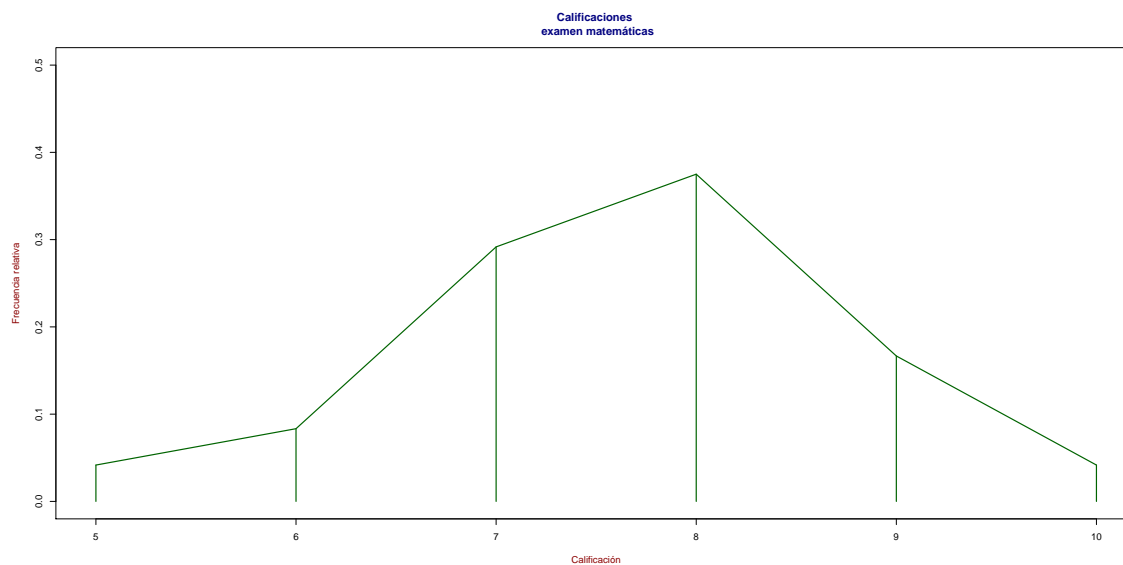
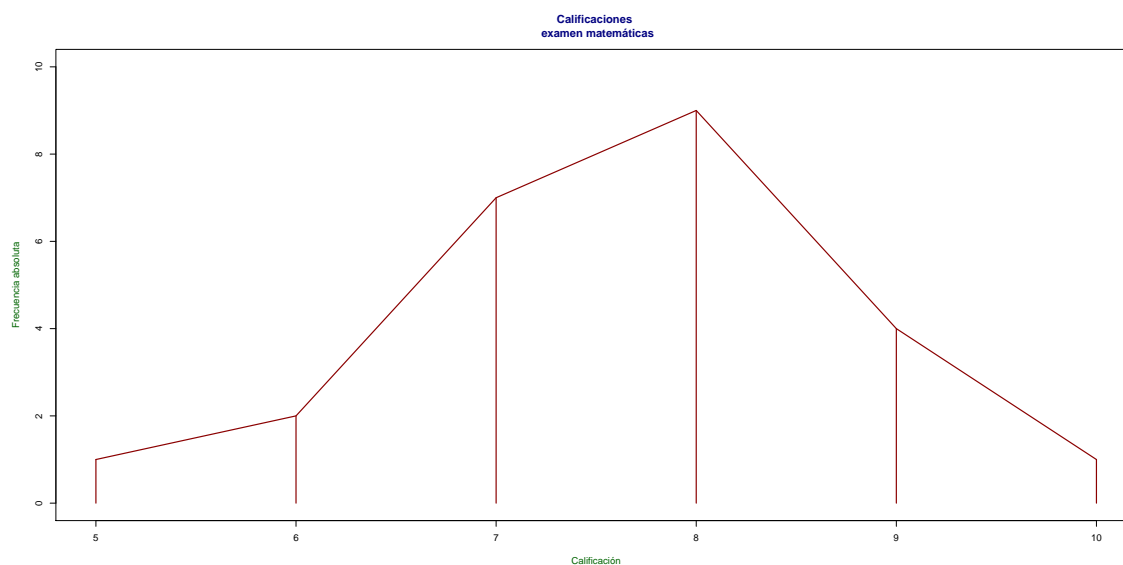
Además de la tabla de frecuencias, se pueden desplegar los resultados que se observan en ella mediante algunas gráficas que facilitan y complementan la lectura de la información. Una de estas herramientas gráficas es el *diagrama de barras* que representa, para cada categorías de la variable (en alguno de los ejes), su frecuencia absoluta o relativa (en el otro eje). Su objetivo es visualizar de manera rápida y clara la frecuencia de cada una de las categorías. Es común tener un diagrama de barras vertical, en el cual el eje horizontal representa las categorías de la variable y el eje vertical sus frecuencias, absolutas o relativas. Si invertimos el orden de estos ejes obtenemos un diagrama horizontal.

La representación de nuestro ejemplo es:



## Polígono de frecuencias

Otra alternativa de representación a la tabla de frecuencias es el *Polígono de frecuencias*, en el que gráficamente se aprecian los distintos valores de nuestra variable representados por puntos y a su vez éstos son enlazados por medio de líneas rectas.



## Diagramas circulares o de pie

Los *diagramas circulares* o de *pie*, son representaciones gráficas ampliamente utilizadas, que muestran la frecuencia relativa de cada categoría como una porción (sector) de un círculo, en la que el ángulo se corresponde con la frecuencia relativa correspondiente. Entonces, la muestra se representa por un círculo de  $360^{\circ}$ , y cada una de las categorías que la componen,

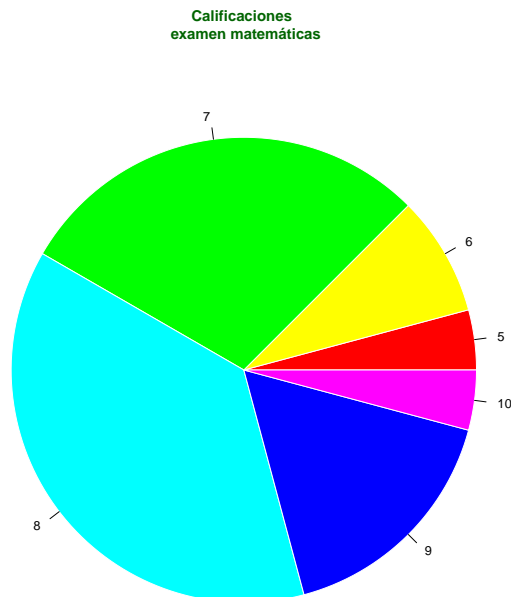
por un sector de éste. La regla de proporcionalidad entre los sectores y el círculo completo se obtiene tomando los ángulos proporcionales a las frecuencias de las categorías. Por ejemplo, si una categoría corresponde al 25 % del total de la muestra, le corresponderá un sector del círculo cuyo ángulo sea de  $90^\circ$ , exactamente el 25 % de  $360^\circ$ . En general, el ángulo correspondiente a cada sector estará determinado por la *regla de tres*

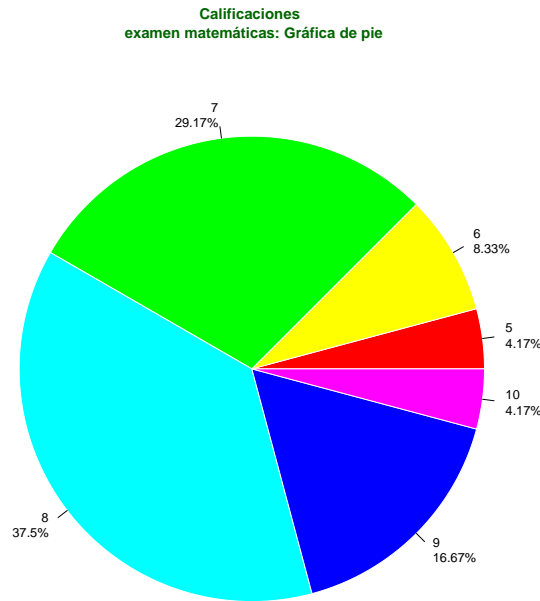
$$\begin{array}{ccc} 360^\circ & \rightarrow & x^0 \\ n & \rightarrow & fa_x \end{array}$$

Donde,  $x^0$  es el valor desconocido en grados de la categoría que queremos representar;  $n$  es el total de la muestra y  $fa_x$  es la frecuencia absoluta de la categoría en cuestión. Entonces, el valor en grados de esta categoría se puede calcular como

$$x^0 = \frac{360^\circ fa_x}{n}$$

Obsérvese que  $\frac{fa_x}{n}$  corresponde a la frecuencia relativa de la categoría en cuestión.





## Distribuciones de frecuencia e histogramas

Comentamos que una buena alternativa para iniciar un análisis descriptivo, era la de contar el número de observaciones en cada categoría cuando se trataba de variables cualitativas, y crear una distribución de frecuencias. Pero, cuando la variables son de tipo cuantitativo (discreta o continua), éstas suelen tomar un número de valores diferentes mucho mayor que las variables cualitativas, en cuyo caso, su representación a través de una distribución de frecuencias sería de poca utilidad, ya que la información estaría poco resumida. Una manera simple de resumir la información de este tipo de variables, es agrupándolas en *intervalos de valores* o simplemente intervalos y, posteriormente, realizar la distribución de frecuencias para poder representar su distribución.

## Histograma

El *histograma* es una de las formas gráficas más comunes de representar datos de tipo cuantitativo, ya sean discretos o continuos. Es útil para resumir variables que toman gran cantidad de valores distintos. El procedimiento usual es el de agrupar los valores en *intervalos o clases*. En principio, podemos tomar una longitud arbitraria de los intervalos y, por lo tanto, también un número arbitrario de ellos. Una vez agrupadas las observaciones en intervalos, se realiza un procedimiento de conteo del número de observaciones en cada uno de ellos, análogo al de



las gráficas de barras. Al hacer estas agrupaciones ya no se dispone de la información original de los sujetos (claramente hay una pérdida de información debido a la agrupación), pero se espera que los rasgos generales de la distribución subyacente de la variable se conserven y se puedan mostrar con claridad.

Aunque los histogramas son parecidos a los diagramas de barras, tienen algunas diferencias importantes. Mientras que en un diagrama de barras la altura de la barra es proporcional a la frecuencia, en el histograma, la frecuencia es la que es proporcional al área del rectángulo. Como ya vimos, en el diagrama, cada barra representa la frecuencia de una categoría de la variable, mientras que en un histograma, cada rectángulo representa las frecuencias de los diversos valores de la variable que pertenecen al intervalo. Entonces, el área del  $i$ -ésimo intervalo es

$$A_i = b_i \cdot a_i = \frac{n_i}{n}$$

con  $n_i$  la frecuencia absoluta en el intervalo  $i$ , y  $n$  el total de la muestra. La base del rectángulo,  $b_i$ , es la *longitud o amplitud del intervalo* y la altura,  $a_i$ , es igual a

$$a_i = \frac{1}{b_i} \frac{n_i}{n}$$

La altura del rectángulo, de acuerdo a esta definición, se conoce como *densidad de frecuencia*.

Dado que histograma es la representación de la distribución subyacente a la variable, debe ser posible determinar o visualizar cuáles son los valores que ocurren con mayor frecuencia, los valores de poca frecuencia, la simetría o asimetría de la distribución, su dispersión, así como la forma de la misma.

## Marca de clase

Uno elemento usual en la construcción de un histograma es la denominada *marca de clase*, que no es más que el punto medio del intervalo, es decir

$$M_i = \frac{LS_i + LI_i}{2}, \quad i = 1, 2, \dots, k$$

Con  $M_i$  la marca de clase del intervalo  $i$ ;  $LS_i$  y  $LI_i$  los respectivos límites superior e inferior de este intervalo. En estos histogramas, la marca de clase se toma como el valor representativo de todos los valores del intervalo. El considerar que un sólo valor representa a todos los de la clase, ayuda a simplificar cálculos aproximados de las medidas de resumen usuales, ya que en lugar de hacerlos con el total de los datos, se realizan únicamente con el total de marcas de clase.

Un problema al construir los histogramas consiste en determinar cuántos intervalos debemos de considerar, y si los intervalos deben ser todos de la misma longitud. Por lo regular, los histogramas se construyen con intervalos de igual longitud, de manera que, o se define el número de intervalos y a partir de ello se deduce, mediante el rango de la variable, la longitud de cada intervalo, o bien, se decide arbitrariamente la longitud del intervalo y después se calcula el número de intervalos. Estas decisiones son, por supuesto, arbitrarias. Una regla de uso frecuente es definir tantos intervalos de la misma longitud, como el entero más grande de  $\sqrt{n}$ .

A manera de ejemplo tenemos 40 datos acerca de los niveles de colesterol en la sangre, cuyos valores oscilan entre 170 y 230. El número de intervalos que se considerarán será de acuerdo a la fórmula  $\sqrt{n}$ , en cuyo caso será  $\sqrt{40} \simeq 6$ , tomando en cuenta que los niveles en los que se “mueve” nuestra variable son aproximadamente 60 ( $230 - 170$ ) y que los intervalos de clase deben ser de la misma longitud, tendremos entonces 6 intervalos de clase con longitud 10, es decir

[170, 180)

[180, 190)

[190, 200)

[200, 210)

[210, 220)

[220, 230).

y las marcas de clase son

$$\frac{170 + 180}{2} = 175$$

$$\frac{180 + 190}{2} = 185$$

$$\frac{190 + 200}{2} = 195$$

$$\frac{200 + 210}{2} = 205$$

$$\frac{210 + 220}{2} = 215$$

$$\frac{220 + 230}{2} = 225.$$

## Tabla de frecuencias para la construcción del histograma

Para construir un histograma, suele presentarse la información en una tabla de frecuencias muy similar a la mostrada para las variables categóricas, sólo que, en este caso, se rempazan las categorías observadas, por los intervalos y se anexa la marca de clase. El formato es

### Cuadro 2

#### Tabla de frecuencias para un histograma

Intervalo	Marca de clase	Frecuencias Absolutas	Frecuencias Relativas	Frecuencias Abs. Acumuladas	Frecuencias Rel. Acumuladas
$[c_i, c_{i+1})$	$M_i$	$n_i$	$f_i$	$N_i$	$F_i$
$[c_1, c_2)$	$M_1$	$n_1$	$\frac{n_1}{n}$	$n_1$	$\frac{n_1}{n}$
$[c_2, c_3)$	$M_2$	$n_2$	$\frac{n_2}{n}$	$n_1 + n_2$	$\frac{n_1+n_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[c_{k-1}, c_k)$	$M_k$	$n_k$	$\frac{n_k}{n}$	$n_1 + n_2 + \dots + n_k$	$\frac{n_1+n_2+\dots+n_k}{n}$

En nuestro ejemplo, la tabla resumen es

Intervalos de clase	Marca de clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Abs. Acumulada	Frecuencia Rel. Acumulada
[170 ,180)	175	3	0.075	3	0.075
[180, 190)	185	7	0.175	10	0.25
[190, 200)	195	13	0.325	23	0.575
[200, 210)	205	8	0.20	31	0.775
[210, 220)	215	5	0.125	36	0.9
[220, 230)	225	4	0.10	40	1

## Resúmenes numéricos de información

Una manera usual de resumir los datos, es a través de algunos valores numéricos que puedan describirlos de forma adecuada y un poco más completa. El objetivo principal de estas medidas de resumen es proporcionar una idea aproximada de su comportamiento, es decir, una idea cercana a la distribución subyacente de estos datos. Estas medidas debieran aportar información sobre el valor del *centro* o *central* de los datos, la *dispersión* que presentan, la *posición* o si presentan alguna especie de *simetría*.

Para este fin, existen diferentes medidas de resumen para *datos numéricos o cuantitativos*, como son:

## Medidas de tendencia central

Estas medidas corresponden a la forma más común de resumir información cuantitativa, y responden a la preguntas estándar del tipo: ¿Cuál es el valor típico de los datos? o ¿Dónde se localiza el centro de los datos?. Las medidas de tendencia central de uso más frecuentes son

### Media

La *media*, *media aritmética* o *media muestral* es la medida de tendencia central más popular, la forma de calcularla es

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

con  $x_1, \dots, x_n$  los valores observados de la variable en la muestra, y  $n$  el tamaño de la muestra.

La media se considera como el valor *típico* o valor *central* de nuestros datos, y, generalmente, se utiliza como un elemento fundamental para caracterizar o representar a todo el conjunto de observaciones.

### Algunas características de la media aritmética

- Es una medida numérica que sólo se puede calcular para datos cuantitativos.

- Es una medida que se ve fuertemente afectada por valores extremos en los datos.

Consideramos por ejemplo, la estatura en centímetros de 12 estudiantes de sexto de primaria, 162, 165, 157, 164, 152, 147, 148, 131, 147, 155, 145, 132; la media es

$$\bar{X} = \frac{162 + 165 + 157 + 164 + 152 + 147 + 148 + 131 + 147 + 155 + 145 + 132}{12}$$

$$\bar{X} = 150.4167$$

## Mediana

La *mediana* es otra medida de tendencia central, es, propiamente, en valor central de los datos. Una vez que se han ordenado los datos en forma creciente, el cálculo de esta medida está dado por

$$\mathbf{Me} = \begin{cases} x_{(\frac{n+1}{2})} & \text{Si } n \text{ es impar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{Si } n \text{ es par} \end{cases}$$

Entonces, si  $n$  es impar, la mediana será el valor de los datos ordenados que ocupa la posición  $(n + 1)/2$ , es decir, el valor central. Mientras que si  $n$  es par, la mediana es *la media aritmética* de los dos valores centrales:  $(n/2)$  y  $(n/2 + 1)$ .

## Algunas características de la mediana

- Es una medida numérica que se puede calcular para datos cuantitativos o cualitativos.
- La mediana no se ve afectada por valores extremos de los datos (es *resistente* contra datos extremos).

Para el caso de las estaturas de los estudiantes de sexto año, como 12 es par, la mediana será la media de los valores que ocupen el lugar 6 y 7 de los datos ordenados, es decir, de 131, 132, 145, 147, 147, **148**, **152**, 155, 157, 162, 164, 165,

$$Me = \frac{148 + 152}{2}, Me = 150$$

## Moda

Otra medida de tendencia central usual es *la moda* que, en un conjunto de datos numéricos, es el valor que más se repite. En este sentido, está acorde al concepto usual que los seres humanos tenemos sobre “algo que está de moda” y que significa que corresponde a una situación muy común entre las personas, algo que la mayoría de ellas está usando o realizando, etc. Su cálculo estriba simplemente en registrar cuál valor en los datos se presenta con mayor frecuencia.

### Algunas características de la moda

- Es una medida que puede calcularse tanto para datos cuantitativos como cualitativos.
- No se ve afectada por valores extremos.
- No es única, puede existir más de una moda (en este caso se dice que la distribución es *multimodal*) en un conjunto de datos.
- Es posible que no exista, situación que ocurre cuando ninguno de los valores se repite o todos se repiten el mismo número de veces.

En nuestro ejemplo el valor que se repite únicamente es el 147, por lo que podemos decir que nuestra variable estatura es unimodal o que sólo tiene una moda.

## Medidas de dispersión

El concepto de variabilidad juega un papel fundamental en la estadística. De hecho, si los datos asociados a un fenómeno no presenta variaciones, la estadística prácticamente no tendría ninguna razón de ser. Por lo tanto, la cualidad más importante de los datos, desde el punto de vista estadístico, es su *variabilidad*. Es claro que los valores de alguna característica varían de sujeto a otro, de un grupo a otro, de una región a otra, incluso, de un momento a otro. Esta variabilidad siempre existe en un conjunto de datos y debe ser considerada para comparar de

manera más eficiente dos conjuntos de datos. De hecho, desde el punto de vista estadístico, es tan importante conocer las medidas de tendencia central, e.g., promedio, como la variabilidad de las observaciones alrededor de él, ya que la validez de este valor típico depende fuertemente de si los datos individuales se concentran o dispersan a su alrededor. Mientras más cercanos estén los datos a su valor promedio, tendremos mayor certeza de que éste valor caracteriza de forma adecuada a todos los datos. Por ejemplo, si consideramos los dos conjuntos de datos

$$48, 49, 50, 51, 52 \quad \text{y} \quad 1, 2, 50, 98, 99$$

Ambos conjuntos tienen igual promedio: 50, e igual mediana: 50, no obstante, es evidente que la variabilidad del segundo conjunto es mucho mayor que la del primero, por lo que las conclusiones que se tomen sobre ambos conjuntos, debe ser distintas. Por ejemplo, si ambos conjuntos son salarios (en miles) de trabajadores, es claro que en el primer conjunto estaríamos hablando de salarios muy homogéneos, tal vez de trabajadores con el mismo puesto o rango dentro de una institución, mientras que en el segundo, se trataría de salarios con una gran disparidad que, quizás, corresponderían a puestos de trabajo totalmente diferentes.

## Varianza y desviación estándar

La medida de variabilidad más común es la *desviación estándar*, que representa la distancia de cualquier punto al centro de los datos. Es aproximadamente la distancia media de cualquier punto al centro de los datos, en este caso, el centro es la media. La forma de calcular esta medida es

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Veamos cuáles son los valores de esta medida de variabilidad en los dos conjuntos de datos que mostramos anteriormente. Para el primer conjunto, cuya media es 50, tenemos

---

<sup>2</sup>Dado que este es un promedio, se debería dividir por el número total de elementos que lo componen,  $n$ , el porqué se divide entre  $n-1$  se debe a razones técnicas.

$$\begin{aligned}
s_1 &= \sqrt{\frac{\sum_{i=1}^5 x_i}{5-1}} = \sqrt{\frac{(48-50)^2 + (49-50)^2 + (50-50)^2 + (51-50)^2 + (52-50)^2}{4}} \\
&= \sqrt{\frac{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2}{4}} \\
&= \sqrt{\frac{4+1+0+1+4}{4}} = 1.581
\end{aligned}$$

Para el segundo conjunto cuya media también es 50, tenemos

$$\begin{aligned}
s_2 &= \sqrt{\frac{\sum_{i=1}^5 x_i}{5-1}} = \sqrt{\frac{(1-50)^2 + (2-50)^2 + (50-50)^2 + (98-50)^2 + (99-50)^2}{4}} \\
&= \sqrt{\frac{(-49)^2 + (-48)^2 + 0^2 + 48^2 + 49^2}{4}} \\
&= \sqrt{\frac{2401 + 2304 + 0 + 2304 + 2401}{4}} = 48.503
\end{aligned}$$

observemos que la desviación estándar de esta segunda población es aproximadamente 30.7 (48.503/1.581) veces más grande que la primera población.

### Características de la desviación estándar

- $s \geq 0$
- El mínimo de la desviación estándar es *cero* que corresponde a la situación donde todos los datos son iguales y, por lo tanto, iguales a su media. Claramente estos datos no presentan variabilidad alguna.
- La desviación estándar se ve afectada por observaciones extremas, ya que se basa en distancias respecto a la media, que es afectada por datos extremos.



- La desviación estándar tiene las mismas unidades que los datos.

Otra medida usual de variabilidad o dispersión es la *varianza*, que simplemente es el cuadrado de la desviación estándar

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Las correspondientes varianzas de los dos conjuntos de datos anteriores son 2.5 para el primer conjunto y 2352.5 para el segundo. Un inconveniente de esta medida es que está dada en unidades cuadradas, lo que no la hace interpretable en las unidades de los datos. Algunas características de la varianza son

- $s^2 \geq 0$
- El mínimo de la varianza es *cero* que corresponde a la situación donde todos los datos son iguales.
- La varianza se ve afectada por observaciones extremas, ya que se basa en distancias respecto a la media, que es afectada por datos extremos.
- La varianza está dada en unidades cuadradas de los datos.

Tanto para la varianza como para la desviación estándar, mientras más cercanas estén a cero, más concentrados estarán los datos alrededor de su media; en contraparte, mientras más grandes sean estas medidas, mayor será la dispersión de los datos respecto a esta media.

## Otras medidas de dispersión

Aunque la desviación estándar y la varianza son las principales medidas estadísticas de dispersión, existen algunas otras que pueden utilizarse para tener un panorama más completo sobre la variabilidad de nuestros datos. Por ejemplo

## Rango

Una de las medidas de variabilidad más simples de calcular es el *rango* de los datos, que se define como

$$\mathbf{R} = X_{(n)} - X_{(1)}$$

y corresponde a la diferencia entre el valor más grande en la muestra (máximo) y el valor más pequeño (mínimo). En nuestros dos conjuntos de datos este rango es

$$\mathbf{R}_1 = 52 - 48 = 4 \quad \mathbf{R}_2 = 99 - 1 = 98$$

nuevamente se observa que la variabilidad del segundo conjunto es mucho mayor que la del primero.

Una limitante muy importante de esta medida de variabilidad es que depende únicamente de las observaciones extremas en los datos, y no considera en su cálculo al resto de ellas. En la práctica, el rango se utiliza cuando requerimos una medida burda o aproximada de la variabilidad de nuestros datos.

## Desviación media

Esta medida está relacionada con el hecho de que una medida de variabilidad debería de considerar las desviaciones, en valor absoluto, de los datos respecto a su valor promedio, por lo que se define como

$$\mathbf{DM} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

En nuestros conjuntos de datos, esta desviación media es

$$\begin{aligned}
\mathbf{DM}_1 &= \frac{\sum_{i=1}^5 |x_i - \bar{x}|}{5} = \frac{|48 - 50| + |49 - 50| + |50 - 50| + |51 - 50| + |52 - 50|}{5} \\
&= \frac{|-2| + |-1| + |0| + |1| + |2|}{5} \\
&= \frac{2 + 1 + 0 + 1 + 2}{5} = 1.2
\end{aligned}$$

$$\begin{aligned}
\mathbf{DM}_2 &= \frac{\sum_{i=1}^5 |x_i - \bar{x}|}{5} = \frac{|1 - 50| + |2 - 50| + |50 - 50| + |98 - 50| + |99 - 50|}{5} \\
&= \frac{|-49| + |-48| + |0| + |48| + |49|}{5} \\
&= \frac{49 + 48 + 0 + 48 + 49}{5} = 38.8
\end{aligned}$$

esta medida también refleja la mayor variabilidad del segundo conjunto de datos.

La desviación media no es una medida de variabilidad muy usual, esencialmente por la dificultad técnica<sup>3</sup> de trabajar con la función valor absoluto, y porque, como ya vimos, la desviación estándar es una medida basada en este mismo principio de distancias respecto al valor medio, que es más conveniente y que no presenta este tipo de problemas técnicos en su manipulación.

## Medidas relativas de dispersión

En ocasiones es necesario comparar la variabilidad de dos conjuntos de datos medidos en escalas diferentes. Por ejemplo, comparar la variabilidad de las calificaciones que obtiene un grupo de estadística, respecto al peso de los estudiantes del grupo. En este caso, es claro que la comparación de la variabilidad de estas dos medidas a través de la desviación estándar o cualquiera de las otras medidas de variabilidad que hemos definido, carece completamente de sentido. Además, la varianza y desviación estándar se ven influenciadas por la magnitud de las escalas de medición, ya que, generalmente, las escalas más grandes tienen varianzas más grandes. Entonces, es necesario tener medias que no dependan ni de las unidades de medición

---

<sup>3</sup>No es diferenciable en cero

de las variables involucradas, ni de la magnitud de sus escalas de medición. Para resolver esta situación se utilizan las llamadas *medidas relativas de dispersión*, la más importante de ellas es *el coeficiente de variación*, definido como

$$\text{C.V.} = \frac{s}{|\bar{x}|} \times 100 \%$$

Al dividir,  $s$ , que es una medida de varibilidad y  $\bar{x}$  medida de centralidad, se elimina el efecto de las unidades de medición, ya que ambas tiene las mismas unidades. Además, dividir por el promedio permite corregir el efecto que tiene la magnitud de la escala de medición sobre la desviación estándar.

## Medidas de posición

La mediana definida como una medida de posición es aquella observación que (cuando los valores se ordenan de menor a mayor) se sitúa en el centro de la muestra. Por lo tanto, la mediana es el valor hasta el que se acumulan el 50 % de los datos. Para describir de manera más completa un conjunto de datos, se suelen calcular y reportar otros valores que acumulan frecuencias diferentes a este 50 %, estos valores se conocen genéricamente como *percentiles* o *centiles* de la distribución. De acuerdo al porcentaje que acumulan, estos percentiles reciben diferentes nombres, a saber

## Cuartiles

Los *cuartiles* son los valores que (con la variable ordenada de menor a mayor) dejan por debajo de su posición el 25 % (primer cuartil), 50 % (segundo cuartil) y 75 % (tercer cuartil) de las frecuencias acumuladas, respectivamente. Los cuartiles dividen a los datos en cuatro grupos con igual número de observaciones.

La forma de calcular estos valores es similar a la de la mediana.

$$\text{Cuartiles} = \begin{cases} Q_1 = x_{\left(\frac{n+1}{4}\right)} \\ Q_2 = x_{\left(\frac{2(n+1)}{4}\right)} \\ Q_3 = x_{\left(\frac{3(n+1)}{4}\right)} \end{cases} \quad \text{Si } n \text{ es impar}$$

$$Cuartiles = \begin{cases} Q_1 = \frac{1}{2} \left( x_{(\frac{n}{4})} + x_{(\frac{n+1}{4})} \right) \\ Q_2 = \frac{1}{2} \left( x_{(\frac{2n}{4})} + x_{(\frac{2(n+1)}{4})} \right) \\ Q_3 = \frac{1}{2} \left( x_{(\frac{3n}{4})} + x_{(\frac{3(n+1)}{4})} \right) \end{cases} \quad \text{Si } n \text{ es par}$$

Es claro que el cuartil 2,  $Q_2$ , corresponde a la mediana.

En el conjunto de 18 datos 122, 126, 133, 140, 145, 149, 150, 157, 162, 166, 175, 177, 177, 183, 188, 199, 212, los cuartiles son:

$$Q_1 = 145$$

ya que  $x_{(4.5)}$  y  $x_{(4.75)}$  es  $x_5$ , por lo que el primer cuartil corresponde al quinto valor (de menor a mayor) que es 145.

$$Q_2 = 159.5$$

debido a que  $x_{(9)}$  y  $x_{(9.5)} \simeq x_{(10)}$  se hace la media de  $x_{(9)}$  y  $x_{(10)}$  que es 159.5.

Por último se tiene  $x_{(13.5)}$  y  $x_{(14.25)}$  por lo que

$$Q_3 = 177$$

.

## Deciles

De forma totalmente análoga se definen los *deciles* que dividen a los datos ordenados en *diez* grupos de igual tamaño. El primer decil deja 10 % de las observaciones debajo de él, el segundo el 20 %, y así sucesivamente hasta el noveno que deja 90 %. Nótese que el decil 5,  $D_5$ , corresponde a la mediana de los datos.

## Percentiles

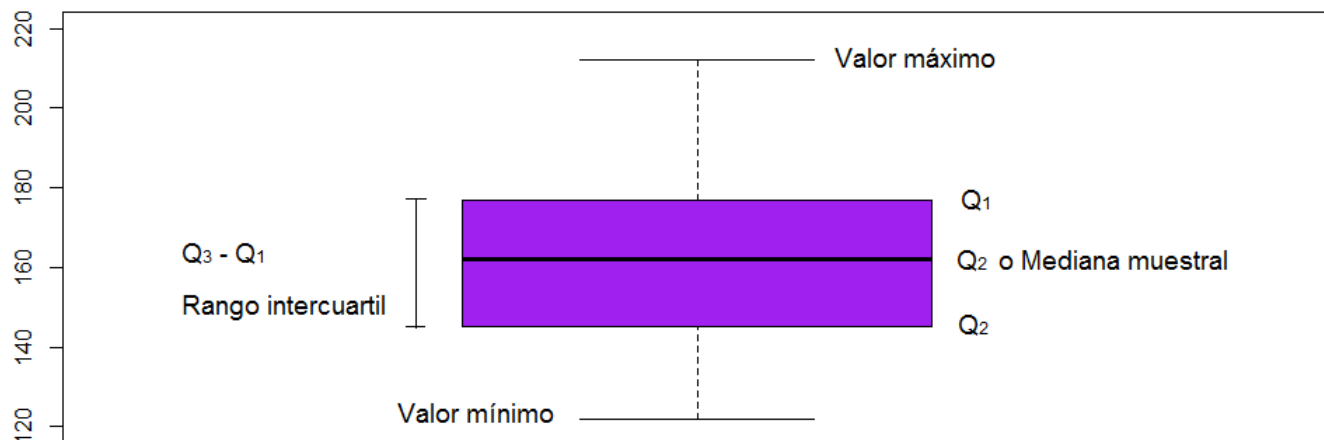
La partición más fina de los datos se realiza a través de los llamados *percentiles* que corresponden a los 99 valores que dividen la serie de datos en 100 partes iguales. Los percentiles dan los valores que dejan el 1 %, 2 %, ..., 99 % de los datos debajo de ellos. Por ejemplo, los

percentiles  $P_{25}$ ,  $P_{50}$  y  $P_{75}$  corresponden al primer cuartil, al segundo cuartil o la mediana y al tercer cuartil, respectivamente.

## Diagrama de caja

Una forma gráfica de representar los cuartiles, valores máximo y mínimo, así como presencia de datos atípicos es por medio de un *diagrama de caja*.

El *boxplot* o diagrama de caja de nuestro ejemplo anterior es



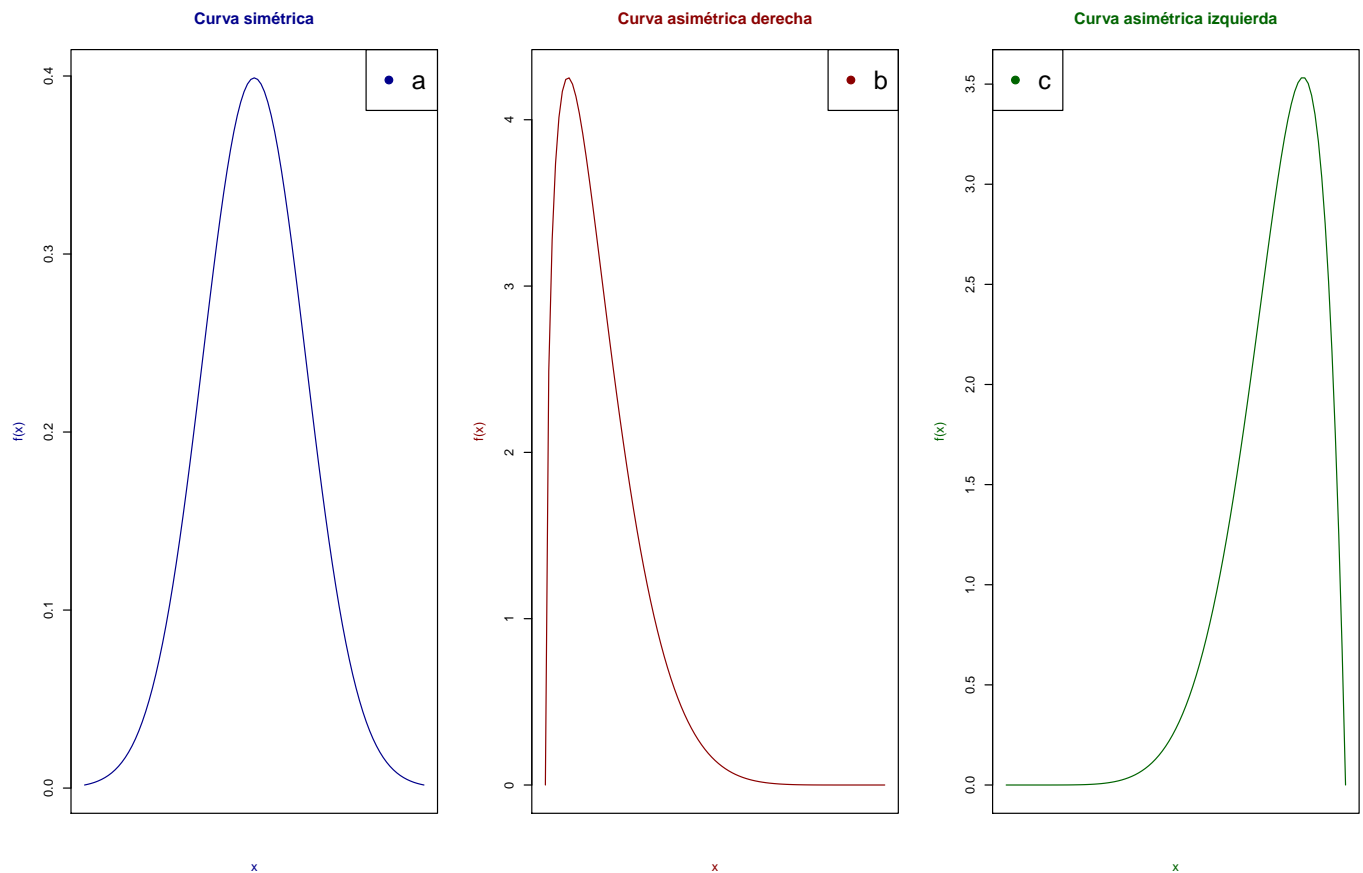
## Medidas de forma

Dos características muy importantes de la forma de la distribución subyacente a los datos, son el grado de *simetría* de la misma y su nivel de *picudez* o *apuntamiento*. Las medidas numéricas que proporcionan información sobre estos dos aspectos de una distribución son: *el coeficiente de asimetría* y *el coeficiente de curtosis*, respectivamente.

## Caracterización de simetría a través de las medidas de tendencia central

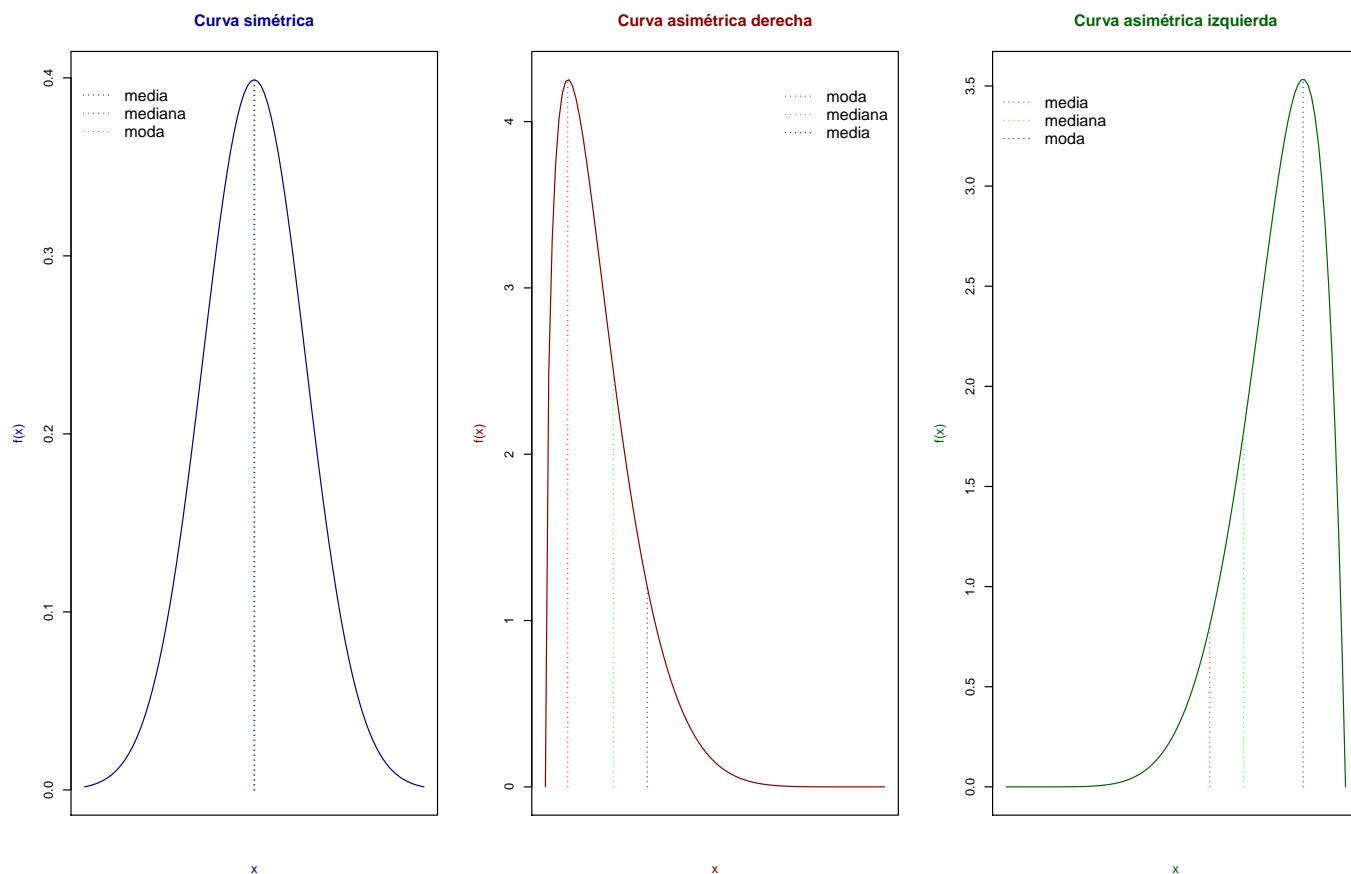
Una primera forma de determinar si la distribución subyacente a nuestros datos es o no simétrica y, si fuera el caso, qué tipo de asimetría presenta, es a través del uso de las medidas de tendencia central: media, mediana y moda. En general, podemos distinguir tres tipos de

simetría en las estas distribuciones. Las distribuciones simétricas (fig a), asimétrica derecha (o con asimetría por la derecha) (fig b) y asimétrica izquierda (o con asimetría por la izquierda) (fig c). Entonces, una distribución es simétrica si los valores de ella están en la misma proporción a derecha e izquierda de su mediana. Es asimétrica por la derecha (tiene una “cola derecha” larga) si posee más frecuencia de valores en ese lado de la distribución, y es asimétrica por la izquierda (tiene una “cola izquierda” larga) si son más frecuentes los valores a su lado izquierdo.



Basándonos en las medidas de tendencia central referidas, una distribución es simétrica si  $media = mediana = moda$ , es asimétrica por la derecha si  $moda \leq mediana \leq media$  y es asimétrica por la izquierda si  $media \leq mediana \leq moda$ . Es decir

$$\left\{ \begin{array}{l} \text{Distribución simétrica : } media = mediana = moda \\ \text{Distribución asimétrica por la derecha : } moda \leq mediana \leq media \\ \text{Distribución asimétrica por la izquierda : } media \leq mediana \leq moda \end{array} \right.$$



## Coeficiente de asimetría

Si la distribución de los datos es simétrica, entonces las observaciones tienden a ubicarse en igual proporción a ambos lados de su valor medio, por lo que cualquier medida que tome en cuenta las desviaciones de los datos sobre este valor medio, podría utilizarse como un indicativo de la simetría o asimetría de la distribución. Entonces, una propuesta inicial para cuantificar la magnitud de estas desviaciones podría ser el promedio de éstas

$$\frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$$

Esta medida parece adecuada, ya que si tuviéramos muchas observaciones por encima de la media, esperaríamos un valor positivo de ella, mientras que una mayor cantidad de valores por debajo de ella, arrojaría un valor negativo. Desafortunadamente, ya que para cualquier conjunto de datos se tiene que



$$\begin{aligned}
\frac{\sum_{i=1}^n (x_i - \bar{x})}{n} &= \frac{1}{n} \left( \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \right) \\
&= \frac{1}{n} \left( \sum_{i=1}^n x_i - n\bar{x} \right) \\
&= \frac{1}{n} \left( \sum_{i=1}^n x_i - n \frac{\sum_{i=1}^n x_i}{n} \right) \\
&= \frac{1}{n} \left( \sum_{i=1}^n x_i - \sum_{i=1}^n x_i \right) = 0
\end{aligned}$$

esta medida no es adecuada para este fin.

Considerar el promedio del cuadrado de estas diferencia, resolvería el problema de la suma cero, pero no proporcionaría la dirección de la asimetría. Es decir, dado que este promedio siempre sería positivo, no podríamos saber si hay más observaciones situadas a la derecha o a la izquierda de esta media. Ya que requerimos una potencia de las desviaciones que conserve el signo de esta diferencia, parece razonable considerar como propuesta

$$\mathbf{m}_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n}$$

Entonces, si la distribución de los datos es simétrica esta medida estará cercana a *cero*, si los datos tienden a presentar más observaciones por encima de su media, se dirá que tiene una *asimetría por la derecha o positiva*, y si la tendencia es a que se observen más datos por debajo de la media, entonces lo que se tiene es una *asimetría por la izquierda o negativa*.

$$\mathbf{m}_3 = \begin{cases} 0 & \text{Distribución simétrica} \\ > 0 & \text{Distribución asimétrica: asimetría (sesgo) por la derecha o positiva} \\ < 0 & \text{Distribución asimétrica: asimetría (sesgo) por la izquierda o negativa} \end{cases}$$

A partir de esta medida se define el coeficiente de asimetría de Fisher como

$$\gamma_1 = \frac{\mathbf{m}_3}{\mathbf{s}^3}$$

Observe que el numerador de esta expresión es un número positivo, por lo que el signo de este coeficiente lo determina  $\mathbf{m}_3$ , y tenemos las mismas condiciones de simetría para  $\gamma_1$ , que las dadas sólo por  $\mathbf{m}_3$  en la tabla anterior. Es decir

$$\gamma_1 = \begin{cases} 0 & \text{Distribución simétrica} \\ > 0 & \text{Distribución asimétrica: asimetría (sesgo) por la derecha o positiva} \\ < 0 & \text{Distribución asimétrica: asimetría (sesgo) por la izquierda o negativa} \end{cases}$$

Otra medida de asimetría es el llamado *coeficiente de asimetría de Pearson*. Esta medida se basa en la relación que se estableció entre las medidas de tendencia central al caracterizar la simetría de una distribución, y se define como

$$\mathbf{A}_P = \frac{\bar{x} - moda}{\mathbf{s}}$$

por lo que tenemos

$$\mathbf{A}_P = \begin{cases} 0 & \text{Distribución simétrica} \\ > 0 & \text{Distribución asimétrica: asimetría (sesgo) por la derecha o positiva} \\ < 0 & \text{Distribución asimétrica: asimetría (sesgo) por la izquierda o negativa} \end{cases}$$

## Curtosis

Esta es una medida de qué tanto las observaciones están acumuladas en la parte central de la distribución; como aplica únicamente para distribuciones simétricas, entonces sirve para ver qué tan concentradas se encuentran las observaciones alrededor de su media. El coeficiente de *curtosis* o simplemente *la curtosis* de una distribución se calcula como

$$\gamma_2 = \frac{\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n}}{s^4} = \frac{\mathbf{m}_4}{s^4}$$

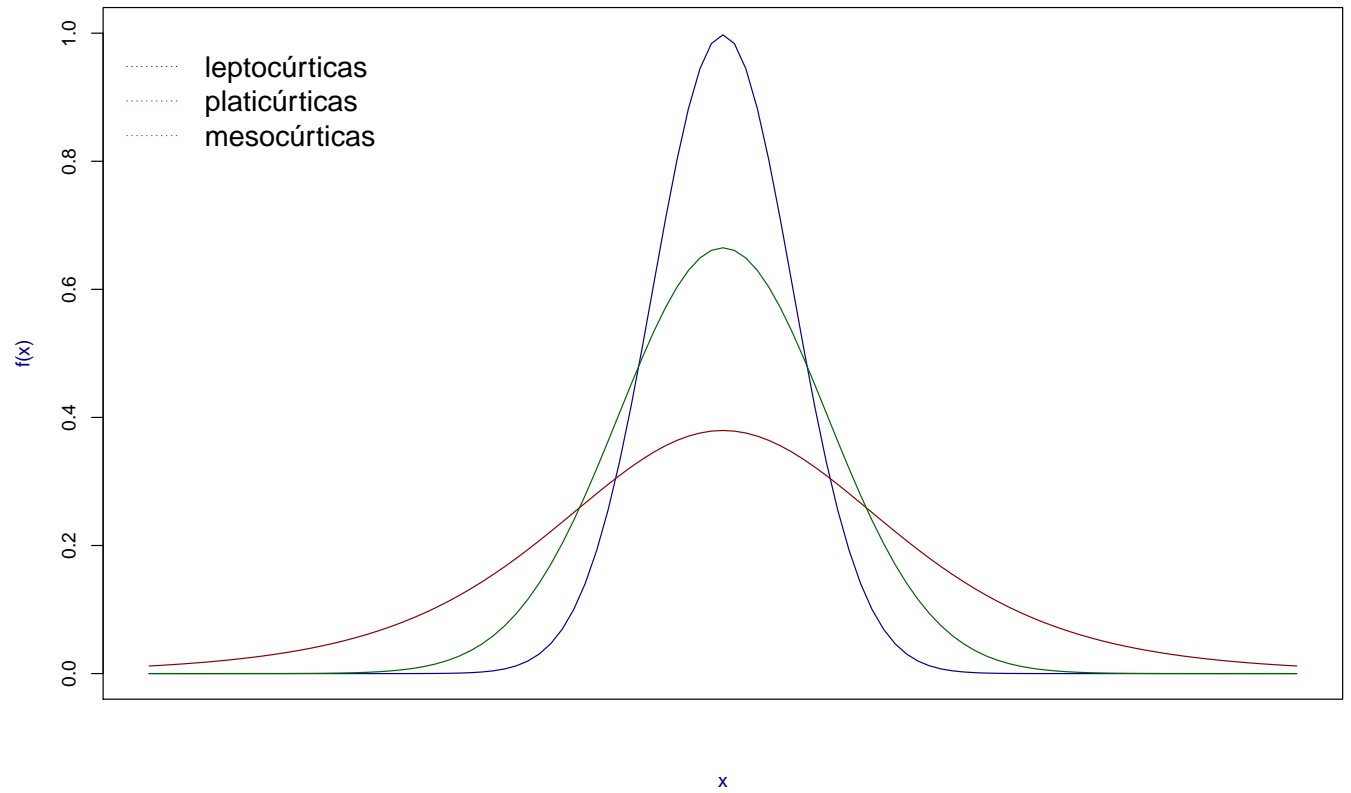
Por lo general, el valor de la curtosis de una distribución se compara con la correspondiente coeficiente de curtosis en la *distribución normal*, que es igual a 3, por lo que se acostumbra reportar, en lugar del coeficiente de curtosis anterior, el llamado *exceso de curtosis*, dado por

$$\gamma_2^* = \gamma_2 - 3$$

que claramente reporta qué tanto difiere la “picudez” de nuestra curva de la de la normal.

Al igual que para la simetría, existen tres tipos de distribuciones que pretendemos identificar y corresponden a las llamadas *curvas mesocúrticas* que tienen la curtosis o nivel de “picudez” similar a la de una distribución normal. El segundo tipo corresponde a aquellas distribuciones que son *más achatadas* que la normal, y reciben el nombre de *platicúrticas*, por ejemplo, una *distribución t* con pocos grados de libertad. Finalmente, las distribuciones más picudas que la normal reciben el nombre de *distribuciones leptocúrticas*. En esencia, las distribuciones platicúrticas tienen una varianza o dispersión más grande que que la normal, lo que implicaría que sus observaciones están más dispersas alrededor de su media que lo que ocurre en esta normal. Por el contrario, en las distribuciones leptocúrticas esta dispersión es menor que en la normal y, por lo tanto, las observaciones están más aglutinadas sobre la media, que lo que ocurre en una normal.

Curvas: mesocúrticas, platicúrticas y leptocúrticas



En términos del coeficiente de curtosis modificado, tenemos que

$$\gamma_2^* = \begin{cases} 0 & \text{Distribución mesocúrtica (normal)} \\ < 0 & \text{Distribución platicúrtica ("achatada")} \\ > 0 & \text{Distribución leptocúrtica ("puntiaguda")} \end{cases}$$

## Resúmenes numéricos para datos agrupados

### Media

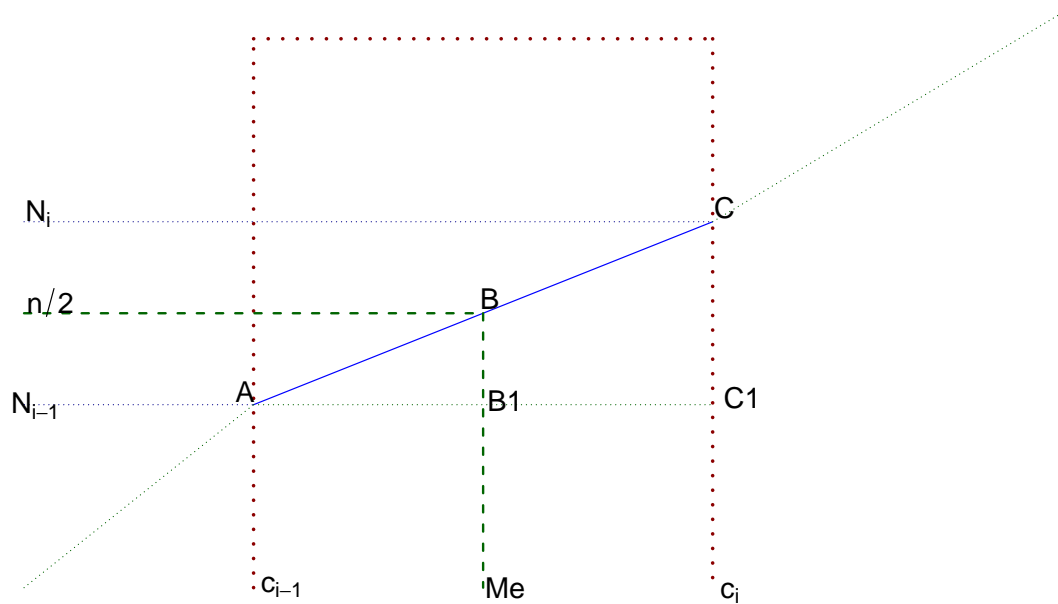
Comentamos que una vez agrupadas las observaciones en intervalos, los cálculos se simplificaban en el sentido de realizar tantas operaciones como intervalos tuvieramos, por lo que el cálculo de la media con datos agrupados es

$$\bar{X} = \frac{\sum_{i=1}^k M_i \cdot n_i}{n} = \sum_{i=1}^k M_i \cdot f_i$$

La interpretación de esta medida es exactamente igual a la que se hace sin agrupar los datos.

### Mediana

El cálculo de la mediana para datos agrupados reviste una complejidad mucho mayor que la de datos sin agrupar. Para realizar su cálculo nos auxiliaremos de la siguiente figura



Asumamos que el rectángulo punteado en rojo, es la barra del histograma que corresponde al intervalo donde se encuentra la mediana de los datos agrupados. Primeramente, supondremos que el valor de la mediana, **Me**, se encuentra en el intervalo  $[c_{i-1}, c_i)$  y, por supuesto, corresponde al punto donde se ha acumulado el 50 % de los datos:  $\frac{n}{2}$ . Entonces, hay que determinar este punto, **Me**, que deja exactamente la mitad de observaciones por debajo de él y la otra mitad por encima de él.

Obsérvese que el triángulo  $\triangle ACC1$  es semejante al triángulo  $\triangle ABB1$ , entonces se tiene que

$$\begin{aligned}\frac{CC1}{AC} &= \frac{BB1}{AB} \Rightarrow \\ \frac{N_i - N_{i-1}}{c_i - c_{i-1}} &= \frac{\frac{n}{2} - N_{i-1}}{\mathbf{Me} - c_{i-1}} \Rightarrow \\ \mathbf{Me} &= c_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} (c_i - c_{i-1})\end{aligned}$$

recordemos que  $N_i = \sum_{j=1}^i n_j$ . Si denotamos como  $\ell = c_i - c_{i-1}$  la longitud de los intervalos, podemos reescribir esta fórmula de la manera usual como

$$\mathbf{Me} = c_{i-1} + \frac{\frac{n}{2} - \sum_{ant} f}{f_{med}} \cdot \ell$$

con  $\sum_{ant} f$  : suma de las frecuencias de las clases anteriores a la clase de la mediana y

$f_{med}$  : frecuencia de la clase de la mediana.

## Moda

El cálculo de esta moda también es mucho más complicado que su equivalente en datos no agrupados. Para su cálculo nos basaremos en la siguiente figura



$$\begin{aligned}
\frac{EP}{RQ} &= \frac{PF}{ST} \Rightarrow \\
\frac{\mathbf{Mo} - L_i}{\Delta_1} &= \frac{U_i - \mathbf{Mo}}{\Delta_2} \Rightarrow \\
\Delta_2 (\mathbf{Mo} - L_i) &= \Delta_1 (U_i - \mathbf{Mo}) \Rightarrow \\
\Delta_2 \mathbf{Mo} - \Delta_2 L_i &= \Delta_1 U_i - \Delta_1 \mathbf{Mo} \Rightarrow \\
\mathbf{Mo} (\Delta_1 + \Delta_2) &= \Delta_1 U_i + \Delta_2 L_i \Rightarrow \\
\mathbf{Mo} &= \frac{\Delta_1 U_i + \Delta_2 L_i}{\Delta_1 + \Delta_2} \quad \text{Ya que } U_i = L_i + \ell \\
\mathbf{Mo} &= \frac{\Delta_1 (L_i + \ell) + \Delta_2 L_i}{\Delta_1 + \Delta_2} \\
&= \frac{(\Delta_1 + \Delta_2) L_i + \Delta_1 \ell}{\Delta_1 + \Delta_2} \\
&= L_i + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \cdot \ell
\end{aligned}$$

Este resultado tiene una interpretación muy interesante: Si se traza una parábola que pase por los tres puntos medios de los techos de los rectángulos de la figura base de la construcción, la abscisa del máximo de esta parábola será igual a la moda que obtuvimos.

## Desviación estándar

La forma de calcular la desviación estándar para datos agrupados es

$$s = \sqrt{\frac{\sum_{i=1}^k n_i (M_i - \bar{X})^2}{n - 1}}$$

y la correspondiente varianza se calcula como

$$s^2 = \frac{\sum_{i=1}^k n_i (M_i - \bar{X})^2}{n - 1}$$



## Medidas de posición

Tal como en los casos de la mediana y moda, cuando las variables son continuas y están agrupadas en intervalos, el cálculo de las medidas de posición es más complicado. Supondremos que el valor a calcular se encuentra en un intervalo dado  $[c_{i-1}, c_i)$  y que debemos determinar el punto que deja exactamente el porcentaje correspondiente de observaciones a un lado y al otro de él. Mediante argumentos geométricos similares a los utilizados en la construcción de la mediana, se puede demostrar que la fórmula para el cálculo de cualquier percentil con datos agrupados es

$$\mathbb{P}_k = c_{i-1} + \frac{n \cdot \frac{k}{100} - N_{i-1}}{n_i} \cdot \ell_i$$

Estos elementos ya han sido definidos anteriormente. En este caso

$N_{i-1} = \sum_{ant} f$ : suma de las frecuencias de las clases anteriores a la clase que contiene al  $k$ -ésimo percentil, y  $n_i$  es la frecuencia absoluta de la clase que contiene a este percentil.

Por ejemplo, para el primer cuartil (percentil 25) se tiene que  $k = 25$  y  $n \cdot \frac{k}{100} = \frac{n}{4}$ , por lo que

$$\mathbb{Q}_1 = \mathbb{P}_{25} = c_{i-1} + \frac{\frac{n}{4} - \sum_{ant} f}{n_i} \cdot \ell_i$$

## Coeficientes de simetría y curtosis

Para datos agrupados, los coeficientes de simetría y curtosis se calculan de manera similar a la de datos sin agrupar, considerando que ahora

$$\mathbf{m}_k = \frac{\sum_{i=1}^k n_i (M_i - \bar{X})^k}{n}$$

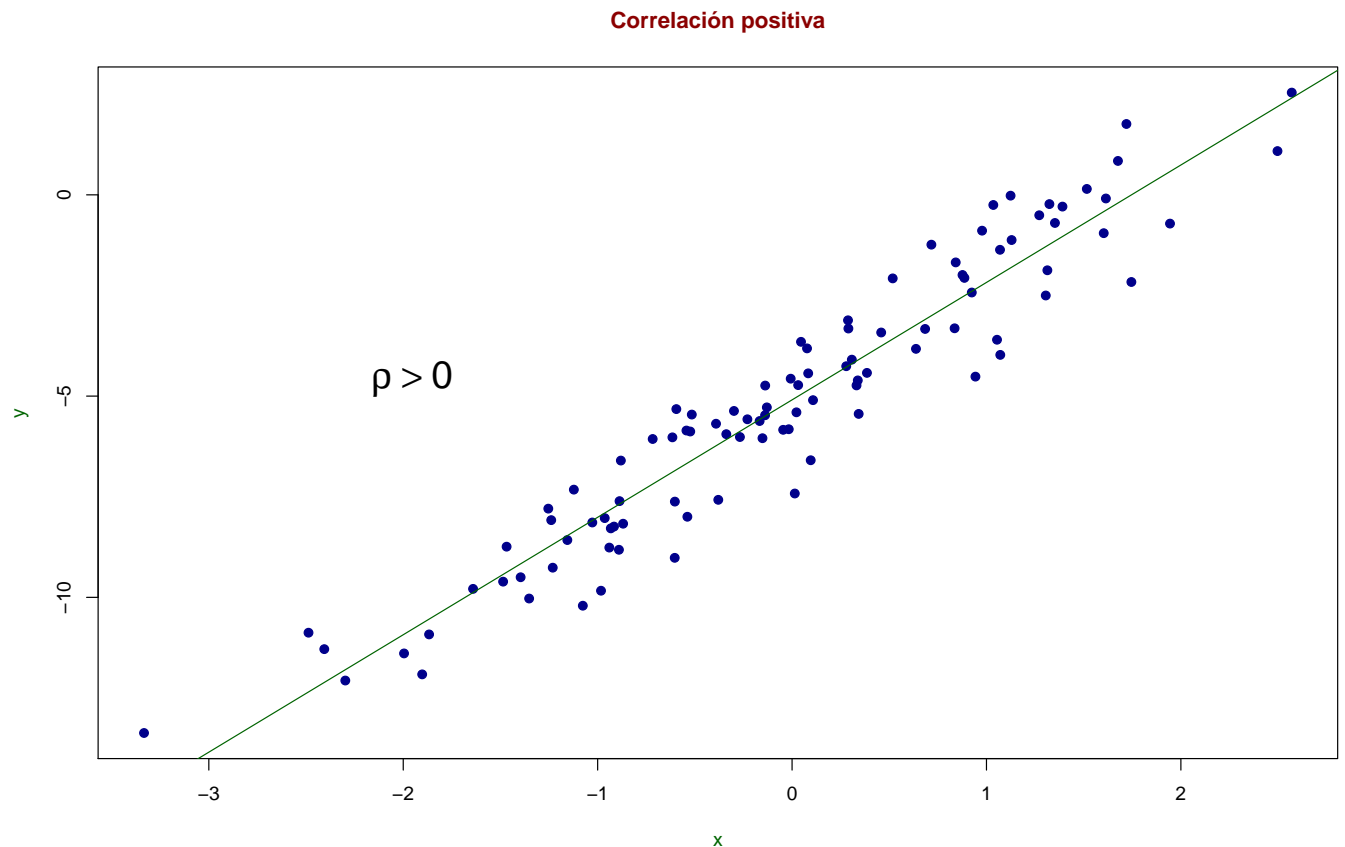
y utilizando también la expresión de la desviación estándar,  $\mathbf{s}$ , para datos agrupados.

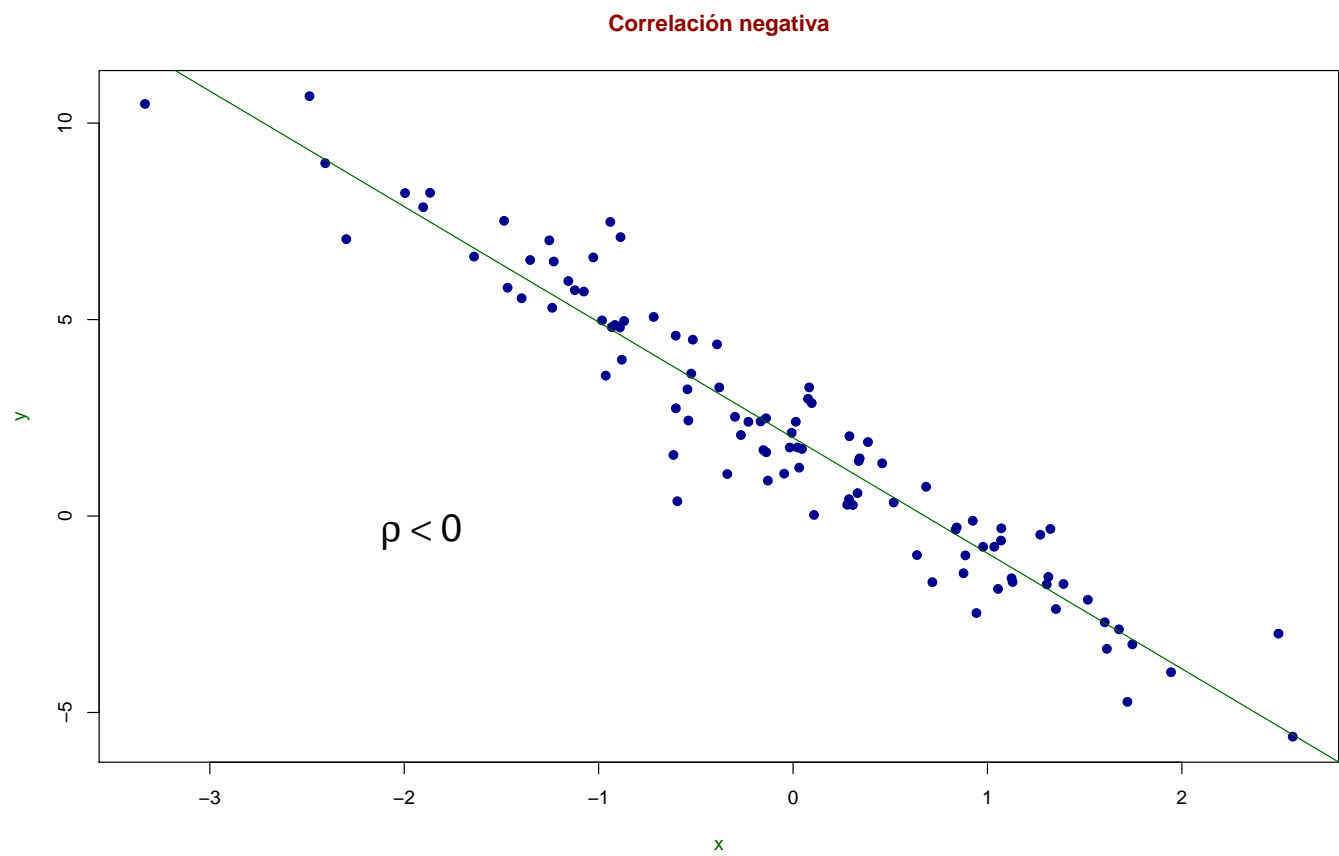
# Medidas de asociación y correlación

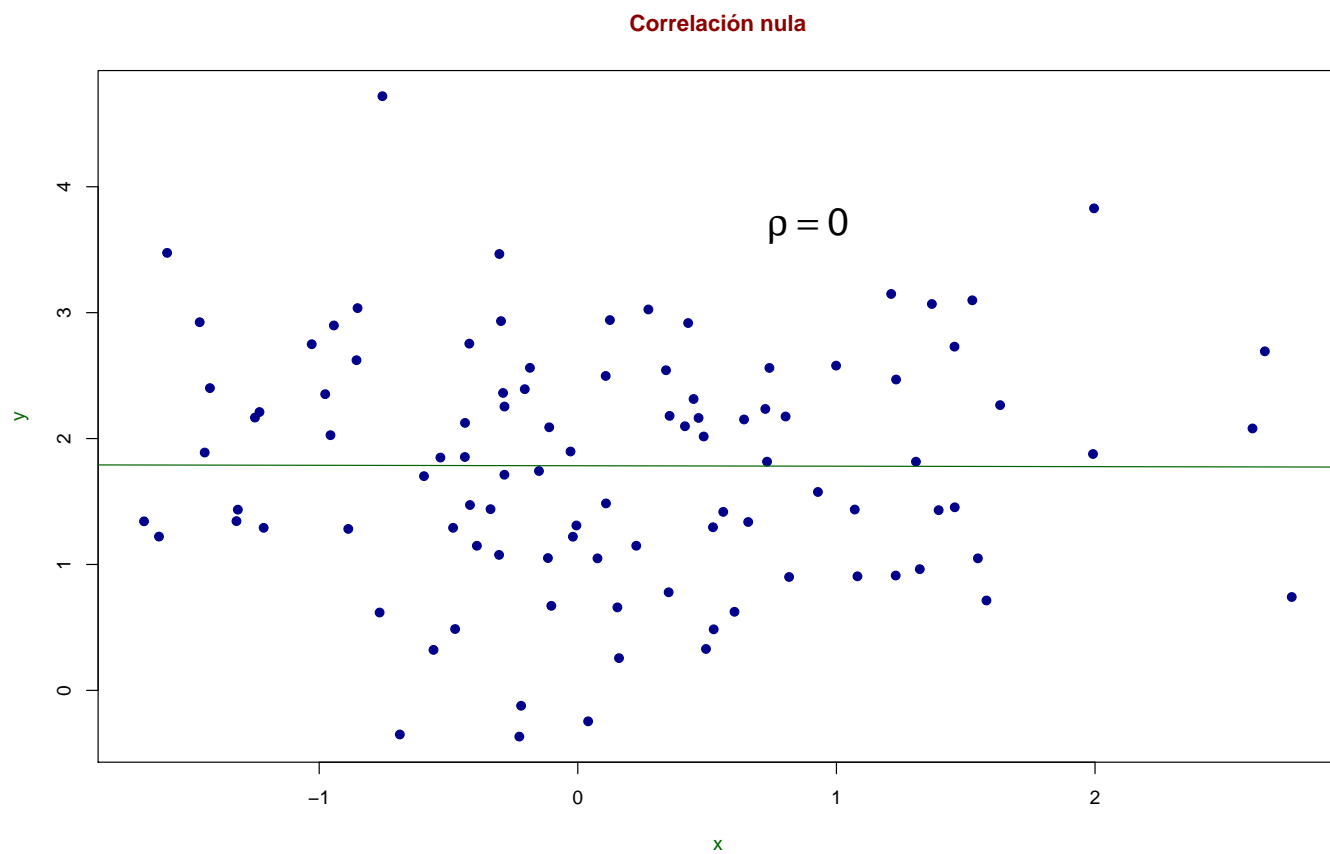
## Coefficiente de correlación lineal

Un coeficiente de correlación proporciona un valor numérico del grado de asociación entre dos variables,  $\mathbf{X}$  y  $\mathbf{Y}$ , i.e., en qué medida valores grandes de una de las variables se asocian con valores grandes de la otra (*correlación positiva*), o valores grandes de una se asocian con valores pequeños de la otra (*correlación negativa*).

Cuando no ocurre ninguna de las situaciones anteriores, decimos que las variables no están relacionadas (*correlación nula*) de manera lineal.







Los coeficientes de correlación fluctúan entre -1 y 1. Los valores positivos indican una asociación directa entre las variables, mientras que valores negativos indican una asociación inversa entre ellas. Un coeficiente cercano a cero, implica poca o nula asociación.

## Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson, *para variables cuantitativas medidas en escala continua*, es un índice que mide el grado de asociación lineal entre dos variables.

Adviértase que se menciona expresamente variables relacionadas linealmente. Esto significa que puede haber variables relacionadas no de forma lineal, en cuyo caso no procede calcular la correlación de Pearson. Cuya expresión es:

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x) \text{Var}(y)}}$$

Una vez calculado el valor del coeficiente de correlación, interesa determinar si tal valor obtenido indica que las variables están relacionadas linealmente en realidad, o tan sólo presentan dicha relación como consecuencia del azar. En otras palabras, nos preguntamos por la significancia de dicho coeficiente de correlación. Para dilucidar esta cuestión, debemos realizar la prueba de hipótesis

$$\mathbb{H}_0 : \rho = 0 \quad \text{vs.} \quad \mathbb{H}_a : \rho \neq 0$$

Cuyo estadístico de prueba es

Si  $\mathbf{X}$  y  $\mathbf{Y}$  son *conjuntamente normales bivariadas*, entonces

$$T = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t_{(n-2)}$$

Y rechazamos la hipótesis nula, a un nivel de significancia  $\alpha$ , si

$$|T| > t_{(1-\alpha/2, n-2)}$$

Si nos interesa probar que el coeficiente de correlación poblacional tiene un valor específico:  
 $\rho = \rho_0 \neq 0$ , entonces la prueba para este valor general es:

$$\mathbb{H}_0 : \rho = \rho_0 \quad vs. \quad \mathbb{H}_a : \rho \neq \rho_0$$



Tenemos que, bajo la hipótesis nula

$$\frac{1}{2} \log \left( \frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right) \xrightarrow{d} N \left( \frac{1}{2} \log \left( \frac{1 + \rho}{1 - \rho} \right), \frac{1}{n-3} \right)$$

Entonces, la prueba se puede realizar considerando la variable estandarizada

$$Z = \frac{\frac{1}{2} \log \left( \frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right) - \frac{1}{2} \log \left( \frac{1 + \rho}{1 - \rho} \right)}{\sqrt{\frac{1}{n-3}}} \underset{a}{\sim} N(0, 1)$$

Y rechazamos  $\mathbb{H}_0$ , a un nivel de significancia  $\alpha$ , sii

$$|Z| > Z_{(1-\alpha/2)}$$

## Coeficiente de correlación de Spearman



Este coeficiente es no paramétrico. Su cálculo es igual a la fórmula del de Pearson, cambiando los valores observados de las variables X y Y por sus respectivos rangos. Esto es

$$\hat{r}_S = \frac{\sum_{i=1}^n (R(x_i) - \bar{R}(x_i)) (R(y_i) - \bar{R}(y_i))}{\sqrt{\sum_{i=1}^n (R(x_i) - \bar{R}(x_i))^2 \sum_{i=1}^n (R(y_i) - \bar{R}(y_i))^2}} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

con  $d_i = R(x_i) - R(y_i)$ . Donde el rango de una observación es el lugar relativo que ocupa al ordenar los datos de menor a mayor. Se le asigna el rango promedio a las observaciones *empatadas*, i.e., se les asigna el promedio de los rangos que les ubieran correspondido si no

estuvieran empatadas.

### Prueba de hipótesis

Nuevamente, deseamos probar si la correlación dada por este coeficiente puede o no considerarse estadísticamente significativa, por lo que nos planteamos las hipótesis

$$\mathbb{H}_0 : r_S = 0 \quad vs. \quad \mathbb{H}_a : r_S \neq 0$$

La estadística de prueba también es similar a la de Pearson. Para  $n > 30$

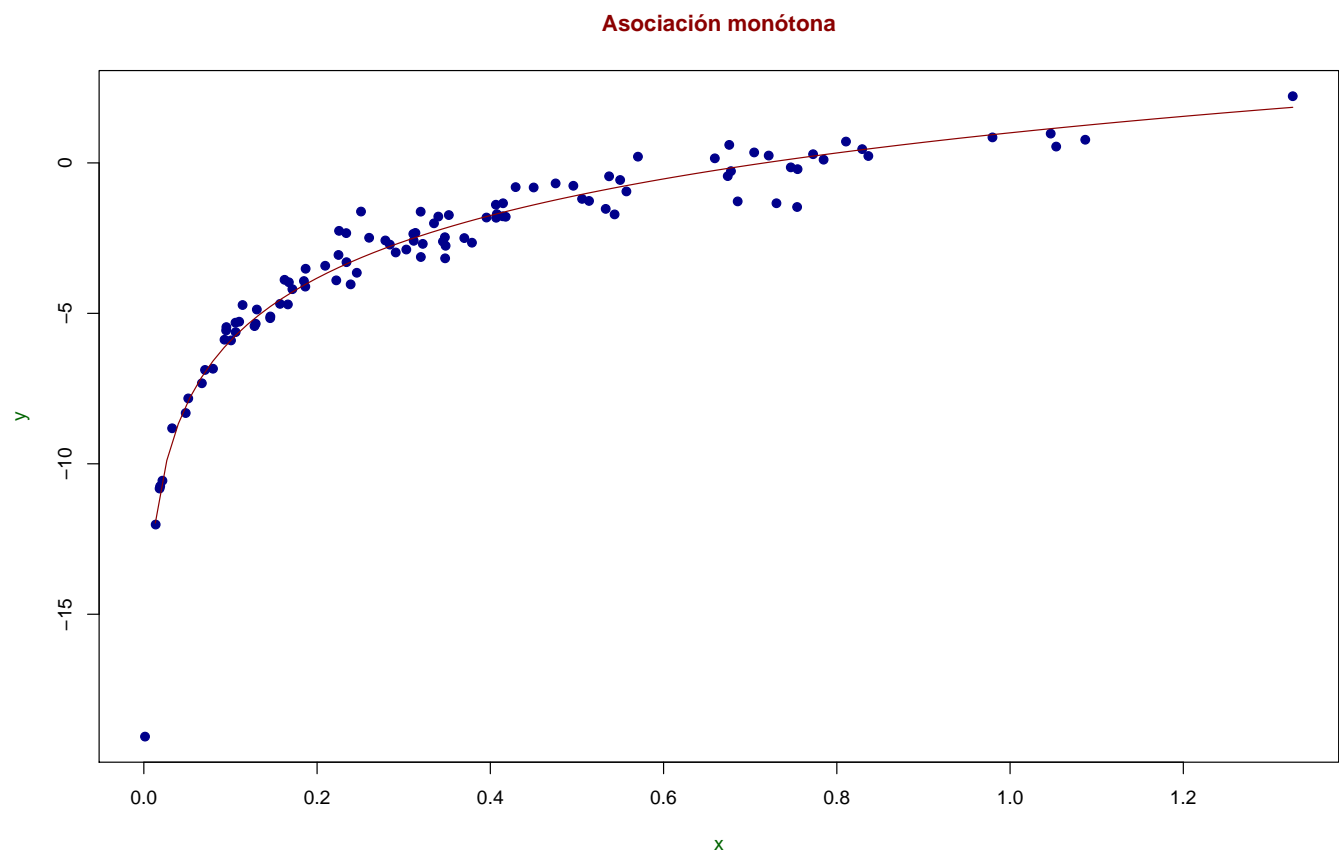
$$T = \frac{\hat{r}_S \sqrt{n-2}}{\sqrt{1-\hat{r}_S^2}} \sim t_{(n-2)}$$

Y rechazamos la hipótesis nula, a un nivel de significancia  $\alpha$ , si

$$|T| > t_{(1-\alpha/2, n-2)}$$

### Muy importante.

Spearman no sólo mide asociación lineal entre las variables, en general mide si existe asociación entre variables *relacionadas de manera monótona*.





## Coeficiente de correlación de Kendall (tau de Kendall)

Supongamos que tenemos un conjunto de parejas de observaciones  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  de dos variables aleatorias,  $\mathbf{X}$  y  $Y$ . Cualquier par de observaciones  $(x_i, y_i)$  y  $(x_j, y_j)$  se dice que son *concordantes* si ocurre que

$$x_i > x_j \Rightarrow y_i > y_j, \quad \text{ó} \quad x_i < x_j \Rightarrow y_i < y_j$$

Y son discordantes si

$$x_i > x_j \Rightarrow y_i < y_j, \quad \text{ó} \quad x_i < x_j \Rightarrow y_i > y_j$$

Si  $x_i = x_j$  ó  $y_i = y_j$  las parejas no son ni concordantes ni discordantes.

El coeficiente de correlación de Kendall es

$$\hat{\tau} = \frac{\# \text{de parejas concordantes} - \# \text{de parejas discordantes}}{\frac{1}{2}n(n-1)}$$

## Prueba de hipótesis

En este caso, también es importante saber si este coeficiente es estadísticamente significativo. Por lo que planteamos el contraste de hipótesis

$$\mathbb{H}_0 : \tau = 0 \quad \text{vs.} \quad \mathbb{H}_a : \tau \neq 0$$

Para muestras grandes, se tiene que

$$Z = \frac{3(n_c - n_d)}{\sqrt{\frac{n(n+1)(2n+5)}{2}}} \stackrel{a}{\sim} N(0, 1)$$

con  $n_c$ : Número de parejas concordantes y  $n_d$ : Número de parejas discordantes. Rechazamos  $\mathbb{H}_0$ , si

$$|T| > t_{(1-\alpha/2, n-2)}$$

## Muy importante

Es muy importante tomar en cuenta que la correlación *no necesariamente indica una relación causal entre las variables involucradas*. Lo que esta medida indica es una relación lineal entre ellas, en el sentido, ya mencionado, de si ambas crecen o decrecen simultáneamente, o si una crece mientras la otra decrece. Para que esta correlación represente una asociación causal, es necesaria la opinión de un *experto del área* de la que provienen los datos, que afirme que esta correlación puede interpretarse en un sentido causal.

## Medidas de asociación en tablas de contingencia

### Tablas de contingencia

Cuando se tienen dos o más variables categóricas, es común desplegar de manera conjunta las relaciones que dan entre ellas, a través de las llamadas *tablas de contingencia* o *tablas de clasificación cruzada*. La distribución conjunta de estas dos variables categóricas determina esta relación. Además de determinar sus distribuciones marginales y condicionales.

Supongamos que tenemos dos variables categóricas  $\mathbf{X}$  y  $\mathbf{Y}$ ,  $\mathbf{X}$  con  $\mathbf{I}$  categorías y  $\mathbf{Y}$  con  $\mathbf{J}$  categorías. Entonces, una clasificación de las respuestas de estas variables tiene  $\mathbf{IJ}$  posibles combinaciones. Las respuestas  $(\mathbf{X}, \mathbf{Y})$  de un sujeto seleccionado aleatoriamente de alguna población tiene una diastribución de probabilidad asociada. Es posible desplegar esta distribución en una tabla rectangular con  $\mathbf{I}$  renglones para las categorías de  $\mathbf{X}$  y  $\mathbf{J}$  columnas para las de  $\mathbf{Y}$ . Las *celdas* de esta tabla representan las  $\mathbf{IJ}$  posibles combinaciones de respuesta. Cuando las celdas contienen frecuencias de conteos muestrales de respuestas, la tabla se conoce como *tabla de contingencia*, término introducido por Karl Pearson (1904). Otra forma de denominarla es *tabla de clasificación cruzada*. Una tabla de contingencia con  $\mathbf{I}$  renglones y  $\mathbf{J}$  columnas se conoce como una tabla  $\mathbf{I} \times \mathbf{J}$ , o como tabla de contingencia de  $\mathbf{I} \times \mathbf{J}$ .

# Distribuciones de probabilidad asociadas a una tabla de contingencia

Consideraremos únicamente una tabla  $2 \times 2$ , porque la generalización para tablas más grandes es inmediata. Para ilustrar estas distribuciones haremos uso de un conjunto de datos sobre la relación entre sexo y ser o no bebedor frecuente. La tabla es la siguiente

	<i>Bebedor Frecuente</i>		
<i>Sexo</i>	SI	NO	Total
Hombres	1630	5550	7180
Mujeres	1684	8232	9916
Total	3314	13782	17096

La primera de nuestras tablas corresponde a la

**Distribución conjunta:** En este caso tenemos dos variables categóricas,  $\mathbf{X}$  (sexo) y  $\mathbf{Y}$  (bebedor frecuente (B.F.)), con  $\mathbf{I} = 2$  y  $\mathbf{J} = 2$  categorías, respectivamente. Su distribución conjunta es la probabilidad de que un sujeto seleccionado aleatoriamente obtenga un valor en el renglón  $i$  y en la columna  $j$  de la tabla. Es decir

$$\pi_{ij} = \mathbb{P}(\mathbf{X} = i, \mathbf{Y} = j), \quad i, j = 1, 2$$

que genera la distribución conjunta dada por la tabla

Distribución conjunta ( $\mathbf{X}, \mathbf{Y}$ )

	<i>B.F.</i>	
<i>Sexo</i>	SI	NO
Hombres	$\pi_{11}$	$\pi_{12}$
Mujeres	$\pi_{21}$	$\pi_{22}$

Sabemos que la forma de estimar estas probabilidades es




$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}, \quad i, j = 1, 2$$

con lo que generamos la tabla

### Distribución conjunta ( $\mathbf{X}, \mathbf{Y}$ )

*B.F.*

<i>Sexo</i>	SI	NO
Hombres	0.095	0.325
Mujeres	0.099	0.482

Las interpretaciones de estas probabilidades son obvias. Por ejemplo, la probabilidad estimada de que una mujer no sea bebedora frecuente es  $\hat{\pi}_{22} = 0.482$ . 

**Distribuciones marginales:** Las distribuciones marginales son las sumas por renglón o columna de las probabilidades conjuntas. En concreto

$$\pi_{i\bullet} = \pi_{i1} + \pi_{i2} = \sum_j \pi_{ij} \quad y \quad \pi_{\bullet j} = \pi_{1j} + \pi_{2j} = \sum_i \pi_{ij}, \quad i, j = 1, 2$$

Con tabla de distribuciones marginales dada por

### Distribuciones marginales ( $\mathbf{X}, \mathbf{Y}$ )

*B.F.*

<i>Sexo</i>	SI	NO	Total
Hombres	$\pi_{11}$	$\pi_{12}$	$\boldsymbol{\pi}_{1\bullet}$
Mujeres	$\pi_{21}$	$\pi_{22}$	$\boldsymbol{\pi}_{2\bullet}$
Total	$\boldsymbol{\pi}_{\bullet 1}$	$\boldsymbol{\pi}_{\bullet 2}$	<b>1</b>

Además, tenemos que las sumas de todas las celdas en la distribución conjunta es igual a la unidad, así como la suma de las probabilidades marginales de las filas o las columnas.

$$\sum_i \pi_{i\bullet} = \sum_j \pi_{\bullet j} = \sum_i \sum_j \pi_{ij} = 1$$

Con nuestros datos, esta tabla queda como

### Distribuciones marginales(**X**,**Y**)

*B.F.*

<i>Sexo</i>	SI	NO	Total
Hombres	0.095	0.325	<b>0.42</b>
Mujeres	0.099	0.482	<b>0.58</b>
Total	<b>0.194</b>	<b>0.806</b>	<b>1</b>

La última de las distribuciones asociada a una tabla de contingencia son las distribuciones condicionales. Definidas como

**Distribuciones condicionales.** Cuando una de las variables es de respuesta (**Y**) y la otra es explicativa (**X**), entonces es conveniente saber cuál es la distribución de probabilidad de la respuesta, **Y**, para cada nivel de la variable explicativa, **X**. Estas distribuciones consisten en probabilidades condicionales de **Y**, dado el nivel de **X**, y se conocen como distribuciones condicionales. Las definimos como

$$\pi_{j|i} = \mathbb{P}(\mathbf{Y} = j | \mathbf{X} = i), \quad i, j = 1, 2$$

El conjunto de probabilidades  $\{\pi_{1|i}, \pi_{2|i}\}$  constituyen la distribución condicional de **Y** en cada categoría  $i$  de **X**. Con tabla de distribuciones condicionales

### Distribuciones condicionales (**X**,**Y**)

*B.F.*

<i>Sexo</i>	SI	NO	Total
Hombres	$\pi_{1 1}$	$\pi_{1 2}$	<b>1</b>
Mujeres	$\pi_{2 1}$	$\pi_{2 2}$	<b>1</b>

Las probabilidades estimadas en esta tabla están dadas por

$$\hat{\pi}_{j|i} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i\bullet}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i\bullet}}{n}} = \frac{n_{ij}}{n_{i\bullet}}, \quad i, j = 1, 2$$

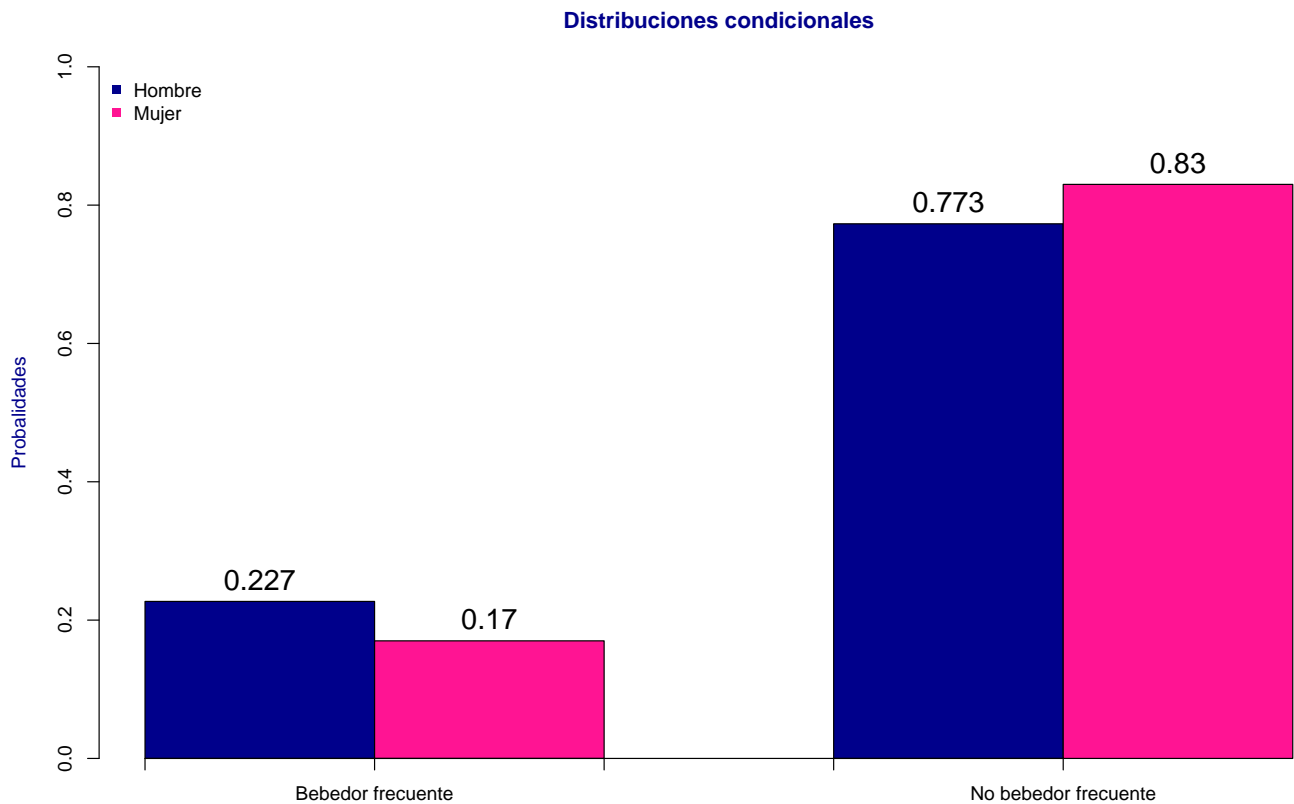
En nuestro caso, esta tabla queda

## Distribuciones condicionales (X,Y)

*B.F.*

<i>Sexo</i>	SI	NO	Total
Hombres	0.2270195	0.7729805	<b>1</b>
Mujeres	0.1698265	0.8301735	<b>1</b>

Es interesante ver que los hombres (condicional hombre) tienen una probabilidad estimada mayor a ser bebedores frecuentes que las mujeres (condicional mujer). Observamos también que las distribuciones condicionales por sexo, no son muy diferentes. No obstante, aún no tenemos elementos para decidir de manera formal si son o no diferentes.



# Medidas de asociación en tablas de contingencia

Ahora definiremos varias medidas de asociación relacionadas con estas tablas de contingencia que hemos presentado.

La pregunta fundamental cuando tenemos dos variables categóricas es si éstas son independientes o no. En nuestro ejemplo, quisiéramos averiguar si el ser bebedor frecuente está relacionado con el sexo. Entonces, lo primero que debemos hacer es verificar si efectivamente estas variables están relacionadas.

## Prueba Ji-cuadrada de independencia

Para realizar esta prueba debemos definir, inicialmente, cuáles son las hipótesis que hay que contrastar para determinar si las variables que conforman nuestra tabla de contingencia son o no independientes.

Recordemos que dentro de la metodología estadística para realizar pruebas de hipótesis, es necesario determinar las llamadas hipótesis nula ( $\mathbb{H}_0$ ) y la hipótesis alternativa ( $\mathbb{H}_a$ ). En este caso de independencia, las enunciamos como

$\mathbb{H}_0$  : Las variables son independientes *vs.*  $\mathbb{H}_a$  : Las variables no son independientes

Ahora debemos *traducir* estas hipótesis a elementos estadísticos para realizar la prueba.

De los cursos básicos de probabilidad, sabemos que si dos variables son independientes, su probabilidad conjunta es el producto de sus probabilidades marginales. En nuestra notación:

$$\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \quad \text{ó} \quad \mathbb{P}(\mathbf{X} = i, \mathbf{Y} = j) = \mathbb{P}(\mathbf{X} = i) \mathbb{P}(\mathbf{Y} = j), \quad i, j = 1, 2$$

Por lo que las hipótesis pueden reescribirse como

$$\begin{aligned} \mathbb{H}_0 : \mathbb{P}(\mathbf{X} = i, \mathbf{Y} = j) &= \mathbb{P}(\mathbf{X} = i) \mathbb{P}(\mathbf{Y} = j) \quad \text{vs.} \\ \mathbb{H}_a : \mathbb{P}(\mathbf{X} = i, \mathbf{Y} = j) &\neq \mathbb{P}(\mathbf{X} = i) \mathbb{P}(\mathbf{Y} = j), \quad p.a. \, i, j = 1, 2 \end{aligned}$$

La estimación de estas probabilidades marginales es



$$\hat{\mathbb{P}}(\mathbf{X} = i) = \frac{n_{i\bullet}}{n}, \quad i = 1, 2$$

$$\hat{\mathbb{P}}(\mathbf{Y} = j) = \frac{n_{\bullet j}}{n}, \quad j = 1, 2$$

La manera de probar la hipótesis de independencia es comparando las frecuencias observadas de la muestra, con las que se esperarían observar si efectivamente estas variables fueran independientes, es decir, comparar lo que realmente observamos con lo que esperamos observar si la hipótesis nula (independencia) es cierta. Esta comparación la realizaremos utilizando la famosa Ji-cuadrada de Pearson. Para esto, necesitamos definir primero los valores esperados de nuestra tabla.


De acuerdo a la definición de valor esperado, para obtener los valores esperados de cada una de las celdas de nuestra tabla de contingencia, debemos multiplicar el número de sujetos en la muestra por la probabilidad de caer en la celda correspondiente, *suponiendo que las variables son independientes*, es decir, bajo la hipótesis nula  $H_0$ . En este caso los valores esperados son:

$$\mathbb{E}_{ij} = n\pi_{i\bullet}\pi_{\bullet j}$$

que podemos estimar (abusando de la notación) como

$$\mathbb{E}_{ij} = n \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}, \quad i, j = 1, 2$$

Con estos valores podemos construir la prueba *Ji-cuadrada de independencia*, comparándolos contra los valores que observamos,  $n_{ij}$ , de la siguiente manera:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - \mathbb{E}_{ij})^2}{\mathbb{E}_{ij}} \stackrel{a}{\sim} \chi_{(1)}^2$$


Presentamos los desarrollos para tablas  $2 \times 2$ , mismos que se pueden extender fácilmente para tablas generales de  $\mathbf{I} \times \mathbf{J}$ . La estadística Ji-cuadrada quedaría en este caso como

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \mathbb{E}_{ij})^2}{\mathbb{E}_{ij}} \stackrel{a}{\sim} \chi_{((I-1)(J-1))}^2$$

Por la forma de la estadística, es obvio que si el valor de la misma es grande, entonces los valores esperados y los observados difieren mucho, lo que implicaría que la hipótesis nula es falsa. De lo contrario, estos valores son similares, lo que implicaría que la hipótesis nula es cierta y las variables son independientes.

## Reglas de uso de la Ji-cuadrada

La distribución Ji-cuadrada es una distribución continua que, en el caso de la prueba Ji-cuadrada de Pearson, aproxima a un proceso discreto, por lo que hay que verificar ciertas reglas de uso para garantizar que esta aproximación sea adecuada en una situación particular.

### Tablas $2 \times 2$

- Si  $n < 20$ , utilizar la prueba exacta de Fisher. Agresti (2002, pag. 91)
- Si  $n \geq 20$ , utilizar la Ji-cuadrada si los valores esperados son mayores o iguales que 5.  
 $\mathbb{E}_{ij} \geq 5$

### Tablas $I \times J$

- Usar la Ji-cuadrada si a lo más el 20 % de las celdas tiene  $\mathbb{E}_{ij} < 5$ , pero ninguna de ellas tiene un valor esperado menor que 1.
- Cuando no se cumpla esta regla, pueden agruparse categorías, siempre y cuando esto tenga sentido.

Una manera de mejorar la aproximación a la Ji-cuadrada en este tipo de tablas, es incluir *la corrección por continuidad de Yates (1934)*. No obstante, ahora es posible, gracias al desarrollo de métodos computacionales más eficientes, calcular la distribución exacta de esta estadística cuando se tienen muestras pequeñas.

## Corrección por continuidad de Yates

Reescribamos la tabla  $2 \times 2$ , de la siguiente manera

Tabla (X,Y)			
	$y_1$	$y_2$	Total
$x_1$	a	b	a+b
$x_2$	c	d	c+d
Total	a+c	b+d	a+b+c+d=n

Puede verificarse que la estadística Ji-cuadrada puede escribirse como

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \stackrel{a}{\sim} \chi_{(1)}^2$$

Yates (1934), argumenta que la distribución  $\chi_{(1)}^2$  proporciona únicamente una aproximación de las probabilidades asociadas a estos datos discretos, y, por lo tanto, los *p-values* basados en la estadística Ji-cuadrada generalmente subestiman los verdaderos p-values. En este contexto, Yates sugiere que la Ji-cuadrada debe ser corregida por continuidad y propone la siguiente corrección

$$\chi^2 = \frac{n \left( |ad - bc| - \frac{n}{2} \right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Para ilustrar el uso de esta estadística, consideremos el ejemplo de la revisión del status de algunas instituciones, por parte de una autoridad reguladora. En concreto, veamos si existe asociación entre la condición de ser una institución de gobierno o no, y tener en regla los documentos. Nuestra tabla es

*Inspección de documentos en instituciones de gobierno y no gobierno*

En regla	Instituciones		
	Gobierno	No gobierno	Total
SI	23	34	57
NO	35	132	224

y queremos probar la hipótesis

$\mathbb{H}_0$  : El tipo de institución y el tener en regla los documentos son independientes vs.

$\mathbb{H}_a$  : El tipo de institución y el tener en regla los documentos no son independientes

Como mencionamos, es necesario construir los valores esperados para aplicar la prueba Ji-cuadrada. Recordemos que, de manera general, se calculan como  $\mathbb{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ ,  $i, j = 1, 2$ .

Mostremos estos cálculos para nuestro ejemplo

$$\begin{aligned}\mathbb{E}_{11} &= \frac{57 * 58}{224} = 14.75893; & \mathbb{E}_{12} &= \frac{57 * 166}{224} = 42.24107 \\ \mathbb{E}_{21} &= \frac{167 * 58}{224} = 43.24107; & \mathbb{E}_{22} &= \frac{167 * 166}{224} = 123.7589\end{aligned}$$

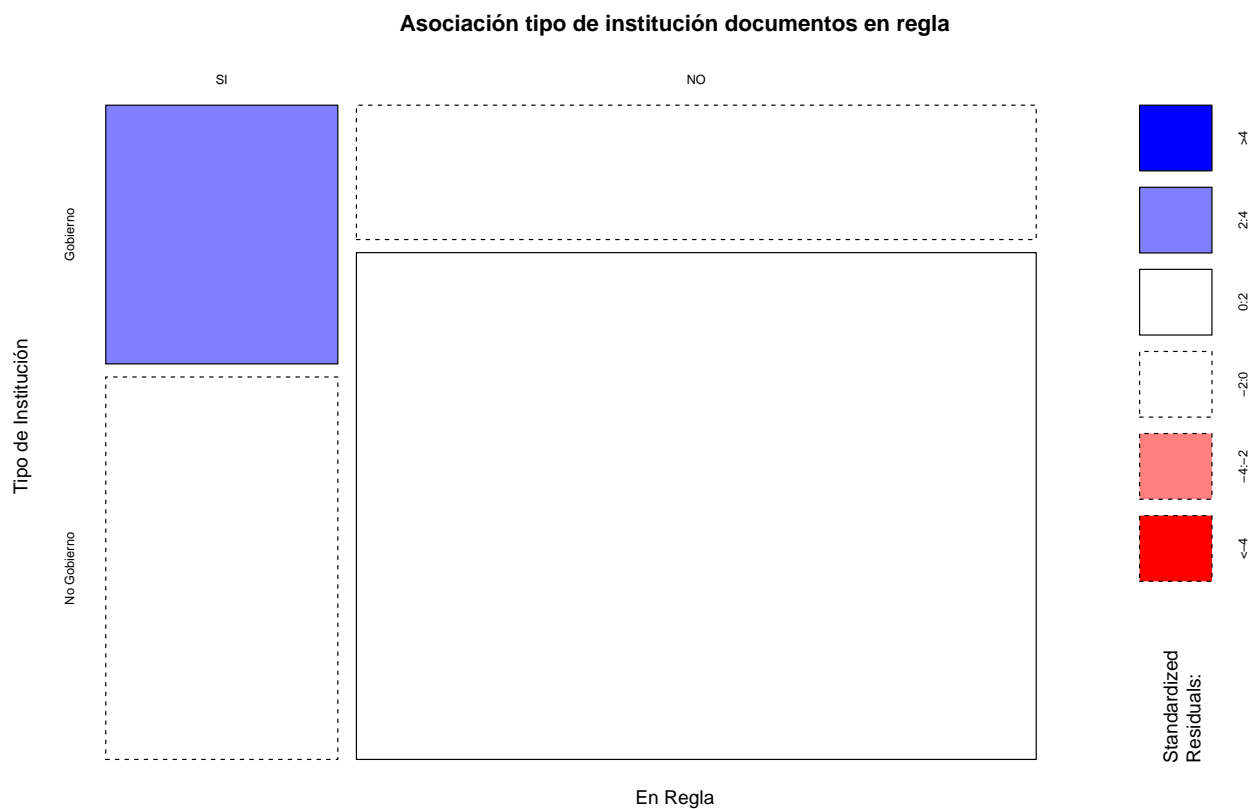
que generan la tabla de valores esperados

Valores Esperados				
<i>Tipo de institución</i>				
<i>En regla</i>	Gobierno	NO Gobierno	Total	
SI	14.75893	42.24107	57	
NO	43.24107	123.7589	167	
Total	58	166	224	

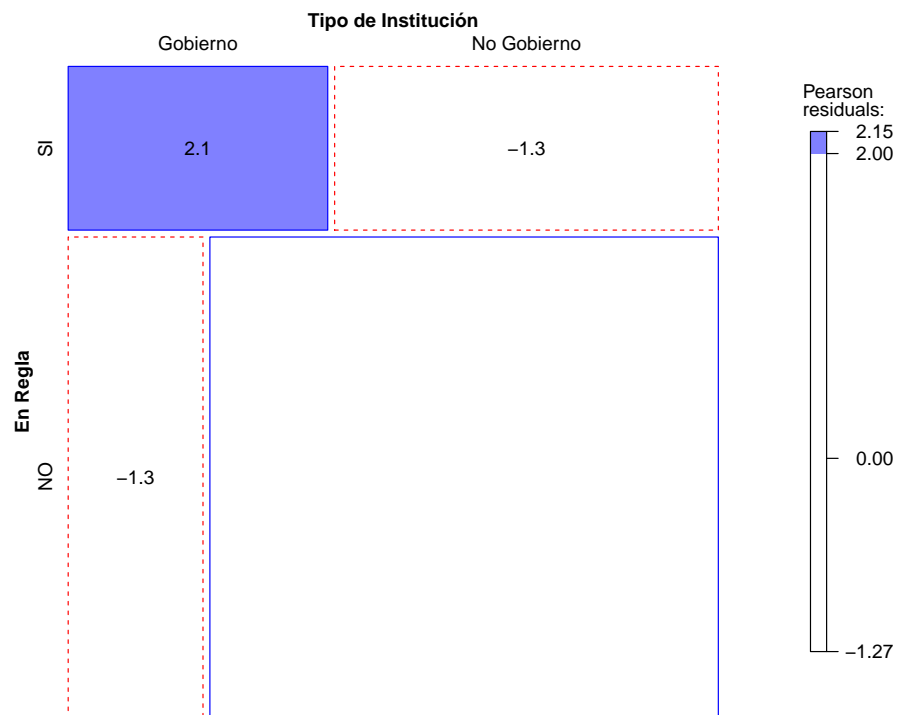
Podemos apreciar que se cumplen de sobra, las condiciones para aplicar la prueba Ji-cuadrada de independencia. Entonces, el cálculo de la misma es

$$\begin{aligned}\chi^2 &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - \mathbb{E}_{ij})^2}{\mathbb{E}_{ij}} = \frac{(23 - 14.75893)^2}{14.75893} + \frac{(34 - 42.24107)^2}{42.24107} + \\ &\quad \frac{(35 - 43.24107)^2}{43.24107} + \frac{(132 - 123.7589)^2}{123.7589} = 7.3488\end{aligned}$$

que proporciona un p-value:  $\mathbb{P}\left(\chi_{(1)}^2 \geq 7.3488\right) = 0.006711$ . Por lo que concluimos que existe asociación entre el tipo de instituciones y estar en regla con los documentos.



## Asociación: Tipo de institución documentos en regla



## Prueba exacta de Fisher

Cuando en una tabla  $2 \times 2$  no se cumplen las condiciones para utilizar la prueba Ji-cuadrada de independencia, tamaño de muestra pequeño o ( $E_{ij} > 5$ ), se puede utilizar la llamada *prueba exacta de Fisher*. La prueba se basa en el hecho de que la probabilidad exacta de observar una tabla con celdas:  $a$ ,  $b$ ,  $c$  y  $d$ , corresponde a una distribución *hipergométrica*. Es importante remarcar que esta prueba asume que los *marginales son fijos*, lo que permite encontrar la distribución de la tabla únicamente a través de una de sus celdas. Nuevamente consideremos la tabla

Tabla (X,Y)			
	$y_1$	$y_2$	Total
$x_1$	a	b	a+b
$x_2$	c	d	c+d
Total	a+c	b+d	a+b+c+d=n

Entonces

$$\mathbb{P}(a, b, c, d) = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!}$$

## Famoso ejemplo de la catadora de té

Ilustraremos el proceso de cómo funciona esta prueba exacta, con el mismo experimento y datos coleccionados por el propio Fisher. El experimento, conocido como *la catadora de té*, consiste en averiguar si una persona puede determinar si en una taza de té se vertió primero la leche y después el té o viceversa. Entonces, se le proporcionó al azar a una catadora *ocho tazas de té*, de las cuales en *cuatro* de ellas se puso primero la leche y en las otras cuatro primero el té. La tabla asociada a este experimento es

Catadora de Té			
	Opinión		
Primero	Leche	Té	Total
Leche	3	1	4
Té	1	3	4
Total	4	4	8

En este caso, las hipótesis a contrastar son

$\mathbb{H}_0$  : No hay asociación entre la opinión de ella y lo que realmente ocurrió: Sólo adivina, vs.

$\mathbb{H}_a$  : Si hay relación: Ella realmente posee esta habilidad para discernir

Entonces, dados los marginales fijos, la probabilidad asociada a estos datos observados se puede encontrar calculando, por medio de una hipergeométrica, la probabilidad asociada a la primera entrada. En concreto

$$\mathbb{P}(a = 3) = \frac{\binom{4}{3} \binom{4}{1}}{\binom{8}{4}} = 0.22857$$



Para determinar si existe o no asociación entre la opinión de la catadora y lo que realmente ocurrió, o si se puede afirmar que ella sólo adivinaba, es necesario encontrar el *p-value* asociado.

El proceso estándar para calcular este p-value es, primeramente, considerar el conjunto de todas las tablas que tienen exactamente los mismos marginales que la tabla observada

*Catadora de Té: Tablas con marginales iguales*

<i>Opinión</i>			
<i>Primero</i>	Leche	Té	Total
Leche	3		4
Té			4
Total	4	4	8

para, posteriormente, calcular este valor como

*p-value* =  $\sum$  Probabilidades de tablas a favor de  $\mathbb{H}_a$ , incluyendo a la de los datos observados

Entonces, observemos que el único valor *más extremo* en la tabla observada, es cuando la primera celda sea  $a = 4$ . En este caso, la probabilidad asociada a este valor es



$$\mathbb{P}(a = 4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.01429$$

de donde el p-value es:  $0.22857+0.01429=0.24286$ , y concluimos que esta mujer sólo está adivinando y no posee ninguna habilidad especial para discernir si se vierte primero la leche o el té. Como dato informativo, el p-value asociado por medio de la Ji-cuadrada es: 0.4795, cercano al doble de la prueba exacta.

## Medidas de asociación

Cuando utilizamos la estadística Ji-cuadrada para probar independencia entre dos variables categóricas, nuestra conclusión es si éstas son o no independientes. En caso de que fueran independientes, afirmamos que no están asociadas o que su grado de asociación es nulo, pero cuando no son independientes, la Ji-cuadrada no proporciona ninguna medida para determinar este grado o fuerza de la asociación, ni tampoco la dirección de la misma. Esto hace necesario la introducción de medidas relacionadas a estas tablas de contingencia, que proporcionen la magnitud de esta asociación y su dirección.

Una forma de medir la asociación entre distintas categorías de estas tablas, es a través de los residuos asociados a la  $\chi^2$ . Ya que esta estadística compara valores esperados *bajo el supuesto de independencia*, con los valores observados, entonces, la diferencia

$$\mathbb{E}_{ij} - n_{ij}$$

debería ser una medida de la falta de independencia, es decir, una medida de *correlación o de asociación*.

Los residuos estandarizados de estas tablas, son

$$e_{ij} = \frac{\mathbb{E}_{ij} - n_{ij}}{\mathbb{E}_{ij}}$$

y los ajustados

$$z_{ij} = \frac{e_{ij}}{\mathbb{E}_{ij} \left(1 - \frac{n_{i\bullet}}{n}\right) \left(1 - \frac{n_{\bullet j}}{n}\right)}$$

Obsérvese que los residuos estandarizados son, de hecho, la raíz cuadrada de cada uno de los elementos de la Ji-cuadrada, y los residuos ajustados, son una modificación de los estandarizados. Esta modificación tiene el objetivo de dar importancia a los elementos de la tabla que pertenecen a renglones o columnas con baja frecuencia. Entonces, su objetivo es que estas diferencias no se vean atenuadas por frecuencias bajas. Además, los residuos ajustados

tienen una distribución aproximadamente normal para muestras grandes, lo que permite determinar cuáles de ellos pueden considerarse grandes.

Con nuestros datos, estas tablas son

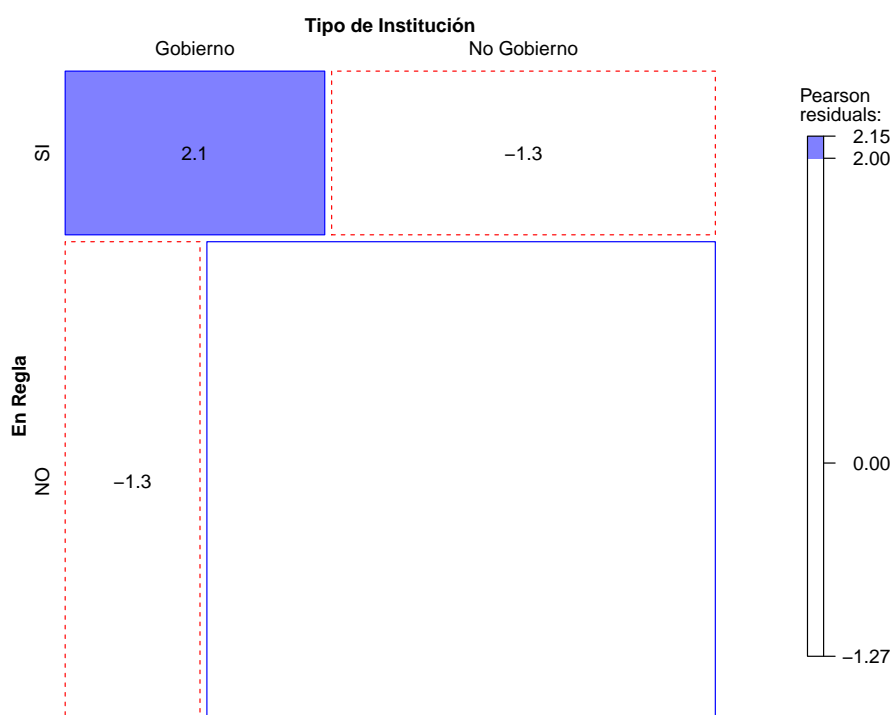
Residuos estandarizados

En regla	<i>Tipo de institución</i>	
	Gobierno	No Gobierno
SI	2.15	-1.27
NO	-1.25	0.74

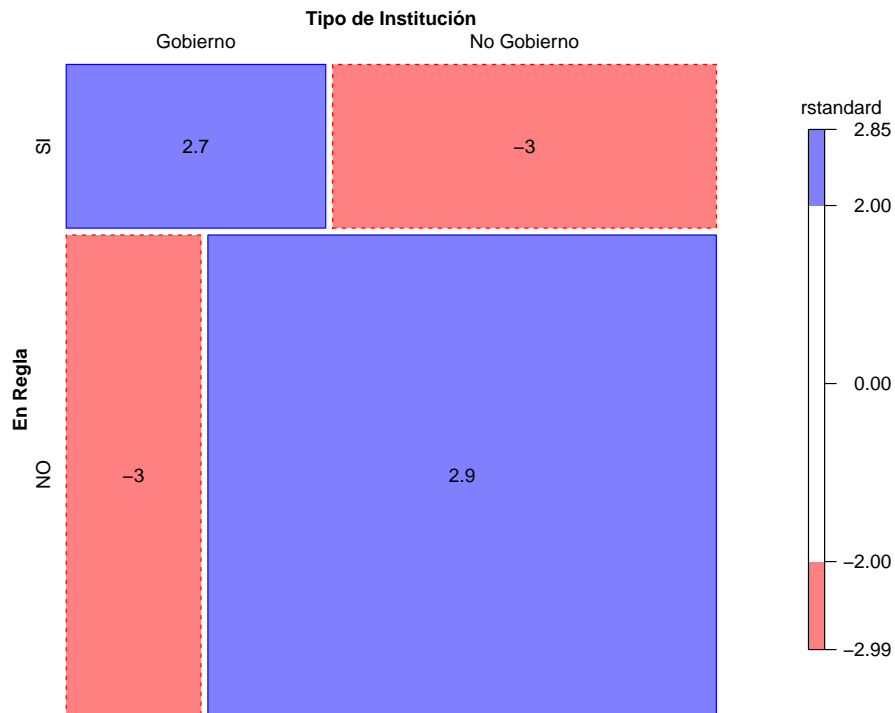
Residuos Ajustados

En regla	<i>Tipo de institución</i>	
	Gobierno	No Gobierno
SI	2.89	-2.89
NO	-2.89	2.89

Asociación: Tipo de institución documentos en regla



### Asociación: Tipo de institución y documentos en regla



Si observamos la tabla de residuos ajustados, podemos ver que, como son normales, los cuatro son estadísticamente significativos al nivel de significancia (0.05,  $Z_{0.975} = 1.96$ ). Y ¿cómo se interpretan estos residuos?

En los casos con residuo positivo tenemos que el valor esperado bajo independencia, es menor que el valor observado, mientras que cuando el residuo es negativo, tenemos valores esperados mayores que los observados. En ambos casos hay una dependencia en cada una de las categorías; entonces, las instituciones gubernamentales están asociados de manera *positiva* con la condición de tener en regla sus documentos y *negativa* con la de no tenerlos, mientras que las instituciones no gubernamentales están asociadas de manera *negativa* con tener en regla sus documentos, pero de manera *positiva* con no tenerlos.

## Fuerza de asociación

Cuando las variables son nominales, existen tres medidas de uso común para medir la fuerza de asociación entre ellas.

### Coefficiente de contingencia

$$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Donde  $\chi^2$  es el valor de la Ji-cuadrada de independencia asociada a la tabla, y  $n$  es el tamaño de la muestra. Inconvenientes

- $0 < CC < 1$
- Valores grandes de este coeficiente dependen del número de sujetos en la tabla:  $n$ .

en nuestro ejemplo

$$CC = \sqrt{\frac{7.3488}{7.3488 + 224}} = 0.18$$

que es una asociación positiva débil. No obstante, obsérvese que un valor moderadamente grande de la muestra (224), tiende a hacer este coeficiente cercano a cero.

### Coefficiente phi ( $\phi$ )

Medida de asociación para variables nominales, de uso exclusivo para tablas  $2 \times 2$ .

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$$

Características

- Depende de las marginales y de las celdas de la tabla.
- $-1 \leq \phi \leq 1$

- $\phi = 1$  sólo si las marginales son iguales.

- $\phi = \sqrt{\frac{\chi^2}{n}}$

con nuestros datos

$$\phi = \frac{23 * 132 - 34 * 35}{\sqrt{57 * 167 * 58 * 166}} = 0.193$$

Mismo comentario sobre la fuerza de asociación entre estas variables.

## La V de Cramér

Este es el último de los coeficientes que mostraremos

$$\mathbf{V} = \sqrt{\frac{\chi^2}{n(k-1)}}$$

con  $k$  el mínimo entre el número de renglones y columnas, i.e.,  $k = \min(I, J)$ .

En tablas  $2 \times 2$ , tenemos que  $k - 1 = 1$  y  $\mathbf{V} = \phi$ . En nuestro caso

$$\mathbf{V} = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{7.3488}{224 * (2-1)}} = 0.181$$

Las tres medidas nos reportan una fuerza de asociación pequeña entre las variables.

En **R** podemos obtener todas estas medidas de la siguiente forma

```
library(vcd)
```

```
assocstats(DE)
```

	$X^2$	df	$(P > X^2)$
--	-------	----	-------------

Likelihood Ratio	7.8676	1	0.0050327
------------------	--------	---	-----------

Pearson	8.3288	1	0.0039021
---------	--------	---	-----------

Phi-Coefficient	: 0.193
-----------------	---------

Contingency Coeff.	: 0.189
--------------------	---------

Cramer's V	: 0.193
------------	---------

Nuestras medidas no coinciden, porque nosotros calculamos la Ji-cuadrada de esta tabla con corrección de Yates, mientras que este proceso automático lo hace sin correccion. La Ji-cuadrada sin corrección es *8.3288*, y las medidas asociadas son:

$$\mathbf{CC} = 0.189 \quad \phi = 0.193 \quad \mathbf{V} = 0.193$$

que son las medidas de la tabla anterior.

# ANÁLISIS DE REGRESIÓN

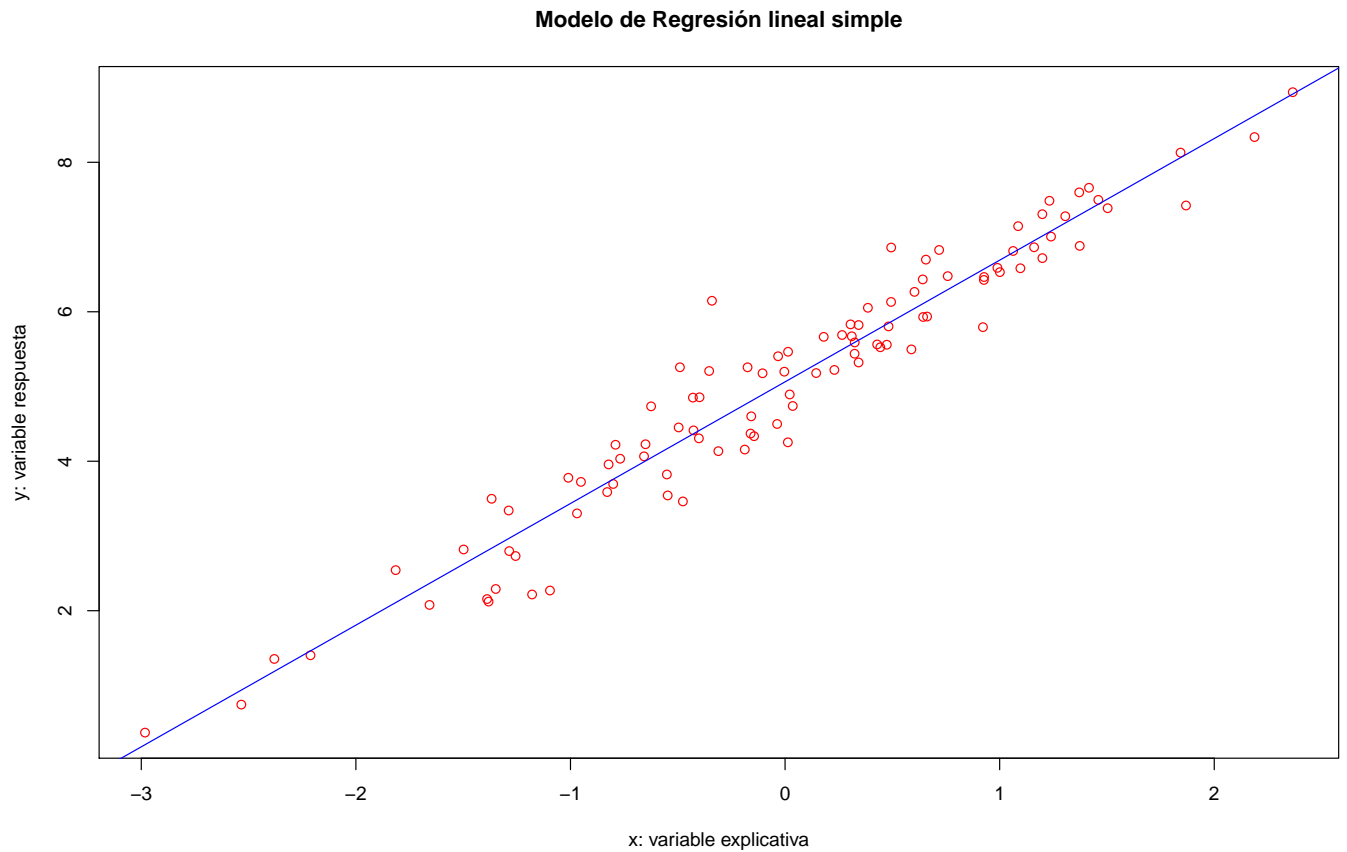
## INTRODUCCIÓN

### Un poco de historia

Los primeros problemas prácticos tipo regresión iniciaron en el siglo XVIII, relacionados con la navegación basada en la Astronomía. Legendre desarrolló el método de *mínimos cuadrados* en 1805. Gauss afirma que él desarrolló este método algunos años antes y demuestra, en 1809, que mínimos cuadrados proporciona una solución óptima cuando los errores se distribuyen normal. Francis Galton acuña el término regresión al utilizar el modelo para explicar el fenómeno de que los hijos de padres altos, tienden a ser altos en su generación, pero no tan altos como lo fueron sus padres en la propia, por lo que hay un efecto de *regresión*. El modelo de regresión lineal es, probablemente, el modelo de su tipo **más conocido** en estadística.

El modelo de regresión se usa para explicar o modelar la relación entre una sola variable  $y$ , llamada *dependiente* o *respuesta*, y una o más variables *predictoras*, *independientes*, *covariantes*, o *explicativas*,  $x_1, x_2, \dots, x_p$ . Si  $p = 1$ , se trata de un modelo de regresión *simple* y si  $p > 1$ , de un modelo de regresión *múltiple*. En este modelo se asume que la variable de respuesta,  $y$ , es **aleatoria** y las variables explicativas son **fijas**, es decir, **no aleatorias**.





La variable de respuesta debe ser continua, pero los regresores pueden tener cualquier escala de medición (continua, discreta o categórica).

## OBJETIVOS DEL ANÁLISIS DE REGRESIÓN

Existen varios objetivos dentro del análisis de regresión, entre otros:

- Determinar el efecto, o relación, entre las variables explicativas y la respuesta.
- Predicción de una observación futura
- Describir de manera general la estructura de los datos

# MODELO DE REGRESIÓN LINEAL SIMPLE

Para este modelo supondremos que nuestra respuesta  $y$  es explicada únicamente por una covariable  $x$ . Entonces, escribimos nuestro modelo como:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2, \dots, n$$

Como podemos observar, se ha propuesto una relación lineal entre la respuesta ( $y$ ) y la variable explicativa ( $x$ ), que es nuestro primer supuesto sobre el modelo: *La relación funcional entre  $x$  y  $y$  es una línea recta.*

Observamos que la relación no es perfecta, ya que se agrega el término de error  $\epsilon$ . Dado que la parte aleatoria del modelo es la respuesta  $y$ , asumimos que al error se le “carga” los errores de medición de esta respuesta, así como las perturbaciones que le pudieran ocasionar los términos omitidos en el modelo. Gauss desarrolló este modelo a partir de la teoría de errores de medición, que es de donde se desprenden los supuestos sobre este término:

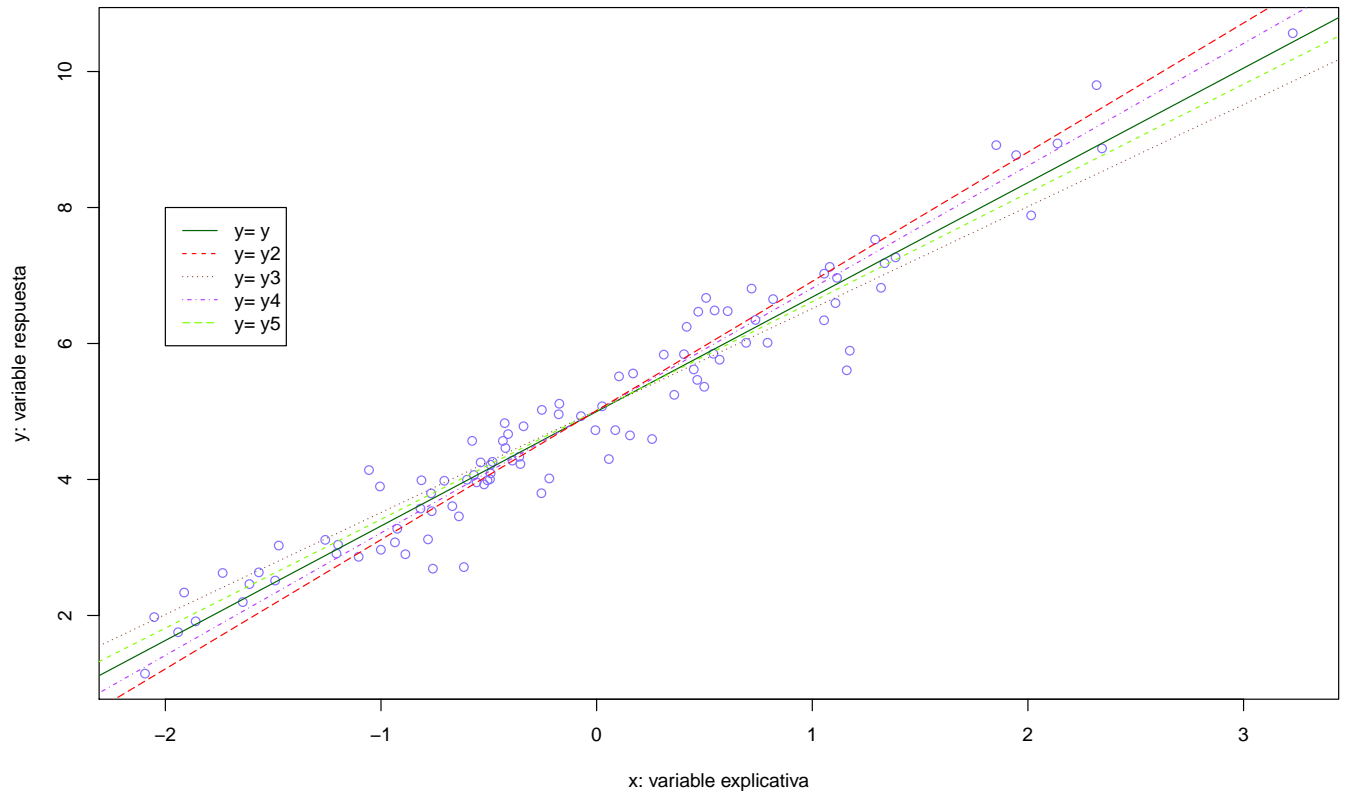
- La esperanza de los errores es cero.  $\mathbb{E}(\epsilon_i) = 0$
- La varianza de los errores es constante.  $\text{Var}(\epsilon_i) = \sigma^2$
- Los errores no están correlacionados.  $\text{Cov}(\epsilon_i, \epsilon_j) = 0, \quad \forall i \neq j$

Por cierto, los errores  $\epsilon_i$  son variables aleatorias no observables.

## Mínimos Cuadrados

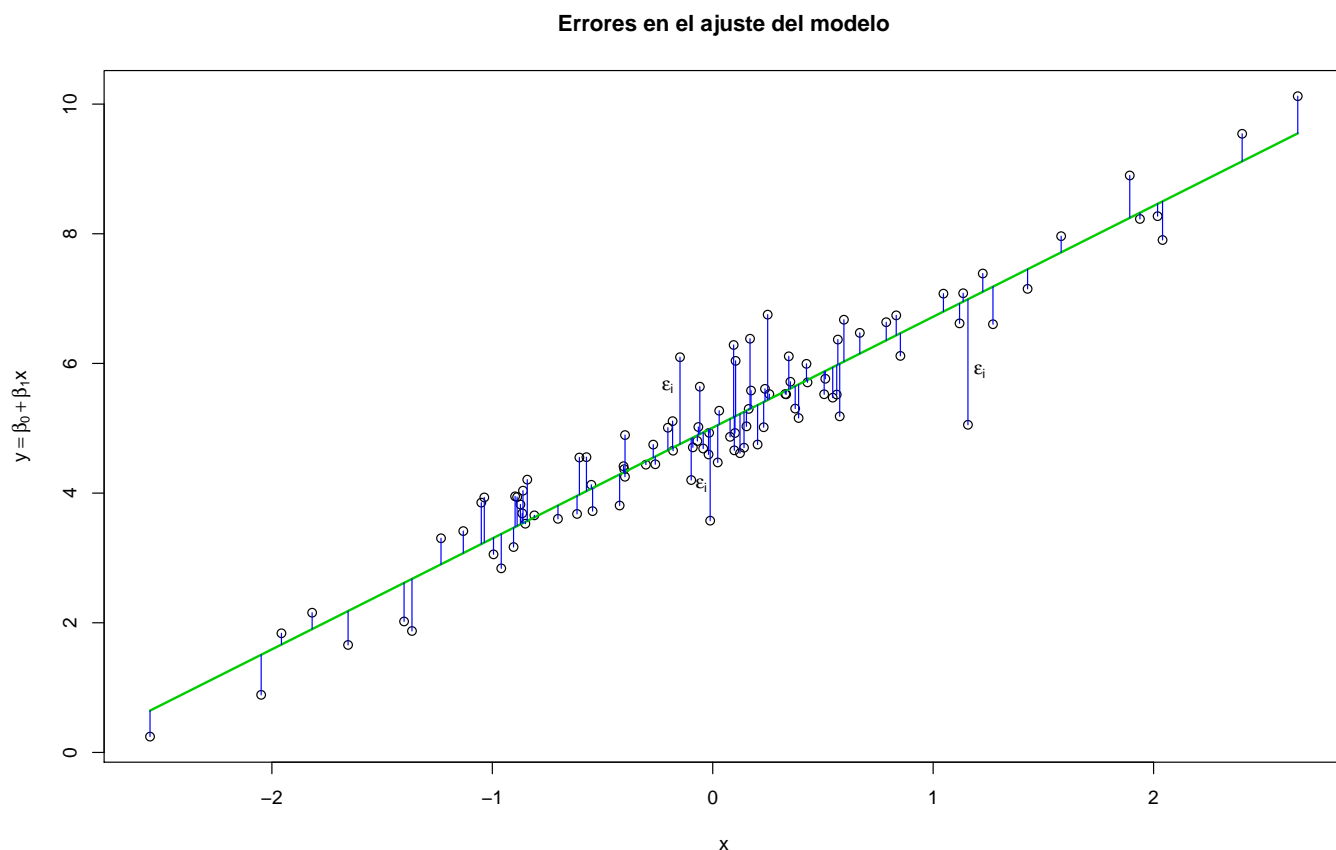
En una situación real, tenemos  $n$  observaciones de nuestra respuesta y de la variable explicativa, que conforman las parejas  $(x_i, y_i)$   $i = 1, 2, \dots, n$ . Entonces, nuestro objetivo será encontrar la recta que mejor ajuste a los datos observados.

### Varios modelos que pueden ajustar



Utilizaremos el método de mínimos cuadrados, que consiste en minimizar la suma de los errores al cuadrado. Concretamente:

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$



Utilizaremos procesos estándar de cálculo diferencial para encontrar los valores de  $\beta_0$  y  $\beta_1$  que minimizan esta función, mismos que serán nuestros estimadores. Las expresiones para estos estimadores son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad y$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

y la recta estimada es

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Una de las desventajas que tiene el método de mínimos cuadrados, es que no se pueden hacer procesos de inferencia sobre los parámetros de interés  $\beta_0$  y  $\beta_1$ ; procesos como intervalos de confianza o pruebas de hipótesis. Para subsanar esta deficiencia, es necesario asumir una

distribución para el error  $\epsilon_i$ . Que, siguiendo la teoría general de errores, se asume que tiene distribución normal, con media cero y varianza  $\sigma^2$ .

$$\epsilon_i \sim N(0, \sigma^2)$$

Este supuesto sobre el error garantiza que las distribuciones de  $y_i$ ,  $\hat{\beta}_0$  y  $\hat{\beta}_1$  sean normales, lo que permite tanto la construcción de intervalos de confianza como las pruebas de hipótesis.

El proceso de inferencia para los parámetros de este modelo, requiere que la varianza sea estimada. El estimador de  $\sigma^2$  es:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

## Prueba de hipótesis fundamental

En el modelo de regresión lineal simple, la prueba de hipótesis más importante es determinar si estadísticamente existe la dependencia entre  $x$  y  $y$ , y no sólo puede considerarse como producto del muestreo (debido al azar). Es decir, realizar la prueba de hipótesis:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_a : \beta_1 \neq 0$$

No rechazar la hipótesis nula, implicaría que la variable  $x$  no ayuda a explicar a  $y$  o bien que, tal vez, la relación entre estas variables **NO ES LINEAL**. En este modelo, esta última explicación es un poco cuestionable, ya que se parte, de inicio, del diagrama de dispersión de los datos.

Si rechazamos la hipótesis nula, implicará que  $x$  es importante para explicar la respuesta  $y$  y que la relación lineal entre ellas puede ser adecuada.

## Interpretación de los parámetros

Cuando se tiene una recta en el sentido determinista, los parámetros  $\beta_0$  y  $\beta_1$  tienen una interpretación muy clara;  $\beta_0$  se interpreta como el valor de  $y$  cuando  $x$  es igual a cero y  $\beta_1$  como el cambio que experimenta la variable de respuesta  $y$  por unidad de cambio en  $x$ . La

interpretación, desde el punto de vista estadístico, de los parámetros estimados en el modelo de regresión es muy similar.  $\hat{\beta}_0$  es el promedio promedio o cambio esperado de la respuesta cuando  $x = 0$  (este parámetro tendrá una interpretación dentro del modelo, si tiene sentido que  $x$  tome el valor cero, de lo contrario, no tiene una interpretación razonable) y  $\hat{\beta}_1$  es el cambio promedio o cambio esperado en  $y$  por unidad de cambio en  $x$ .

# El modelo de regresión lineal múltiple

La mayoría de los fenómenos reales son multicausales, por esta razón, un modelo de regresión más acorde a estudios reales es el modelo de regresión lineal múltiple, que es la generalización del modelo simple. En este modelo supondremos que la variable de respuesta,  $y$ , puede explicarse a través de una colección de  $k$  covariables  $x_1, \dots, x_k$ .

El modelo se escribe de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} \quad i = 1, 2, \dots, n$$

En notación matricial

$$\underline{y} = \mathbf{X}\underline{\beta} + \underline{\epsilon}$$

con

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Se asume que las  $y$ 's están medidas en escala continua.



Al igual que en el caso simple, los parámetros del modelo se pueden estimar por mínimos cuadrados, con el inconveniente de que no se pueden realizar inferencias sobre ellos. Nuevamente, para poder hacer intervalos de confianza y pruebas de hipótesis sobre los verdaderos parámetros hay que suponer que el vector de errores  $\underline{\epsilon}$  se distribuye normal, en este caso multivariada, es decir:

$$\underline{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$$

Esta estructura del error permite tener las mismas propiedades distribucionales que en regresión simple, es decir,  $y_i$  se distribuye normal y cada  $\hat{\beta}_i$   $i=0,1,\dots,k$  también tiene distribución normal, facilitando las inferencias sobre cada parámetro y la construcción de intervalos de

predicción para las  $y$ 's.

Las expresiones para estimar los parámetros involucrados en el modelo son:

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\underline{y} \quad (\text{Ecuaciones Normales}) \text{ y}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p}$$

donde  $p = k + 1$  es el número total de parámetros en el modelo.

Tanto en el modelo simple como en el múltiple, la variación total de las  $y$ 's se puede descomponer en una parte que explica el modelo, i.e., los  $k$  regresores o variables explicativas y otra no explicada por estas variables, llamada error.

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{Total}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{Regresión}} + \underbrace{\sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\text{Error}}$$

Esta descomposición ayuda para realizar la importante prueba de hipótesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad \text{vs.} \quad H_a : \beta_i \neq 0 \quad \text{p.a. } i = 1, 2, \dots, k$$

misma que se realiza a través del cociente entre los errores cuadráticos medios

$$F_0 = \frac{SS_R/k}{SS_E/(n - k - 1)} = \frac{MS_R}{MS_E} \sim F_{(k, n-k-1)}$$

Esta estadística se desprende de la tabla de *análisis de varianza*, que es muy similar a la tabla *ANOVA* que se utiliza para hacer pruebas de hipótesis. En este caso, la tabla es

Fuente variación	Grados libertad (g.l.)	Suma cuadrados	Cuadrados medios	F
Regresión	k	$SS_R$	$MS_R = SS_R/k$	
Error	n-k-1	$SS_E$	$MS_E = SS_E/(n-k-1)$	$F = \frac{MS_R}{MS_E}$
Total	n-1	$S_{yy}$		



Por lo general, esta estadística rechaza la hipótesis nula, ya que de lo contrario, implicaría que **NINGUNA DE LAS VARIABLES EXPLICATIVAS CONTRIBUYE A EXPLICAR LA RESPUESTA**,  $y$ . Como se puede observar en la hipótesis alternativa, el rechazar  $H_0$  sólo implica que al menos uno de los regresores contribuye significativamente a explicar a  $y$ , pero no implica que **TODOS** contribuyan ni tampoco dice cuál o cuáles contribuyen, por esta razón, una salida estándar de regresión múltiple tiene pruebas individuales sobre la significancia de cada regresor en el modelo. El estadístico para hacer tanto los contrastes de hipótesis como los intervalos de confianza individuales, es

$$t = \frac{\hat{\beta}_i - \beta_{0i}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_i)}} \sim t_{(n-p)}$$

Podemos apreciar que los contrastes de hipótesis se pueden hacer contra cualquier valor particular del parámetro  $\beta_{0i}$ , en general. No obstante, en las pruebas estándar sobre los parámetros de un modelo, este valor particular es 0, ya que se intenta determinar si la variable asociada al  $i$ -ésimo parámetro es *estadísticamente significativa* para explicar la respuesta. Por lo que el estadístico para este caso es:

$$t = \frac{\hat{\beta}_i - 0}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_i)}} = \frac{\hat{\beta}_i}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_i)}} \sim t_{(n-p)}$$

De este estadístico se desprenden también los intervalos de confianza para cada parámetro

$$\hat{\beta}_i - t_{(n-p, 1-\alpha/2)} \sqrt{\widehat{\text{Var}}(\hat{\beta}_i)} \leq \beta_i \leq \hat{\beta}_i + t_{(n-p, 1-\alpha/2)} \sqrt{\widehat{\text{Var}}(\hat{\beta}_i)} \quad i = 0, 1, \dots, k$$

## Interpretación de los parámetros

La interpretación de cada parámetro es similar a la del coeficiente de regresión  $\hat{\beta}_1$  en el modelo simple, anexando la frase: “manteniendo constantes el resto de las variables”. Esto es,  $\hat{\beta}_i$  es el cambio promedio o cambio esperado en  $y$  por unidad de cambio en  $x_i$ , sin considerar cambio alguno en ninguna de las otras variables dentro del modelo, es decir, suponiendo que estas otras variables permanecen fijas. Esta interpretación es similar a la que se hace de la derivada parcial en un modelo determinista. Nuevamente, la interpretación de  $\hat{\beta}_0$  estará sujeta a la posibilidad de que, en este caso, **TODAS** las variables puedan tomar el valor **CERO**.

# PREDICCIÓN

Uno de los usos más frecuentes del modelo de regresión es el de predecir un valor de la respuesta para un valor particular de las covariables en el modelo. Si la predicción se realiza para un valor de las covariables *dentro* del rango de observación de las mismas, se tratará de una *interpolación*, y si se realiza para un valor *fuera* de este rango, hablaremos de una *extrapolación*. En cualquiera de los dos casos, estaremos interesados en dos tipos de predicciones

- Predicción de la respuesta media:  $y_0 = \mathbb{E}(y|\mathbf{X}_0)$
- Predicción de una nueva observación:  $y_0$

En ambos casos, la estimación puntual es la misma

$$\hat{y}_0 = \mathbf{X}_0' \hat{\beta}$$

lo que difiere es el intervalo de predicción. Para la respuesta media es

$$\hat{y}_0 - t_{(n-p, 1-\alpha/2)} \sqrt{\hat{\sigma}^2 \mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0} \leq y_0 \leq \hat{y}_0 + t_{(n-p, 1-\alpha/2)} \sqrt{\hat{\sigma}^2 \mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0}$$

y para predecir una observación

$$\hat{y}_0 - t_{(n-p, 1-\alpha/2)} \sqrt{\hat{\sigma}^2 \left(1 + \mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0\right)} \leq y_0 \leq \hat{y}_0 + t_{(n-p, 1-\alpha/2)} \sqrt{\hat{\sigma}^2 \left(1 + \mathbf{X}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_0\right)}$$

¿Se ve la “diferencia”?

# Evaluación del ajuste del modelo

## Coefficiente de determinación


Un primer elemento de juicio sobre el modelo de regresión lo constituye el coeficiente de determinación  $R^2$ , que es la proporción de variabilidad de las  $y$ 's que es explicada por las  $x$ 's y que se escribe como:

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}}$$

Una  $R^2$  cercana a **UNO** implicaría que mucha de la variabilidad de la respuesta es explicada por el conjunto de regresores incluidos en el modelo ( $SS_R \approx S_{yy}$ ).

## Evaluación de los supuestos

Los dos modelos de regresión presentados, el simple y el múltiple, se construyeron sobre los supuestos de:

- La relación funcional entre la variable de respuesta  $y$  y cada regresor  $x_j$  es lineal  $j=1,2,\dots,k$ .
- La esperanza de los errores es cero.  $\mathbb{E}(\epsilon_i) = 0$  
- La varianza de los errores es constante.  $\text{Var}(\epsilon_i) = \sigma^2$
- Los errores no están correlacionados.  $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad i \neq j$
- Los errores tienen distribución normal con media cero y varianza  $\sigma^2$ .

Entonces, para garantizar que el modelo es adecuado, es indispensable verificar estos supuestos.

## Residuos

Los elementos más importantes para verificar estos supuestos son los residuos, definidos como:

$$e_i = y_i - \hat{y}_i$$

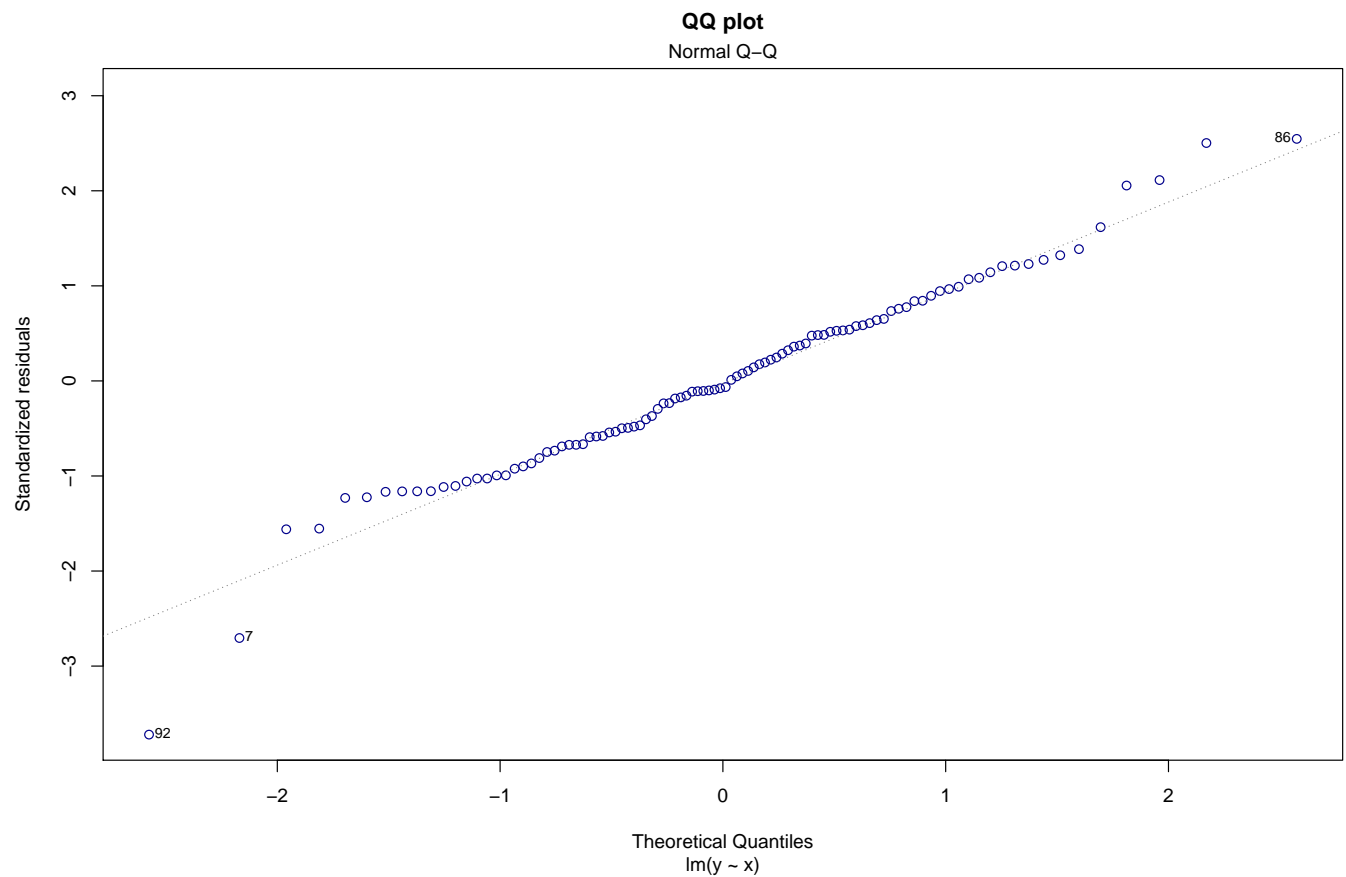
Dado que los errores verdaderos  $\epsilon_i$  no son observables, podemos considerar a los residuos como “la realización” de estos errores; así que cualquier propiedad asumida sobre los errores verdaderos, deberá verificarse a través de los residuos. En la literatura de regresión lineal existen cuatro tipos de residuos, a saber:

- **Residuo crudo  $e_i$ .** Tiene la desventaja de que depende de la escala de la respuesta. No es fácil determinar qué tan grande es grande
- **Residuo estandarizado.** Es el residuo crudo dividido por la raíz cuadrada del estimador de la varianza
- **Residuo estudentizado interno.** Es el residuo crudo dividido entre la raíz cuadrada de su varianza
- **Residuo estudentizado externo.** Es el residuo crudo dividido entre la raíz cuadrada de su varianza calculada sin tomar en cuenta al propio individuo.

Estos residuos se utilizan en los distintos procedimientos para evaluar los supuestos y lo adecuado del ajuste del modelo. La mayoría de las pruebas conocidas para la verificación de los supuestos, son pruebas gráficas.

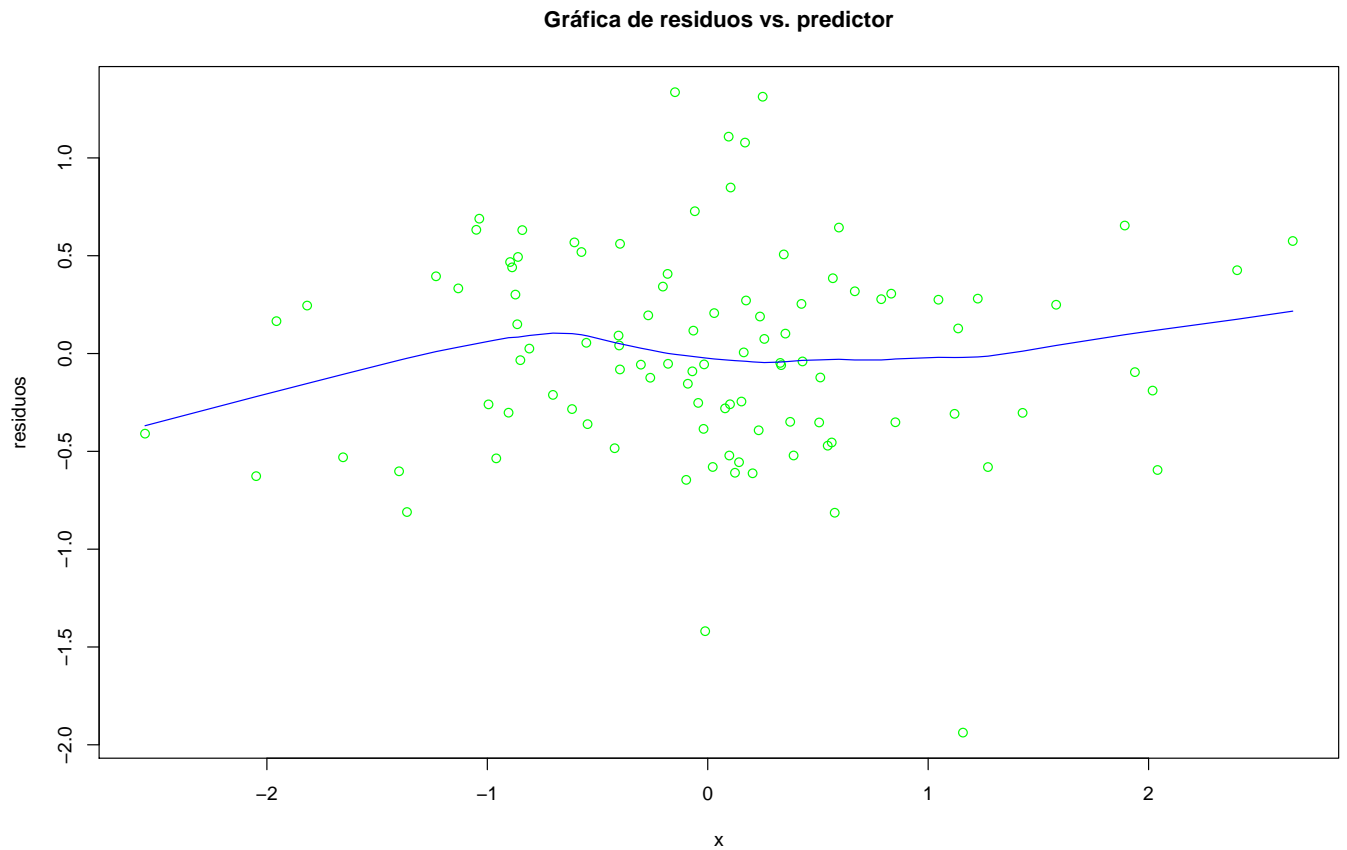
## Normalidad de los residuos

Indudablemente, la prueba más importante es sobre la normalidad de los errores, misma que se juzga a través de la gráfica conocida como *QQ plot* o *QQ Norm*, que grafica los cuantiles teóricos de una normal (eje x) vs. los cuantiles asociados a los residuos. Entonces, si los residuos realmente provienen de una normal, la gráfica debe mostrar una línea a  $45^\circ$  que pasa por el origen de coordenadas. Fuertes desviaciones de esta línea darían evidencia de que los errores no se distribuyen normal.



## Linealidad de los predictores

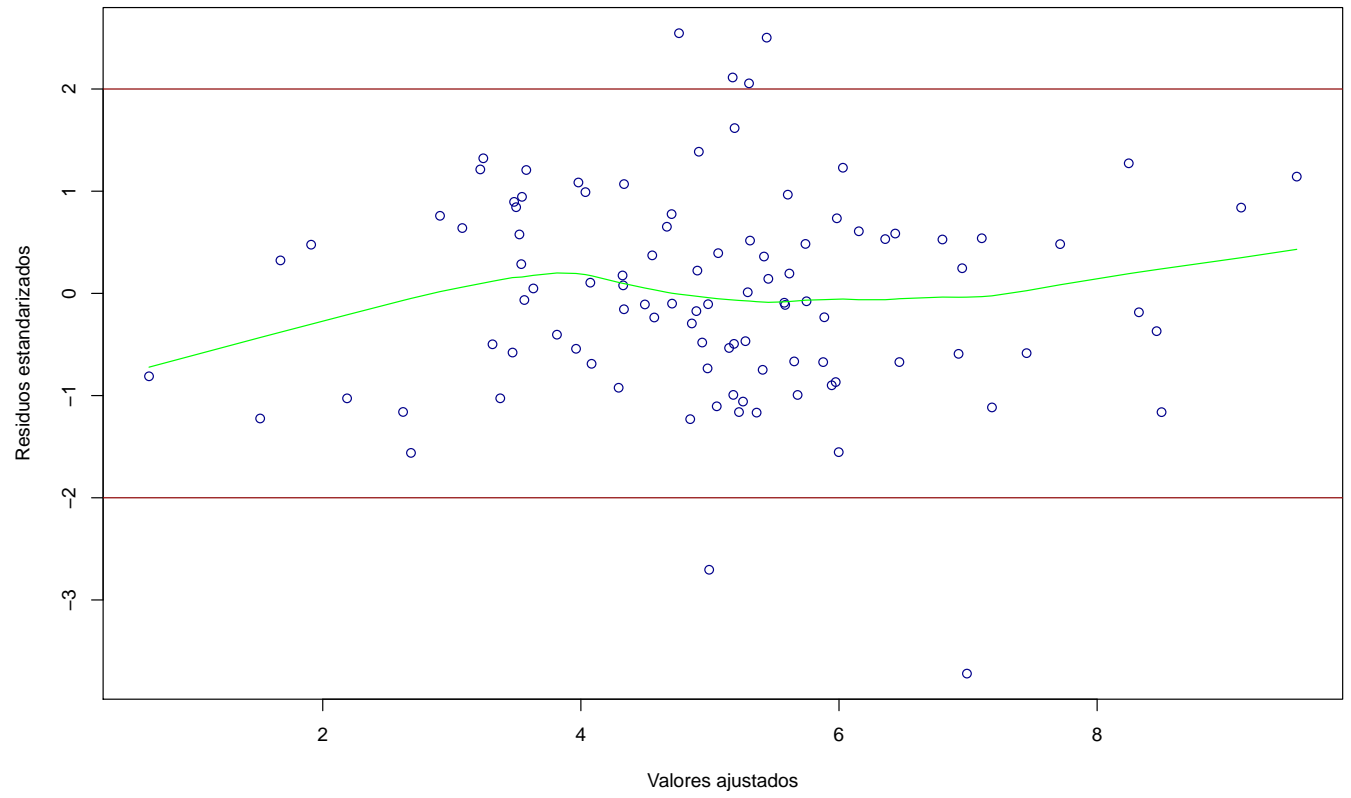
La manera estándar de evaluar la linealidad de las variables explicativas es a través de la gráfica de cada una de ellas contra los residuos. Si la variable en cuestión ingresa al modelo de manera lineal, esta gráfica debe mostrar un patrón totalmente aleatorio entre los puntos dispuestos en ella. Cuando la variable explicativa es politómica, este tipo de gráficas son poco ilustrativas en este sentido.



## Supuestos sobre los errores

Si la gráfica entre los valores ajustados y los residuos estandarizados, muestra un patrón aleatorio, es simétrica alrededor del cero y los puntos están comprendidos entre los valores -2 y 2, entonces se tendrá evidencia que los errores tienen media cero, varianza constante y no están correlacionados.

Gráfica de valores ajustados vs. residuos estandarizados



## Diagnóstico del modelo

Los métodos mostrados hasta ahora, permiten evaluar el modelo de manera global y no por cada observación dentro del mismo. Dado que una observación puede resultar determinante sobre alguna(s) característica(s) del modelo, es conveniente verificar el impacto que cada observación pueda tener en los distintos aspectos del modelo. Las estadísticas para evaluar el impacto que tiene una observación sobre *todo el vector de parámetros, alguno de los regresores y sobre los valores predichos*, se basan en la misma idea, que consiste en cuantificar el cambio en la característica de interés con y sin la observación que se está evaluando. Antes de presentar las estadísticas que servirán para hacer este diagnóstico, introduciremos un elemento que es común a ellas, la llamada *palanca* (leverage) de una observación. Recordemos que el ajuste del modelo se expresaba como:

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\underline{y} \Rightarrow \underline{\hat{y}} = \mathbf{X}\underline{\hat{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\underline{y} = \mathbf{H}\underline{y}$$

con  $\mathbf{H}$  conocida como la *matriz sombrero*, ya que

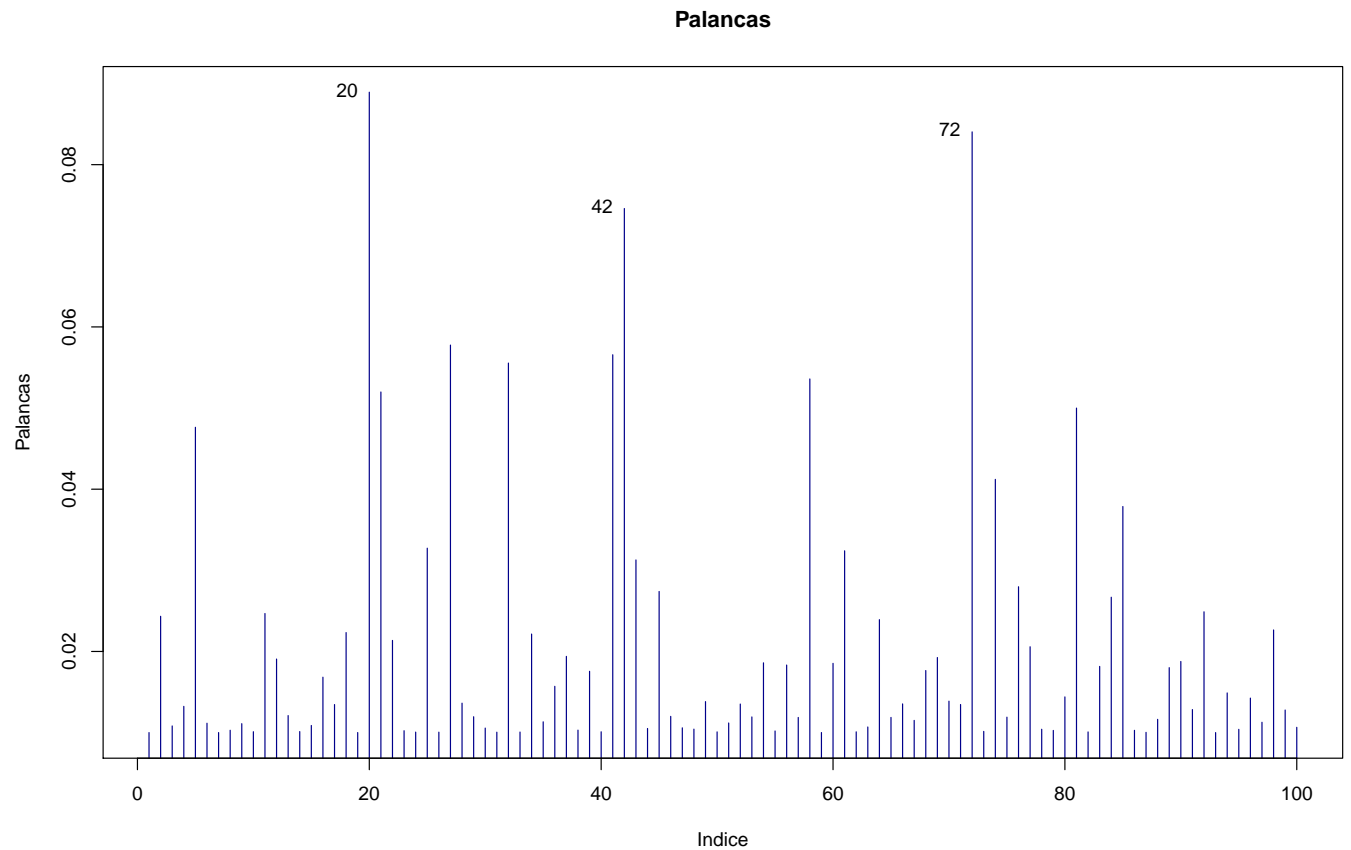
$$\underline{\hat{y}} = \mathbf{H}\underline{y}$$

Un resultado fundamental sobre esta matriz sombrero es

$$\mathbb{V}ar(\underline{\mathbf{e}}) = (\mathbf{I} - \mathbf{H})\sigma^2 \Rightarrow \mathbb{V}ar(e_i) = (1 - h_i)\sigma^2$$

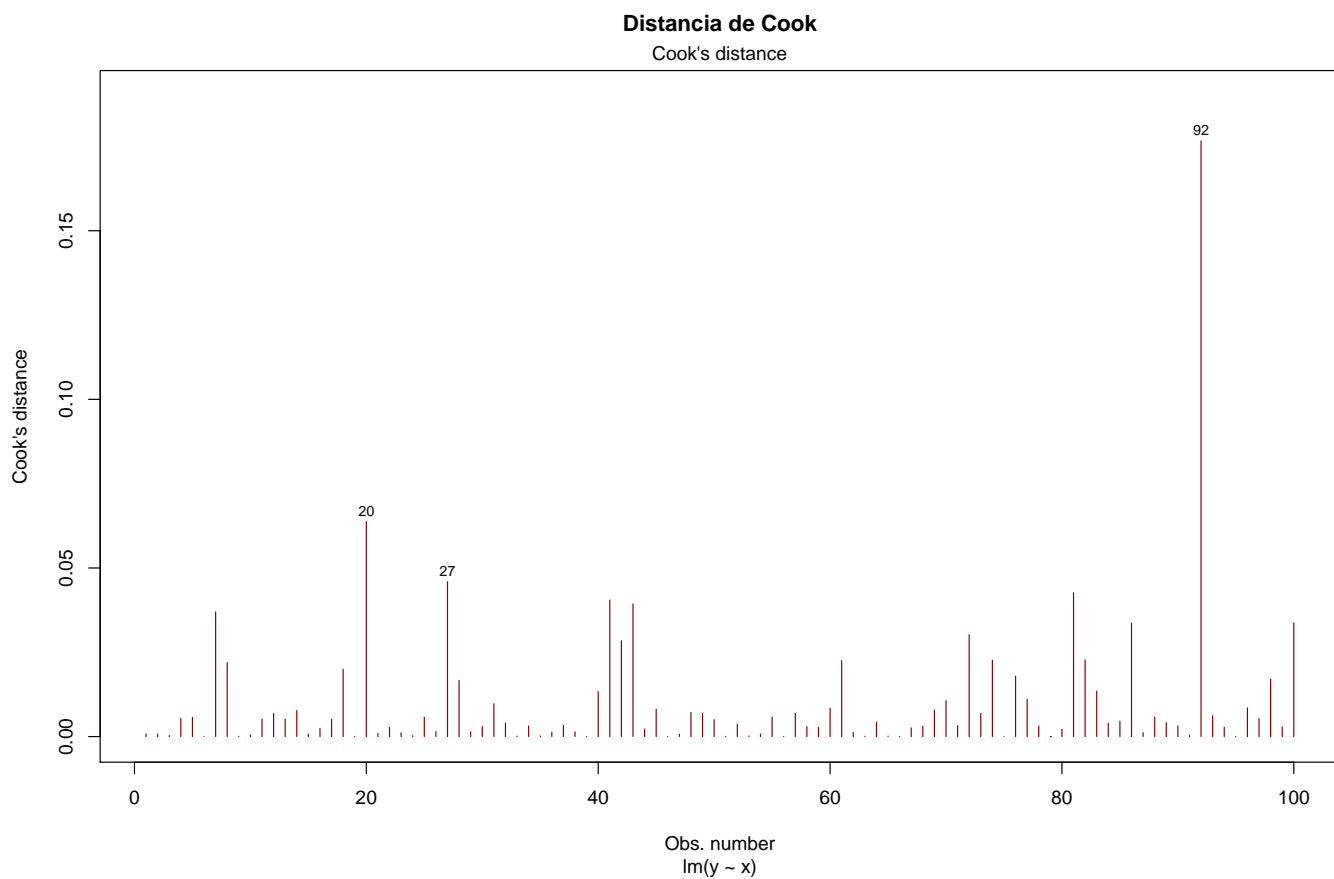
con  $h_i$  el  $i$ -ésimo elemento de la diagonal de la matriz  $\mathbf{H}$ . Observemos que esta palanca sólo depende de  $\mathbf{X}$ , entonces, una observación con una palanca,  $h_i$ , grande, es aquella con valores extremos en alguna(s) de su(s) covariable(s). Ya que el promedio de las  $h_i$ 's es  $p/n$ , consideraremos una observación con *palanca grande* si su palanca es mayor a  $2p/n$ . En este sentido,  $h_i$  corresponde a la *Distancia de Mahalanobis* de  $\mathbf{X}$  definida como  $(\mathbf{X} - \bar{\mathbf{X}})' \hat{\Sigma}^{-1}(\mathbf{X} - \bar{\mathbf{X}})$ .





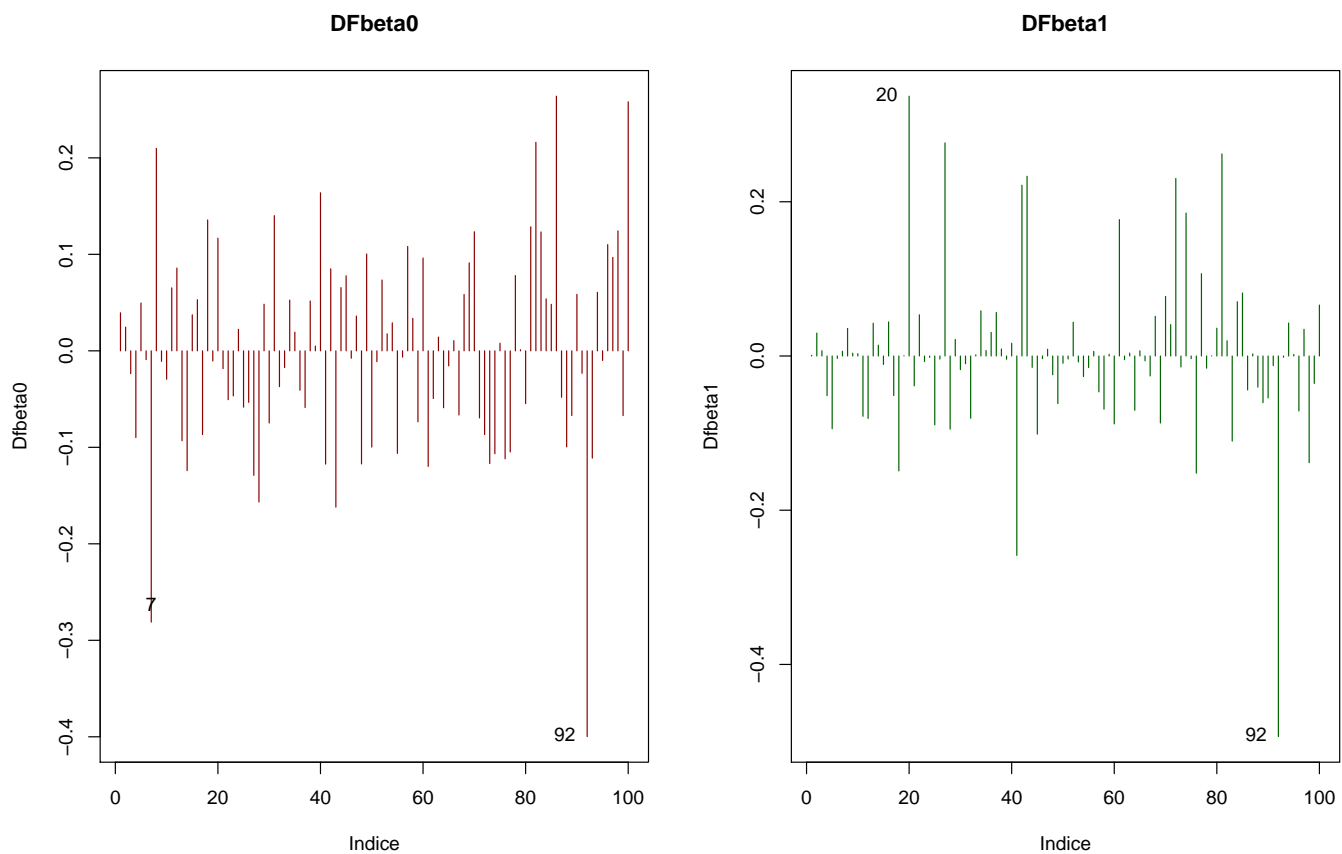
La dependencia de las estadísticas para el diagnóstico de las observaciones, estriba en que sus cálculos dependen de los valores de la palanca de cada individuo. Estas estadísticas son:

- **Distancia de Cook:** Sirve para determinar si una observación es influyente en **TODO EL VECTOR DE PARÁMETROS**. Se considera que una observación es influyente en este caso, si su Distancia de Cook sobrepasa el valor de **UNO**.



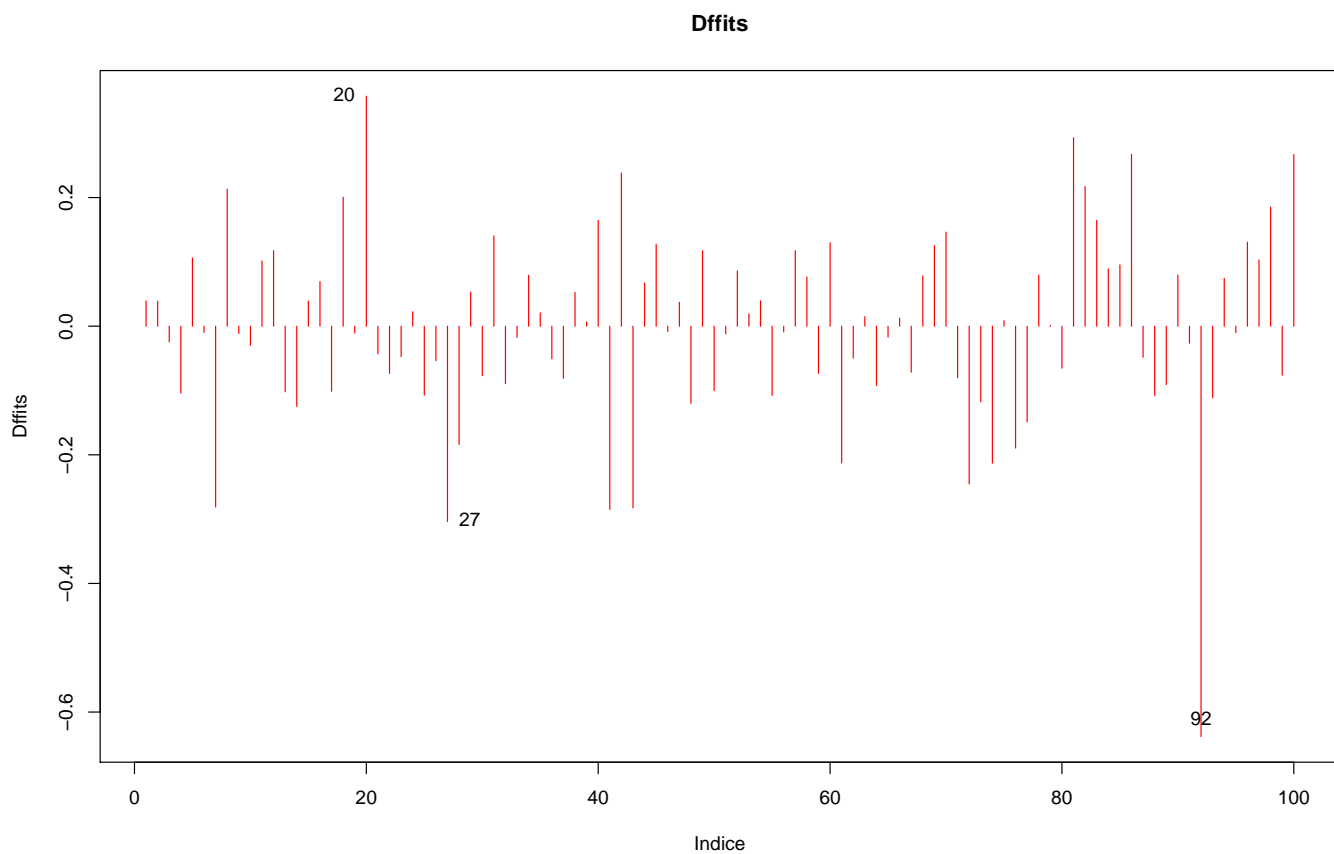
- **Dfbetas:** Sirven para determinar si una observación es influyente en alguno de los coeficientes de regresión. Hay un dfbeta por cada parámetro dentro del modelo, incluido, por supuesto, el de la ordenada al origen. La regla de corte es que *la observación  $i$  es influyente en el  $j$ -ésimo coeficiente de regresión* si:

$$|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$$



- **Dffits**: Se utilizan para determinar si una observación es influyente en la predicción de  $y$ . Se dice que la  $i$ -ésima observación es influyente para predecir  $y$ , si:

$$|DFITTS_i| > 2\sqrt{\frac{p}{n}}$$



con  $p$  el número de parámetros en el modelo (en nuestro caso,  $p=k+1$ ).

# Multicolinealidad

El modelo de regresión lineal múltiple, se construye bajo el supuesto de que los regresores son *ortogonales*, i.e., son independientes. Desafortunadamente, en la mayoría de las aplicaciones el conjunto de regresores no es ortogonal. Algunas veces, esta falta de ortogonalidad no es seria; sin embargo, en algunas otras los regresores están muy cercanos a una perfecta relación lineal, en tales casos las inferencias realizadas a través del modelo de regresión lineal pueden ser erróneas. Cuando hay una cercana dependencia lineal entre los regresores, se dice que estamos en presencia de un problema de *multicolinealidad*.

## Efectos de la multicolinealidad

- Varianzas de los coeficientes estimados son muy grandes. Las implicaciones de este hecho son:
- Los estimadores calculados de distintas sub muestras de la misma población, pueden ser muy diferentes.
- La significancia de algún regresor se puede ver afectada (volverse no significativo) por que su varianza es más grande de lo que debería ser en realidad o por la correlación de la variable con el resto dentro del modelo.
- Es común que algún signo de un parámetro cambie, haciendo ilógica su interpretación dentro del modelo.

## ¿Cómo detectar la multicolinealidad?

**Matriz de correlación.** Examinar las correlaciones entre pares de variables

$$r_{ij} \quad i, j = 1, 2, \dots, k \quad i \neq j$$

Sin embargo, cuando más de dos regresores están linealmente relacionados, puede ocurrir que ninguna de las correlaciones entre cada par de variables, sea grande.

**Factor de inflación de la varianza**

$$VIF_j = (1 - R_j^2)^{-1} \quad j = 1, 2, \dots, k$$

Donde  $R_j^2$  es el coeficiente de determinación del modelo de regresión realizado entre el j-ésimo regresor,  $x_j$  (tomado como variable de respuesta) y el resto de los regresores  $x_i$   $i \neq j$

Experiencias prácticas indican que si algunos de los VIF's excede a 10, su coeficiente asociado es pobremente estimado por el modelo debido a multicolinealidad

### Análisis del eigensistema

Basado en los eigenvalores de la matriz  $\mathbf{X}'\mathbf{X}$ ,  $\lambda_1, \lambda_2, \dots, \lambda_k$

**Número de condición.** (*Multicolinealidad global*)

$$K = \frac{\lambda_{max}}{\lambda_{min}}$$

Si el número de condición es menor que 100, no existen problemas serios de multicolinealidad. Si está entre 100 y 1000 existe de moderada a fuerte multicolinealidad y si excede a 1000, hay severa multicolinealidad.

### El índice de condición

$$k_j = \frac{\lambda_{max}}{\lambda_j}$$

Si el índice de condición es menor que 10, no hay ningún problema. Si está entre 10 y 30, hay moderada multicolinealidad, y si es mayor que 30, existe una fuerte colinealidad en la j-ésima variable  $j=1,2,\dots,k$ , en el modelo.

### IMPORTANTE

En algunos paquetes estos índices se presentan aplicando la raíz cuadrada a su expresión:

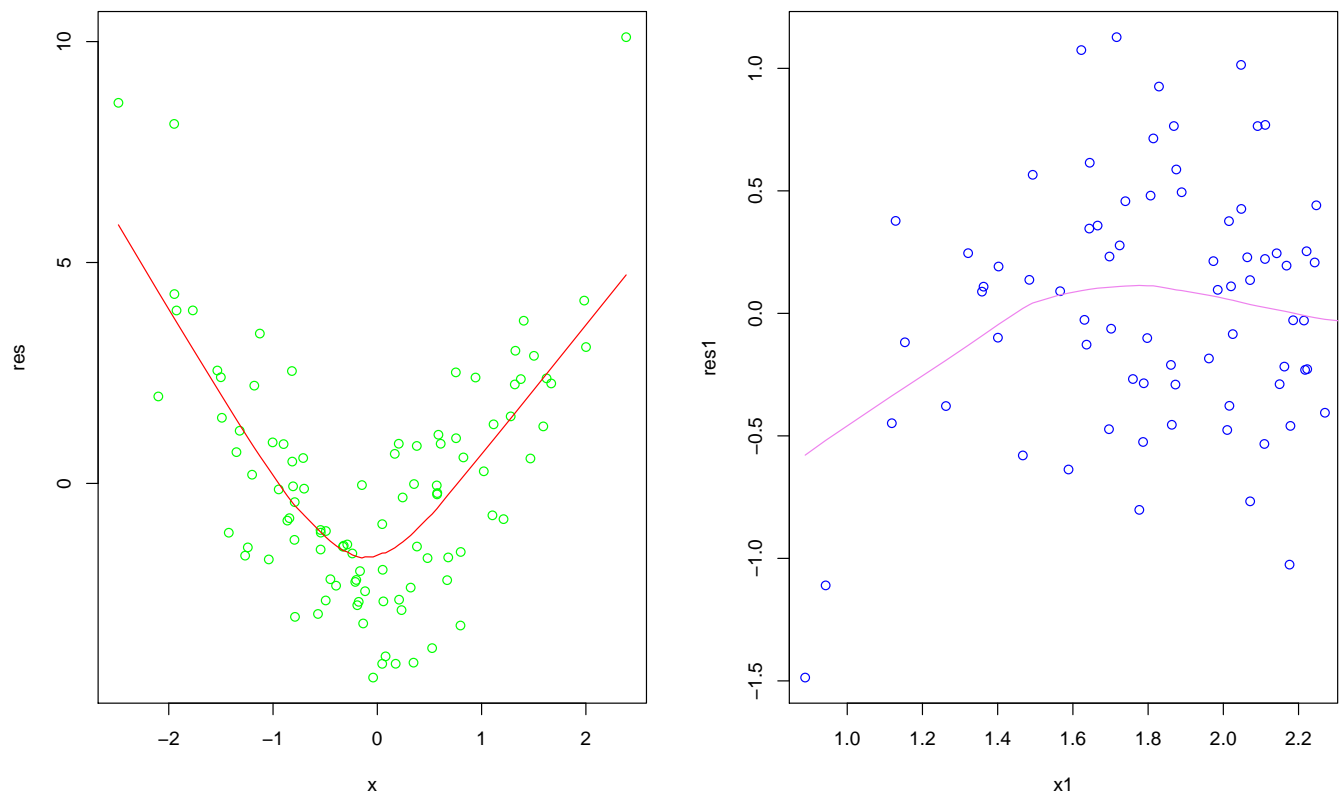
$$K = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \quad \text{y} \quad k_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}} \quad j = 1, 2, \dots, k$$

Entonces, hay que extraer raíz a los puntos de corte de los criterios correspondientes.

# Transformaciones para lograr linealidad

Un supuesto importante en el modelo de regresión es el que considera que debe existir una relación funcional lineal entre cada regresor  $X_j$ ,  $j = 1, 2, \dots, k$  y la respuesta  $y$ . Pero, qué debemos hacer si no se cumple esta relación lineal de la respuesta con alguno(s) de los regresor(es)?

Ya dijimos que este supuesto se evalúa realizando la gráfica de dispersión entre los residuos del modelo y los valores de la variable en cuestión. Cuando no hay una asociación lineal entre la respuesta y la covariable, generalmente este diagrama de dispersión muestra un patrón (tendencia) que sugiere qué tipo de transformación se debería hacer a la covariable para lograr linealidad con la respuesta. Debe quedar claro que la transformación puede realizarse a la variable explicativa o a la variable de respuesta. A muchos investigadores no le gusta transformar la respuesta porque argumentan que pierden “interpretabilidad” del modelo. Aunque esto puede ser cierto, existen transformaciones de la respuesta que pueden “regresarse” para interpretar el modelo con la respuesta original.



Un problema asociado a esta identificación por parte del usuario, es que debe tener experiencia para asociar estas formas a una función analítica específica; hecho no necesariamente cierto. Por lo tanto, requiere de alguna herramienta técnica que pudiera auxiliarlo en esta labor.

## Transformación Box-Cox

Un buen auxiliar, en el caso de que se crea que es necesario transformar la respuesta,  $y$ , es usar la llamada *transformación Box-Cox*.

La transformación Box-Cox de la respuesta,  $y$ , es una función que sirve para *normalizar* la distribución del error, estabilizar la varianza de este error y mejorar la relación lineal entre  $y$  y las  $X$ 's. Se define como

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda}, & \text{Si } \lambda \neq 1 \\ \ln(y_i), & \text{Si } \lambda = 0 \end{cases}$$

Por lo que el modelo de regresión queda como

$$y_i^{(\lambda)} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad i = 1, 2, \dots, n$$

La siguiente tabla muestra el rango de valores de  $\lambda$  que estarían asociados a una transformación analítica común.

Relación $\lambda$ y transformaciones analíticas		
Rango $\lambda$	Transformación analítica asociada	Nombre
$(-2.5, -1.5]$	$\frac{1}{y^2}$	Inversa cuadrada
$(-1.5, -0.75]$	$\frac{1}{y}$	Recíproca
$(-0.75, -0.25]$	$\frac{1}{\sqrt{y}}$	Inversa raíz cuadrada
$(-0.25, 0.25]$	$\ln(y)$	Logaritmo natural
$(0.25, 0.75]$	$\sqrt{y}$	Raíz cuadrada
$(0.75, 1.25]$	$y$	Idéntica o ninguna
$(1.5, 2.5)$	$y^2$	Cuadrada



## Transformación Box-Tidwell

Box y Tidwell implementan un proceso iterativo para encontrar la mejor transformación de las variables predictoras en el modelo de regresión lineal. En este caso, el modelo queda como

$$y_i = \beta_0 + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_k X_{ik}^{\gamma_k} + \epsilon_i, \quad i = 1, 2, \dots, n$$

# Selección de variables

## Introducción

El problema de la selección de variables en la regresión lineal continúa siendo objeto de atención de muchos especialistas (Montgomery and Peck, 1982; Ronchetti and Staudte, 1994). El propósito general del mismo es establecer un modelo de regresión lineal para la variable respuesta,  $y$ , en términos de ciertas variables predictoras  $X_1, X_2, \dots, X_k$  (o funciones de éstas), tratando de conciliar dos criterios contrapuestos:

- Por una parte, incluir tantas variables predictoras en el modelo como sea posible, para que éste sea útil para propósitos predictivos, y
- Por otra, incluir el menor número posible de predictores (principio de parsimonia) para disminuir el error del modelo, además de los costos de obtención de información.

De esta manera, el problema se convierte en *buscar un balance entre simplicidad y ajuste* y es precisamente en lo que consiste el proceso de *seleccionar el mejor modelo de regresión*.

Al tratar de dar solución a este problema se corren dos riesgos; por una parte, el de incluir variables irrelevantes, y por otra, el de omitir alguna importante. Hay que tener en cuenta la dificultad esencial que constituye el desconocimiento de la varianza de las observaciones, lo que implica la necesidad de juicios subjetivos. Así, puede decirse que no existe un procedimiento único para seleccionar el mejor modelo de regresión y que se continúa investigando con el fin de brindar nuevos procedimientos para la solución de este problema.

## Métodos computacionales de selección

En la práctica, la selección del subconjunto de variables explicativas de los modelos de regresión se deja en manos de procedimientos más o menos automáticos. Los procedimientos más usuales son los siguientes:

**Método Backward:** Se inicia incluyendo en el modelo a todas las variables disponibles y se van eliminando del modelo de una en una según su capacidad explicativa. En concreto, la primera variable que se elimina es aquella que presenta un menor coeficiente de correlación parcial con la variable dependiente (o equivalentemente, un menor valor del estadístico  $t$ ) y así sucesivamente hasta llegar al punto en que ninguna de las variables que permanecen en el modelo es menor que el criterio estadístico que determina su salida. De manera esquemática este método procede de la siguiente manera

- Se incluyen todos los regresores en el modelo
- Se calculan *las  $F$ -parciales* de cada regresor

$$F_{\text{parcial}} = \frac{SS_R(X_j | X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_k)}{SS_E(X_1, X_2, \dots, X_k)}, \text{ (coincide con la } \mathbf{F} \text{ para probar cada regresor)}$$

- El candidato a salir del modelo es aquél que tenga la  $\mathbf{F}$  parcial más pequeña.
- Si esta  $\mathbf{F}$  es menor que el criterio estadístico para salir:  $\mathbf{F}_{OUT} = \mathbf{F}_{(1, n-p, \alpha)}$  el regresor sale del modelo; de lo contrario, permanece en él.

**Método forward:** Se parte inicialmente de un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas a la que presente un mayor coeficiente de correlación (en valor absoluto) con la variable dependiente. En los pasos sucesivos se va incorporando al modelo aquella variable que presenta un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento se detiene cuando ninguna de las variables por ingresar al modelo, cumple con el criterio estadístico de inclusión. De forma esquemática el método opera de la siguiente manera

- Inicialmente no hay regresores, únicamente el intercepto.

- El primer regresor candidato a ingresar al modelo, digamos  $X_1$ , es aquél que tenga la mayor correlación con la respuesta,  $y$ . También es el regresor que produce el mayor valor de  $\mathbf{F}$  para probar su significancia estadística.
- El segundo regresor a ingresar al modelo, es el que tenga la correlación más alta con  $y$ , después de ajustar por el efecto de  $X_1$  en  $y$ . Llamemos a esta correlación, *correlación parcial*.
- Supongamos que en este segundo paso, el regresor con mayor correlación es  $X_2$ . Entonces, la **F-parcial** más grande es

$$\mathbf{F}_{\text{parcial}} = \frac{SS_R(X_2|X_1)}{MS_E(X_1, X_2)}$$

- Si el valor de esta  $\mathbf{F}$  es más grande que el valor de inclusión, entonces  $X_2$  se adiciona al modelo, de lo contrario “*no*”.
- Para la siguiente variable, digamos  $X_3$ , la  $\mathbf{F}$  parcial es

$$\mathbf{F} = \frac{SS_R(X_3|X_1, X_2)}{MS_E(X_1, X_2, X_3)}$$

y así sucesivamente. Usualmente el valor de inclusión es  $\mathbf{F}_{IN} = \mathbf{F}_{(1, n-p, \alpha)}$ .

**Método stepwise:** Es uno de los más empleados y consiste en una combinación de los dos anteriores. En el primer paso se procede como en el método forward pero, a diferencia de éste en el que cuando una variable entra en el modelo ya no vuelve a salir, en el procedimiento stepwise es posible que la inclusión de una nueva variable haga que otra que ya estaba en el modelo resulte redundante y sea “expulsada” de él. Este procedimiento continúa hasta que ninguna de las variables por entrar al modelo, cumple con el criterio estadístico para hacerlo,  $\mathbf{F}_{IN}$ , ni las variables dentro del modelo, cumplen con el criterio estadístico para abandonarlo,  $\mathbf{F}_{OUT}$ .

## Inconvenientes de los métodos automáticos de selección

- Los métodos de selección automáticos pueden producir modelos finales distintos, pese a que hayan considerado el mismo conjunto de variables para su selección.

- Ninguno garantiza *plausibilidad dentro del área de aplicación*. Dado que la selección se realiza atendiendo criterios estadísticos, no necesariamente, el modelo debe ser plausible, es decir, lógico dentro del área de aplicación. Podría generarse un modelo cuya estructura sea contradictoria de acuerdo al conocimiento del área. Este es, sin duda, uno de lo inconvenientes más importante de esta selección automática.

# Análisis Multivariado

## Introducción

Los datos multivariados se presentan cuando el investigador recaba varias variables sobre cada “unidad” en su muestra. La mayoría de los conjuntos de datos que se colectan para una investigación son multivariados. Aunque algunas veces tiene sentido estudiar por separado cada una de las variables, en la mayoría de los casos no. En el común de las situaciones, las variables están relacionadas de tal manera que si se analizan por separado, no se revela la estructura completa de los datos. En la gran mayoría de los conjuntos de datos multivariados, todas las variables necesitan analizarse de manera simultánea para descubrir patrones y características esenciales de la información que contienen. El análisis multivariado incluye métodos que son totalmente descriptivos y otros que son inferenciales. El objetivo principal es revelar la estructura de los datos, eliminando el “ruido” de los mismos.

Un aspecto muy importante a considerar en los datos multivariados, es que, por lo general, las variables que los componen tienen diferentes escalas de medición, hecho que se debe considerar al momento de realizar el análisis estadístico.

## Estructura de los datos multivariados

### Matriz de datos

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix}$$

Donde cada vector  $\mathbf{x}'_j$ , es un vector columna,  $p \times 1$ , que representa los valores de las  $p$  variables sobre el individuo  $j$ . Y  $x_{jk}$  es el valor de la  $k$ -ésima variable ( $k=1,2,\dots,p$ ) del  $j$ -ésimo individuo ( $j=1,2,\dots,n$ ).

## Resumen mediante descripciones numéricas

En una extensión simple de los procesos descriptivos que se realizan con una muestra, podemos hacer los correspondientes resúmenes numéricos para cada una de las variables involucradas en el análisis.

- Resúmenes univariados, respetando la escala de medición de cada variable
- **Vector de medias**

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)'$$

$$\text{con } \bar{x}_k = \frac{1}{n} \sum_{j=1}^n x_{jk}, \quad k = 1, 2, \dots, p.$$

- **Matriz de Varianza-Covarianza**

$$\mathbf{S}^2 = \begin{pmatrix} s_{11}^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22}^2 & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp}^2 \end{pmatrix}$$

$$\text{con las varianzas muestrales } s_{kk}^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2, k = 1, 2, \dots, p, \text{ y}$$

$$\text{las covarianzas muestrales } s_{ik} = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k), i \neq k = 1, 2, \dots, p$$

- **Matriz de correlación**

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}$$

$$\text{con las correlaciones muestrales } r_{ik} = \frac{s_{ik}}{s_{ii}s_{kk}}, i \neq k = 1, 2, \dots, p$$

## Algunas características de las correlaciones

- $-1 \leq r_{ik} \leq 1$
- $r_{ik}$  es una medida de la fuerza de la asociación lineal entre las variables involucradas
- $r_{ik}$  es invariante ante cambios de escala
- $r_{ik}$  usualmente se refiere a la correlación de *Pearson*. Para medidas generales de correlación (incluida la no lineal), se pueden utilizar la *tau de Kendall* o *rho de Spearman*.

## Representación matricial

- **Media muestral:**  $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)'$
- **Matriz de varianza-covarianza muestral:**  $\mathbf{S} = [s_{ik}]$
- **Matriz de correlación muestral:**  $\mathbf{R} = [r_{ij}]$ , con  $r_{ii} = 1$



# Resumen mediante descripciones gráficas

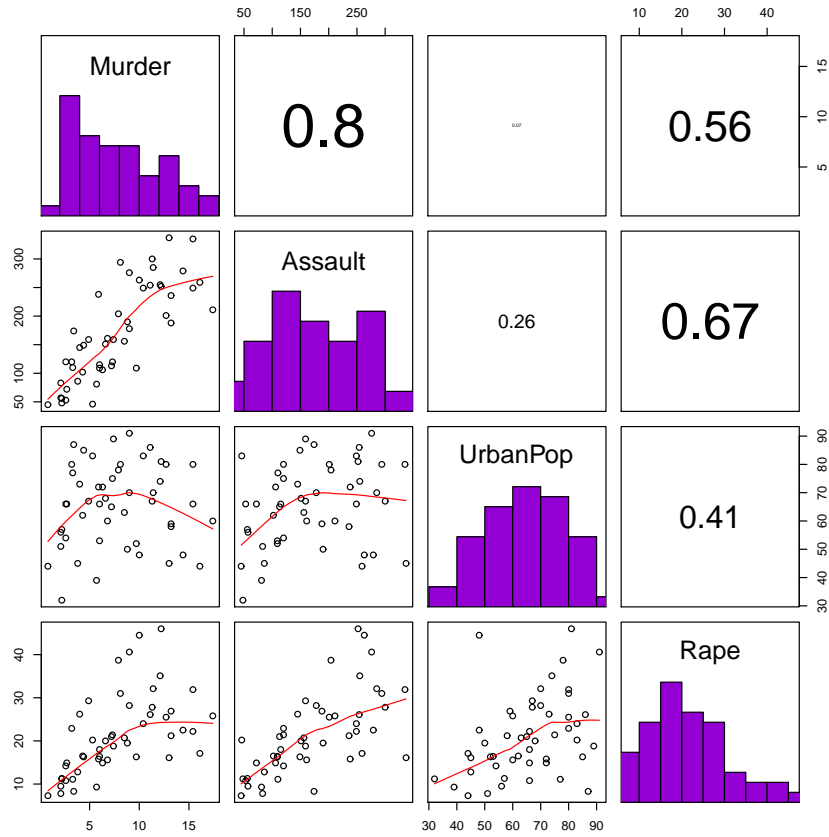
Una manera natural en estadística de mostrar la información contenida en un conjunto de datos, es a través de algunas representaciones gráficas de los mismos. Similar al análisis univariado estándar, se pueden hacer las representaciones gráficas que se considere necesarias, para cada variable. Pero, dada la naturaleza multivarida de nuestros datos, es más conveniente realizar estas representaciones tratando de involucrar a todas las variables de manera simultánea. El problema para graficar datos multivariados, es su dimensión.

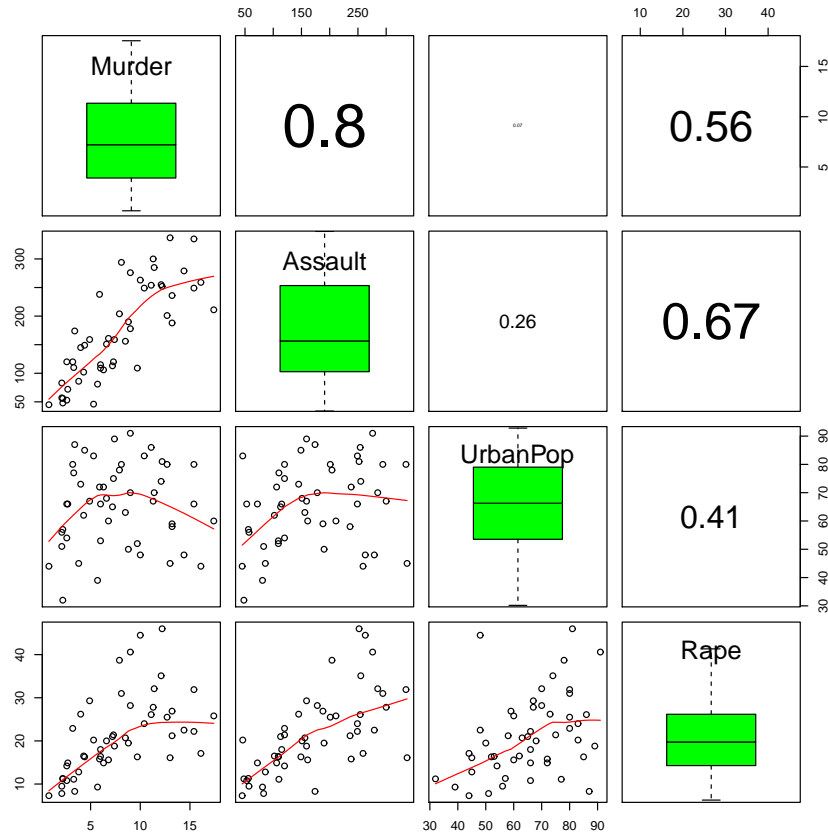
Existen diversas técnicas gráficas para desplegar datos multivariados. La finalidad esencial de éstas es tratar de identificar grupos similiares de sujetos, observaciones atípicas, dispersión de las variables, correlación entre ellas, etc.

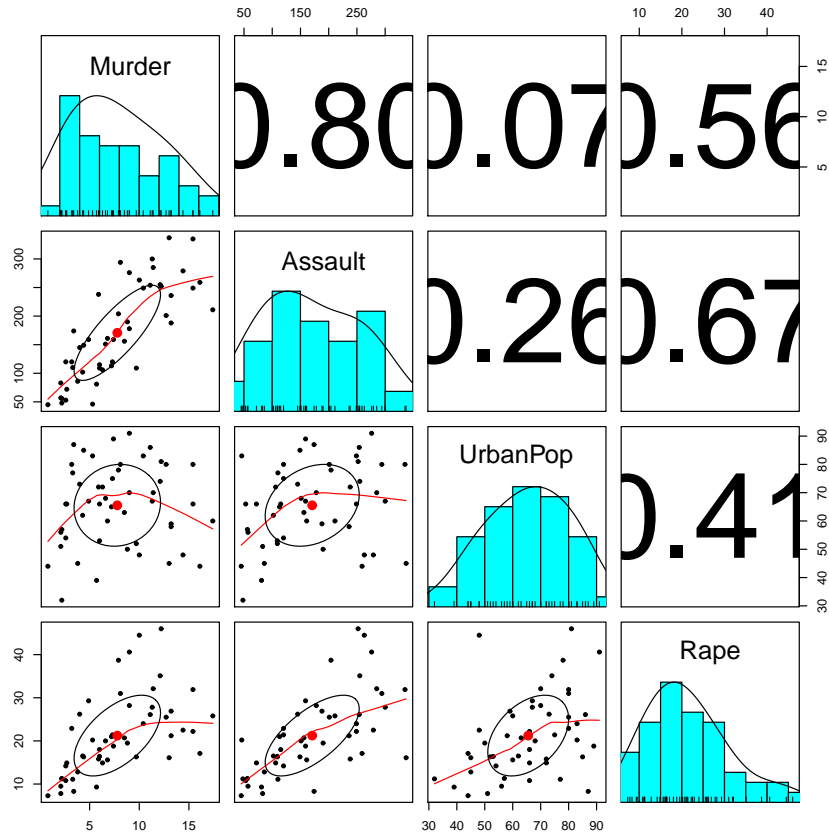
El uso de diagramas y gráficas ahorra tiempo, ya que las características esenciales de grandes volúmenes de datos estadísticos puede apreciarse de un solo vistazo.

## Gráfica de la matriz de datos

Una procedimiento útil para iniciar una exploración de las variables en datos multivariados, es desplegar gráficas de dispersión entre pares de variables contenidas en la matriz de datos. Dijimos que para que un análisis multivariado tenga sentido, debemos tener una *fuerte* correlación entre las variables involucradas. Una gráfica que es útil para estos propósitos y que proporciona información adicional, se obtiene con el comando *pairs* de **R**. Los datos pertenecen a la base en **R**, *USArrests* que reporta el número de arrestos por asesinatos (Murder), asaltos (Assault), y violaciones (Rape), además del porcentaje de población urbana (Urban Pop) de los 50 estados que constituyen los Estados Unidos de América



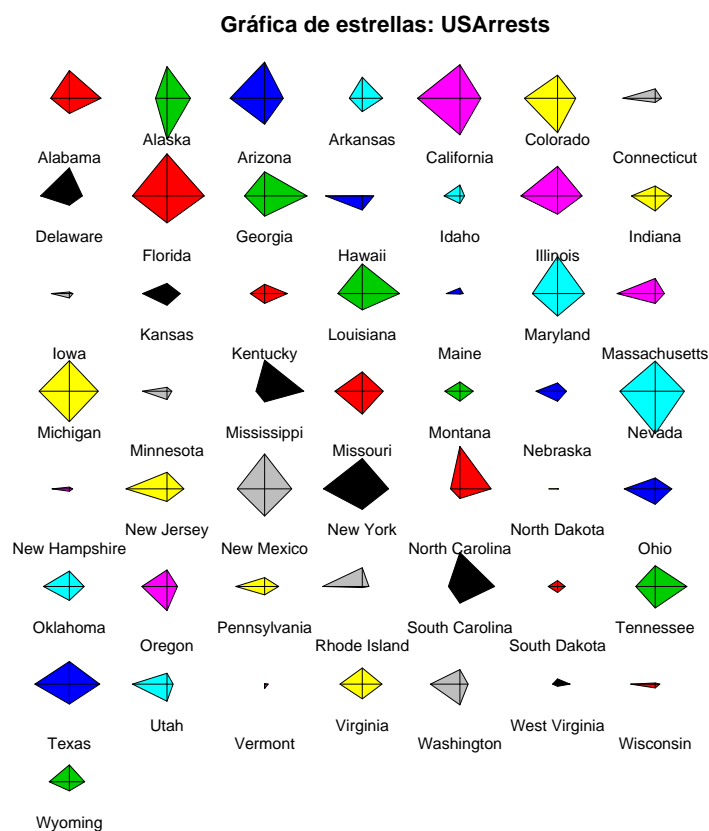




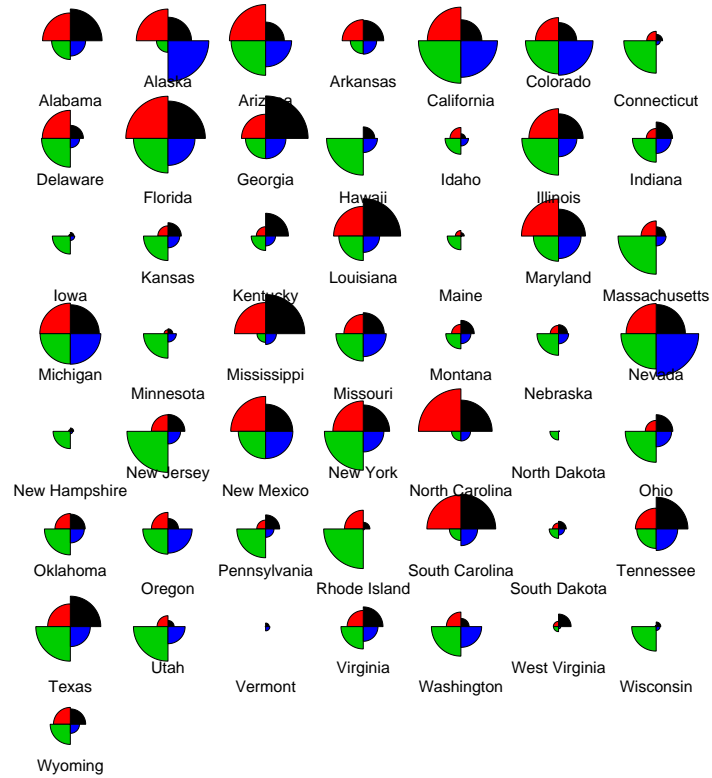
La gráfica anterior presenta características de la forma de la densidad de la variable (histograma y densidad tipo kernel) y de la correlación entre el grupo de variables. Pero no sería útil para descubrir qué estados son similares de acuerdo a este grupo de variables medidas. Para ello, recurriremos a algunas técnicas que intentan resumir todas las variables en una sola gráfica.

## Diagramas de estrellas

Cada individuo se representa en una estrella, con tantos rayos o ejes como variables posea su vector de observaciones. Cada eje representa el valor de la variable re-escalada de manera independiente entre variables. Para re-escalar se utilizan todos los datos. En todas las estrellas se usa siempre el mismo eje para representar la misma variable. El eje  $j$  en la estrella del individuo  $i$  depende de  $x_{ij}$  (en valor absoluto o relativo)



### Gráfica de estrellas: USArrests

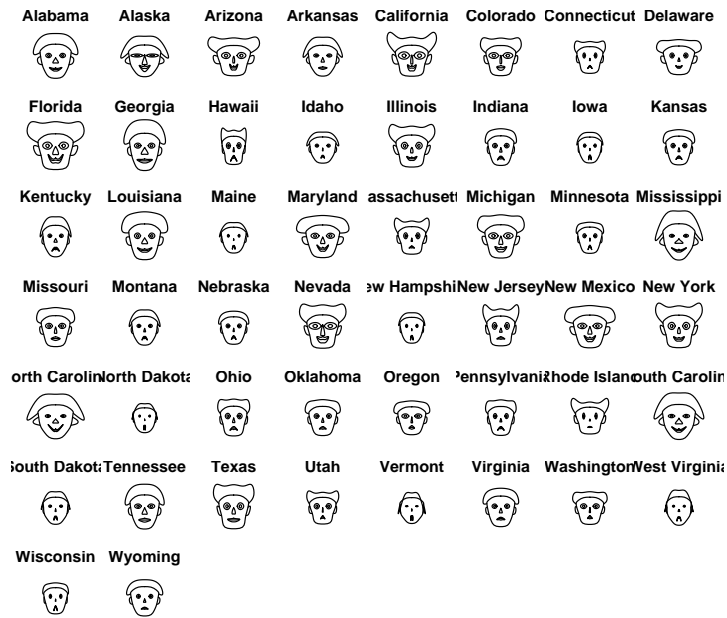


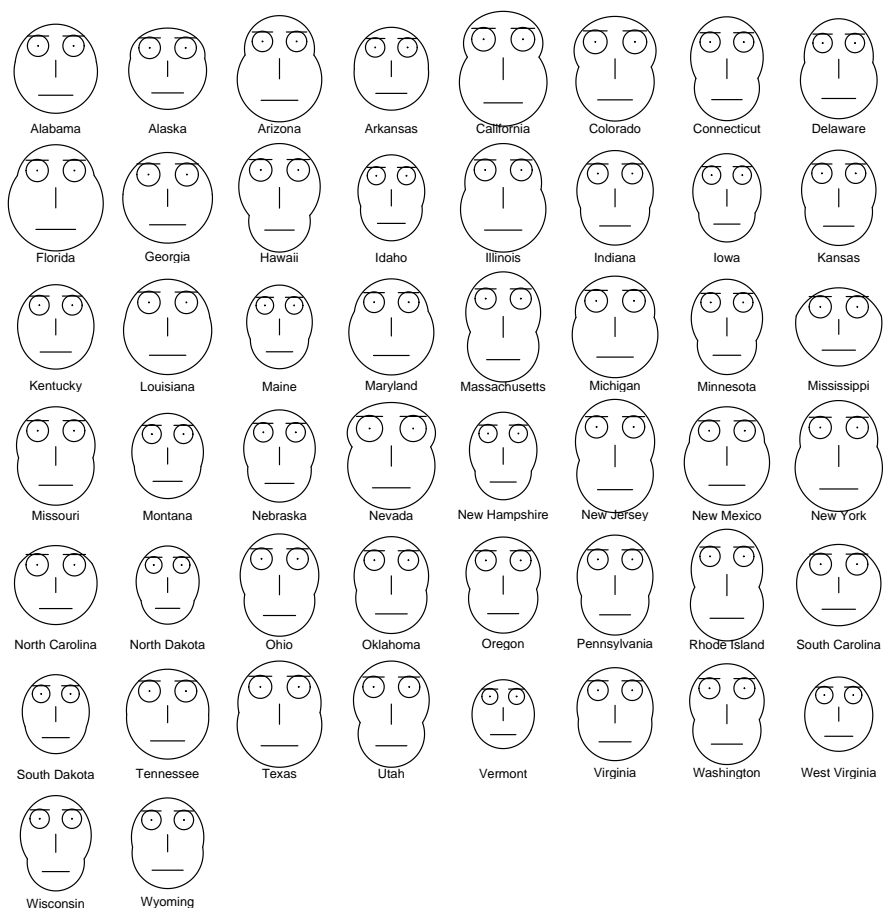
## Caritas de Chernoff

El objetivo en esta técnica es asociar el valor de cada variable con alguna característica de una cara humana. Las variables están asociadas con seis aspectos básicos de la carita: *forma de la cara, la boca, la nariz, los ojos, las cejas y las orejas*. Cuando el número de variables es grande, algunas de ellas estarán asociadas con varios aspectos relacionados con los anteriores: *Amplitud de la cara, longitud de las cejas, altura de la cara, separación de los ojos, posición de las pupilas, longitud de la nariz, ancho de la nariz, diámetro de las orejas, nivel de las orejas, longitud de la boca, inclinación de los ojos, altura de las cejas*, etc. Bernard Flury ideó, con base al trabajo de Chernoff, duplicar la cantidad de variables para representar la carita, dejando de lado la simetría, i.e., del lado izquierdo del rostro es posible graficar 18 variables y otras tantas del lado derecho.

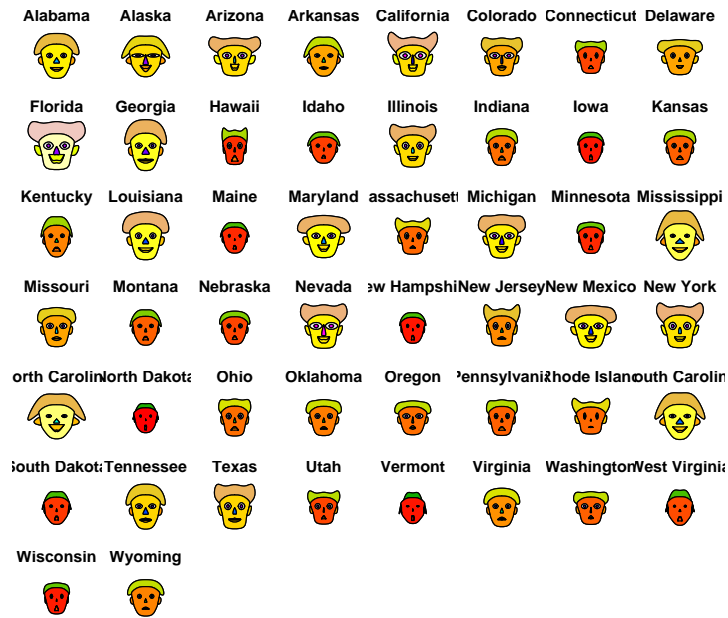


# Caritas de Chernoff: USArrests





### Caritas de Chernoff: USArrests



## Curvas de Andrew

Supongamos que cada individuo tiene  $p$  variables medidas  $(X_{i1}, X_{i2}, \dots, X_{ip})$ . Se define la función

$$f_{X_i} = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\sin(t) + X_{i3}\cos(t) + X_{i4}\sin(2t) + X_{i5}\cos(2t) + \dots \quad -\pi < t < \pi$$

Algunas propiedades interesantes de estas curvas

i) Preserva medias, i.e.

$$f_{\bar{X}} = \frac{1}{n} \sum_{i=1}^n f_{X_i}(t)$$

ii) Preserva distancias

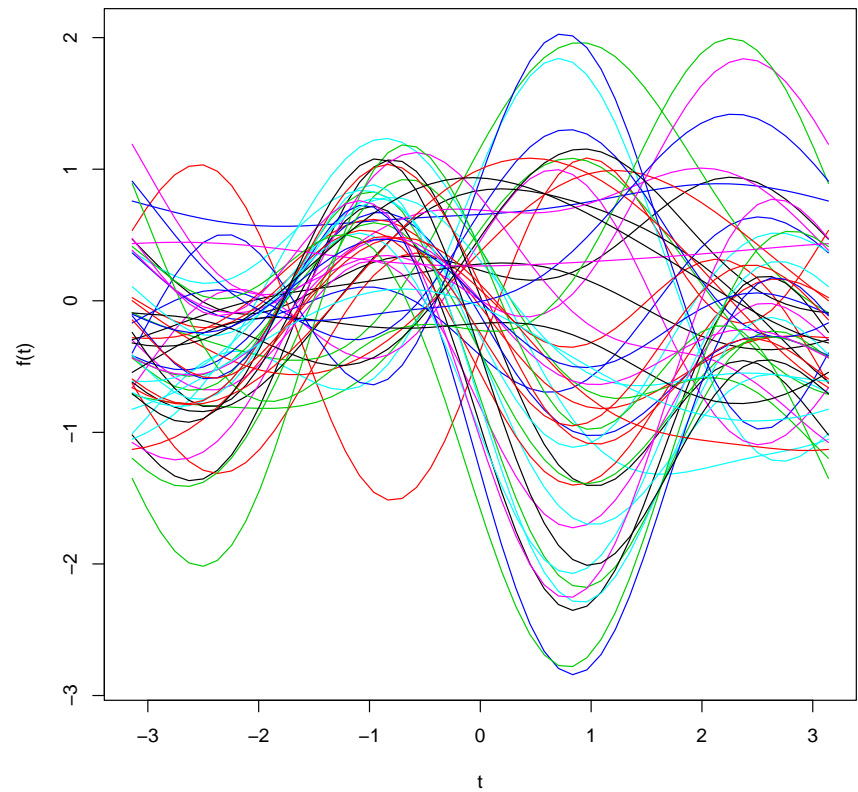
$$\|f_{X_i}(t) - f_{X_j}(t)\|^2 = \int_{-\pi}^{\pi} (f_{X_i}(t) - f_{X_j}(t))^2 dt = \pi \sum_{k=1}^p (X_{ik} - X_{jk})^2$$

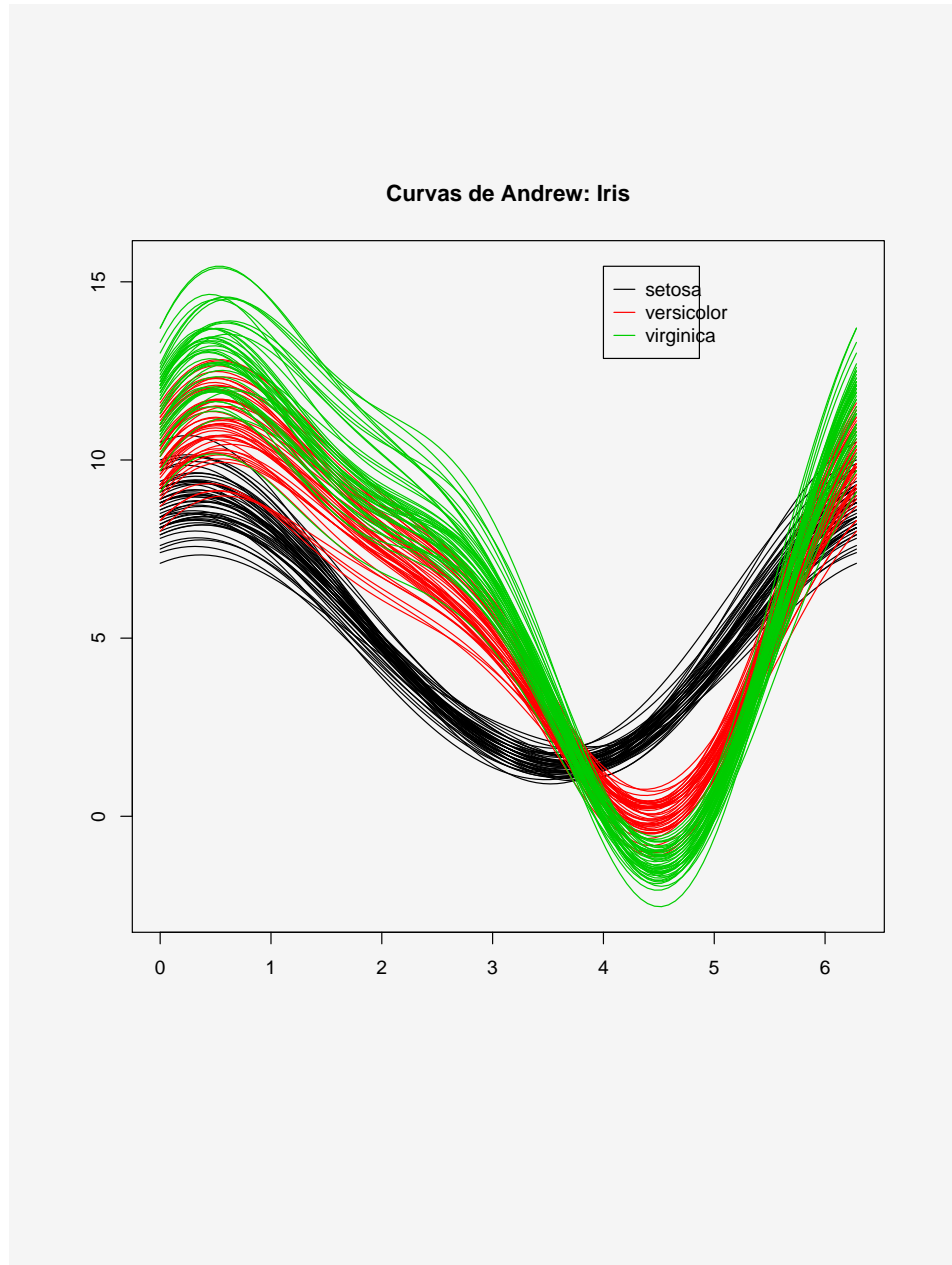
Por lo tanto, si los sujetos  $X_i, X_j$ , están cerca, las respectivas curvas lo estarán también.

En esta representación gráfica, el orden de las variables juega un papel importante. Si la dimensión de  $\mathbf{X}$  es muy alta, las últimas variables tendrán una contribución pequeña. Por lo que se recomienda ordenar las variables de manera que las variables “más importantes” aparezcan al principio (por ejemplo, aquellas que discriminan mejor los posibles subgrupos presentes en los datos). También es recomendable no incluir demasiadas observaciones (curvas) en una sola gráfica.

En este tipo de gráficas, las observaciones atípicas aparecen como curvas aisladas que se distinguen claramente de las demás.

Curvas Andrews: USArrests





**Nota:** Cada una de estas técnicas se vuelve inadecuada si el número de sujetos es muy grande.

Estas no son las únicas técnicas de representación gráfica de datos multivariados, existen otras como

- Gráficas de perfiles
- Parallel coordinates plot

# Análisis de Componentes Principales

## Introducción



El objetivo principal de la mayoría de las técnicas numéricas de análisis multivariado, es reducir la dimensión de nuestros datos. Por supuesto, si esta reducción se puede hacer a 2 ó 3 dimensiones, se tiene la posibilidad de una visión gráfica de los mismos. Obvio, *siempre* es posible hacer la reducción a este número de dimensiones, pero es importante juzgar si estas pocas dimensiones son suficientes para resumir la información contenida en todas las variables.

El *análisis de componentes principales* tiene este objetivo: dadas  $n$  observaciones de  $p$  variables, se analiza si es posible representar adecuadamente esta información con un número menor ( $q \ll p$ ) de *variables construidas como combinaciones lineales de las originales*, llamadas *componentes principales*. Esta técnica se debe a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901).

En concreto, los objetivos del análisis de componentes principales son:

- Reducir la dimensión de los datos ( $q \ll p$ )
- Generar nuevas variables: Componentes principales

*¿Para qué?*

- Explorar datos multivariados
- Encontrar agrupaciones
- Encontrar datos atípicos
- Como auxiliar para combatir la multicolinealidad en los modelos de regresión

*¿Qué hace?*

Forma nuevas variables llamadas *Componentes Principales* (c.p.) con las siguientes características:

- 1) No están correlacionadas (bajo el supuesto de distribución normal, son independientes)
- 2) La primera c.p. explica la mayor cantidad de varianza de los datos, que sea posible

3) Cada componente subsecuente explica la mayor cantidad de la variabilidad restante de los datos, que sea posible.

Las componentes son de la forma:

$$Z_i = a_i' X = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \quad i = 1, 2, \dots, p \quad \text{ó}$$

$$Z_i = a_i'(X - \mu) \text{ (centradas)}$$

Es decir, son combinaciones lineales de las  $p$  variables.

Para la primer componente, el objetivo es construir esta combinación lineal, de tal manera que la varianza de ella sea máxima. Por supuesto, suena a resolver un problema de maximización. Entonces, el problema consiste en encontrar el vector  $a_1$ , que haga máxima la varianza de esta primer componente. Para garantizar la unicidad de la solución, forzaremos el procedimiento a que  $a_1$  sea de *norma uno* ( $\|a_1\| = 1$ ).

En concreto, debe elegirse  $a_1$ , un vector de norma uno,  $\|a_1\| = a_1' a_1 = 1$ , de tal manera que:

$$\mathbb{V}(Z_1) = \mathbb{V}(a_1' X) = a_1' \mathbb{V}(X) a_1 = a_1' \Sigma a_1 \quad \text{sea máxima}$$

Bajo esta restricción, el problema se transforma a encontrar un máximo con restricciones, para lo que utilizaremos la técnica de los *multiplicadores de Lagrange*.

### Deducción de la construcción de las componentes

El problema se plantea de la siguiente manera. Maximizar

$$F(a) = \mathbb{V}(Z) = \mathbb{V}(a' X) = a' \mathbb{V}(X) a = a' \Sigma a$$

$$\text{s.a. } \lambda \|a\| = \lambda a' a = 1$$

Que genera la función

$$F(a) = a' \Sigma a - (\lambda a' a - 1)$$

Derivando respecto al vector  $a$ , obtenemos

$$\frac{\partial F(a)}{\partial a} = 2\Sigma a - 2\lambda a = 0$$



cuya solución está dada por la igualdad

$$\Sigma a = \lambda a$$

que, como vimos en el repaso de los conceptos de álgebra lineal, implica que  $a$  es un eigenvector de la matriz  $\Sigma$  y  $\lambda$  el eigenvalor correspondiente a este eigenvector.

Para determinar cuál valor propio de  $\Sigma$  es el que corresponde a la solución de la ecuación anterior, multipliquemos por la izquierda por  $a'$ , dicha ecuación

$$a' \Sigma a = \lambda a' a \Rightarrow a' \Sigma a = \lambda$$

y observamos, entonces, que  $\mathbb{V}(Z) = \lambda$ , y como esta cantidad es la que deseamos maximizar, entonces  $\lambda$  es el eigenvalor más grande de la matriz  $\Sigma$  con  $a$  el eigenvector asociado a este eigenvalor, llamémoslos  $\lambda_1$  y  $a_1$ , respectivamente.

La siguiente componente debe cumplir con las condiciones de tener la mayor varianza del remanente, una vez calculada la primera, y no estar correlacionada con ésta. Obsérvese que esta última condición se obtiene si los correspondientes vectores, digamos  $a_1$  y  $a_2$  son ortogonales, y como pediremos que  $a_2$  sea también de norma uno, entonces serán ortonormales. Una manera de garantizar que esta segunda componente es la de mayor varianza posible, después de la primera, es que la suma de estas dos varianzas sea máxima. Entonces el problema se puede plantear de la siguiente manera. Maximizar

$$F(a_1, a_2) = a_1' \Sigma a_1 + a_2' \Sigma a_2$$

s.a  $\lambda_1 a_1' a_2 = 1$  ,  $\lambda_2 a_2' a_2 = 1$  y  $\mu a_1' a_2 = 0$

Derivando esta función respecto a los vectores  $a_1$  y  $a_2$ , tenemos

$$\begin{aligned} \frac{\partial F(a_1, a_2)}{\partial a_1} &= 2\Sigma a_1 - 2\lambda_1 a_1 + \mu a_2 = 0 \\ \frac{\partial F(a_1, a_2)}{\partial a_2} &= 2\Sigma a_2 - 2\lambda_2 a_2 + \mu a_1 = 0 \end{aligned}$$

Multiplicando la parcial respecto a  $a_1$  por  $a_1'$  por la izquierda y recordando que  $a_1' a_2 = 0$ , porque son ortonormales, tenemos

$$a_1' \Sigma a_1 = \lambda_1 \Rightarrow a_1 a_1' \Sigma a_1 = \lambda_1 a_1 \Rightarrow \Sigma a_1 = \lambda_1 a_1$$

De manera similar, multiplicando la parcial respecto a  $a_2$  por  $a_2'$  por la izquierda y recordando que  $a_2' a_1 = 0$ , porque son ortonormales, tenemos

$$a_2' \Sigma a_2 = \lambda_2 \Rightarrow a_2 a_2' \Sigma a_2 = \lambda_2 a_2 \Rightarrow \Sigma a_2 = \lambda_2 a_2$$

que implica que  $a_1$  y  $a_2$  deben ser eigenvectores de  $\Sigma$ . Tomando estos vectores propios de norma uno y sustituyendo en la función objetivo, obtenemos

$$\lambda_1 a_1' a_1 + \lambda_2 a_2' a_2 - \lambda_1 (a_1' a_1 - 1) - \lambda_2 (a_2' a_2 - 1) - \mu a_1' a_2 = \lambda_1 + \lambda_2$$

Por lo que es claro que  $\lambda_1$  y  $\lambda_2$  deben ser los dos eigenvalores más grandes de la matriz  $\Sigma$  y  $a_1$  y  $a_2$  sus correspondientes eigenvectores.

De manera general, la  $j$ -ésima componente principal será

$$Z_j = a_j' X \quad j = 1, 2, \dots, p \quad \text{con } a_j \text{ el eigenvector de la matriz } \Sigma \text{ asociado al eigenvalor } \lambda_j$$

y  $\lambda_1 > \lambda_2 > \dots > \lambda_p$ .

## Propiedades de los componentes principales

Los componentes principales como variables derivadas de las originales, tienen las siguientes propiedades:

- *Conservan la variabilidad original de los datos:* En el sentido de que la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales.

Por construcción tenemos que

$$\mathbb{V}(Z_1) = \lambda_1, \mathbb{V}(Z_2) = \lambda_2, \text{ etc.}$$

y además se tiene también que  $\text{Cov}(Z_1, Z_2) = 0$ . En general  $\text{Cov}(Z_i, Z_j) = 0$  para toda  $i \neq j$ ,  $i, j = 1, 2, \dots, p$ . Entonces

$$\text{traza}(\Sigma) = \sum_{i=1}^p \mathbb{V}(X_i) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \text{Var}(Z_i)$$

Las nuevas variables  $Z_i$  tienen conjuntamente la misma variabilidad que las variables originales, la suma de varianzas es la misma, pero su estructura o constitución es muy diferente.

- La proporción de la varianza total explicada por una componente, es el cociente entre su varianza (el valor propio asociado al vector propio que la define), y la suma de los valores propios de la matriz. Por esta razón se dice que el  $i$ -ésimo componente principal explica una proporción de varianza igual a:

$$\frac{\lambda_i}{\sum_{i=1}^p \lambda_i}$$

y los primeros  $r$  de ellos

$$\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p \lambda_i} \quad r \leq p$$

- Las covarianza entre el vector de variables originales  $X$  y la  $i$ -ésima componente principal  $Z_i$ , es:

$$\text{Cov}(X, Z_i) = \text{Cov}(X, a_i' X) = a_i' \text{Cov}(X, X) = a_i' \Sigma = a_i' \lambda_i = \Sigma a_i = \lambda_i a_i \quad i = 1, 2, \dots, p$$

Es decir

$$\mathbb{C}ov(X, Z_i) = \mathbb{C}ov(X_1, X_2, \dots, X_p, Z_i) = \lambda_i a_i = \lambda_i (a_{i1}, a_{i2}, \dots, a_{ip})$$

Entonces, la covarianza entre la  $i$ -ésima componente y la  $j$ -ésima variable es:

$$\mathbb{C}ov(X_j, Z_i) = \lambda_i a_{ij}$$

Como  $\mathbb{V}(X_j) = \sigma_{jj}^2$  y  $\mathbb{V}(Z_i) = \lambda_i$ , entonces tenemos que:

$$\mathbb{C}or(X_j, Z_i) = \frac{\mathbb{C}ov(X_j, Z_i)}{\sqrt{\mathbb{V}(X_j) \mathbb{V}(Z_i)}} = \frac{\lambda_i a_{ij}}{\sqrt{\sigma_{jj}^2 \lambda_i}} = \frac{\sqrt{\lambda_i} a_{ij}}{\sigma_{jj}}$$

El peso que tiene la variable  $j$  en la componente  $i$ , está dado por  $a_{ij}$ . El tamaño relativo de las  $a_{ij}$ 's reflejan la contribución relativa de cada variable en la componente. Para interpretar, en el contexto de los datos, una componente, debemos analizar el patrón de las  $a_{ij}$  de cada componente.

Si utilizamos la matriz de correlación para realizar el análisis de *c.p.*, como  $\sigma_{jj}^2 = 1$ , entonces

$$a_{ij}^* = \sqrt{\lambda_j} a_{ij}$$

se interpreta como el coeficiente de correlación entre la variable  $j$  y el componente  $i$ . Esta es una de las interpretaciones particularmente más usuales.

## Componentes muestrales

Como sabemos,  $\Sigma$  es desconocida, pero podemos estimarla con  $\mathbf{S}$  la matriz de varianza-covarianza muestral, que es un estimador con muy buenas propiedades estadísticas. Entonces, con datos reales, el análisis de componentes principales se realiza con esta matriz y se obtienen los estimadores

$$\hat{\lambda}_i \quad y \quad \hat{a}_i$$

## ¿Matriz de varianza-covarianza o de correlación?

¿Cuándo una, cuándo otra?

### Varianza-covarianza

- Variables medidas en las mismas unidades o, por lo menos, en unidades comparables
- Varianzas de tamaño semejante.

Si las variables no están medidas en las mismas unidades, entonces cualquier cambio en la escala de medición en una o más variables tendrá un efecto sobre las *c.p.* Por ejemplo, supongamos que una variable que se midió originalmente en pies, se cambió a pulgadas. Esto significa que la varianza de la variable se incrementará en  $12^2 = 144$ . Ya que *c.p.* se basa en la varianza, esta variable tendría una mayor influencia sobre los *c.p.* cuando se mide en pulgadas que en pies.

Si una variable tiene una varianza mucho mayor que las demás, dominará el primer componente principal, sin importar la estructura de covarianza de las variables.

Si no se tienen las condiciones para realizar un análisis de *c.p.* con la matriz de varianza-covarianza, se recomienda hacerlo con la matriz de correlación.

Aplicar análisis de *c.p.* a la matriz de correlación, es equivalente a aplicarlo a datos estandarizados (“puntajes *z*”), en lugar de los datos crudos. Realizar el análisis de *c.p.* con la matriz de correlación, implica intrínsecamente asumir que todas las variables tienen igual importancia dentro del análisis, supuesto que no siempre puede ser cierto.

Pueden presentarse situaciones en donde las variables no estén en unidades comparables y en las que el investigador considere que tienen una importancia distinta. Algunos paquetes estadísticos permiten asignar pesos a las variables. Entonces se procedería a estandarizar las variables y posteriormente asignar pesos mayores a aquéllas que el investigador considere más importantes.

## Análisis de *c.p.* con la matriz de correlación

Estandarizar los datos, hacer análisis de *c.p.* utilizando la matriz de correlación en lugar de la de varianza-covarianza.

**Importante:** El análisis de *c.p.* transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas. Si las variables originales no están correlacionadas o están muy poco correlacionadas esta técnica no tiene ninguna utilidad y la dimensión real de los datos es la misma que el número de variables medidas.

## ¿Cómo decidir cuántas componentes es apropiado considerar?

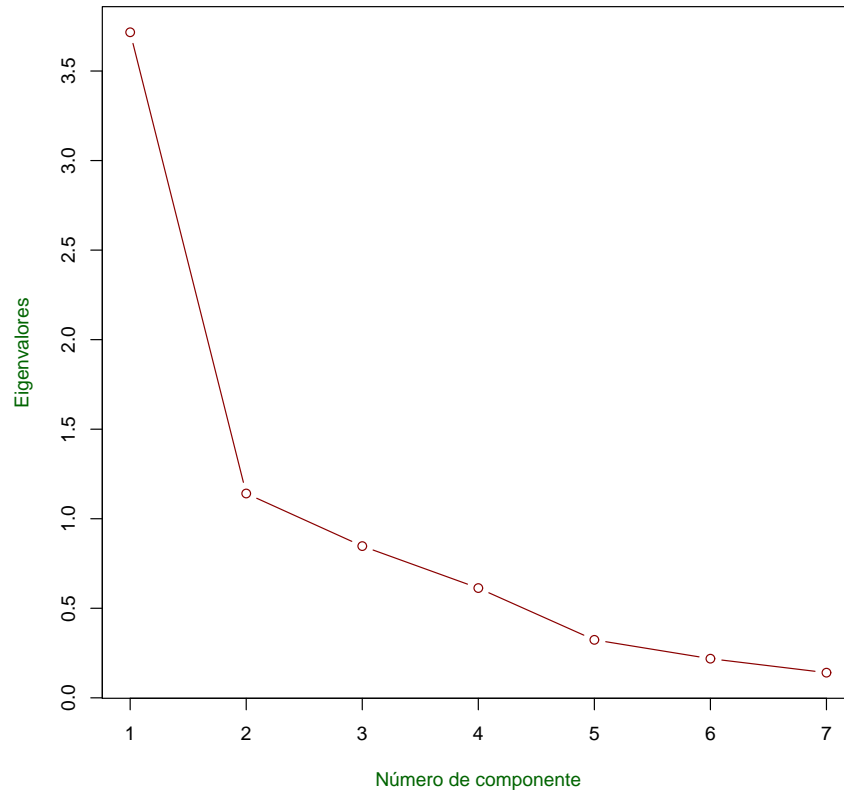
- Porcentaje de varianza explicada requerido (Matriz de varianza-covarianza)
- Porcentaje requerido  $\gamma * 100\%$  de la variabilidad total.

Encontrar el número de componentes que cubra este requerimiento. Este criterio depende de la población bajo estudio y del investigador.

**Gráfica de codo (SCREE).** Cuando los puntos en la gráfica tienden a nivelarse (horizontalmente), los eigenvalores están lo suficientemente cercanos a cero y pueden ignorarse. Entonces, elegir el número de componentes igual al número de eigenvalores antes de que la gráfica se nivele.

Desafortunadamente, mientras más componentes se requiere, menos útiles resultan cada una.

Gráfica de codo



## Matriz de correlación

- Los criterios mostrados para la matriz de varianza-covarianza.
- Uno más. Considerar el número de componentes cuyo eigenvalor sea mayor que uno.

## Puntajes factoriales

Dado que se han generado  $p$  componentes principales a partir de las  $p$  variables originales, es claro que cada uno de los individuos en nuestra matriz de información, tiene asociados *un valor por cada componente principal*, mismo que se calcula de la siguiente manera

$$\mathbf{Z}_i = \mathbf{A}' \mathbf{X}_i, \quad i = 1, 2, \dots, p$$

que proporcionan las coordenadas de la observación  $\mathbf{X}_i$  en el nuevo sistema de ejes generado por las *c.p.*

$$z_{ij} = \mathbf{a}'_j \mathbf{X}_i = \sum_{k=1}^p a_{jk} x_{ik}$$

es el valor de la  $j$ -ésima componente para el  $i$ -ésimo individuo.

Entonces, podemos representar un individuo en el plano, mediante la pareja  $(z_{i1}, z_{i2})$ .

Ya que uno de los usos comunes de esta técnica es identificar individuos similares, es importante tener en cuenta que *las c.p. preservan la distancia entre las observaciones*, como mostraremos en seguida.

Denotemos por  $\mathbf{Z}_i$  : Vector de c.p. del individuo  $\mathbf{X}_i$  y por  $\mathbf{Z}_j$  : Vector de c.p. del individuo  $\mathbf{X}_j$ . Entonces, se trata de mostrar que la distancia entre estas componentes es igual a la distancia entre los vectores originales de los sujetos.



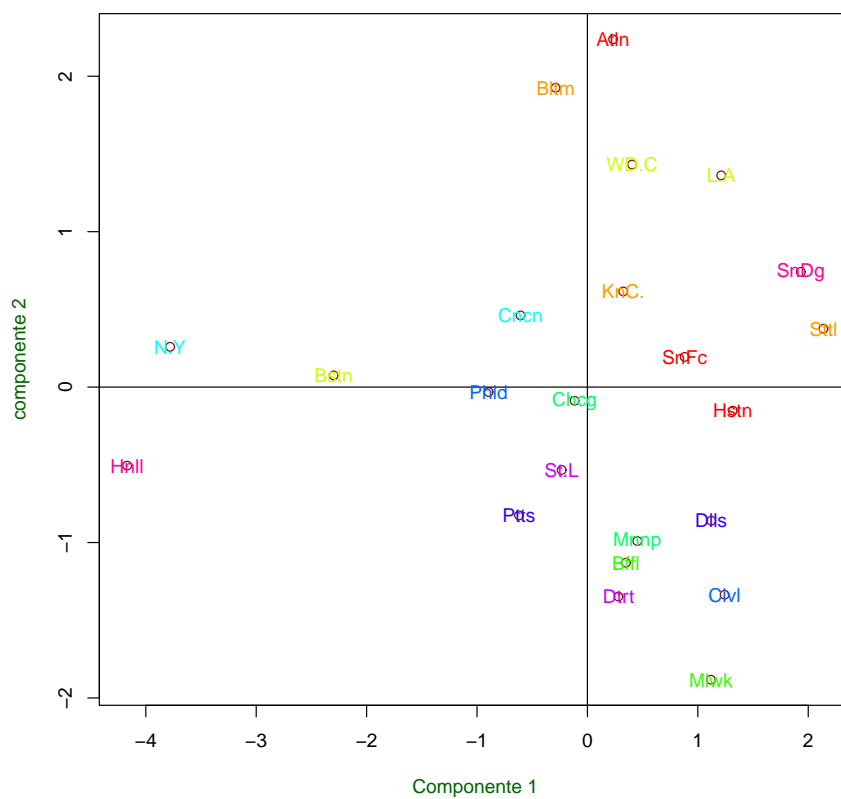
$$\begin{aligned}
\|\mathbf{Z}_i - \mathbf{Z}_j\|^2 &= (\mathbf{Z}_i - \mathbf{Z}_j)' (\mathbf{Z}_i - \mathbf{Z}_j) \\
&= (\mathbf{A}' \mathbf{X}_i - \mathbf{A}' \mathbf{X}_j)' (\mathbf{A}' \mathbf{X}_i - \mathbf{A}' \mathbf{X}_j) \\
&= (\mathbf{A}' (\mathbf{X}_i - \mathbf{X}_j))' (\mathbf{A}' (\mathbf{X}_i - \mathbf{X}_j)) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{A} \mathbf{A}' (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{A} \mathbf{A}^{-1} (\mathbf{X}_i - \mathbf{X}_j) \quad (\mathbf{A} \text{ es ortogonal}) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{I}_p (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)' (\mathbf{X}_i - \mathbf{X}_j) \\
&= \|\mathbf{X}_i - \mathbf{X}_j\|^2
\end{aligned}$$

Observación. Esta distancia se conserva en el espacio original de los vectores, que es de dimensión  $p$ . Si sólo tomamos pocas componentes (2 ó 3) para representar las observaciones, entonces

$$\|\mathbf{X}_i - \mathbf{X}_j\|^2 \approx \|\mathbf{Z}_i^* - \mathbf{Z}_j^*\|^2$$

con  $\mathbf{Z}^*$  un vector de dimensión 2 ó 3, únicamente. Esta aproximación será adecuada si estas pocas dimensiones explican un alto porcentaje de la varianza total de los datos.

Representación gráfica con dos componentes



## Aplicación de c.p. con variables medidas en diversas escalas

El análisis de c.p. se realiza, generalmente, utilizando variables continuas; no obstante, existen aplicaciones donde se presentan diversas escalas de medición en las variables. Una manera generalizada de abordar esta situación, es realizar el análisis ignorando la escala de medición, i.e., suponiendo que todas provienen de una escala de intervalo. En este caso, la correlación entre cualquier par de variables, es la de Pearson. El hecho de no respetar la escala de cada variable, propicia que las correlaciones sean más pequeñas de lo debido, lo que, para una técnica basada en la asociación entre las variables, resulta poco deseable. Otra alternativa es construir variables *dummy's* con las variables medidas en escalas nominal y ordinal. Este procedimiento tiene la desventaja de incrementar el número de variables dentro del análisis (hay que recordar que si una variable nominal u ordinal tiene  $k$  categorías, entonces genera un número igual de variables dummy's). Este incremento de dimensión repercutirá en el hecho de que tendremos menos posibilidades de poder representar nuestros datos en pocas dimensiones, i.e., tendremos poca varianza explicada por unas cuantas dimensiones.

Una forma alternativa de enfrentar este problema, es utilizando la *matriz de correlaciones policóricas*. En esta matriz se utiliza un tipo de correlación de acuerdo a la escala de medición de las dos variables en cuestión. La siguiente tabla muestra las correlaciones que se sugiere calcular.

<b>Escala de medición</b>	Continua	Ordinal	Dicotómica
Continua	Pearson	Policórica	Punto biserial
Ordinal		Policórica	Policórica
Dicotómica			Tetracórica

Una vez calculada esta matriz, el análisis de c.p. se lleva a cabo utilizándola para realizar todos los procesos de cálculo.

# BIPLOTS



Podemos dividir el análisis de datos multivariados en un análisis que se centre en la estructura de asociación entre las variables, y uno basado en las relaciones entre las observaciones (los sujetos). Es deseable tener una técnica que nos permita mostrar las relaciones entre las variables, entre los sujetos y entre ambos. El *biplot* es una representación bidimensional de la matriz de datos  $\mathbf{X}$  en la que tanto los renglones (sujetos) como las columnas (variables) se representan a través de puntos. La representación se basa en la *descomposición en valor singular* de la matriz de datos.

## Descomposición en valor singular

Sea  $\mathbf{X}_{n \times p}$  una matriz. Esta matriz se puede escribir como el producto de una matriz de columnas ortogonales ( $n \times n$ ), una matriz diagonal ( $n \times p$ ) con elementos no negativos y una matriz ortogonal ( $p \times p$ ). En concreto, la descomposición en valor singular es

$$\mathbf{X}_{n \times p} = \mathbf{U}_{n \times n} \Sigma_{n \times p} \mathbf{V}'_{p \times p}$$

Con

- $\mathbf{U}$  es ortogonal, i.e.,  $\mathbf{U}'\mathbf{U} = \mathbf{I}$
- $\mathbf{V}$  es ortogonal, i.e.,  $\mathbf{V}'\mathbf{V} = \mathbf{I}$  y
- $\Sigma$  es diagonal.

Hagamos uso de esta descomposición para representar a los individuos y las variables de nuestros datos. Es claro que para lograr una buena representación de los individuos y de las variables en pocas dimensiones, debemos suponer que podemos reconstruir la matriz de datos considerando sólo unas cuantas dimensiones. En concreto, debemos suponer que

$$\mathbf{X} \approx \sum_{j=1}^q \lambda_j^{1/2} \mathbf{u}_j \mathbf{v}_j' = \mathbf{U}_q \Sigma_q \mathbf{V}_q'$$

para la representación bidimensional, pediríamos  $q = 2$ . Ya que  $\Sigma_q$  es una matriz diagonal, la podemos asociar a la matriz  $\mathbf{U}$  a  $\mathbf{V}$  o a ambas a la vez. Por ejemplo, podemos definir

$$\mathbf{G}_q = \mathbf{U}_q \Sigma_q^{1-c} \quad y \quad \mathbf{H}_q' = \Sigma_q^c \mathbf{V}_q'$$

$0 \leq c \leq 1$ . Para cada valor de  $c$  que elijamos, tenemos

$$\mathbf{X} = \mathbf{G}_q \mathbf{H}_q = \mathbf{U}_q \Sigma_q^{1-c} \Sigma_q^c \mathbf{V}_q'$$

El exponente  $c$  se puede elegir de varias maneras. Las elecciones habituales son  $c = 0$ ,  $c = \frac{1}{2}$  y  $c = 1$

Sea  $\mathbf{g}_i$  el  $i$ -ésimo renglón de  $\mathbf{G}$  y  $\mathbf{h}_j$  el  $j$ -ésimo renglón de  $\mathbf{H}$  (por tanto, la  $j$ -ésima columna de  $\mathbf{H}'$ ). Si  $q=2$ , los  $n+p$  vectores  $\mathbf{g}_i$  y  $\mathbf{h}_j$  pueden representarse en el plano, dando lugar a la representación conocida como *biplot*. Los puntos  $\mathbf{g}_i$  representan observaciones, y los puntos  $\mathbf{h}_j$  representan variables.

Utilizando esta descomposición para construir el biplot, definimos los elementos de la descomposición de  $\mathbf{X}$  como

$$\mathbf{X} = \mathbf{G}\mathbf{H}', \quad \text{con} \quad \mathbf{G} = \mathbf{U} \quad \text{y} \quad \mathbf{H}' = \mathbf{L}\mathbf{A}'$$

Esta definición implica tomar  $c=1$  en la representación general de los biplots. Si denotamos por  $\mathbf{g}'_i, i = 1, 2, \dots, n$  y  $\mathbf{h}'_j, j = 1, 2, \dots, p$  los renglones de  $\mathbf{G}$  y  $\mathbf{H}$ , respectivamente. Entonces, el elemento  $(i,j)$  de  $\mathbf{X}$  se puede escribir como

$$x_{ij} = \mathbf{g}'_i \mathbf{h}_j$$

## Interpretación

Dada la descomposición en valor singular de  $\mathbf{X}$

$$\mathbf{X} = \mathbf{G}\mathbf{H}', \quad \text{con} \quad \mathbf{G} = \mathbf{U} \quad \text{y} \quad \mathbf{H}' = \mathbf{L}\mathbf{A}'$$

los elementos de  $\mathbf{G}$  representan a los individuos con

$$\|\mathbf{g}_i - \mathbf{g}_j\|^2 \propto \delta_{ij}^2 = (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{S}^{-1} (\mathbf{X}_i - \mathbf{X}_j)$$

Los elementos de  $\mathbf{H}$  representan a las variables, con las siguientes características

- $Var(\mathbf{X}_j) = \mathbf{h}'_j \mathbf{h}_j \|h_j\|^2, j=1,2,\dots,p$

- $Cov(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{h}_i' \mathbf{h}_j$
- $Corr(\mathbf{X}_i, \mathbf{X}_j) = \frac{\mathbf{h}_i' \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|}$

Entonces el **Biplot** es una representación gráfica bidimensional de los individuos y las variables, a través de los vectores  $\mathbf{g}$  y  $\mathbf{h}$ , suponiendo que esta representación en dos dimensiones es una buena aproximación. Es decir que

$$x_{ij} \approx g_i^* h_j^*$$

Con  $g^*$  y  $h^*$  vectores en  $\mathbb{R}^2$ . Entonces, el biplot se construye graficando a los individuos como puntos  $\mathbf{g}_i^* = (\ell_1^{1/2} u_{1i}, \ell_2^{1/2} u_{2i})$  y los  $p$  vectores, cuyo punto final se encuentra en  $\mathbf{h}_j^* = (\ell_1^{1/2} a_{1j}, \ell_2^{1/2} a_{2j})$ .

Ahora sí estamos en posibilidad de hacer la interpretación del biplot.

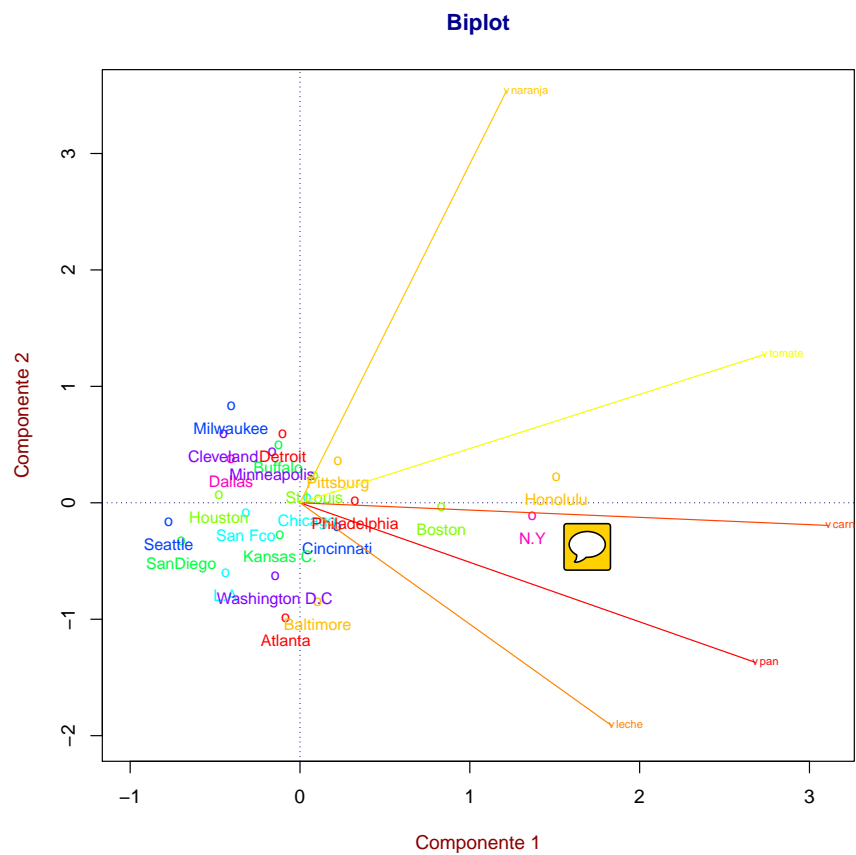
- Individuos semejantes representarán puntos cercanos en la gráfica
- Variables cuyo ángulo entre los vectores que las representan sea pequeño, serán variables con una fuerte correlación, ya que  $\cos(\theta)$  es una función decreciente de  $0^0$  a  $90^0$  y  $\cos(0^0) = 1$  (los vectores son colineales) y  $\cos(90^0) = 0$  (los vectores son ortogonales).

Colineales  $\Rightarrow$  corr=1, ortogonales  $\Rightarrow$  corr=0.

- Finalmente, ya que escribimos a los elementos de la matriz  $\mathbf{X}$  como

$$x_{ij} \approx \mathbf{g}_i' \mathbf{h}_j = \|\mathbf{g}_i\| \|\mathbf{h}_j\| \cos(\theta_{ij})$$

que es la proyección de la observación  $i$  en la variable  $j$ . Para apreciar la magnitud del registro de un individuo en una variable, hay que proyectar el punto que representa al individuo sobre el vector que representa la variable, mientras más pequeña sea esta proyección, más grande será la magnitud del registro del individuo en la variable.



# ANÁLISIS DISCRIMINANTE

## INTRODUCCIÓN

El problema de discriminación o clasificación es habitual en muchas áreas de la actividad humana, que van desde un diagnóstico médico hasta los sistemas que posibilitan la concesión de un crédito bancario o de reconocimiento de falsas obras de arte (pinturas o escritos).

El problema de *discriminar* aparece en muchas situaciones en que es necesario clasificar elementos con información incompleta. Por ejemplo, los sistemas automáticos de concesión de créditos (*credit scoring*) implementados en muchas instituciones financieras o bancarias, deben utilizar algunas variables de los individuos sujetos al crédito, tales como : nivel de ingresos, historial crediticio, antigüedad en el trabajo, patrimonio, edo. civil, etc., para decidir si el sujeto es o no confiable para otorgarle dicho crédito. En ingeniería este problema se conoce con el nombre de reconocimiento de patrones (pattern recognition), para diseñar máquinas capaces de realizar clasificaciones de manera automática. Por ejemplo, reconocer voces y sonidos, clasificar billetes o monedas, reconocer caracteres escritos en una pantalla de una computadora o clasificar cartas según el distrito postal. Otros ejemplos de aplicaciones del análisis discriminante son: asignar la autoría de un texto escrito de procedencia desconocida a uno de entre varios autores por las frecuencias de uso de palabras; asignar una partitura musical o un cuadro a un artista; determinar una declaración de impuestos como potencialmente fraudulenta o no; determinar una empresa como en riesgo de quiebra o no; un paciente como enfermo de cáncer o no; en Biología se presenta en la llamada taxonomía de especies, que consiste en asignar diversos individuos en *taxones*, etc.

El nombre del análisis discriminante como *técnica de clasificación supervisada*, proviene del hecho que conocemos una muestra de elementos bien clasificados (nuestra muestra) que sirve de pauta o modelo para la clasificación de futuras observaciones.

## Planteamiento estadístico del problema

Desde el punto de vista estadístico, el análisis discriminante tiene los siguientes elementos.

- Se dispone de un conjunto de elementos que pueden provenir de dos o más poblaciones distintas.
- En cada elemento se ha observado un vector aleatorio de dimensión  $p$ :  $\mathbf{X} = (x_1, x_2, \dots, x_p)$  de características de los individuos que, suponemos, son potencialmente distintas en las poblaciones,



i.e., pueden ayudar a discriminar entre estas poblaciones.

## Objetivos del análisis discriminante

- **Discriminación:** Describir las características que diferencian a los distintos grupos conocidos de una población. Para encontrar factores discriminantes cuyos valores numéricos sean tales que separen a los grupos lo más posible.
- **Clasificación:** Asignar nuevos sujetos a un grupo, de entre dos o más. Derivar una regla que pueda usarse para asignar de forma *óptima* un individuo a un grupo de los ya conocidos.

**Nota hitórica:** La primera aplicación del análisis discriminante consistió en clasificar los restos de un cráneo descubierto en una excavación, como humano, utilizando la distribución de medidas físicas para los cráneos humanos y los de antropoides (Def. diccionario: Que se parece al ser humano en sus características externas | antropomorfo).

## En resumen

### ¿Qué es el análisis discriminante?

- Es una técnica estadística *de reducción de dimensión*, cuyo objetivo es maximizar la separación entre los datos de  $p \gg 2$  ó 3 dimensiones, cuando se realice esta reducción de dimensión a 2 ó 3.

### ¿Para qué?

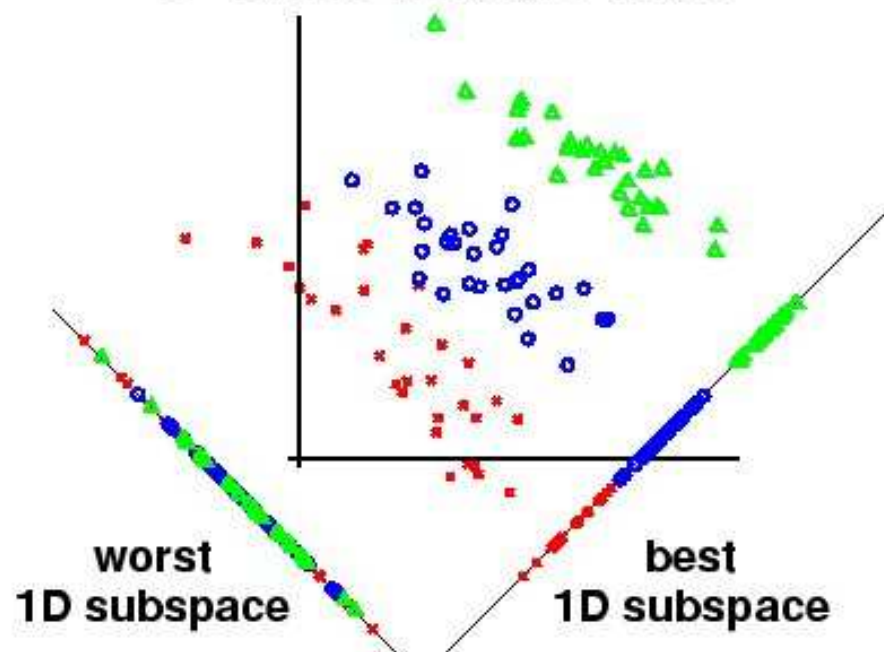
- Identificación de las características de los grupos
- Variables que discriminan entre los grupos
- Selección de las variables discriminantes
- Clasificación de nuevos individuos en los grupos ya existentes

**IMPORTANTE:** Para implementar esta técnica, *los grupos deben estar definidos de antemano*. Esta agrupación podría ser producto de algunos de los métodos multivariados para este fin, como *cluster* o *componentes principales*, producto de una agrupación natural o de la experiencia del usuario.

## Variables canónicas discriminantes: Dos grupos

Mencionamos que el análisis discriminante es una técnica de reducción de dimension, reducción que sabemos se logra proyectando nuestras observaciones, originalmente en dimensión  $p$ , a un espacio de dimensión menor, idealmente 2 ó 3 para poderlas visualizar gráficamente. En este caso, esta proyección debiera ser tal que logre la mayor separación posible de los grupos en el espacio donde se proyectan. El ejemplo de la gráfica siguiente muestra que la elección del plano de proyección (en este caso una recta) no es trivial, por lo que se requieren de elementos técnicos para su determinación.

### 3-class feature data



Podemos observar que la proyección sobre el plano en  $\mathbb{R}^1$ : la línea recta del lado izquierdo de la gráfica, no posibilita la separación de los tres grupos de observaciones. Por el contrario, una proyección de estos datos sobre el plano en  $\mathbb{R}^1$ , representado por la línea recta del lado derecho, logra una muy buena separación de los grupos en este espacio reducido. En este caso de dos grupos, el problema se transforma en elegir, de todas las posibles líneas rectas, aquella que maximice la separación de estas proyecciones, que son valores escalares.

## Las funciones lineales discriminantes de Fisher

La función lineal discriminante para dos grupos fue deducida por primera vez por Fisher, a través de un razonamiento intuitivo. El criterio propuesto por Fisher es encontrar una variable escalar, que sea tal que maximice la distancia entre los datos proyectados.

$$Y = \mathbf{a}' \mathbf{X}$$

Como tenemos sólo dos poblaciones, entonces necesitamos una única función lineal discriminante

$$Y = \mathbf{a}' \mathbf{X} = a_1 X_1 + \dots + a_p X_p$$

Entonces, de manera general, tenemos el siguiente planteamiento.

Dos poblaciones:  $\pi_1$  y  $\pi_2$ , donde cada uno de los individuos que las componen tiene un vector de  $p$  variables medidas  $\mathbf{X}' = (X_1, \dots, X_p)$ , con  $\mathbf{X}_1$  y  $\mathbf{X}_2$ , las matrices de datos de los sujetos en cada uno de los dos grupos, respectivamente.

Entonces, una vez proyectados los datos originales  $\mathbf{X}' = (X_1, \dots, X_p)$  a través de las funciones lineales (combinaciones lineales). Tenemos

- Todos los puntos (sujetos)  $(\mathbf{X}_1, \mathbf{X}_2)$  son proyectados (mapeados) sobre el plano,  $Y$ .
- Hay que elegir a  $Y$  de tal manera que logremos la mayor separación entre los grupos proyectados.

Pero, para encontrar un vector que proporcione una “buena proyección”, en el sentido que digamos, necesitamos definir una medida de separación entre estas proyecciones. Una buena alternativa podría ser elegir como nuestra función objetivo, la distancia entre las medias proyectadas por estas funciones lineales es decir

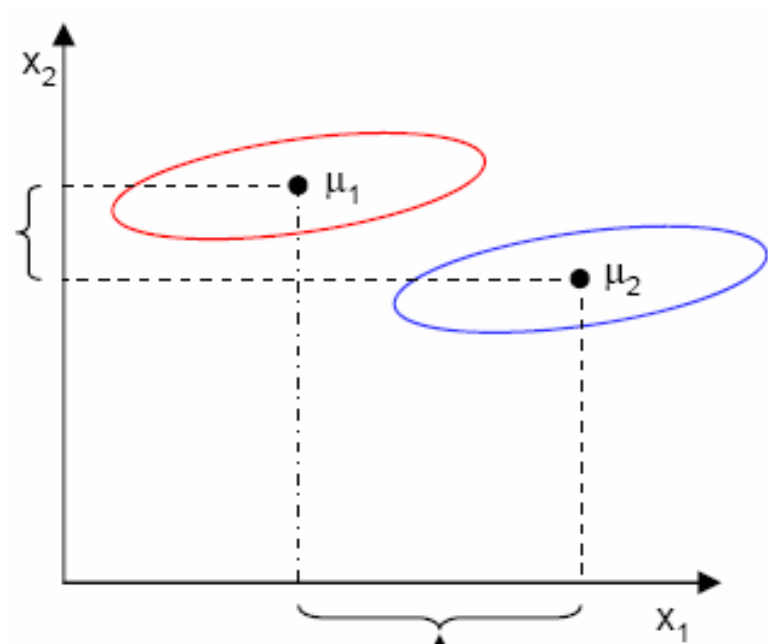
$$J(\mathbf{a}) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |\mathbf{a}' \mu_1 - \mathbf{a}' \mu_2| = |\mathbf{a}' (\mu_1 - \mu_2)|$$

con

$$\mu_1 = \mathbf{E}(\mathbf{X}|\pi_1) : \text{media de } \mathbf{X} \text{ en la población 1}$$

$$\mu_2 = \mathbf{E}(\mathbf{X}|\pi_2) : \text{media de } \mathbf{X} \text{ en la población 2}$$

Sin embargo, la distancia entre las medias proyectadas de cada grupo, no es una muy buena medida, ya que no toma en cuenta la variabilidad (varianza o desviación estándar) dentro de estos grupos. En la gráfica siguiente se muestra que aunque existe mayor separación de las medias proyectando sobre el eje horizontal, se logra una mejor separación de los grupos, proyectando sobre el eje vertical.



La solución propuesta por Fisher para salvar esta dificultad, fue maximizar una función que presente esta diferencia de medias, pero normalizada (escalada) por una medida de la variabilidad dentro de los grupos.

Para cada grupo, esta variabilidad es equivalente a la varianza del grupo proyectado

$$\tilde{S}_i^2 = \sum_{Y \in G_i} (Y - \tilde{\mu}_i)^2, \quad i = 1, 2$$

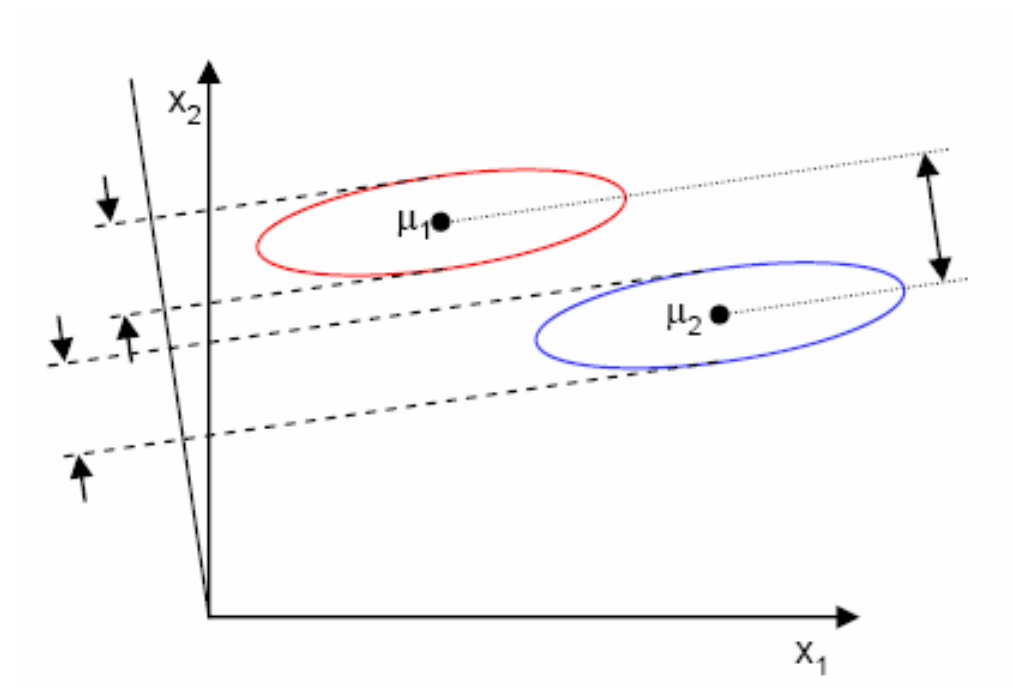
Entonces,  $\tilde{S}_i$  mide la *variabilidad dentro del grupo i* después de que ha sido proyectado en el plano  $Y$ .

Por lo tanto,  $\tilde{S}_1^2 + \tilde{S}_2^2$  mide la variabilidad dentro de los dos grupos, una vez realizada la proyección, denominada *variabilidad dentro de grupos* de las muestras proyectadas.

Entonces, la función lineal discriminante de Fisher, se define como la función lineal:  $Y = \mathbf{a}' \mathbf{X}$  que maximiza la función objetivo

$$J(\mathbf{a}) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$





Para encontrar el vector,  $\mathbf{a}^*$ , que maximice esta expresión, es necesario escribirla de forma explícita como función de  $\mathbf{a}$ .

Definamos la variabilidad en y dentro de los grupos en el espacio original  $\mathbf{X}$ , como

$$S_i = \sum_{x \in G_i} (\mathbf{X} - \mu_i) (\mathbf{X} - \mu_i)', \quad i = 1, 2 \text{ y}$$

$$S_w = S_1 + S_2$$

Donde  $S_i$  es la matriz de varianza-covarianza del grupo  $i$ , y  $S_w$  la matriz de dispersión dentro de grupos.

Ahora, regresemos a estas mismas definiciones, pero con las observaciones proyectadas en el plano  $Y$ . Y tenemos

$$\begin{aligned} \tilde{S}_i^2 &= \sum_{Y \in G_i} (Y - \tilde{\mu}_i)^2 = \sum_{Y \in G_i} (\mathbf{a}' \mathbf{X} - \mathbf{a}' \mu_i)^2 \\ &= \sum_{x \in G_i} \mathbf{a}' (\mathbf{X} - \mu_i) (\mathbf{X} - \mu_i)' \mathbf{a} \\ &= \mathbf{a}' S_i \mathbf{a} \end{aligned}$$

y

$$\begin{aligned} \tilde{S}_1^2 + \tilde{S}_2^2 &= \mathbf{a}' S_1 \mathbf{a} + \mathbf{a}' S_2 \mathbf{a} \\ &= \mathbf{a}' (S_1 + S_2) \mathbf{a} \\ &= \mathbf{a}' S_w \mathbf{a} \\ &= \tilde{S}_w \end{aligned}$$

Con  $\tilde{S}_w$  la matriz de dispersión dentro de grupos proyectados.

De modo similar, las medias proyectadas en el espacio  $Y$ , pueden escribirse en términos de las medias en el espacio original, de la siguiente manera

$$\begin{aligned}
(\tilde{\mu}_1 - \tilde{\mu}_2)^2 &= \left( \mathbf{a}' \mu_1 - \mathbf{a}' \mu_2 \right)^2 \\
&= \mathbf{a}' (\mu_1 - \mu_2) (\mu_1 - \mu_2)' \mathbf{a} \\
&= \mathbf{a}' S_B \mathbf{a} \\
&= \tilde{S}_B
\end{aligned}$$

La matriz  $S_B$  se conoce como *la matriz de dispersión entre los grupos*, mientras que  $\tilde{S}_B$  es la matriz de dispersión entre grupos de las muestras proyectadas.

Ya que  $\tilde{S}_B$  es el producto interno entre dos vectores, es de rango a lo más *uno*.

Finalmente, podemos expresar el criterio de Fisher en términos de las dos matrices de dispersion,  $S_w$  y  $S_B$ , como

$$J(\mathbf{a}) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{S}_1^2 + \tilde{S}_2^2} = \frac{\mathbf{a}' S_B \mathbf{a}}{\mathbf{a}' S_w \mathbf{a}}$$

Una vez planteada la función objetivo, lo que resta es derivarla respecto a  $\mathbf{a}$ , para encontrar el máximo de ella. Este procedimiento se realiza, por supuesto, utilizando técnicas de cálculo vectorial. Realizando el proceso de maximización, encontramos que el vector  $\mathbf{a}^*$  que maximiza esta expresión es

$$\mathbf{a}^* = S_w^{-1} (\mu_1 - \mu_2)$$

## Función lineal discriminante estimada

Para utilizar el discriminante con datos muestrales, es necesario estimar esta función lineal discriminante a través de los datos observados, recordando que estos datos son los que se generan una vez proyectados a través de la función discriminante. Entonces, las matrices que necesitamos son

$$\begin{aligned}
\sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n \left( \mathbf{a}' (y_i - \bar{y}) \right)^2 = \mathbf{a}' \mathbf{T} \mathbf{a} : \text{Suma de cuadrados totales o varianza total, y} \\
\sum_{g=1}^G n_g (\bar{X}_g - \bar{X})^2 &= \sum_{g=1}^G n_g \left( \mathbf{a}' (\bar{y}_g - \bar{y}) \right)^2 = \mathbf{a}' \mathbf{E} \mathbf{a} : \text{Suma de cuadrados entre grupos o varianza entre grupos}
\end{aligned}$$

$\hat{\mu}_1 = \bar{\mathbf{X}}_1, \hat{\mu}_2 = \bar{\mathbf{X}}_2$  Las medias en los grupos 1 y 2, y

$\mathbf{S}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} (X_{1j} - \bar{\mathbf{X}}_1) (X_{1j} - \bar{\mathbf{X}}_1)' , \mathbf{S}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (X_{2j} - \bar{\mathbf{X}}_2) (X_{2j} - \bar{\mathbf{X}}_2)'$ , varianzas muestrales en los grupos

$\mathbf{S} = \mathbf{S}_{pool} = \frac{(n_1 - 1) \mathbf{S}_1 + (n_2 - 1) \mathbf{S}_2}{n_1 + n_2 - 2}$  : La varianza conjunta de los grupos

El supuesto de matrices de varianza-covarianza iguales dentro de los dos grupos, es fundamental y hace que la matriz de varianza-covarianza total se estime como un *pool* de las correspondientes matrices de cada grupo.

Entonces, la función lineal estimada queda como

$$\hat{y} = \hat{\mathbf{a}}' \mathbf{X} = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \mathbf{S}_{pool}^{-1} \mathbf{X}$$

## Urgente: Un ejemplo

49 adultos mayores del sexo masculino participaron en un estudio interdisciplinario sobre su condición humana, y fueron clasificados en dos grupos: “factor senil presente” y “factor senil ausente”, basados en una intensiva evaluación psicológica. Los siguientes resultados son el resultado de cuatro pruebas realizadas a estos sujetos.

	No Senil (n=37)		Senil (n = 12)	
Prueba	$\bar{\mathbf{X}}$	S.D.	$\bar{\mathbf{X}}$	S.D.
Información	12.566	3.387	8.750	3.251
Similaridades	9.486	3.380	5.333	4.271
Aritmética	11.514	3.363	8.500	3.631
Pintura	7.973	1.922	4.750	3.571

$$\mathbf{S}_{Senil} = \begin{pmatrix} 11.47 & 8.55 & 6.39 & 2.07 \\ 8.55 & 11.42 & 5.49 & 0.29 \\ 6.39 & 5.49 & 11.31 & 1.82 \\ 2.07 & 0.29 & 1.82 & 3.69 \end{pmatrix}, \mathbf{S}_{NoSenil} = \begin{pmatrix} 10.57 & 10.45 & 9.68 & 7.66 \\ 10.45 & 18.24 & 12.09 & 8.91 \\ 9.68 & 12.09 & 13.18 & 5.32 \\ 7.66 & 8.91 & 5.32 & 12.75 \end{pmatrix},$$

$$\mathbf{S}_{pool} = \begin{pmatrix} 11.26 & 9.00 & 7.16 & 3.38 \\ 9.00 & 13.02 & 7.04 & 2.31 \\ 7.16 & 7.04 & 11.75 & 2.64 \\ 3.38 & 2.31 & 2.64 & 5.81 \end{pmatrix}$$

$$\begin{aligned}
(\bar{X}_{NoSenil} - \bar{X}_{Senil})' &= (3.82, 4.15, 3.01, 3.23) \\
\mathbf{a}' &= (\bar{X}_{NoSenil} - \bar{X}_{Senil})' \mathbf{S}_{pool}^{-1} \\
&= (3.82, 4.15, 3.01, 3.23) \begin{pmatrix} 0.249 & -0.127 & -0.060 & 0.066 \\ -0.127 & 0.180 & -0.034 & 0.0182 \\ -0.060 & -0.034 & 0.146 & -0.017 \\ 0.066 & 0.0182 & -0.017 & 0.211 \end{pmatrix} \\
&= (0.02453159, 0.2162928, 0.01043125, 0.4510016)
\end{aligned}$$

Las medias de los grupos proyectados son

$$\begin{aligned}
\bar{y}_{NoSenil} &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) \begin{pmatrix} 12.566 \\ 9.486 \\ 11.514 \\ 7.973 \end{pmatrix} = 6.07 \\
\bar{y}_{Senil} &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) \begin{pmatrix} 8.750 \\ 5.333 \\ 8.500 \\ 4.750 \end{pmatrix} = 3.59
\end{aligned}$$

La función lineal discriminante para cada sujeto es

$$\begin{aligned}
y_j = \mathbf{a}' X_j &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) (X_{1j}, X_{2j}, X_{3j}, X_{4j})' \\
&= 0.02X_{1j} + 0.22X_{2j} + 0.01X_{3j} + 0.45X_{4j}
\end{aligned}$$

## Clasificación

Una manera muy simple de utilizar esta función lineal,  $Y$ , para clasificar una nueva observación,  $X_0$ , a alguno de los grupos es

1.- Calcular la proyección en el plano  $Y$ , de esta observación

$$y_0 = (\mu_2 - \mu_1)' S_w^{-1} X_0$$

2.- Encontrar el punto medio de las medias de los grupos proyectadas  $\tilde{\mu}_1$  y  $\tilde{\mu}_2$ .

$$\begin{aligned}
m &= \frac{1}{2} (\tilde{\mu}_1 + \tilde{\mu}_2) \\
&= \frac{1}{2} (\mathbf{a}' \mu_1 + \mathbf{a}' \mu_2) \\
&= \frac{1}{2} (\mu_2 - \mu_1)' S_w^{-1} (\mu_2 + \mu_1)
\end{aligned}$$

3.- Regla de clasificación

Asignar  $X_0$  al grupo 1 ( $\pi_1$ ) si  $y_0 = (\mu_2 - \mu_1)' S_w^{-1} X_0 \geq m$ , y

Asignar  $X_0$  al grupo 2 ( $\pi_2$ ) si  $y_0 = (\mu_2 - \mu_1)' S_w^{-1} X_0 < m$

o bien si

$$y_0 - m = (\mu_2 - \mu_1)' S_w^{-1} X_0 \geq 0 \text{ ó } < 0$$

## Estimación de la regla de clasificación

La regla de clasificación estimada queda como

Asignar  $X_0$  al grupo 1 ( $\pi_1$ ) si  $y_0 = (\bar{X}_2 - \bar{X}_1)' S_{pool}^{-1} X_0 \geq m$ , y

Asignar  $X_0$  al grupo 2 ( $\pi_2$ ) si  $y_0 = (\bar{X}_2 - \bar{X}_1)' S_{pool}^{-1} X_0 < m$

En nuestro caso

$$m = \frac{1}{2} (6.07 + 3.59) = 4.83$$

Entonces, un nuevo individuo se asignaría al grupo: *No senil* si  $y_0$ , su puntaje dado por la proyección de sus valores en el plano  $Y$ , es mayor que 4.83, y se asignaría al grupo *Senil* si es menor a 4.83.

Por ejemplo, ¿a qué grupo asignaríamos a un individuo que tiene el siguiente vector de observaciones:  $X_0 = (8.150, 6.001, 9.050, 4.510)$ ?

Notemos primeramente que este vector está cercano a las medias del grupo *senil*: (8.750, 5.333, 8.500, 4.750), entonces, debería de clasificarse en ese grupo. Calculemos su proyección al plano  $Y$ , i.e., calculemos

$$\begin{aligned} y_0 = \mathbf{a}' X_0 &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) (X_{10}, X_{20}, X_{30}, X_{40})' \\ &= 0.02X_{10} + 0.22X_{20} + 0.01X_{30} + 0.45X_{40} \\ &= (0.02453159, 0.2162928, 0.01043125, 0.4510016) * (8.150, 6.001, 9.050, 4.510) = 3.626326 \end{aligned}$$

Si consideramos ahora un sujeto con valores más cercanos a las medias del grupo *No senil*: (12.566, 9.486, 11.514, 7.973), digamos,  $X_0 = (11.950, 10.00, 10.73, 8.103)$ , debería clasificarse como No senil. Su proyección es: 6.222474; que corrobora nuestra especulación.

## Discriminante clásico

Se denomina discriminante clásico al discriminante que asume poblaciones normales multivariadas para cada uno de los grupos. Es decir, se supone que cada población tiene función de densidad de probabilidad, dada por

$$f_i(\mathbf{X}) = \frac{(2\pi)^{-p/2}}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i) \right\}, \quad i=1,2,\dots,G$$

## Discriminante clásico con dos grupos

En el caso de que tengamos dos grupos con probabilidades a priori de pertenencia a cada uno de ellos  $\pi_1$  y  $\pi_2$ , respectivamente ( $\pi_1 + \pi_2 = 1$ ). Entonces, para clasificar a un nuevo individuo,  $\mathbf{x}_0$ , por ejemplo, al grupo 2, sólo debemos comparar sus densidades, y lo asignamos al grupo 2 si

$$\pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

si las probabilidades iniciales son iguales, entonces lo asignamos a dicho grupo si

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

Bajo el supuesto de que las densidades sean normales de dimensión  $p$ , tenemos

$$\begin{aligned} \frac{\pi_2}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) \right\} &> \frac{\pi_1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) \right\} \implies \\ \log(\pi_2) - \frac{1}{2} (\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) &> \log(\pi_1) - \frac{1}{2} (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) \\ (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) &> (\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) - 2 \log \left( \frac{\pi_2}{\pi_1} \right) \end{aligned}$$

si denotamos como  $D_i^2$  el cuadrado de la distancia de Mahalanobis entre el punto observado,  $\mathbf{x}$ , y la media de la población  $i=1,2$ , tenemos

$$D_i^2 = (\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i)$$

si suponemos probabilidades iniciales iguales, entonces la regla que se obtiene para clasificar  $\mathbf{x}$  en el grupo 2 es: Clasificar esta observación en el grupo 2 si



$$D_1^2 > D_2^2$$

es decir, clasificar la observación en el grupo cuyas medias estén más próximas, según la distancia de Mahalanobis cuadrada.

### Interpretación de la regla anterior

Desarrollemos las siguientes expresiones

$$\begin{aligned} (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) &= \mathbf{x}' \Sigma^{-1} \mathbf{x} - 2\mu_1' \Sigma^{-1} \mathbf{x} + \mu_1' \Sigma^{-1} \mu_1, \text{ y} \\ (\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) &= \mathbf{x}' \Sigma^{-1} \mathbf{x} - 2\mu_2' \Sigma^{-1} \mathbf{x} + \mu_2' \Sigma^{-1} \mu_2 \end{aligned}$$

entonces, la regla divide al conjunto de valores posibles de  $\mathbf{x}$ , en dos regiones cuya frontera es

$$-2\mu_1' \Sigma^{-1} \mathbf{x} + \mu_1' \Sigma^{-1} \mu_1 = -2\mu_2' \Sigma^{-1} \mathbf{x} + \mu_2' \Sigma^{-1} \mu_2$$

que es equivalente, como función de  $\mathbf{x}$  a

$$(\mu_2 - \mu_1)' \Sigma^{-1} \mathbf{x} = (\mu_2 - \mu_1)' \Sigma^{-1} \left( \frac{\mu_2 + \mu_1}{2} \right)$$

Observemos que el hecho de suponer matriz de varianzas covarianzas iguales entre los grupos, permite el agrupamiento de los términos de esta manera. Si denotamos por

$$\mathbf{a}' = (\mu_2 - \mu_1)' \Sigma^{-1}$$

entonces, la frontera entre las dos regiones de clasificación para  $\pi_1$  y  $\pi_2$  puede escribirse como

$$\mathbf{a}' \mathbf{x} = \mathbf{a}' \left( \frac{\mu_2 + \mu_1}{2} \right)$$

que es la ecuación de un hiperplano. También equivalente a

$$\begin{aligned} 2\mathbf{a}' \mathbf{x} &= \mathbf{a}' (\mu_1 + \mu_2) \\ \mathbf{a}' \mathbf{x} - \mathbf{a}' \mu_1 &= \mathbf{a}' \mu_2 - \mathbf{a}' \mathbf{x} (*) \end{aligned}$$

Se puede demostrar que esta regla equivale a proyectar el punto  $\mathbf{x}$  que queremos clasificar y las medias de ambas poblaciones sobre la función lineal discriminante, y después asignar el punto a aquella población de cuya media se encuentre más próxima en la proyección. Situación que habíamos visto anteriormente.

Esta última ecuación indica que el procedimiento para clasificar un elemento  $X_0$  puede resumirse como sigue:

- Calcular el vector  $\mathbf{a}'$ , mediante la expresión correspondiente
- Construir la función lineal discriminante

$$Y = \mathbf{a}' \mathbf{X} = a_1 X_1 + \dots + a_p X_p$$

- Calcular la proyección en el plano  $Y$ ,  $Y_0 = \mathbf{a}' X_0$ , del individuo  $X_0 = (X_{10}, \dots, X_{p0})$ , y el valor de las medias proyectadas de las poblaciones,  $\tilde{\mu}_i = \mathbf{a}' \mu_i$ . Clasificar esta observación en aquella población donde la distancia,  $|Y_0 - \tilde{\mu}_i|$ , sea mínima.

Obsérvese que

$$\mathbb{E}(Y|\pi_i) = \tilde{\mu}_i = \mathbf{a}' \mu_i, \quad i = 1, 2$$

Entonces, la regla de decisión que se desprende de (\*), equivale a clasificar la observación en el grupo  $\pi_2$ , sí

$$|Y - \tilde{\mu}_1| > |Y - \tilde{\mu}_2|$$

Esta variable aleatoria  $Y$  tiene varianza dada por

$$\begin{aligned} \mathbb{V}(Y) &= \mathbb{V}(\mathbf{a}' \mathbf{X}) = \mathbf{a}' \mathbb{V}(\mathbf{X}) \mathbf{a} = \mathbf{a}' \Sigma \mathbf{a} = (\mu_2 - \mu_1)' \Sigma^{-1} \Sigma \Sigma^{-1} (\mu_2 - \mu_1) \\ &= (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1) = D^2 \end{aligned}$$

y el cuadrado de la distancia que es un escalar, entre las medias proyectadas es la distancia de Mahalanobis entre los vectores de medias originales:

$$(\tilde{\mu}_2 - \tilde{\mu}_1)^2 = \left( \mathbf{a}' (\mu_2 - \mu_1) \right)^2 = (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1) = D^2$$

## Funciones lineales discriminantes para varios grupos

El enfoque de Fisher puede generalizarse para encontrar las funciones lineales que tengan máximo poder discriminante para clasificar nuevos elementos entre  $G > 2$  poblaciones. La manera de hacerlo es semejante al caso de dos grupos, sólo que ahora se tienen  $k=\min(G-1,p)$  funciones discriminantes. Es decir

$$\begin{aligned}Y_1 &= \mathbf{a}'_1 X \\Y_2 &= \mathbf{a}'_2 X \\&\vdots \\Y_k &= \mathbf{a}'_k X\end{aligned}$$

Entonces, en este caso, el proceso de clasificación es como sigue:

- Proyectamos las medias de cada grupo. Esto es, obtenemos

$$\tilde{\mu}_i = \mathbf{E}(Y|\pi_i) = \mathbf{E}(\mathbf{a}'\mathbf{X}|\pi_i) = \mathbf{a}'\mathbf{E}(\mathbf{X}|\pi_i) = \mathbf{a}'\mu_i \quad i=1,2,\dots,G$$

- Proyectamos el vector de covariables del sujeto a clasificar,  $X_0$ , y obtenemos  $y_0$  su proyección sobre el espacio  $Y$ .
- Clasificamos el punto en aquella población de cuya media se encuentre más cercana.

Las distancias se miden con la distancia euclídeana en el espacio de las variables canónicas,  $y$ . Es decir, clasificaremos al sujeto,  $X_0$ , en la población  $i$  si:

$$(y_0 - \tilde{\mu}_i)'(y_0 - \tilde{\mu}_i) = \min_g (y_0 - \tilde{\mu}_g)'(y_0 - \tilde{\mu}_g)$$

Como tenemos varios grupos, la separación entre las medias la mediremos por el cociente entre la variabilidad entre grupos, y la variabilidad dentro de los grupos. Este es el criterio habitual para comparar varias medias en el análisis de la varianza y genera el estadístico *F de Fisher*. De hecho, lo que estamos haciendo es plantear un análisis de varianza en el espacio de proyección,  $Y$ .

Nuevamente, para obtener las variables lineales discriminantes, comenzamos buscando un vector de proyección,  $\mathbf{a}$ , de norma uno, tal que los grupos de observaciones proyectados sobre él tengan separación relativa máxima. La proyección de la media de las observaciones del grupo  $g$  en esta dirección corresponde al escalar:

$$\bar{\mu}_g = \mathbf{a}' \bar{\mathbf{X}}_g$$

Con la correspondiente proyección para la media de todos los datos, dada por

$$\bar{\mu} = \mathbf{a}' \bar{\mathbf{X}}$$

ambas medias proyectadas son vectores de dimensión  $p \times 1$ , sólo que la primera es para los individuos en el grupo  $g$ ,  $g=1,2,\dots,k$ , y la segunda es para todos los datos, sin importar la pertenencia a algún grupo.

Entonces, tomando como medida de la distancia entre las medias de los grupos proyectadas:  $\bar{\mu}_1, \bar{\mu}_2, \dots, \bar{\mu}_k$ , la varianza total dentro de grupos es

$$\sum_{g=1}^k n_g (\bar{\mu}_g - \bar{\mu})^2$$

que debemos comparar contra la varianza dentro de grupos o variabilidad total, dada por

$$\sum_i \sum_g (y_{ig} - \bar{\mu}_g)^2$$

El proceso para encontrar las funciones lineales se realiza mediante el cociente de las varianzas entre grupos y total (idéntico al procedimiento ANOVA, sólo que aquí estas varianzas se obtienen con los elementos proyectados).

$$\frac{\sum_{g=1}^k n_g (\bar{\mu}_g - \bar{\mu})^2}{\sum_i \sum_g (y_{ig} - \bar{\mu}_g)^2}$$

Ahora, expresemos este criterio en función de los datos originales. La suma de cuadrados *dentro de grupos*, para los puntos proyectados, es:

$$\sum_{i=1}^{n_g} \sum_{g=1}^k (y_{ig} - \bar{\mu}_g)^2 = \sum_{i=1}^{n_g} \sum_{g=1}^k \mathbf{a}' (\mathbf{X}_{ig} - \bar{X}_g) (\mathbf{X}_{ig} - \bar{X}_g)' \mathbf{a} = \mathbf{a}' \mathbf{W} \mathbf{a}$$

con  $\mathbf{W}$ , dada por

$$\sum_{i=1}^{n_g} \sum_{g=1}^k (\mathbf{X}_{ig} - \bar{X}_g) (\mathbf{X}_{ig} - \bar{X}_g)'$$

Esta matriz tiene dimensiones  $p \times p$  y, en general, es de rango  $p$ , asumiendo que  $n - k \geq p$ . Estima la variabilidad de los datos respecto a las medias de su grupo.

Por otro lado, la suma de cuadrados *entre grupos*, para los puntos proyectados está dada por

$$\sum_{g=1}^k n_g (\bar{\mu}_g - \bar{\mu})^2 = \sum_{g=1}^k n_g \mathbf{a}' (\bar{X}_g - \bar{X}) (\bar{X}_g - \bar{X})' \mathbf{a} = \mathbf{a}' \mathbf{B} \mathbf{a}$$

Es decir, la matriz  $\mathbf{W}$  corresponde a las diferencias dentro de grupos (withing) y la matriz  $\mathbf{B}$  las diferencias entre grupos (between).

Entonces, la cantidad a maximizar para encontrar las funciones lineales discriminantes es

$$\mathbf{J} = \frac{\mathbf{a}' \mathbf{B} \mathbf{a}}{\mathbf{a}' \mathbf{W} \mathbf{a}}$$

Realizando el proceso usual, tenemos

$$\frac{2\mathbf{B}\mathbf{a}(\mathbf{a}'\mathbf{W}\mathbf{a}) - (\mathbf{a}'\mathbf{B}\mathbf{a})\mathbf{W}\mathbf{a}}{(\mathbf{a}'\mathbf{W}\mathbf{a})^2} = 0$$

$$\mathbf{B}\mathbf{a} = \mathbf{W}\mathbf{a} \frac{(\mathbf{a}'\mathbf{B}\mathbf{a})}{(\mathbf{a}'\mathbf{W}\mathbf{a})}$$

$$\mathbf{B}\mathbf{a} = \mathbf{J}\mathbf{W}\mathbf{a}$$

Suponiendo que  $\mathbf{W}$  tiene inversa, i.e., es no singular, y observando que  $\mathbf{J}$  es un escalar, que denotaremos como  $\lambda$ , entonces, obtenemos el sistema

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$$

lo que implica que  $\mathbf{a}$  debe ser un vector propio de la matriz  $\mathbf{W}^{-1}\mathbf{B}$  y  $\lambda$  su valor propio asociado. Como el objetivo es maximizar  $\lambda = \mathbf{J}$ , que corresponde a la versión de la ANOVA en el espacio de proyección,  $Y$ , entonces  $\mathbf{a}$  debe ser el vector propio asociado al valor propio más grande de la matriz  $\mathbf{W}^{-1}\mathbf{B}$ , que llamemos  $\mathbf{a}_1$ . Con este vector construiríamos la primer función lineal discriminante

$$Y_1 = \mathbf{a}_1' \mathbf{X}$$

Por construcción, esta función discriminante debe tener el mayor poder para discriminar entre los grupos. La segunda de estas funciones debe tener el mayor poder de discriminación restante, una vez construida la primer función discriminante, y debe ser ortogonal a la primera

$$Y_2 = \mathbf{a}_2' \mathbf{X}, \quad Y_1 \perp Y_2 \Rightarrow \mathbf{a}_1 \perp \mathbf{a}_2$$

de forma análoga a la construcción de la primer función discriminante, se puede demostrar que el poder de discriminación de esta segunda función se maximiza si  $\mathbf{a}_2$  es el correspondiente vector propio asociado al segundo valor propio más grande de la matriz  $\mathbf{W}^{-1}\mathbf{B}$ . En general, se tiene que si  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$  son los vectores propios de  $\mathbf{W}^{-1}\mathbf{B}$ , asociados a los valores propios  $\lambda_1, \lambda_2, \dots, \lambda_k$ , con  $\lambda_1 > \lambda_2 > \dots > \lambda_k$ , entonces las funciones lineales

$$Y_i = \mathbf{a}_i' \mathbf{X}, \quad i = 1, 2, \dots, k$$

proporcionan máxima separación entre los  $G$  grupos proyectados. Además son ortogonales entre ellas.

## Estimación

Supongamos que

- $\mathbf{X}_i$  es una matriz de datos  $n_i \times p$  del grupo  $i=1, \dots, G$

- $\bar{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij}$  que es un estimador de  $\mu_i$

- $\mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i) (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)'$  y

- $\bar{\mathbf{X}} = \left( \frac{1}{\sum_i n_i} \right) \sum_{i=1}^G n_i \bar{\mathbf{X}}_i = \left( \frac{1}{\sum_i n_i} \right) \sum_{i=1}^G \sum_{j=1}^{n_i} \mathbf{X}_{ij}$  que estima a  $\bar{\mu}$
- $\mathbf{S}_{pool} = \sum_{g=1}^G \frac{n_g - 1}{n - G} \mathbf{S}_g$  es la matriz de covarianza común a los  $G$  grupos.

La estimación de  $\mathbf{B}$ , la correspondiente versión muestral de la suma de cuadrados entre grupos, es

$$\hat{\mathbf{B}} = \sum_{i=1}^G (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}) (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})'$$

con la correspondiente estimación de  $\mathbf{W}$ , la suma de cuadrados dentro de grupos, dada por

$$\hat{\mathbf{W}} = \sum_{i=1}^G \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i) (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)'$$

## Discriminante clásico para $G > 2$ grupos

Nuevamente, la idea para generalizar el procedimiento a  $G$  poblaciones normales es similar al anterior con dos poblaciones. En este caso, asignaremos el sujeto con covariables  $\mathbf{X}$  al grupo  $g = 1, 2, \dots, G$  sii

$$\pi_g f_g(\mathbf{x}) > \pi_j f_j(\mathbf{x}) \quad \forall g \neq j \quad g, j = 1, 2, \dots, G$$

Si las probabilidades a priori de pertenencia a cada grupo son iguales, y las matrices de varianza y covarianza son iguales entre los grupos, la condición anterior es equivalente a calcular la distancia de Mahalanobis del punto observado,  $\mathbf{X}$ , al centriode (vector de medias) de cada población y clasificarlo en la población que haga mínima esta distancia. Al realizar el proceso que es semejante al caso de dos grupos obtenemos

Las funciones lineales discriminantes tienen las siguientes características:

- $Y_1$  es la combinación lineal que proporciona el mayor poder de discriminación entre los grupos, y está asociada al valor característico más grande de  $\mathbf{W}^{-1}\mathbf{E}$ .
- $Y_2$  es la combinación lineal que proporciona el mayor poder discriminador entre los grupos, después de  $Y_1$ , y es *ortogonal* a  $Y_1$ . Esta función está asociada con el segundo valor característico más grande de  $\mathbf{W}^{-1}\mathbf{E}$ .

Y así sucesivamente. El número máximo de funciones que se puede construir es

$$k = \min(G-1, p).$$

En un proceso de análisis multivariado que lleva inmersa una reducción de dimensión, es muy importante determinar qué tan bien se reproducen los datos en las pocas dimensiones que se consideren para realizar su análisis. Una de las medidas más comunes para determinar lo adecuado de esta reducción de dimension, es el total de varianza explicada por las funciones lineales discriminantes. Una buena representación se logra si la varianza retenida por estas pocas dimensiones está cercana al 100 %. El total de la variata explicada por las primeras  $m \leq k$  funciones lineales discriminantes es:

$$\sum_{i=1}^m \lambda_i \quad \text{y}$$



$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^k \lambda_i} \times 100 \% \quad \text{Porcentaje que explican las primeras } m \text{ funciones discriminantes}$$

## Análisis de las funciones lineales discriminantes

**Centroides de los grupos.** La media de los puntajes que arrojen las evaluaciones de cada individuo en estas funciones lineales discriminantes. Debería ser una medida inicial de qué tan separados están los grupos *proyectados*. Si la discriminación es buena, deberíamos observar centroides muy alejados uno de otro.

**$\Lambda$  de Wilks.** Esta estadística sirve para determinar el poder discriminante de cada una de las funciones discriminantes. Se determina de forma secuencial el número de estas funciones que debemos considerar, a través de la estadística

$$\Lambda = \frac{\text{Suma de cuadrados dentro de grupos}}{\text{Suma de cuadrados totales}} = \frac{|\mathbf{B}|}{|\mathbf{T}|} = \frac{|\mathbf{B}|}{|\mathbf{W} + \mathbf{B}|}$$

Si la discriminación lograda es buena, entonces la varianza dentro de los grupos será pequeña y la varianza entre grupos será grande. Por lo tanto,  $\Lambda$  estará cercana a cero.

Para este fin, es preferible utilizar el estadístico  $\mathbf{V}$  de Barlett, que es una función de  $\Lambda$  y tiene distribución asintótica  $\chi^2$ . El procedimiento es, inicialmente, considerar sólo una función discriminante, realizar la prueba y, si ésta es significativa, querrá decir que es pertinente la incorporación de otra función discriminante, de lo contrario, será indicativo de que con el número actual de funciones se tiene el máximo poder discriminante; equivalentemente, que la inclusión de otra función no aporta nada a la discriminación entre los grupos.

Las hipótesis a probar mediante este procedimiento son

$\mathbf{H}_0$  :  $k$  funciones lineales son suficientes para discriminar vs.

$\mathbf{H}_a$  : son necesarias más de  $k$  funciones  $k = 1, 2, \dots, \min(G - 1, p)$

La manera de determinar la *importancia relativa de las variables dentro de las funciones discriminantes*, es a través de sus coeficientes estandarizados. La razón es que éstos ya están libres de unidades y son comparables. *La variable que posea el coeficiente estandarizado más grande en valor absoluto*, será la que tiene un *poder discriminante mayor*.

### Coeficientes de correlación o de estructura

$Corr(X_i, \mathbf{Y}_g)$ : Correlación lineal entre cada una de las variables y cada una de las funciones lineales. Si esta correlación es grande (cercana a uno en valor absoluto) indica una relación lineal fuerte entre la variable y la función, por tanto, la variable tiene una contribución importante para discriminar entre los grupos. Si está cercana a cero, no tiene poder discriminatorio entre los grupos.

**Tasa de error de clasificación:** Un elemento muy importante, que determina qué tan bien clasifica nuestro discriminante a las observaciones en la población, es la *tasa de error de clasificación*. Si las covariables utilizadas realmente discriminan a los grupos en la población, esta tasa debe ser pequeña, de lo contrario, será grande y concluiremos que las variables utilizadas, no tienen poder de discriminación entre los grupos en la población.

Cuando se hace esta clasificación con la misma muestra que se utilizó para construir el discriminante, generalmente se logra una tasa de error de clasificación “artificialmente” baja. Una forma más honesta de calcular esta tasa, es a través de la llamada *clasificación cruzada*, que no es más que eliminar uno por uno a las observaciones en la muestra, y utilizar el discriminante para asignarlas a algunos de los grupos; por lo regular, este procedimiento genera tasas de error más elevadas, pero más realistas.

## Discriminante cuadrático

Supongamos que las poblaciones son normales, pero que, como ocurre regularmente, no existe igualdad de varianzas; en el caso de dos grupos,  $\Sigma_1 \neq \Sigma_2$ . Entonces, la regla de clasificación, bajo el supuesto de probabilidades a priori iguales, es:

$$\mathbb{Q}(\mathbf{X}) = \frac{1}{2} (\Sigma_2^{-1} - \Sigma_1^{-1}) \mathbf{X} + \mathbf{X}' (\Sigma_2^{-1} \mu_1 - \Sigma_1^{-1} \mu_2) + \frac{1}{2} \mu_2' \Sigma_2^{-1} \mu_2 - \frac{1}{2} \mu_1' \Sigma_1^{-1} \mu_1 + \frac{1}{2} \log |\Sigma_2| - \frac{1}{2} \log |\Sigma_1|$$

Observemos que el término  $\mu_i' \Sigma_i^{-1} \mu_i$ ,  $i = 1, 2$ , no puede cancelarse y origina términos de grado 2, ya sean cuadráticos o cruzados, lo que justifica el nombre de discriminante cuadrático.

Esta regla es equivalente a asignar a un individuo  $\mathbf{X}_0$  al grupo donde se minimice la función

$$\min_{j \in (1,2)} \left[ \frac{1}{2} \log |\Sigma_j| + \frac{1}{2} (\mathbf{X}_0 - \mu_j)' \Sigma_j^{-1} (\mathbf{X}_0 - \mu_j) \right]$$

Para el caso de  $G > 2$  grupos, y suponiendo que las matrices de varianza-covarianza no son iguales, la regla se extiende trivialmente como: asignar a un individuo  $\mathbf{X}_0$  al grupo donde se minimice la función

$$\min_{j \in (1, \dots, G)} \left[ \frac{1}{2} \log |\Sigma_j| + \frac{1}{2} (\mathbf{X}_0 - \mu_j)' \Sigma_j^{-1} (\mathbf{X}_0 - \mu_j) \right]$$