Resultados de la encuesta realizada en el proyecto

PI2_12_035

La Estadística como herramienta de trabajo en Ciencias del Mar y CC Ambientales: relación interdisciplinar y desarrollo de material didáctico

Julio 2012

Universidad de Cádiz

1. Preámbulo

Para conocer las experiencias y opiniones de los alumnos en este proyecto, se ha realizado una encuesta, cuyos resultados recoge el presente informe.

La encuesta ha sido enviada por correo electrónico a los alumnos de aquellas asignaturas que se recogían en este proyecto, cuyo listado se consiguió a través del campus virtual de dichas asignaturas, que incluía 257 registros.

Una vez cumplimentadas las encuestas, éstas se añadían, de forma anónima, en sendos registros de una base de datos para su posterior tratamiento. A aquellas personas que no cumplimentaron la encuesta, en una primera instancia, se les remitió hasta un máximo de dos recordatorios en las dos semanas siguientes del envío inicial.

En todo el proceso se han garantizado los preceptos recogidos en la Ley Orgánica de Protección de Datos (LOPD).

2. Descripción del cuestionario

El cuestionario se divide en varios bloques, donde las preguntas han sido codificadas en una escala de Likert $1-5$, correspondiéndose el valor 1 con un nivel mínimo y el 5 con un nivel máximo de satisfacción.

El cuestionario es el siguiente:

Aspectos generales de organización	1	2	3	4	5
Instalaciones donde se desarrollan las clases	O	O	O	O	O
Duración de las clases	O	O	O	O	O
Calendario y Horario de la asignatura	O	O	O	O	O
Docentes de la asignatura	1	2	3	4	5
Dominio de la materia	O	O	O	O	O
Claridad en la comunicación	O	O	O	O	O
Aspectos globales	1	2	3	4	5
Guiones de trabajo	O	O	O	O	O
Medios y recursos disponibles	O	O	O	O	O
Datos y problemas analizados relacionados con la titulación	O	O	O	O	O
Utilidad de la asignatura para formación académica	O	O	O	O	O
Utilidad asignatura para la formación investigadora	O	O	O	O	O
Valoración general de la asignatura	O	O	O	O	O
Desarrollo de la asignatura	1	2	3	4	5
Método expositivo	O	O	O	O	O
Estudio de casos	O	O	O	O	O
Resolución de ejercicios	O	O	O	O	O
Aprendizaje basado en problemas	O	O	O	O	O
Proyectos tutorizados	O	O	O	O	O
Participación del grupo	1	2	3	4	5
Constancia en la asignatura	O	O	O	O	O
Implicación y motivación de los participantes	O	O	O	O	O

En relación a las aplicaciones informáticas, todas las herramientas empleadas tienen licencia libre (GNU-GPL):

- La plataforma usada ha sido apache sobre ubuntu-linux
- La edición y elaboración de informes se ha realizado con LaTeX
- La herramienta elegida para la administración automatizada del cuestionario ha sido `limesurvey`
- El análisis de datos se ha hecho con el software estadístico `R`

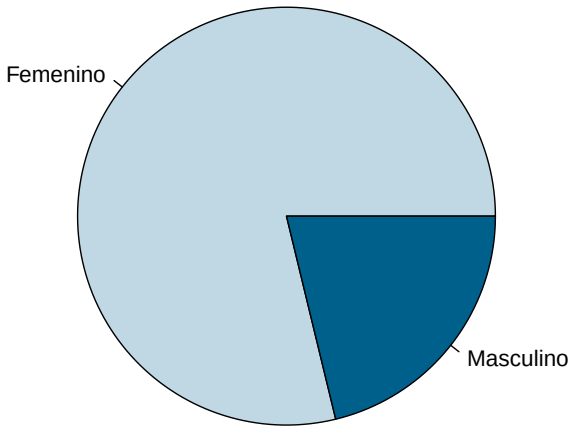
La encuesta ha sido cumplimentada en su totalidad por un total de 33 personas, lo que representa un 12.84 % del total de 257 posibles participantes. Ninguno declinó explícitamente participar en el estudio.

3. Caracterización de perfiles de los que han contestado la encuesta

En este apartado se caracteriza a los encuestados en función del sexo y de la titulación a la que pertenecen.

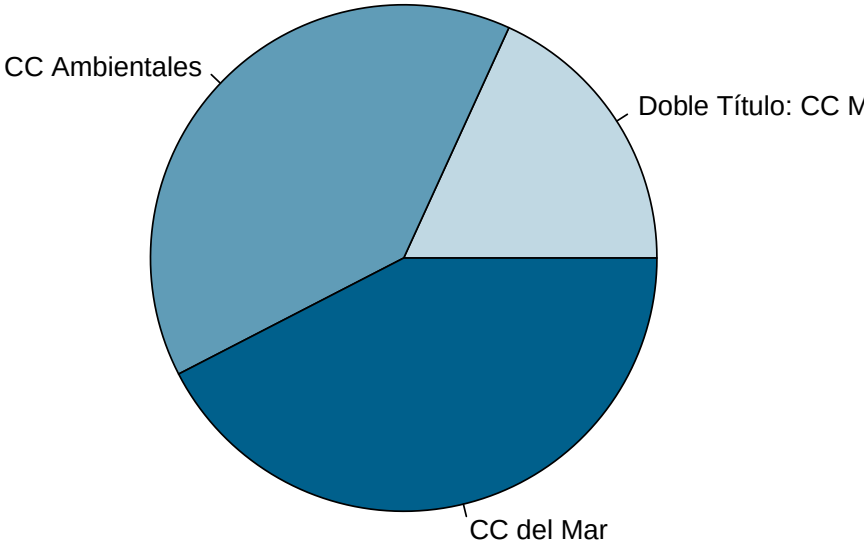
Sexo

	n	%
Femenino	26	78.8
Masculino	7	21.2



Titulación

	n	%
Doble título: Ciencias del Mar y Ambientales	6	18.2
Grado en Ciencias Ambientales	13	39.4
Grado en Ciencias del Mar	14	42.4



3.1. Resultados del cuestionario

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	33	4.18	0.95
Duración de las clases	33	3.88	1.14
Calendario y Horario de la asignatura	33	3.55	1.30
Docentes de la asignatura			
Dominio de la materia	33	4.27	1.07
Claridad en la comunicación	33	3.58	1.15
Aspectos globales			
Guiones de trabajo	33	3.39	1.25
Medios y recursos disponibles	33	3.88	1.02
Datos y problemas analizados relacionados con la titulación	33	3.67	1.31
Utilidad de la asignatura para formación académica	33	3.58	1.44
Utilidad asignatura para la formación investigadora	32	3.81	1.20
Valoración general de la asignatura	33	3.52	1.15
Desarrollo de la asignatura			
Método expositivo	33	3.64	1.37
Estudio de casos	32	3.47	1.16
Resolución de ejercicios	33	3.18	1.38
Aprendizaje basado en problemas	33	3.03	1.40
Proyectos tutorizados	32	3.41	1.32
Participación del grupo			
Constancia en la asignatura	33	3.76	1.15
Implicación y motivación de los participantes	33	3.18	1.33

4. Resultados según titulación

4.1. Doble título: Ciencias del Mar y Ambientales

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	6	4.00	1.10
Duración de las clases	6	3.17	1.47
Calendario y Horario de la asignatura	6	3.33	1.63
Docentes de la asignatura			
Dominio de la materia	6	4.67	0.82
Claridad en la comunicación	6	3.83	1.33
Aspectos globales			
Guiones de trabajo	6	3.00	1.26
Medios y recursos disponibles	6	4.17	0.98
Datos y problemas analizados relacionados con la titulación	6	3.83	1.33
Utilidad de la asignatura para formación académica	6	4.17	1.33
Utilidad asignatura para la formación investigadora	6	4.67	0.52
Valoración general de la asignatura	6	3.83	0.75
Desarrollo de la asignatura			
Método expositivo	6	4.17	1.17
Estudio de casos	6	4.00	1.26
Resolución de ejercicios	6	3.50	1.22
Aprendizaje basado en problemas	6	3.00	1.67
Proyectos tutorizados	6	4.33	0.82
Participación del grupo			
Constancia en la asignatura	6	4.67	0.82
Implicación y motivación de los participantes	6	3.83	1.47

4.2. Grado en Ciencias Ambientales

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	13	4.15	1.21
Duración de las clases	13	3.92	1.32
Calendario y Horario de la asignatura	13	3.31	1.49
Docentes de la asignatura			
Dominio de la materia	13	4.38	1.26
Claridad en la comunicación	13	3.77	1.30
Aspectos globales			
Guiones de trabajo	13	3.85	1.34
Medios y recursos disponibles	13	4.00	1.15
Datos y problemas analizados relacionados con la titulación	13	3.85	1.34
Utilidad de la asignatura para formación académica	13	3.69	1.60
Utilidad asignatura para la formación investigadora	12	3.67	1.37
Valoración general de la asignatura	13	3.69	1.32
Desarrollo de la asignatura			
Método expositivo	13	4.00	1.22
Estudio de casos	12	3.42	1.38
Resolución de ejercicios	13	3.15	1.41
Aprendizaje basado en problemas	13	3.31	1.44
Proyectos tutorizados	13	3.69	1.49
Participación del grupo			
Constancia en la asignatura	13	3.69	1.44
Implicación y motivación de los participantes	13	3.15	1.52

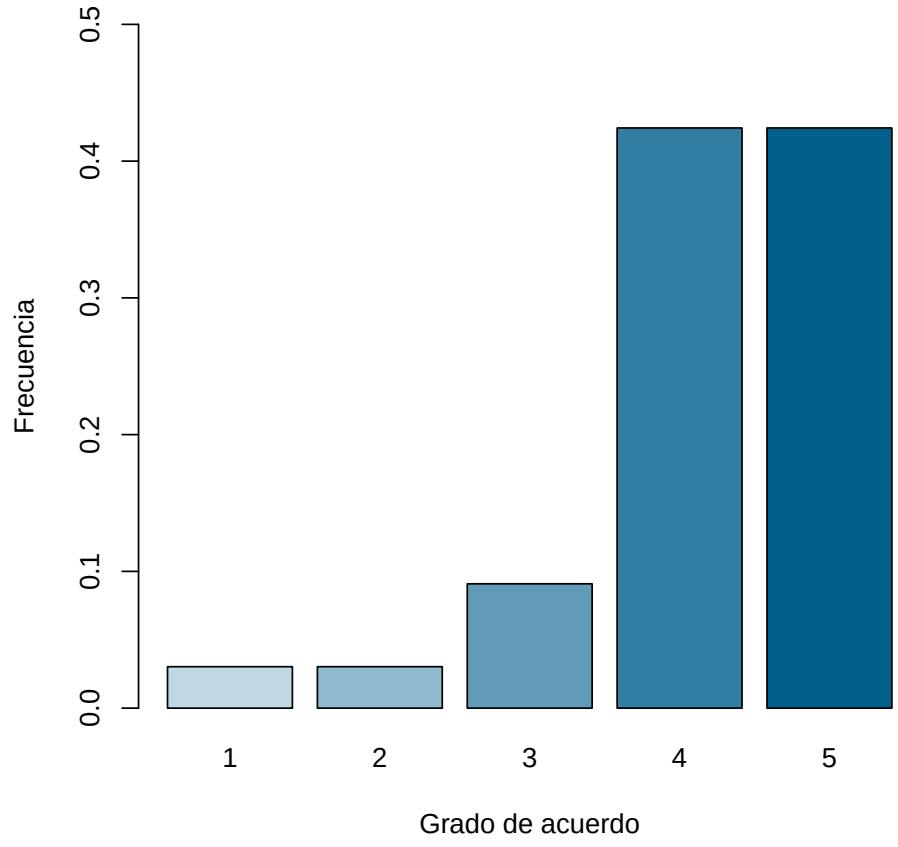
4.3. Grado en Ciencias del Mar

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	14	4.29	0.61
Duración de las clases	14	4.14	0.66
Calendario y Horario de la asignatura	14	3.86	0.95
Docentes de la asignatura			
Dominio de la materia	14	4.00	0.96
Claridad en la comunicación	14	3.29	0.91
Aspectos globales			
Guiones de trabajo	14	3.14	1.10
Medios y recursos disponibles	14	3.64	0.93
Datos y problemas analizados relacionados con la titulación	14	3.43	1.34
Utilidad de la asignatura para formación académica	14	3.21	1.31
Utilidad asignatura para la formación investigadora	14	3.57	1.16
Valoración general de la asignatura	14	3.21	1.12
Desarrollo de la asignatura			
Método expositivo	14	3.07	1.44
Estudio de casos	14	3.29	0.91
Resolución de ejercicios	14	3.07	1.49
Aprendizaje basado en problemas	14	2.79	1.31
Proyectos tutorizados	13	2.69	0.95
Participación del grupo			
Constancia en la asignatura	14	3.43	0.76
Implicación y motivación de los participantes	14	2.93	1.07

5. Aspectos generales de organización

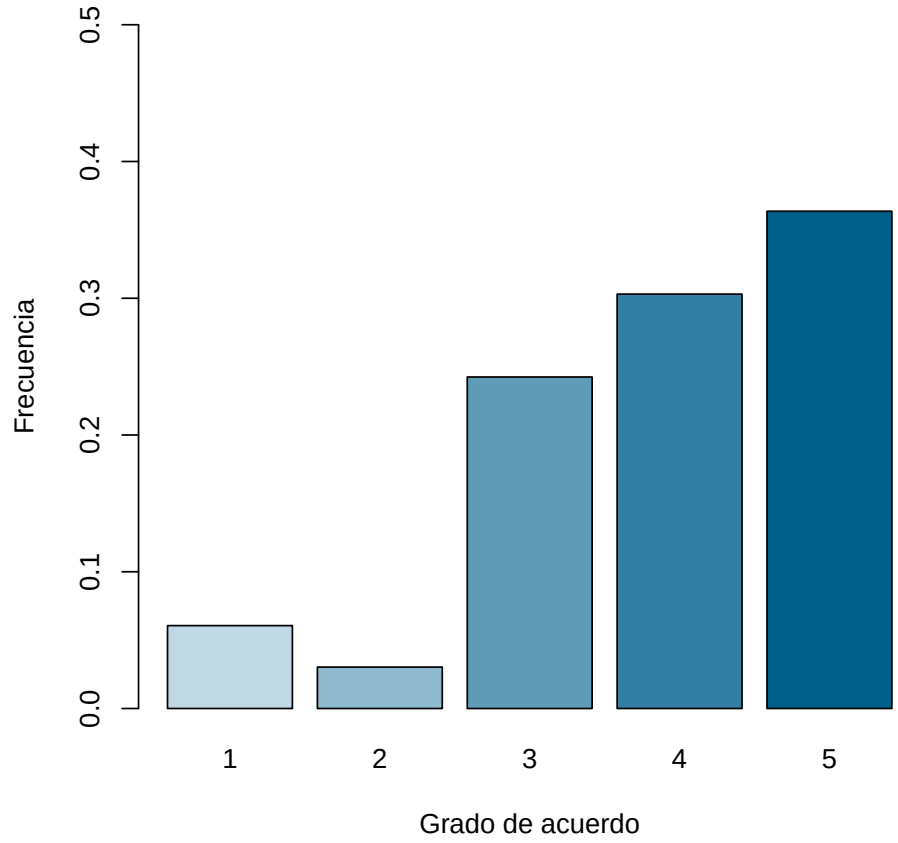
Instalaciones donde se desarrollan las clases

Cuenta	33.00
Mínimo	1.00
Media	4.18
Mediana	4.00
Máximo	5.00
Des. Típica	0.95



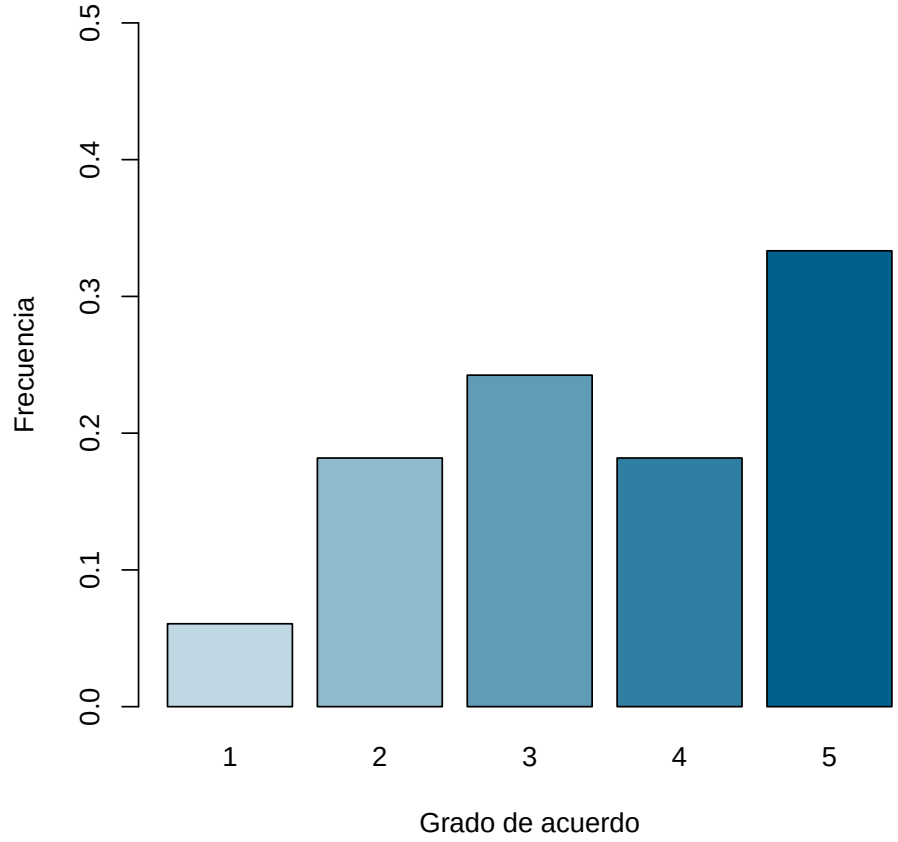
Duración de las clases

Cuenta	33.00
Mínimo	1.00
Media	3.88
Mediana	4.00
Máximo	5.00
Des. Típica	1.14



Calendario y Horario de la asignatura

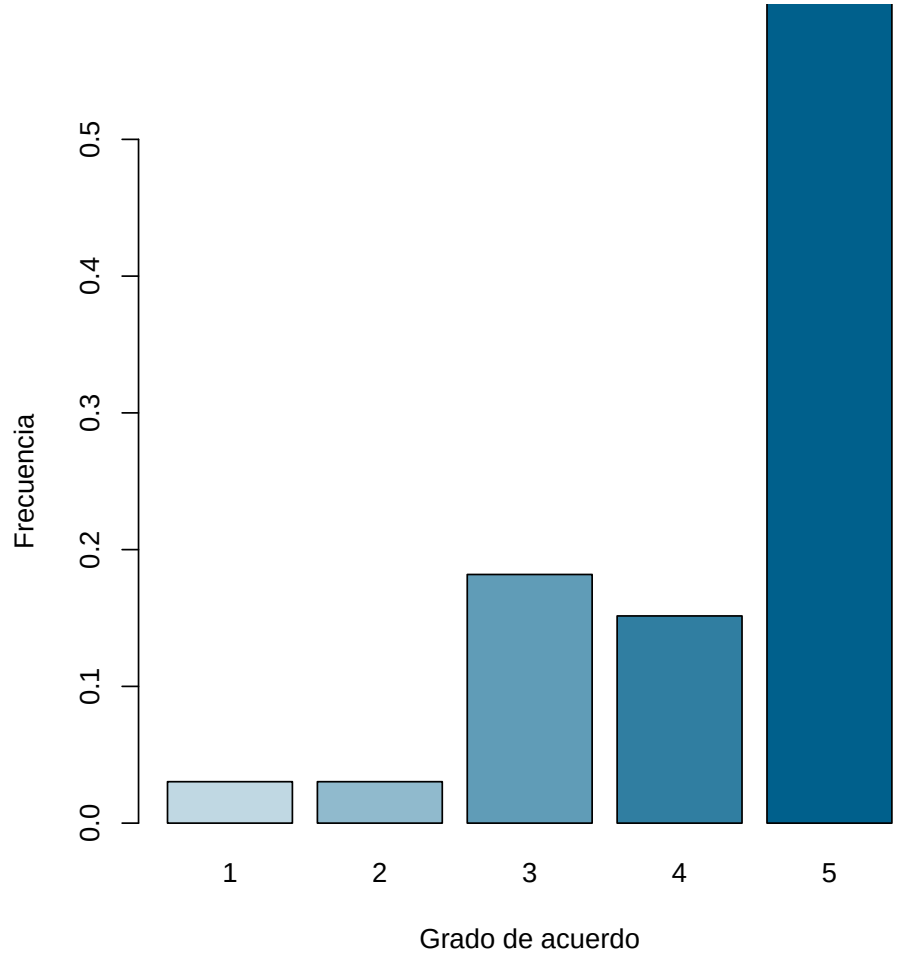
Cuenta	33.00
Mínimo	1.00
Media	3.55
Mediana	4.00
Máximo	5.00
Des. Típica	1.30



6. Docentes de la asignatura

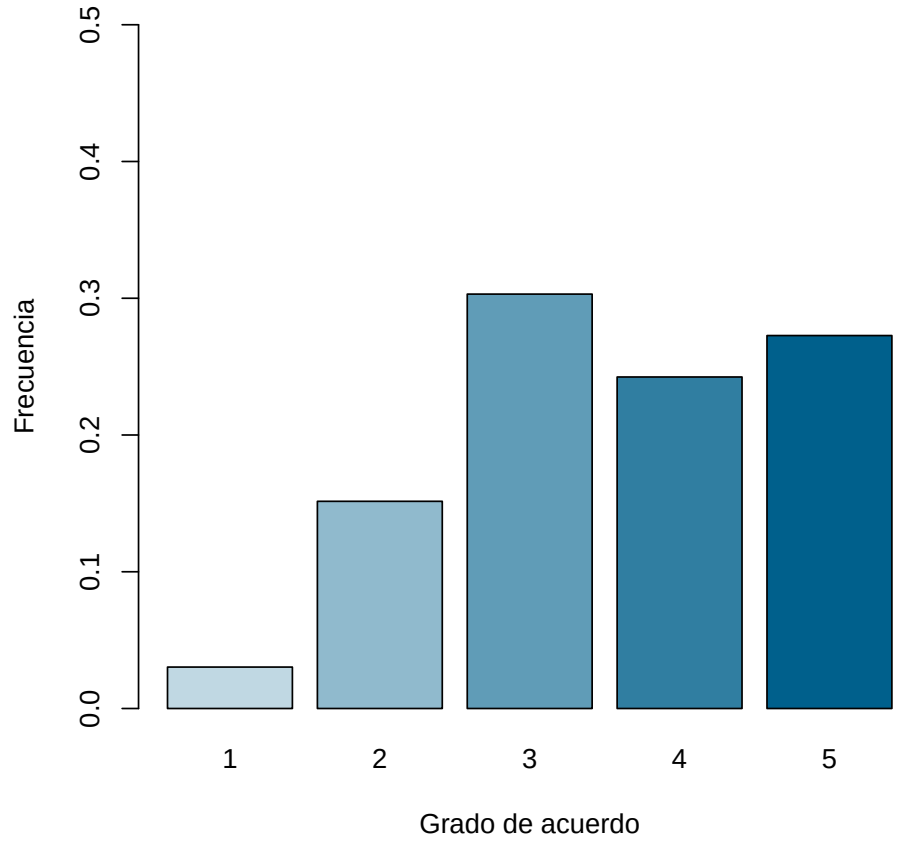
Dominio de la materia

Cuenta	33.00
Mínimo	1.00
Media	4.27
Mediana	5.00
Máximo	5.00
Des. Típica	1.07



Claridad en la comunicación

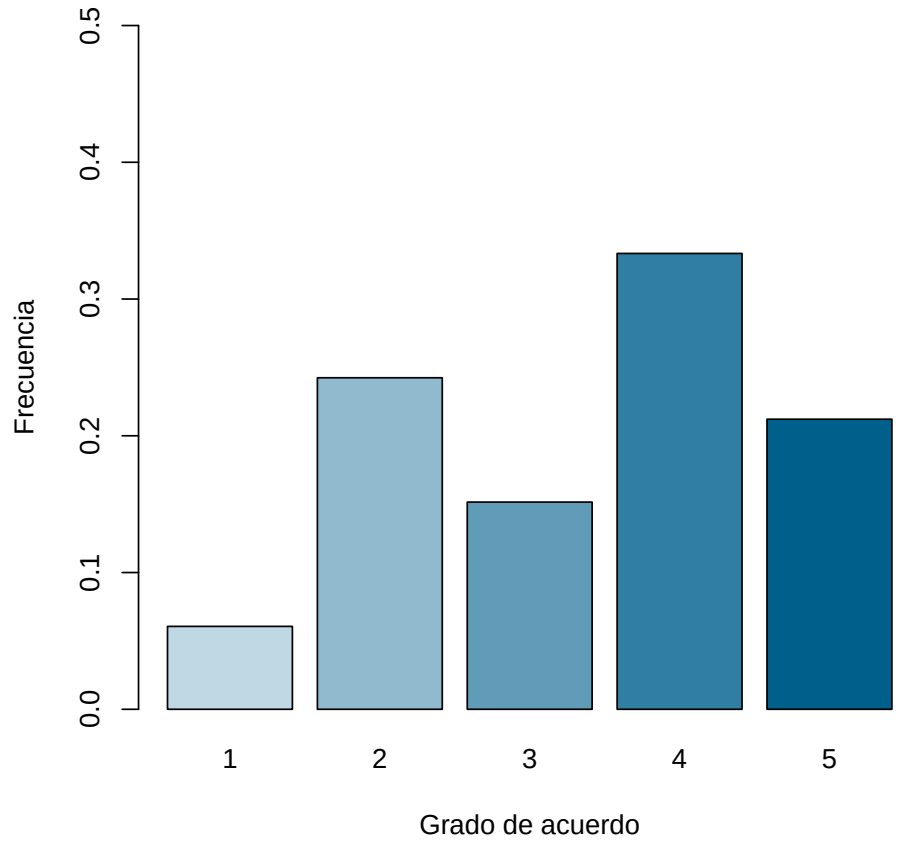
Cuenta	33.00
Mínimo	1.00
Media	3.58
Mediana	4.00
Máximo	5.00
Des. Típica	1.15



7. Aspectos globales

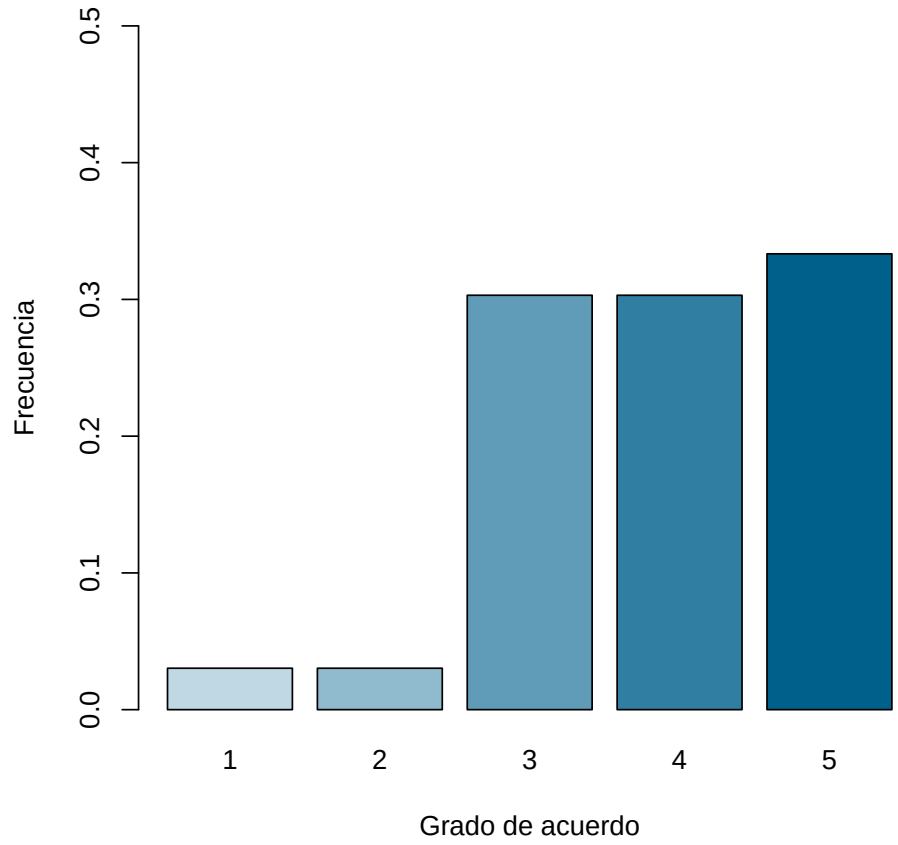
Guiones de trabajo

Cuenta	33.00
Mínimo	1.00
Media	3.39
Mediana	4.00
Máximo	5.00
Des. Típica	1.25



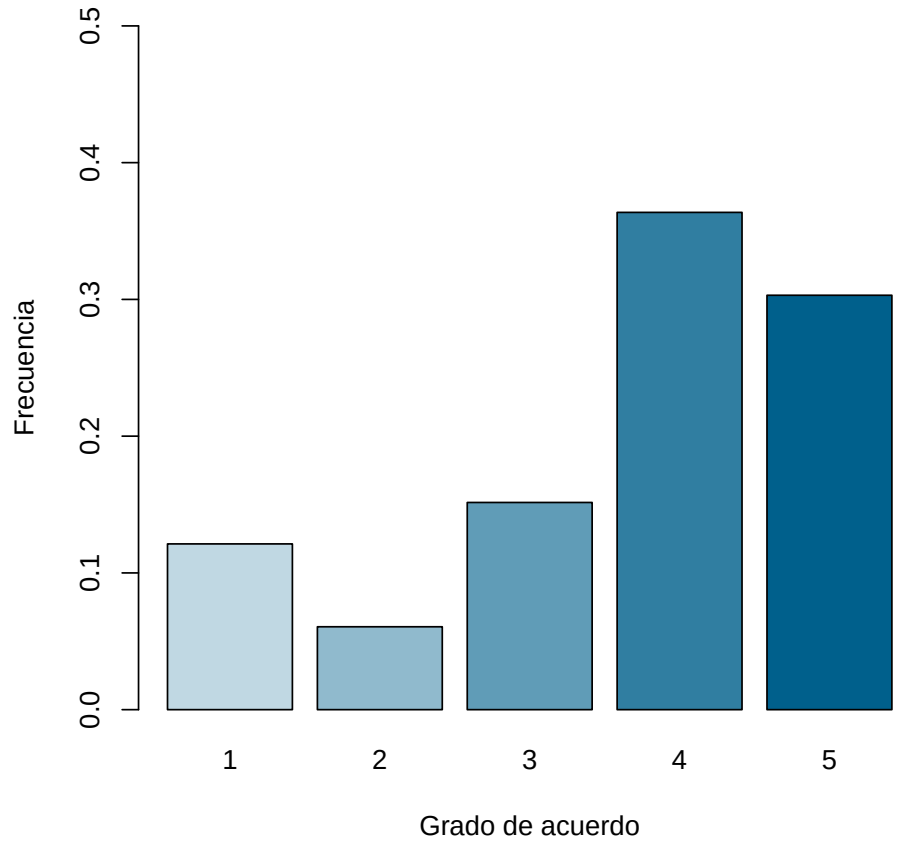
Medios y recursos disponibles

Cuenta	33.00
Mínimo	1.00
Media	3.88
Mediana	4.00
Máximo	5.00
Des. Típica	1.02



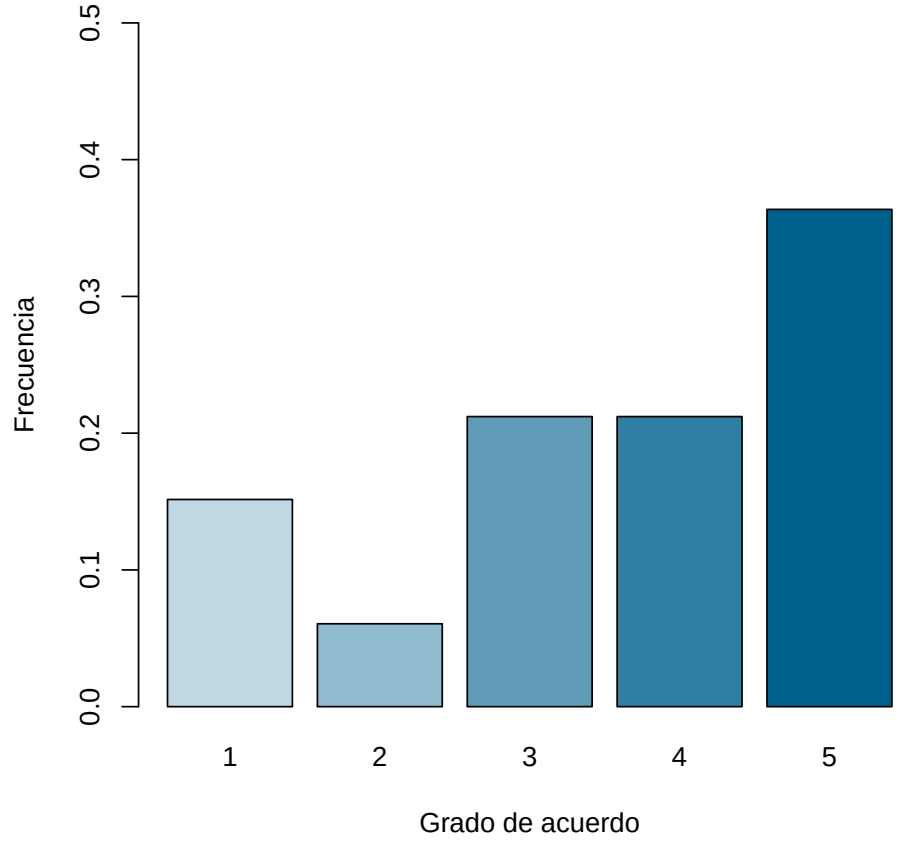
Datos y problemas analizados relacionados con la titulación

Cuenta	33.00
Mínimo	1.00
Media	3.67
Mediana	4.00
Máximo	5.00
Des. Típica	1.31



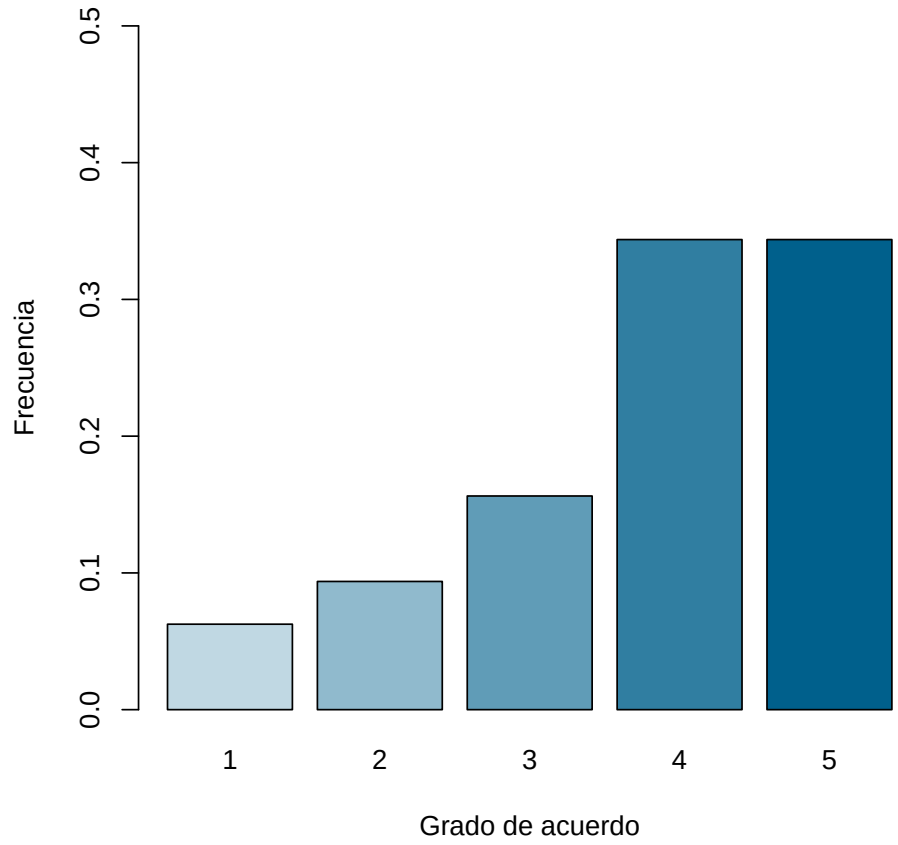
Utilidad de la asignatura para formación académica

Cuenta	33.00
Mínimo	1.00
Media	3.58
Mediana	4.00
Máximo	5.00
Des. Típica	1.44



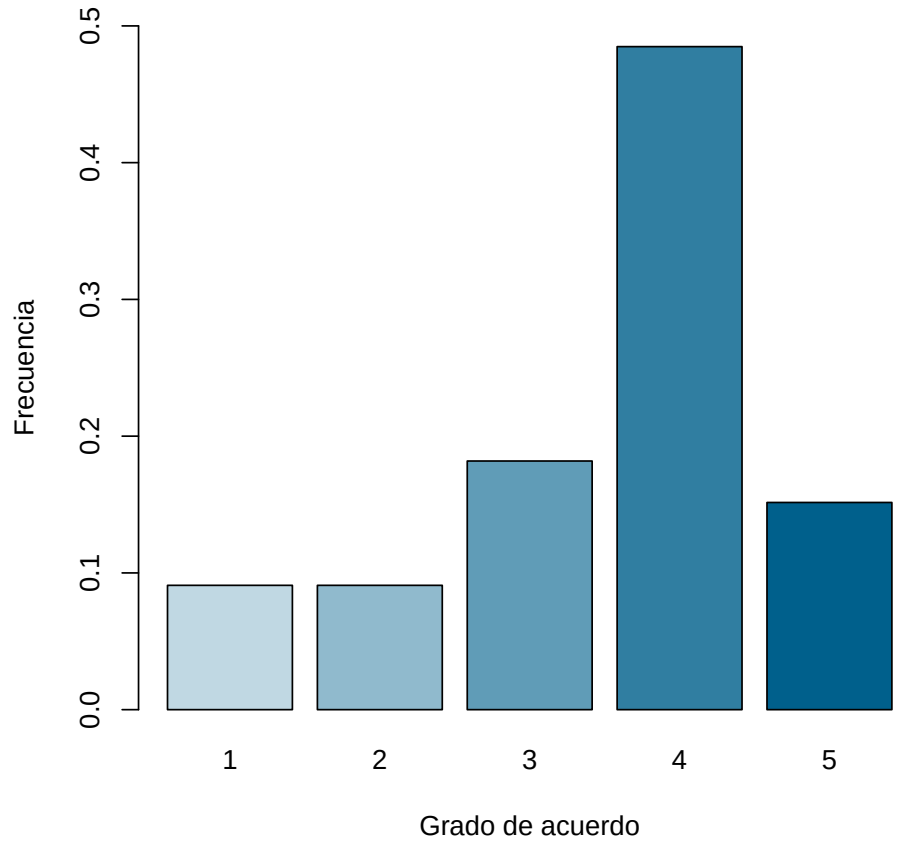
Utilidad de la asignatura para la formación investigadora

Cuenta	32.00
Mínimo	1.00
Media	3.81
Mediana	4.00
Máximo	5.00
Des. Típica	1.20



Valoración general de la asignatura

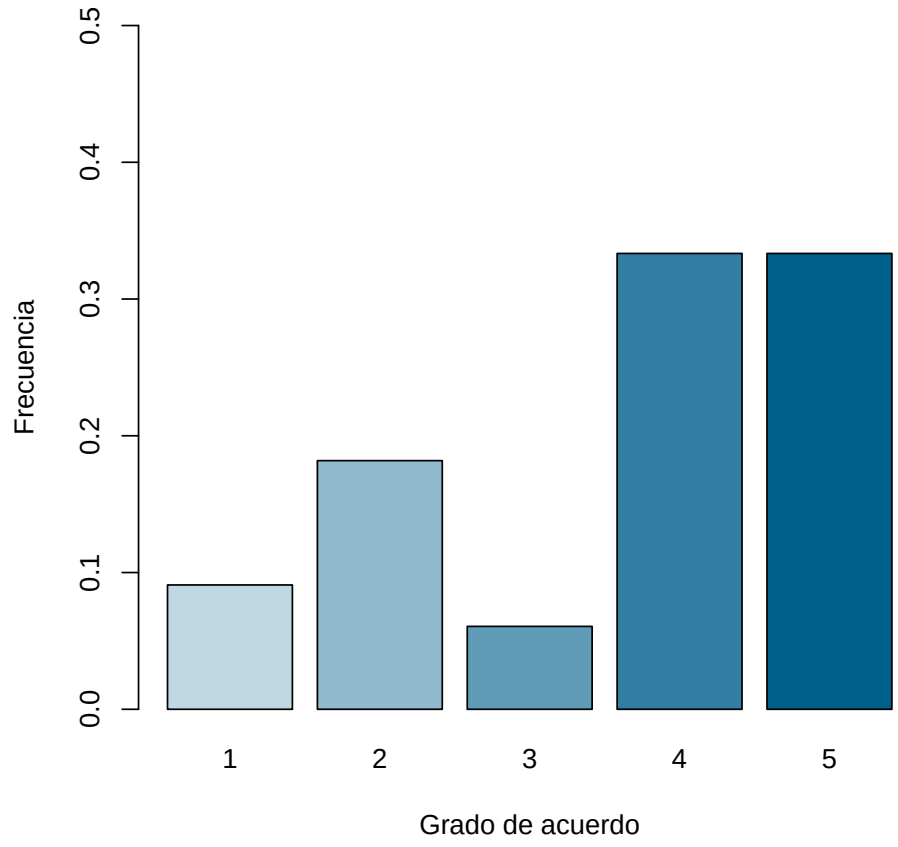
Cuenta	33.00
Mínimo	1.00
Media	3.52
Mediana	4.00
Máximo	5.00
Des. Típica	1.15



8. Desarrollo de la asignatura

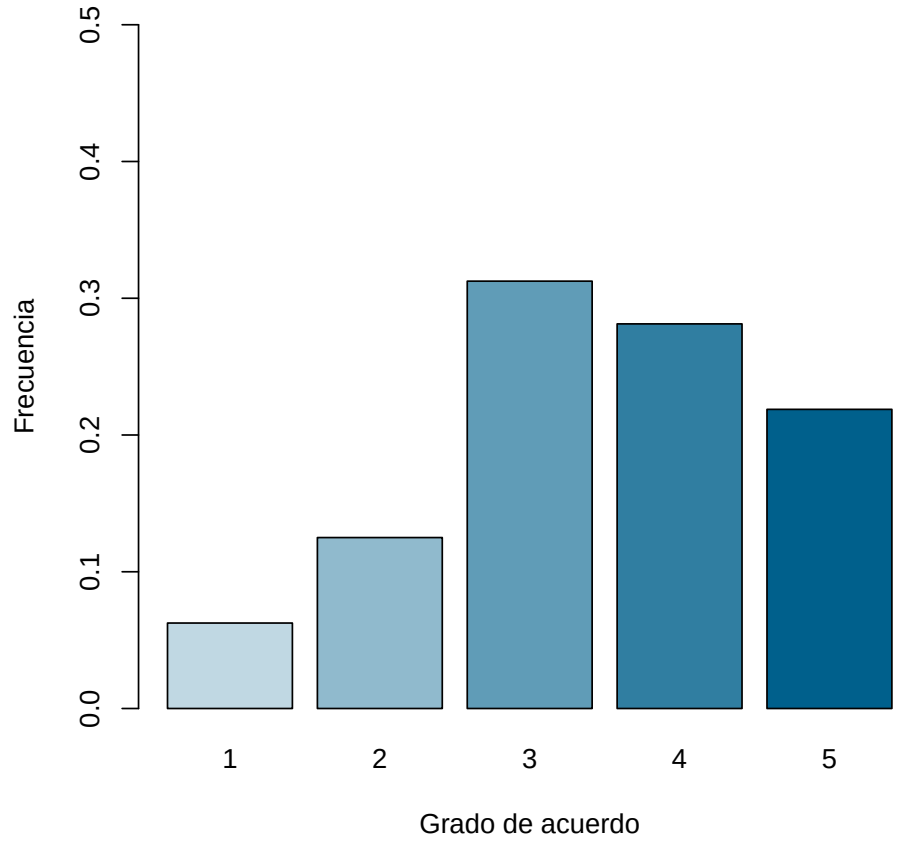
Método expositivo

Cuenta	33.00
Mínimo	1.00
Media	3.64
Mediana	4.00
Máximo	5.00
Des. Típica	1.37



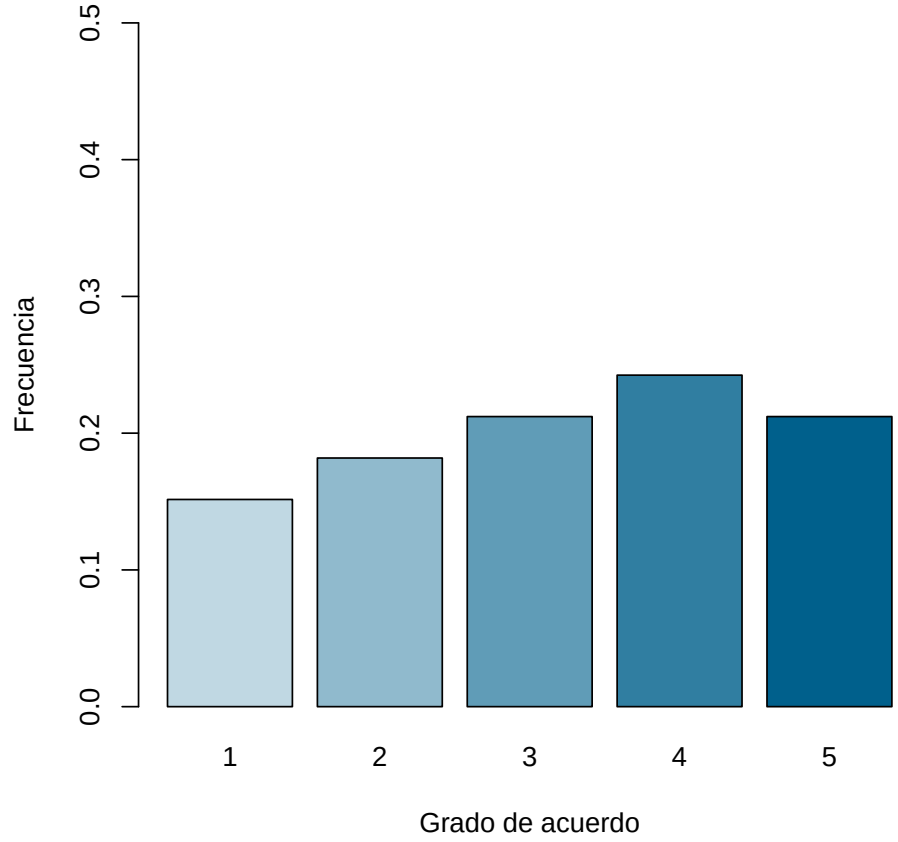
Estudio de casos

Cuenta	32.00
Mínimo	1.00
Media	3.47
Mediana	3.50
Máximo	5.00
Des. Típica	1.16



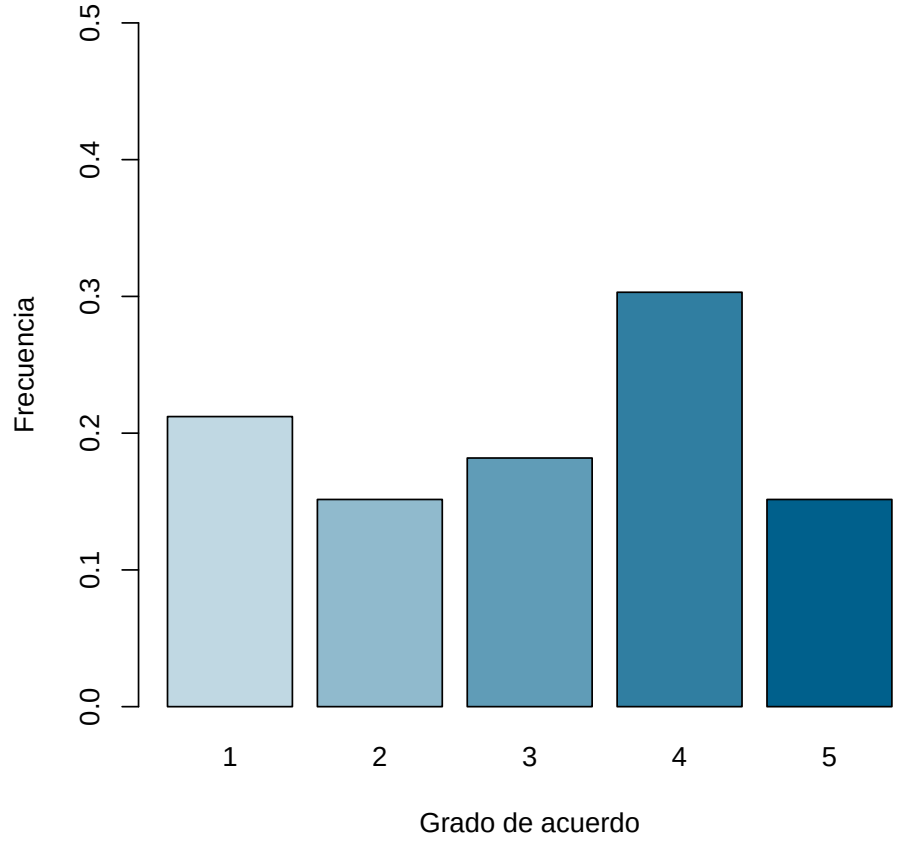
Resolución de ejercicios

Cuenta	33.00
Mínimo	1.00
Media	3.18
Mediana	3.00
Máximo	5.00
Des. Típica	1.38



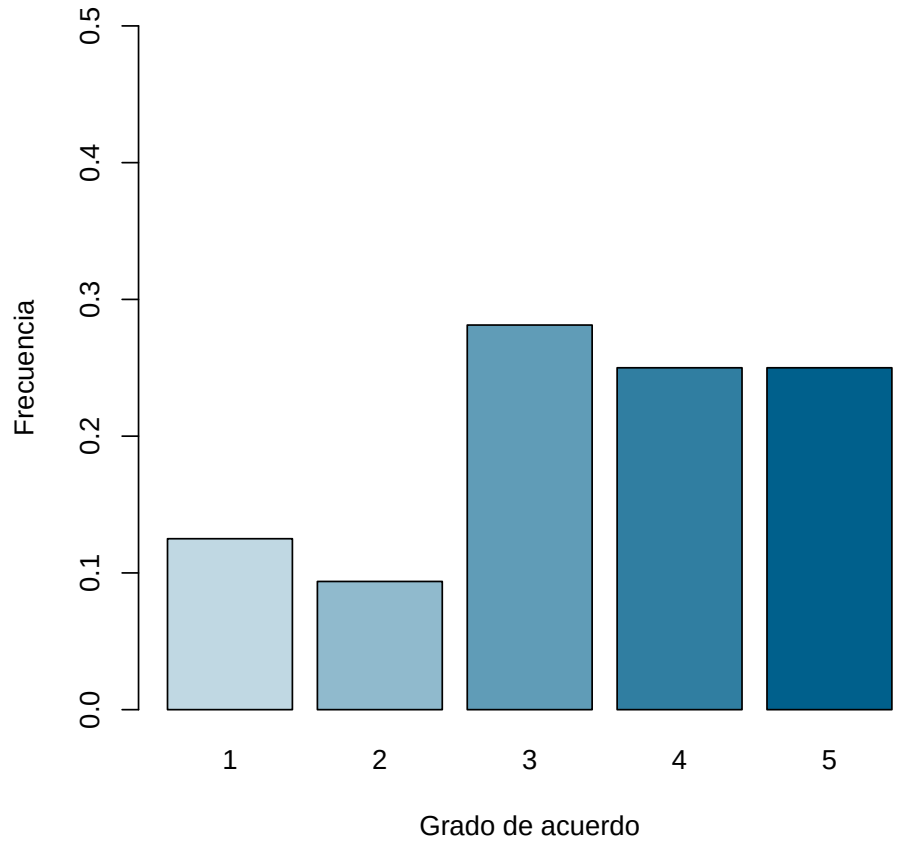
Aprendizaje basado en problemas

Cuenta	33.00
Mínimo	1.00
Media	3.03
Mediana	3.00
Máximo	5.00
Des. Típica	1.40



Proyectos tutorizados

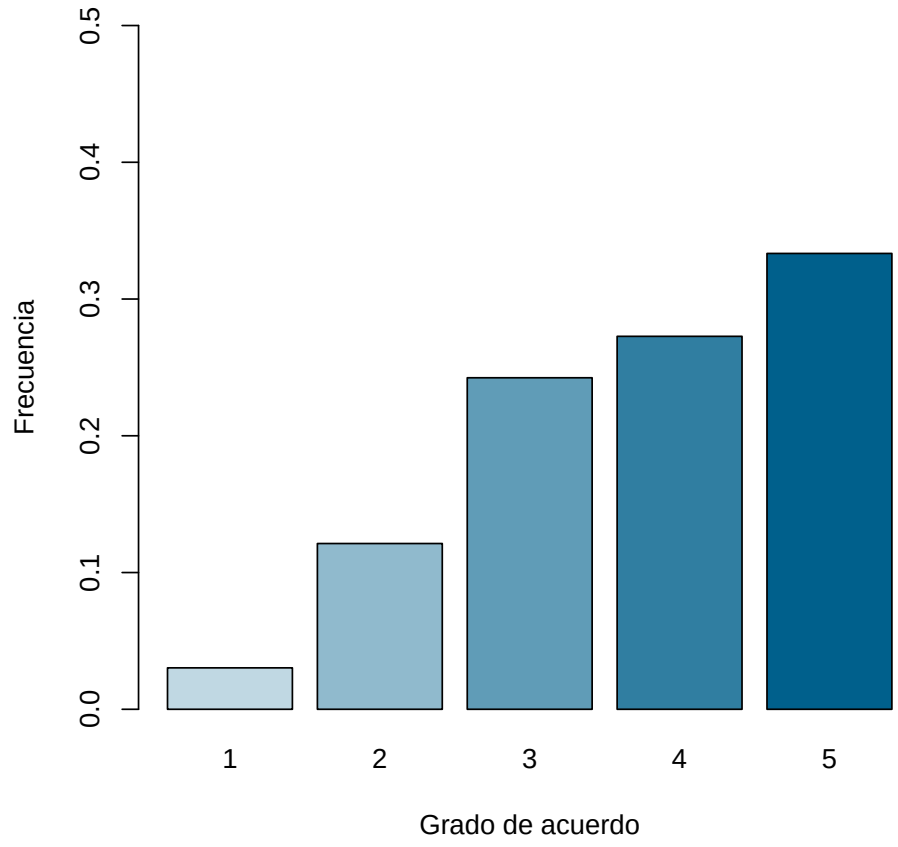
Cuenta	32.00
Mínimo	1.00
Media	3.41
Mediana	3.50
Máximo	5.00
Des. Típica	1.32



9. Participación del grupo

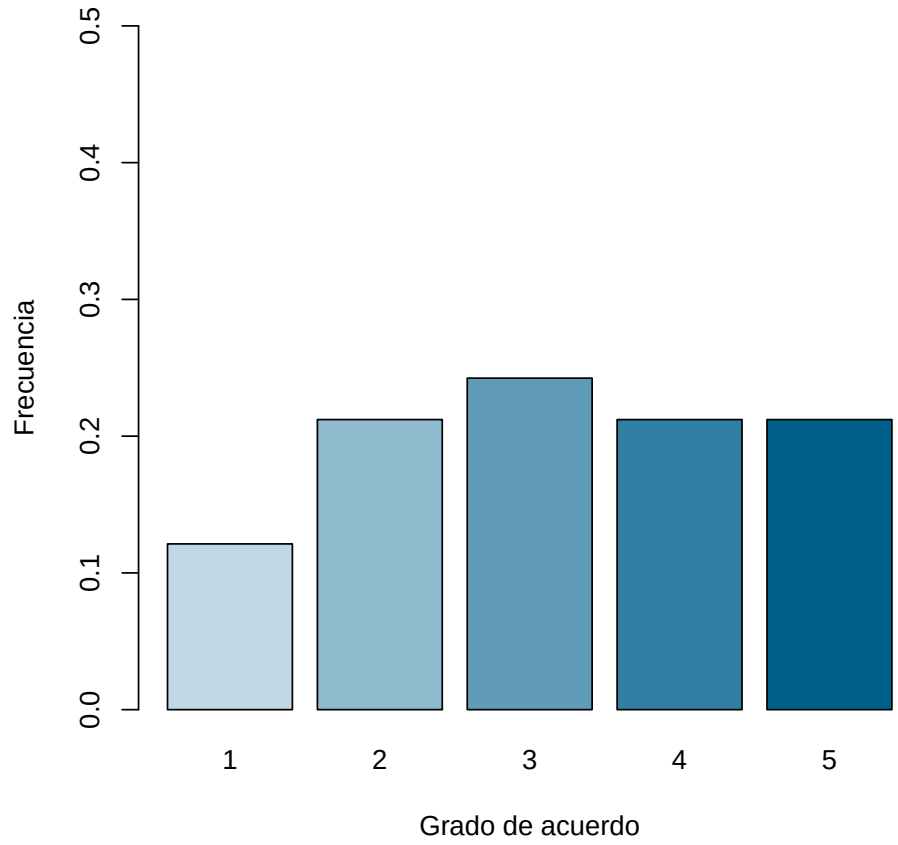
Constancia en la asignatura

Cuenta	33.00
Mínimo	1.00
Media	3.76
Mediana	4.00
Máximo	5.00
Des. Típica	1.15



Implicación y motivación de los participantes

Cuenta	33.00
Mínimo	1.00
Media	3.18
Mediana	3.00
Máximo	5.00
Des. Típica	1.33



MANUAL DE TÉCNICAS MULTIVARIANTES Y SERIES TEMPORALES CON R

En este documento quedan recogidas cuatro de las técnicas que recogen el manual a modo de ejemplo:

1.- Herramientas básicas del análisis multivariante

2.- Gráficos multivariantes

3.- Componentes Principales

4.- Análisis de Correspondencias

UNIVERSIDAD DE CADIZ
Departamento de Estadística e I.O.
Grupo de investigación TeLoyDisRen

Herramientas matemáticas en el Análisis Multivariante

1. Introducción

En este capítulo se estudiarán las herramientas básicas del Análisis Multivariante relacionados con aspectos algebráicos, métricos y geométricos. Como ha quedado de manifiesto en el anterior capítulo, la matriz X_{nxp} de individuos-variables es la unidad básica de información en un problema multivariable; debemos pues familiarizarnos con su manejo, establecer criterios para evaluar las diferencias entre los individuos y entre las variables y conocer técnicas de representación gráfica de los datos.

2. Conceptos algebráicos y geométricos

La matriz X_{nxp} de individuos-variables, puede verse como un conjunto de n vectores-individuos en un espacio de k dimensiones de variables o de k vectores-variables en un espacio de n dimensiones de individuos. Desde un punto de vista abstracto los roles de los individuos y las variables pueden ser intercambiados sin más que trasponer la matriz X_{nxp}, obteniéndose ${X'}_{pxn}$

2.1. Vectores

Como se ha puesto de manifiesto, las filas y columnas de la matriz de datos son vectores que contienen la información relativa a cada uno de los individuos y de las variables del estudio. Los elementos de una fila o columna son las coordenadas o componentes del vector. A continuación se repasarán algunos conceptos relacionados con las operaciones de vectores que tendrán su interpretación en función de individuos o variables.

Un vector $\mathbf{x}$ en un espacio de n dimensiones, $\mathbb{R}^n$ no es sino un segmento orientado, caracterizado por su tamaño y dirección. Si el vector tiene todas sus coordenadas iguales, $(c, c, \ldots, c)$ se dice constante y puede expresarse como $c(1, 1, \ldots, 1) = c\mathbf{1}$. Desde la óptica de los individuos, un vector constante no dice nada, sin embargo, en términos de variables, un vector constante se traduciría en que la característica medida toma el mismo valor para todos los individuos y, por tanto, tendría varianza cero, no discriminaría entre unos individuos y otros y no aportaría información al estudio, por lo que debería eliminarse.

Si se consideran variables, las operaciones de suma o diferencia de vectores y la multiplicación de un vector por un escalar generan, respectivamente, una nueva variable suma(diferencia) o un cambio de escala.

Se define el **producto escalar**, o interno, de dos vectores $\mathbf{x}$ e $\mathbf{y}$ de di-

mensión n como el escalar

$$\mathbf{x'y} = \mathbf{y'x} = \sum_{i=1}^{n} x_i y_i$$

La raiz cuadrada del producto escalar de un vector por sí mismo se conoce como norma cuadrática del vector y coincide con la longitud del mismo,

$$||\mathbf{x}|| = \sqrt{\mathbf{x'x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Desde un punto de vista estadístico, la **varianza** de una variable–vector $\mathbf{x}$ se puede expresar en función de la norma cuadrática como,

$$S_x{}^2 = \frac{||\mathbf{x} - \bar{x}\mathbf{1}||^2}{n}$$

Ejemplo 2.1

Dados los vectores $\mathbf{x} \equiv (1, -3, 4)$ *e* $\boldsymbol{y} \equiv (1, 1, -1)$, *el producto escalar se obtendría como*

```
> x <- c(1, -3, 4)
> y <- c(1, 1, -1)
> x %*% y

     [,1]
[1,]   -6
```

y la longitud de $\mathbf{x}$ *es igual a*

```
> sqrt(x %*% x)

         [,1]
[1,] 5.09902
```

El producto escalar es una medida del grado de ortogonalidad entre dos vectores. En función del ángulo que forman, el producto escalar puede expresarse como

$$\mathbf{x'y} = cos\theta||\mathbf{x}||||\mathbf{y}||$$

cuando el ángulo es de 90^0 vale cero, mientras que cuando el ángulo es de 0^0 toma su máximo valor, igual al producto de las normas de los vectores. En su forma más general, este resultado se conoce como la `desigualdad de Cauchy-Schwarz`,

$$|\mathbf{x'y}| \leq ||\mathbf{x}||||\mathbf{y}||$$

Ejemplo 2.2

Para obtener el ángulo, en grados, que forman los vectores $\mathbf{x}$ *e* $\mathbf{y}$ *se procedería...*

```
> coseno <- (x %*% y)/(sqrt(x %*% x) * sqrt(y %*% y))
> (acos(coseno) * 180)/pi

         [,1]
[1,] 132.7941
```

Dos variables son ortogonales cuando no comparten información, mientras que su producto escalar toma un valor máximo cuando están en la misma dirección, en cuyo caso los valores de una variable se obtienen multiplicando las valores de la otra por un escalar. La **covarianza** entre dos variables, S_{xy}, no es sino el producto escalar normalizado entre los respectivos vectores que las representan,

$$\frac{(\mathbf{x}-\bar{x}\mathbf{1})'(\mathbf{y}-\bar{y}\mathbf{1})}{n} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n} = S_{xy}$$

Para obtener el **coeficiente de correlación lineal**, r_{xy}, basta con dividir la covarianza por el producto de las desviaciones típicas, que en términos de producto escalar sería

$$r_{xy} = \frac{\frac{(\mathbf{x}-\bar{x}\mathbf{1})'(\mathbf{y}-\bar{y}\mathbf{1})}{n}}{\frac{||\mathbf{x}-\bar{x}\mathbf{1}||}{\sqrt{n}}\frac{||\mathbf{y}-\bar{y}\mathbf{1}||}{\sqrt{n}}} = \frac{(\mathbf{x}-\bar{x}\mathbf{1})'(\mathbf{y}-\bar{y}\mathbf{1})}{||\mathbf{x}-\bar{x}\mathbf{1}||||\mathbf{y}-\bar{y}\mathbf{1}||} = cos\theta$$

Si las variables $\mathbf{x}$ e $\mathbf{y}$ están estandarizadas, con media cero y desviación típica 1, entonces, puesto que $\mathbf{x}$ e $\mathbf{y}$ son unitarios, el coeficiente de correlación es igual a

$$r_{xy} = \mathbf{x}'\mathbf{y} = cos\theta$$

En todo caso, la correlación entre las variables es igual al coseno del ángulo que forman. El coeficiente de correlación indica la cantidad de información

que comparten las variables, tomando valores entre -1 y 1. Cuando el coeficiente vale cero, las variables son ortogonales, mientras que cuando cuando vale 1 o -1 ambas variables proporcionan la misma información, aunque en el segundo caso de forma inversa, y consecuentemente una de las variables debería ser eliminada de la matriz de datos.

Ejemplo 2.3

Dadas las variables $\mathbf{x} = (1, 3, 4, 4, 5, 7)$ *e* $\mathbf{y} = (16, 10, 12, 4, 8, 10)$, *que representan a las mediciones de dos características en seis individuos, el coeficiente de correlación se obtendría, utilizando el producto escalar, como*

```
> x <- c(1, 3, 4, 4, 5, 7)
> x1 <- x - mean(x)
> y <- c(16, 10, 12, 4, 8, 10)
> y1 <- y - mean(y)
> datos <- data.frame(x, y)
> r <- (x1 %*% y1)/(sqrt(x1 %*% x1) * sqrt(y1 %*% y1))
> r

     [,1]
[1,] -0.5
```

Además del producto interno, se puede definir el **producto externo** entre dos vectores $\mathbf{x}$ e $\mathbf{y}$ de la forma

$$\mathbf{x}\mathbf{y}' = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \begin{pmatrix} y_1, & y_2, & \dots, & y_n \end{pmatrix} = \begin{pmatrix} x_1y_1 & \dots & x_1y_n \\ \dots & \dots & \dots \\ x_ny_1 & \dots & x_ny_n \end{pmatrix}$$

2.2. Combinaciones lineales de vectores

Si se consideran los vectores $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_p}$ de un espacio $\mathbb{R}^n$ y los escalares $u_1, u_2, \dots, u_p$, la combinación lineal $u_1\mathbf{x_1} + u_2\mathbf{x_2} + \dots + u_p\mathbf{x_p}$ genera un nuevo vector del espacio. Si existe un conjunto de escalares $a_1, a_2, \dots, a_p$, no todos nulos, tal que $a_1\mathbf{x_1} + a_2\mathbf{x_2} + \dots + a_p\mathbf{x_p} = \mathbf{0}$ donde $\mathbf{0}$ es el vector nulo, $(0, 0, \dots, 0)$, se dice que los vectores son linealmente dependientes. En esta situación es posible expresar cualquier vector con coeficiente $a_i \neq 0$ como combinación lineal del resto.

Si los vectores $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ de $\mathbb{R}^n$ con $p < n$ son linealmente independientes, el conjunto de todas las combinaciones lineales de dichos vectores generan un subespacio de dimensión p, que llamaremos E_p, siendo $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ la `base` de dicho subespacio. Cualquier vector $\mathbf{y}$ que no pertenece a E_p es ortogonal a todo vector $\mathbf{x}$ del subespacio y se encuentra en lo que se denomina complemento ortogonal del subespacio E_p, de dimensión $n - p$.

Desde un punto de vista estadístico, si se tiene una matriz nxp de individuos-variables, las p variables estarán definidas en un espacio de dimensión n. En general, las p variables no serán ortogonales e incluso es posible que alguna de ellas se pueda obtener como combinación lineal del resto, con lo que el subespacio generado por el conjunto de variables será de dimensión menor que p. Además, si los productos escalares entre las variables son altos, ello implicará que sus correlaciones son altas y que el subespacio generado ocupe "poco espacio". Se volverá sobre los conceptos de dependencia e independencia lineal cuando se vean las matrices y determinantes y cuando se aborden las **técnicas de reducción de la dimensión**.

2.3. Matrices

Ya se ha comentado el significado estadístico de una matriz de datos de individuos-variables. Se verán en esta sección algunas propiedades y características de las matrices.

Ejemplo 2.4

Para crear en R la matriz `A`$=\begin{pmatrix} 1 & 7 & 3 \\ 2 & 1 & 1 \\ 3 & 8 & 5 \end{pmatrix}$

se procede...

```
> A <- matrix(c(1, 2, 3, 7, 1, 8, 3, 1, 5), 3, 3)
> A

     [,1] [,2] [,3]
[1,]    1    7    3
[2,]    2    1    1
[3,]    3    8    5
```

Si se considera la matriz de datos $\mathbf{X_{nxp}}$ es posible cambiar filas por columnas, obteniéndose la matriz $\mathbf{X'_{pxn}}$ o $\mathbf{X^T_{pxn}}$, dicha matriz se denomina `tras-`

`puesta`. La trasposición cambia los roles de los individuos por el de las variables, ello lleva a lo que se conoce como análisis en **modo Q**, en contraposisión a la forma natural de tratar los datos, conocida como **modo R**. Los vectores son matrices columnas de dimensión genérica `nx1` y cuando se trasponen se convierten en matrices filas de dimensión `1xn`.

Ejemplo 2.5

Para obtener la traspuesta de A...

```
> tA <- t(A)
> tA

     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    7    1    8
[3,]    3    1    5
```

2.3.1. Operaciones con matrices

La suma(diferencia) de matrices, siempre que sean de la misma dimensión, se realiza de forma obvia elemento a elemento. La operación suma(diferencia) es commutativa y además $(\mathbf{A} \pm \mathbf{B})' = \mathbf{A}' \pm \mathbf{B}'$. Observe además que $\mathbf{A} + \mathbf{A} = \mathbf{2A}$.

Ejemplo 2.6

Para multiplicar la matriz A por 5...

```
> 5 * A

     [,1] [,2] [,3]
[1,]    5   35   15
[2,]   10    5    5
[3,]   15   40   25
```

Sistemas de ecuaciones La multiplicación de un escalar por una matriz se obtiene multiplicando cada elemento de la matriz por el escalar. Muchas

de las técnicas multivariantes se basan en la resolución de un sistema de ecuaciones del tipo

$$\begin{gathered} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p = k_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p = k_2 \\ \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{np}x_n = k_n \end{gathered}$$

Los términos de la parte izquierda de las ecuaciones se obtienen como productos escalares de las filas de la matriz $\mathbf{A} = [a_{ij}]$ por el vector incógnita $\mathbf{x}$, mientras que la parte derecha son escalares. Si se considera el vector de escalares $\mathbf{k}$, de forma compacta, el sistema de ecuaciones puede expresarse como

$$\mathbf{Ax} = \mathbf{k}$$

Ejemplo 2.7

Para resolver el sistema...

$$\begin{gathered} x + 7y + 3z = -3 \\ 2x + y + z = 0 \\ 3x + 8y + 5z = 1 \end{gathered}$$

se utiliza la instrucción `solve()`*. Ya que se ha hecho coincidir la matriz de coeficientes con la matriz A, previamente creada, se tendría...*

```
> solve(A, c(-3, 0, 1))

[1] -1 -2  4
```

Matriz inversa La instrucción **solve** aplicada sobre una matriz cuadrada permite calcular su inversa.

```
> solve(A)

           [,1]       [,2]       [,3]
[1,]  0.2307692  0.8461538 -0.3076923
[2,]  0.5384615  0.3076923 -0.3846154
[3,] -1.0000000 -1.0000000  1.0000000
```

Multiplicación de matrices En general, el producto de una matriz por un vector del espacio $\mathbb{R}^n$ es otro vector del mismo espacio $\mathbf{Ax} = \mathbf{y}$. En realidad, como se verá más adelante, las transformaciones de vectores resultan de multiplicar éstos por matrices. No hay ningún problema en generalizar la idea anterior, y en lugar de multiplicar la matriz $\mathbf{A}$ por un vector se podría multiplicar por otra matriz, con la única restricción de que el número de columnas de la primera matriz coincida con el número de filas de la segunda, así se tendría

$$\mathbf{A_{nxp}B_{pxm}} = \mathbf{C_{nxm}}, \qquad c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}$$

Ejemplo 2.8

Si se multiplica la traspuesta de A por A se obtiene

```
> tA %*% A

     [,1] [,2] [,3]
[1,]   14   33   20
[2,]   33  114   62
[3,]   20   62   35
```

Esta multiplicación recibe el nombre de matricial, aunque no es la única que se puede definir.

Matrices especiales Antes de continuar es conveniente definir algunos tipos de matrices especiales.

- Matriz cuadrada: Es la que tiene el mismo número de filas que de columnas, se caracteriza con una única dimensión $\mathbf{A_n}$.
- Matriz unidad. Es una matriz cuadrada que tiene unos en la diagonal principal y ceros fuera de ella. Si tiene dimensión n se nota $\mathbf{I_n}$.
- Matriz diagonal Es una matriz cuadrada que solo tiene elementos distintos de cero en la diagonal.
- Matriz nula. Todos sus elementos son ceros.

- Matriz simétrica. Es una matriz cuadrada tal que $a_{ij} = a_{ji}, \quad \forall i, j$. En este caso $\mathbf{A}' = \mathbf{A}$.

La función `apply` aplica una función sobre las filas o columnas de una matriz.

Ejemplo 2.9

```
> x <- c(175, 65, 27, 167, 58, 33, 184, 77, 31)
> X <- matrix(x, 3, 3)
> colnames(X) <- c("Estatura", "Peso", "Edad")
> rownames(X) <- c("Pepe", "Juan", "Luis")
> apply(X, 1, sum)

Pepe Juan Luis
 526  200   91

> apply(X, 2, mean)[2]

Peso
  86
```

Devuelve la suma de las tres medidas de cada persona y la media del `Peso`.

Algunas otras funciones sobre matrices son:

`nrow()` devuelve el número de filas de una matriz.

`ncol()` devuelve el número de columnas de una matriz.

`diag(entero)` devuelve la matriz identidad del orden especificado.

`diag(vector)` devuelve la matriz cuya diagonal es el vector dado.

`diag(matriz)` devuelve un vector con la diagonal de la matriz dada.

`cbind(matriz, matriz)` une dos matrices horizontalmente.

`rbind(matriz, matriz)` une dos matrices verticalmente.

Ejemplo 2.10

Para definir una matriz unidad de orden 4 se procedería...

```
> diag(4)

     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

Si se desea construir una matriz diagonal, bastaría indicar los elementos de dicha diagonal en un vector...

```
> diag(c(3, 2, 4))

     [,1] [,2] [,3]
[1,]    3    0    0
[2,]    0    2    0
[3,]    0    0    4
```

2.3.2. Suma de Cuadrados y Productos Cruzados

Si se considera $\mathbf{X_{nxp}}$ como el vector de los vectores filas e $\mathbf{Y_{pxm}}$ como el vector de vectores columna, puede verse el producto matricial como la matriz que contiene todos los productos escalares de los vectores de $\mathbf{X}$ por los de $\mathbf{Y}$.

$$\begin{pmatrix} \mathbf{x_1'} \\ \mathbf{x_2'} \\ \cdots \\ \mathbf{x_n'} \end{pmatrix} \begin{pmatrix} \mathbf{y_1} & \mathbf{y_2} & \cdots & \mathbf{y_m} \end{pmatrix} = \begin{pmatrix} \mathbf{x_1'y_1} & \cdots & \mathbf{x_1'y_m} \\ \mathbf{x_2'y_1} & \cdots & \mathbf{x_2'y_m} \\ \cdots & \cdots & \cdots \\ \mathbf{x_n'y_1} & \cdots & \mathbf{x_n'y_m} \end{pmatrix}$$

Las matrices cuadradas y simétricas tienen un especial interés en estadística, ya que esa es la estructura de la **matriz de varianzas-covarianzas** o de la **matriz de correlaciones**. De hecho, si partimos de la matriz de datos $\mathbf{X_{nxp}}$, el producto $\mathbf{X^T X}$ es una matriz cuadrada de dimensión p, dicho producto se conoce como **Suma de cuadrados y productos cruzados(SCPC)**.

$$\begin{pmatrix} \mathbf{x_1'} \\ \mathbf{x_2'} \\ \dots \\ \mathbf{x_p'} \end{pmatrix} \begin{pmatrix} \mathbf{x_1} & \mathbf{x_2} & \dots & \mathbf{x_p} \end{pmatrix} = \begin{pmatrix} \mathbf{x_1'x_1} & \dots & \mathbf{x_1'x_p} \\ \mathbf{x_2'x_1} & \dots & \mathbf{x_2'x_p} \\ \dots & \dots & \dots \\ \mathbf{x_p'x_1} & \dots & \mathbf{x_p'x_p} \end{pmatrix}$$

Si se le resta a cada variable su media y se divide por n, la SCPC coincide con la matriz de varianzas-covarianzas. Por otra parte, si previamente se tipifican las p variables, la SCPC es igual a la matriz de correlaciones.

Ejemplo 2.11

Si se considera la matriz A definida por A<-matrix(c(1,2,3,7,1,8,3,1,5),3,3), la función ***crossprod(A)*** *calcula* $\mathbf{A^T A}$

```
> A <- matrix(c(1, 2, 3, 7, 1, 8, 3, 1, 5), 3, 3)
> crossprod(A)

     [,1] [,2] [,3]
[1,]   14   33   20
[2,]   33  114   62
[3,]   20   62   35
```

Por otra parte, si se tiene otra matriz B<-matrix(c(1,3,5,2,1,4,1,0,7),3,3), la función ***crossprod(A,B)*** *calcula* $\mathbf{A^T B}$ *de forma más eficiente que si calculara el producto matricial transponiendo previamente la primera matriz*

```
> B <- matrix(c(1, 3, 5, 2, 1, 4, 1, 0, 7), 3, 3)
> crossprod(A, B)

     [,1] [,2] [,3]
[1,]   22   16   22
[2,]   50   47   63
[3,]   31   27   38
```

Ejemplo 2.12

Sea la distribución bidimensional:

x	1	3	4	2	7	8	11	15	17	31	35
y	12	17	19	13	25	27	33	41	43	50	55

Si se quieren obtener los coeficientes de la recta de ajuste,$y = \beta_0 + \beta_1 x$, en R; dado que $X^T X\hat{\beta} = X^T y$ se procedería

```
> X <- cbind(1, c(1, 3, 4, 2, 7, 8, 11, 15, 17, 31, 35))
> y <- c(12, 17, 19, 13, 25, 27, 33, 41, 43, 50, 55)
> solve(crossprod(X), crossprod(X, y))

          [,1]
[1,] 15.277171
[2,]  1.245904
```

La multiplicación de matrices se comporta en cierto sentido como el producto de escalares, y así, las propiedades distributivas y asociativas se verifican igualmente, además se verifican otras propiedades destacables:

- $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AB} + \mathbf{AC}$
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- $\mathbf{A_{nxn}I_n} = \mathbf{I_nA_{nxn}} = \mathbf{A}$
- $(\mathbf{AB})^{\mathbf{T}} = \mathbf{B^T A^T}$

2.3.3. Rango de una matriz

El **Rango** de una matriz $\mathbf{A_{nxp}}$ es el número máximo de vectores linealmente independientes, si, como suele ocurrir en una matriz de datos es $n > p$, es decir hay más individuos que variables, el rango de $\mathbf{A}$ será como mucho igual a p, ya que las columnas representan a p vectores en $\mathbb{R}^n$, que podrían ser linealmente independientes, mientras que las filas, al ser n vectores de un espacio $\mathbb{R}^p$, no pueden ser linealmente independientes. Algunas propiedades del rango son:

- $rango((\mathbf{A} + \mathbf{B})) \leq rango(\mathbf{A}) + rango(\mathbf{B})$
- $rango(\mathbf{AB}) \leq min(rango(\mathbf{A}), rango(\mathbf{B}))$

2.3.4. Traza de una matriz

La **traza de una matriz** cuadrada es simplemente la suma de los elementos de su diagonal, la denotamos por $tr(\mathbf{A})$. Si los productos tienen sentido, algunas propiedades de la traza son:

- $tr((\mathbf{AB})) = tr((\mathbf{BA}))$
- $tr((\mathbf{ABC})) = tr((\mathbf{BAC})) = tr((\mathbf{CBA}))$

Ejemplo 2.13

La traza de la matriz $\mathbf{A}$ *se obtiene como...*

```
> sum(diag(A))

[1] 7
```

2.3.5. Determinante de una matriz

El determinante de una matriz arroja información sobre el espacio que ocupan p variables(vectores) en relación al espacio total $\mathbb{R}^p$. De alguna manera, si el espacio que ocupan los p vectores fuera nulo, lo que indicaría que al menos uno de ellos es combinación lineal del resto, no se podría dividir por dicha matriz, ya que se estaría dividiendo por cero. Para ilustrar esta medida se aplicará en primer lugar sobre una matriz diagonal de dimensión 2.

$$\begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}$$

Si se representan ambos vectores en el plano se obtendrá la figura 1. Se observa que el área que ocupan los vectores es igual a $a_{11}a_{22}$. En general, para una matriz de dimensión 2, el determinante se obtendría como $a_{11}a_{22} - a_{12}a_{21}$. La figura 2 representa esta situación.

Ejemplo 2.14

Para calcular el determinante de la matriz A en ***R*** *se utiliza la función* ***det***

```
> det(A)

[1] -13
```

Cuando el determinante es igual a cero, ello indica que el conjunto de los p vectores que conforman la matriz caben en un subespacio de dimensión

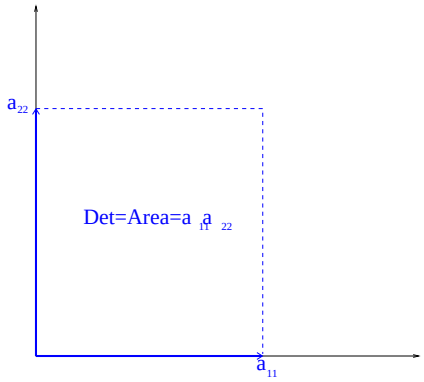


Figura 1: Determinante de dos vectores ortogonales

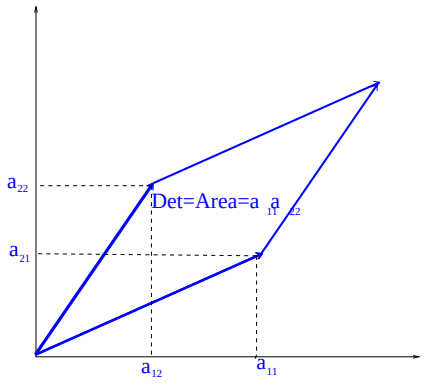


Figura 2: Determinante de dos vectores

$p-1$, o lo que es lo mismo, que al menos uno de los vectores es combinación lineal del resto.

2.4. Diagonalización de matrices

Supongamos que A es una matriz cuadrada de orden n, se dice que el vector $v = (v_1, v_2, \ldots, v_n)^T$ es un autovector de A si

$$Av = \lambda v$$

donde λ es un escalar que recibe el nombre de autovalor o valor propio asociado a v. La justificación del nombre radica en el hecho de que la matriz A transforma al vector v en otro proporcional a sí mismo. En el capítulo de Componentes Principales se da una interpretación estadística de los conceptos de autovalor y autovector.

Los autovalores de A se calculan resolviendo la denominada **ecuación característica**

$$det(A - \lambda I) = 0$$

Para obtener los autovalores y autovectores de una matriz en **R** se utiliza la función **eigen**

Ejemplo 2.15

```
> autov <- eigen(A)
> autov$values

[1]  9.0510086 -2.6028321  0.5518235

> autov$vectors

           [,1]       [,2]       [,3]
[1,] -0.5086938 -0.8253899 -0.3553996
[2,] -0.2294371  0.5195221 -0.3488862
[3,] -0.8298128 -0.2209713  0.8671618
```

3. Distancias, Disimilaridades y Similaridades

Uno de los primeros y principales problemas que hay que resolver en Análisis Multivariante es calcular la semejanza o diferencia existente entre dos elementos de la matriz datos, bien individuos(filas) o variables(columnas). Estas diferencias o parecidos, en función de la naturaleza de las variables, podrían alcanzar el rango de **distancia** o ser simplemente una **similaridad** o una **disimilaridad**, es decir cumplir menos requisitos de los exigidos a una distancia. En general, cuando exista una métrica para las variables se calcularán discrepancias en términos de distancia, mientras que cuando no se de esta métrica, sobre todo con variables dicotómicas, se trabajará sobre parecidos en términos de similaridad. Las propiedades exigibles a una similaridad son:

1. No negatividad, $s(i,j) \geq 0 \quad \forall i,j$
2. $s(i,i) = 0 \quad \forall i$
3. Simetría, $s(i,j) = s(j,i) \quad \forall i,j$

Si s es una similaridad, su opuesto $d = 1 - s$ es una disimilaridad. Para el caso de variables métricas, los individuos pueden ser representados en un espacio euclídeo y entre cada par de ellos puede obtenerse una medidad de discrepancia que verifica las propiedades de una distancia:

1. No negatividad, $d(i,j) \geq 0 \quad \forall i,j$
2. $d(i,i) = 0 \forall i$
3. Simetría, $d(i,j) = d(j,i) \quad \forall i,j$
4. Desigualdad triangular $d(i,j) \leq d(i,k) + d(k,j) \quad \forall i,j,k$
5. $d(i,j) = 0 \Leftrightarrow i \equiv j$

Es evidente que una distancia verifica las propiedades de una disimilaridad. La elección de la disimilaridad va a depender de la naturaleza de las variables consideradas, pudiéndose distinguir entre variables métricas, cualitativas (factores) ordenadas o no ordenadas y, como caso especial de estas últimas, las dicotómicas. Lo habitual en cualquier problema multivariante es tener una matriz con variables de distinto tipo, aunque generalmente la técnica empleada asignará roles en función de la naturaleza de las variables,

no siendo habitual tener que medir diferencias entre individuos sobre un conjunto de variables heterogéneas. Dado el conjunto de variables sobre las que se decida medir las disimilaridades entre individuos se tendrá una medición para cada par de ellos, de forma que toda esta información se podrá organizar en una matriz cuadrada y simétrica de orden n(número de individuos). Muchas técnicas multivariables parten del análisis de la matriz de distancias o de disimilaridades.

Desde la óptica de las variables también se podrá establecer la diferencia o parecido entre cada par de ellas, obteniéndose una matriz cuadrada de orden p(número de variables). El caso más habitual es la **matriz de varianzas-covarianzas**, o su transformada normalizada la **matriz de correlaciones**, de un conjunto de variables continuas. En ambos casos, se trata de matrices de semejanzas entre variables.

3.1. Cuestiones a tener en cuenta

A la hora de medir disimilaridades entre individuos se han de adoptar algunas precauciones. La principal es que la aportación de cada variable estará en función de su variabilidad o rango de variación, y así, si se trata de variables continuas, aquellas con varianza alta aportarán mucho más que aquellas que la tengan pequeña; podría decirse que la desviación típica de una variable representa su factor de escala. Por tanto, una solución para soslayar este problema es tipificar previamente las variables, es decir transformar su factor de escala a la unidad, $\sigma = 1$. Un segundo problema que podría plantearse es que las variables medidas estuvieran fuertemente correlacionadas entre sí, de forma que se estuviera midiendo varias veces lo mismo. La solución en este caso pasa por eliminar la información redundante, bien eliminando las variables altamente correlacionadas con el resto, o bien mediante la transformación de las variables originales en otras ortogonales, empleando en este último caso técnicas de reducción de la dimensión.

4. Disimilaridades y distancias más usadas

En este epígrafe describiremos un conjunto de medidas que podrán emplearse en función de las distintas situaciones que se pueden presentar. Clasificaremos dichas medidas en función de la naturaleza de las variables consideradas.

4.1. Datos cuantitativos

4.1.1. Distancias entre los individuos i y j

1. **Distancia euclídea**

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

2. **Distancia euclídea con pesos**

$$d_{ij} = \sqrt{\sum_{k=1}^{p}\omega_k^2(x_{ik} - x_{jk})^2}$$

Donde ω_k refleja la contribución de cada variable: en general, los pesos se fijan en función de la variabilidad de la variable, siendo los más utilizados:

- $\omega_k = \sigma_k^{-1}$, obteniéndose la Distancia de Mahalanobis
- $\omega_k = (max(x_k) - min(x_k))^{-1}$

3. **Distancia de Minkowski**

$$d_{ij} = (\sum_{k=1}^{p} \mid x_{ik} - x_{jk} \mid^{\lambda})^{\frac{1}{\lambda}}$$

4. **Distancia de Canberra**

$$d_{ij} = \sum_{k=1}^{p} \frac{\mid x_{ik} - x_{jk} \mid}{\mid x_{ik} + x_{jk} \mid}$$

5. **Distancia de Czekanowski**

$$d_{ij} = 1 - \frac{2\sum_{k=1}^{p} min(x_{ik}, x_{jk})}{\sum_{k=1}^{p}(x_{ik} + x_{jk})}$$

4.1.2. Distancias entre las variables k y l

1. $d_{kl} = 1 - r_{kl}$, siendo r_{kl} el coeficiente de correlación entre las variables k y l, siempre que sea positivo.

2. $d_{kl} = 1 - r_{kl}^2$

3.

$$d_{kl} = 1 - (\frac{\sum_{i=1}^{n} x_{ik}x_{il}}{\sqrt{\sum_{i=1}^{n} x_{ik}^2 \sum_{i=1}^{n} x_{il}^2}})^2$$

4.1.3. Distancias entre las poblaciones I y J

1. **Distancia de Mahalanobis**. Distancia al cuadrado entre los centroides de las poblaciones I y J.

2. **Distancia promedio**. Media aritmética de las distancias entre cada individuo de la población I con cada individuo de la población J.

3. **Vecino más próximo**. $min_{ij}d(i,j), i \in I, j \in J$

4. **Vecino más alejado**. $max_{ij}d(i,j), i \in I, j \in J$

4.2. Datos dicotómicos

4.2.1. Disimilaridades entre los individuos i y j

La naturaleza de las variables consideradas y los objetivos de la investigación han dado lugar a una gran variedad de medidas de disimilaridad para variables dicotómicas. Las medidas de disimilaridad sobre variables dicotómicas pueden obtenerse sobre variables del tipo presencia-ausencia de alguna característica o sobre variables bi-estado. Las medidas sobre variables dicotómicas de presencia-ausencia, vienen a medir índices de coexistencia de individuos, de similitud entre individuos o de similitud entre variables. Genéricamente, se consideran a los individuos como entidades o muestras y a las variables en términos de presencia–ausencia de características o de presencia–ausencia de entidades en una localización, abreviadamente taxones o lugares. Los dos esquemas generales de datos para este tipo de situaciones se obtienen al considerar la información conjunta de las objetos i y j organizada en una tabla de doble entrada...

$muestra\ i \backslash\backslash muestra\ j$	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$p=a+b+c+d$

$lugar\ i \backslash\backslash lugar\ j$	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Donde a representa el número de coincidencias en 1, b y c el número de discrepancias y d el número de coincidencias en 0. Se verifica que $a + b + c + d = p$. En función del tipo de índice que se quiera construir, se le exigirá condiciones adicionales a la medida de similaridad. Así, si se pretende construir un índice de coexistencia, la medida debería ser independiente de d, puesto que de otra manera se podrían tener altos grados de similitud entre entidades que, en realidad, casi nunca ocurren. Además, la disimilaridad debería ser cero sí y solo sí $a = 0$. Como se ha comentado anteriormente, la casuística hace que hayan sido mucho los autores que han propuesto medidas para variables dicotómicas. A continuación se describirán las más usuales.

1. **Coeficiente de emparejamiento simple**

$$d_{ij} = 1 - \frac{a+d}{a+b+c+d} = \frac{b+c}{a+b+c+d}$$

   Esta medida considera de igual manera el doble cero (d) que el doble 1 (a). El coeficiente está acotado en el intervalo $[0, 1]$

2. **Russel y Rao**

$$d_{ij} = \frac{a}{a+b+c+d}$$

   Esta medida no toma el valor 1 cuando mide la similitud entre un individuo y sí mismo, salvo que $d = 0$.

3. **Coeficiente de Sokal y Sneath**

$$d_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$$

4. **Coeficiente de Jaccard**

$$d_{ij} = \frac{b+c}{a+b+c}$$

   Este coeficiente excluye los dobles ceros.

5. **Coeficiente de Dice-Sorensen**

$$d_{ij} = \frac{b+c}{2a+b+c}$$

   Es una variante del coeficiente de Jaccard

5. Cálculo de distancias con R

Se considera la matriz de datos siguiente, que contiene 13 variables de distinta naturaleza. Se utilizarán algunas de las variables para calcular algunas distancias con **R**.

Entidad	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}
1	1	49	0	2	3	7	0	6	2	15	1	1	52,46
2	0	59	0	2	2	6	3	7	2	17	1	1	60,83
3	0	49	2	0	6	8	4	9	0	11	1	0	48,70
4	0	58	0	2	1	1	1	7	3	14	1	1	57,33
5	1	49	0	2	1	0	0	3	0	13	0	1	64,26
6	1	43	0	2	0	2	0	4	0	11	1	0	53,90
7	0	46	1	1	4	7	0	8	1	16	1	1	52,45
8	0	59	0	2	7	6	3	4	0	14	1	1	66,18
9	0	53	0	1	4	8	0	1	2	14	1	0	49,54
10	0	51	1	0	4	3	0	7	1	11	1	1	53,88
11	0	50	2	0	10	0	0	3	1	21	0	0	58,46
12	1	50	1	1	3	8	4	2	0	16	1	1	50,22

```
> distancias <- data.frame(structure(list(V1 = c(1, 0, 0, 0, 1,
+     1, 0, 0, 0, 0, 0, 1), V2 = c(49, 59, 49, 58, 49, 43, 46,
+     59, 53, 51, 50, 50), V3 = c(0, 0, 2, 0, 0, 0, 1, 0, 0, 1,
+     2, 1), V4 = c(2, 2, 0, 2, 2, 2, 1, 2, 1, 0, 0, 1), V5 = c(3,
+     2, 6, 1, 1, 0, 4, 7, 4, 4, 10, 3), V6 = c(7, 6, 8, 1, 0,
+     2, 7, 6, 8, 3, 0, 8), V7 = c(0, 3, 4, 1, 0, 0, 0, 3, 0, 0,
+     0, 4), V8 = c(6, 7, 9, 7, 3, 4, 8, 4, 1, 7, 3, 2), V9 = c(2,
+     2, 0, 3, 0, 0, 1, 0, 2, 1, 1, 0), V10 = c(15, 17, 11, 14,
+     13, 11, 16, 14, 14, 11, 21, 16), V11 = c(1, 1, 1, 1, 0, 1,
+     1, 1, 1, 1, 0, 1), V12 = c(1, 1, 0, 1, 1, 0, 1, 1, 0, 1,
+     0, 1), V13 = c(52.46, 60.83, 48.7, 57.33, 64.26, 53.9, 52.45,
+     66.18, 49.54, 53.88, 58.46, 50.22)), .Names = c("V1", "V2",
+     "V3", "V4", "V5", "V6", "V7", "V8", "V9", "V10", "V11", "V12",
+     "V13"), row.names = c(NA, 12L)))
```

5.1. La función dist

La función **dist** del paquete **stats** calcula la matriz de distancias entre las filas de una matriz de datos en función de las variables que se especifiquen. La sintaxis de la función es:

dist(x, method = ëuclidean, diag = FALSE, upper = FALSE, p = 2)

donde:

- **x** es la matriz de datos
- **method** es la distancia a usar. Las opciones son euclidean, maximum, manhattan, canberra, binary y minkowski
- **diag**, si es TRUE imprime además la diagonal principal
- **upper**, si es TRUE imprime además el triángulo superior
- **p** es la potencia de la distancia de Minkowski

Ejemplo 5.1

Para calcular la matriz de distancias euclídeas entre las entidades de la matriz de datos en función de las variables V_2, V_10 y V_13, se procedería:

```
> dist(distancias[, c(2, 10, 13)])

           1         2         3         4         5         6         7
2  13.193063
3   5.489772 16.826672
4  10.281872  4.716991 12.824855
5  11.968291 11.303314 15.688008 11.402846
6   7.353475 18.439764  7.939773 15.676891 12.137941
7   3.162293 15.499174  6.932712 13.107799 12.548948  6.008536
8  17.007010  6.133718 20.360511  8.906318 10.231637 20.391135 19.013493
9   5.052366 13.132559  5.070069  9.256571 15.286543 11.314133  7.840159
10  4.692164 12.177951  5.552693  8.360771 10.758457  8.000025  7.214215
11  8.544004 10.130000 14.009197 10.690037  9.931767 13.030487  8.781805
12  2.649075 13.948910  5.320752 10.888163 14.391720  9.356410  4.579618
           8         9        10        11
2
3
4
5
6
7
8
9  17.688686
```

```
10 14.976315  5.642304
11 13.769473 11.728870 11.044293
12 18.431538  3.669114  6.276591  9.638340
```

Si se desea la matriz completa de distancias euclídeas entre las cinco primeras entidades, en función de las variables V_2, V_10 y V_13:

```
> dist(distancias[1:5, c(2, 10, 13)], diag = T, upper = T)

          1         2         3         4         5
1  0.000000 13.193063  5.489772 10.281872 11.968291
2 13.193063  0.000000 16.826672  4.716991 11.303314
3  5.489772 16.826672  0.000000 12.824855 15.688008
4 10.281872  4.716991 12.824855  0.000000 11.402846
5 11.968291 11.303314 15.688008 11.402846  0.000000
```

Ejemplo 5.2

Para calcular la matriz de distancias de Minkowski con $p = 3$ entre las cinco primeras entidades de la matriz de datos **??***distancias, en función de las variables V_2, V_10 y V_13, se procedería:*

```
> dist(distancias[1:5, c(2, 10, 13)], method = "minkowski", p = 3)

          1         2         3         4
2 11.682352
3  4.893165 14.423730
4  9.455941  4.138386 11.183521
5 11.819121 10.336402 15.571006 10.205138
```

y la distancia de Canberra para los mismos elementos:

```
> dist(distancias[1:5, c(2, 10, 13)], method = "canberra")

          1         2         3         4
2 0.2289738
3 0.1910150 0.4176242
4 0.1629523 0.1349421 0.2855042
5 0.1725252 0.2533462 0.2210812 0.1781440
```

6. Ejercicios propuestos

1. Usando el producto escalar calcule el ángulo que forman los vectores-variables $x \equiv (1, -1, 3, 7, 0)$ e $y \equiv (0, 1, 1, 1, 4)$

2. ¿Podría decir cuál es el `coeficiente de correlación lineal` de las variables x e y del ejercicio anterior?¿Y su `covarianza`?

3. ¿Qué angulo forman los vectores $x \equiv (3, 7)$ e $y \equiv (1, 4)$?¿Cuál es su coeficiente de correlación?

4. ¿Cuál es el coeficiente de correlación de los vectores $x \equiv (5, 1)$ e $y \equiv (1, 4)$?

5. ¿Qué conclusiones saca de los resultados de los dos ejercicios anteriores?

6. Defina en R la matriz `A`= $\begin{pmatrix} 2 & 3 & 5 & 1 \\ 2 & 4 & 1 & 1 \\ 3 & 1 & 5 & 2 \\ 4 & 6 & 2 & 3 \end{pmatrix}$ y a continuación:

   - Obtenga su traspuesta y su inversa
   - Calcule su determinante

7. Resuelva el siguiente sistema de ecuaciones con R:

$$\begin{aligned} x + y + z &= 3 \\ 2x - y - z &= 0 \\ 3x + 4y + 5z &= 1 \end{aligned}$$

8. Sobre el fichero alimentos.dat de la carpeta Datos obtenga las distancias euclídeas sobre los 10 primeros individuos; hágalo tanto sobre los datos originales como sobre los tipificados. Utilice otras distancias sobre los mismos datos.

UNIVERSIDAD DE CADIZ
Departamento de Estadística e I.O.
Grupo de investigación TeLoyDisRen

Gráficos Multivariantes con R

1. Introducción

En Estatística juegan un papel fundamental los gráficos, éstos permiten obtener una visión global de los datos con poco esfuerzo. Existen multitud de gráficos que dependiendo de su tipología, pueden ser adecuados para su uso con datos unidimensionales o bidimensionales, pero que no son fácilmente extensibles a situaciones donde se trabaja en espacios de dimensión mayor o igual a tres. En estas situaciones se buscan criterios que permitan trasladar e interpretar en el plano o en el espacio la información contenida en los datos, de forma que puedan apreciarse algunas de sus características generales. Algunas técnicas multivariantes vienen acompañadas de representaciones gráficas propias, que ayudan a comprobar el cumplimiento de requisitos básicos exigidos a los datos y a interpretar los resultados de su aplicación. Sin embargo, existen otro tipo de representaciones gráficas genéricas, no ligadas a una técnica particular, que permiten apreciar, como hemos dicho, ciertas características generales: regularidades, singularidades, similitudes entre individuos y variables, etc. A ellas vamos a dedicar este capítulo. Las técnicas que se describirán, –Diagramas de dispersión múltiple, curvas de Andrews, caras de Chernov y gráficos de estrella– son aplicables a variables cuantitativas.

2. Curvas de Andrews

Propuesta por D.F. Andrews en 1972 esta representación gráfica construye una curva para cada individuo a partir de los valores en cada una de las variables utilizando una serie de Fourier finita. Sean las variables $X = [X_1, X_2, ..., X_p]$, cada observación multivariante $x_i = (x_{i1}, .., x_{ip})$ se transforma en la siguiente curva:

$f_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}sin(t) + x_{i3}cos(t) + x_{i4}sin(2t) + x_{i5}cos(2t) + \ldots$

con $-\pi \leq t \leq \pi$.

De esta manera es posible representar un objeto multidimensional en el plano. Individuos semejantes se deben corresponder con curvas de comportamiento parecido, mientras que aquellos que no lo sean diferirán en alguna parte de sus respectivas curvas. Por tanto, un conjunto de curvas parecidas caracterizará un subgrupo homogéneo de individuos. Además, podemos interpretar las curvas que difieren contundentemente del resto como individuos extremos o anómalos.

En la representación de Andrews, el orden de las variables en la matriz de datos juega un papel fundamental, puesto que sus valores son los coeficientes de una serie de Fourier. Ello hace aconsejable probar distintas ordenaciones, al objeto de encontrar aquella que proporcione una información más clara. Incluso, previo a la obtención de las curvas, se pueden utilizar técnicas de síntesis y reducción de la dimensión de la matriz de datos, como el `Análisis de componentes principales` que se verá en el próximo capítulo.

Esta técnica no es recomendable para un conjunto grande de observaciones, pues en ese caso se produce una mala representación al superponerse demasiadas curvas en el gráfico. Por otro lado, si se quiere destacar la pertenencia de los individuos a clases determinadas por un factor, puede ser interesante pintar las curvas en distintos colores en función de su clase de pertenencia.

2.1. Propiedades

Las curvas de Andrews poseen las siguientes propiedades:

I Conserva distancias euclídeas: individuos cercanos en el espacio p-dimensional original se corresponden con curvas cercanas en la gráfica para todos los valores de t.

II Preserva las relaciones lineales: Si A está alineado entre B y C, entonces $f_A(t)$ lo está entre $f_B(t)$ y $f_C(t)$.

III Conserva medias y varianzas.

2.2. Curvas de Andrews en R

Actualmente este procedimiento no está implementado en ninguno de los paquetes de la distribución de `R`, por lo que hay que introducir el código en el editor. Concretamente, hay que cargar dos funciones: `andrews.function` y `andrews.curves`.

```
andrews.function <- function (xs, no.pts=101){
    n <- length(xs)
    xpts <- seq(0, 2*pi, length=no.pts)
    ypts <- c()
    for (p in xpts) {
        y <- xs[1]
        for (i in 2:n) {
             if (i %% 2 == 1) { y <- y + xs[i]*sin((i %/% 2)*p) }
             else             { y <- y + xs[i]*cos((i %/% 2)*p) }
        }
        ypts <- c(ypts, y)
    }
    return(ypts)}

andrews.curves <- function(xdf, cls, npts=101, title="Classes") {
    n <- nrow(xdf)
    clss <- as.factor(cls)
    xpts <- seq(0, 2*pi, length=npts)
    X <- xpts
    for (i in 1:n) {
        xi <- unname(unlist(xdf[i, ]))
        ys <- andrews.function(xi, npts)
        X <- cbind(X, ys)
    }
    ymin <- min(X[, 2:(n+1)])
    ymax <- max(X[, 2:(n+1)])
```

```
    plot(0, 0, type="n", xlim=c(0, 2*pi), ylim=c(ymin, ymax),
         main="Curvas de Andrews", xlab="t", ylab="f(t)")

    clrs <- as.integer(clss)
    for (i in 2:(n+1)) {
        lines(X[, 1], X[, i], col=clrs[i-1])
    }
    legend(4, ymax, levels(clss), col=c(1:nlevels(clss)), lty=1)
    # return(X)
}
```

Ejemplo 1 *Se considera el fichero de datos de las* ***Flores de iris*** *(Fisher, 1936), aunque dado que su número de individuos, 150, es excesivo, representaremos las curvas de Andrews de los 10 primeros individuos de cada una de las tres especies. Para ello, se ejecuta en* ***R*** *el código:*

```
library(datasets, pos=4)
data(iris, package="datasets")
iris_10 <- iris[-c(11:50,61:100,111:150),]
#se eliminan los casos que no interesan
old <- par(bg="whitesmoke")
andrews.curves(iris_10[,1:4], iris_10[,5])
par(old)
```

Se observa como la especie setosa posee un comportamiento muy diferente a las otras dos, y como dentro de cada especie la forma de las curvas es muy parecida. Figura 1

3. Caras de Chernov

Chernov plantea una representación gráfica en la que a cada individuo le corresponde una carita cuyas facciones se determinan según los valores que toma en cada una de las variables utilizadas. A individuos con valores parecidos le corresponden caras similares, y viceversa. Una de las críticas que se le hace a esta representación gráfica es el grado de subjetividad a la hora de apreciar caras similares o disimilares, en este punto habrá que tener en cuenta que la mayoría de las representaciones gráficas en mayor o menor medida tienen una cierta carga de subjetividad.

Ejemplo 2 *Para 6 variables económicas (deflactor implícito de precios (1954=100), producto nacional bruto, número de personas en las fuerzas armadas, población no institucionalizada* ≥ 14 *años, número de personas empleadas) estudiadas entre los años 1947 y 1962. Se puede observar un comportamiento parecido en los años 51, 52 y 53, como contrapunto a lo ocurrido en 1949 y 1954. Figura 2*

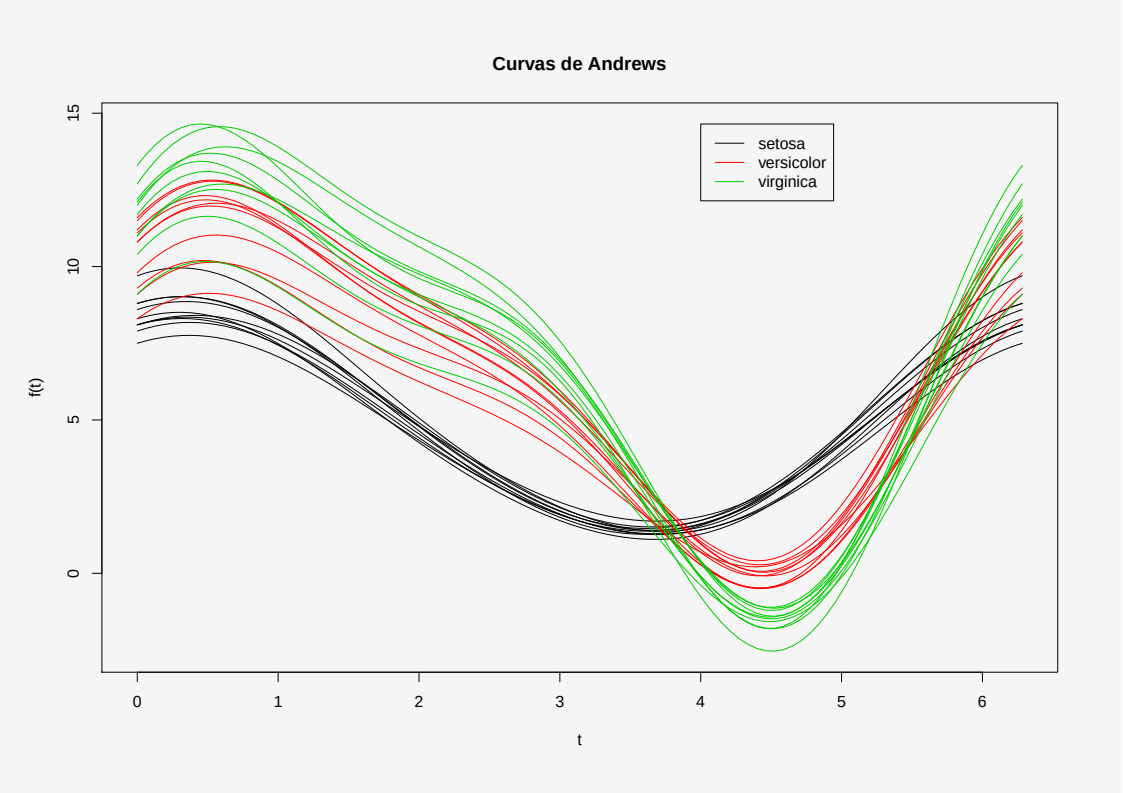


Figura 1: Curvas de Andrews de flores de Iris

3.1. Caras de Chernoff con R

En **R** se pueden obtener las `caras de Chernov` utilizando dos funciones distintas, `faces` y `faces2`. La diferencia entre ellas radica en la forma de construir las caras y en la asignación de las variables a la construcción de las distintas facciones. Ambas se encuentran en el paquete `TeachingDemos`,

Para cargar el paquete, se ejecuta...

```
library(TeachingDemos)
```

La orden `faces` permite crear el diagrama de caras a partir de una matriz de variables numéricas. Las características de la cara se corresponden con el número de columna de la siguiente manera:

Por defecto estas medidas se normalizan según el número de observaciones, por lo que dados `n` individuos si se representan todos menos uno en un gráfico y en un segundo gráfico se representan todos, se tiene que las caras de las observaciones cambian de un gráfico a otro. Esto es un inconveniente si se desea realizar la tarea de clasificación de un individuo según los rasgos faciales del diagrama.

Ejemplo 3 *Se considera el fichero de datos que recoge la composición de la leche de 22 mamíferos,* ***`leche_mamiferos.rda`****. Las variables consideradas, en el orden que aparecen en la matriz, son:* ***`agua`****,* ***`proteina`****,* ***`grasa`*** *y* ***`lactosa`****. Se carga el fichero...*

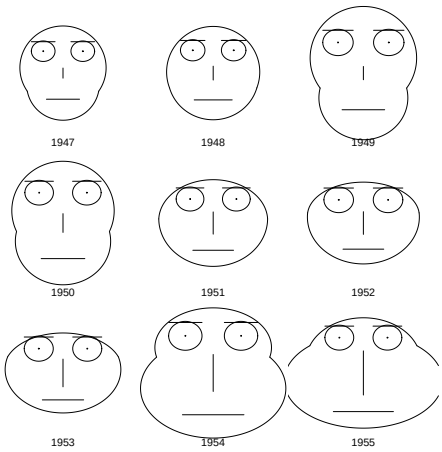


Figura 2: Caras de Chernov de datos económicos

Columna	Rasgo de la cara	Columna	Rasgo de la cara
1	altura de la cara	9	altura del cabello
2	ancho de la cara	10	ancho del cabello
3	forma de la cara	11	estilo de cabello
4	altura de la boca	12	altura de la nariz
5	ancho de boca	13	ancho de la nariz
6	curva de sonrisa	14	ancho de las orejas
7	altura de los ojos	15	altura de las orejas
8	ancho de los ojos		

```
load(".../leche_mamiferos.rda")
```

Utilizando la orden `faces2`

```
faces2(leche_mamiferos[,2:5], main="Leche mamíferos", labels =Datos$Animal)
```

se obtiene la Figura 3. Entre otras, se observan similitudes en la composición de la leche entre Yegua, Burra, Cebra, Mula y Llama, otro grupo estaría formado por Foca y Delfina.

4. Gráficos de estrella

En este tipo de gráfico se trata de representar a cada individuo mediante una estrella, de cuyo centro parten una serie de rayos o ejes (tantos como variables utilicemos en el estudio) manteniendo igual espacio entre uno y otro sobre un círculo o porción de círculo, a partir de un rayo arbitrario, el ángulo entre éste y el j-ésimo rayo es:

$$\theta_j = \frac{2\pi(j-1)}{p}, \forall j = 1, 2, ..., p$$

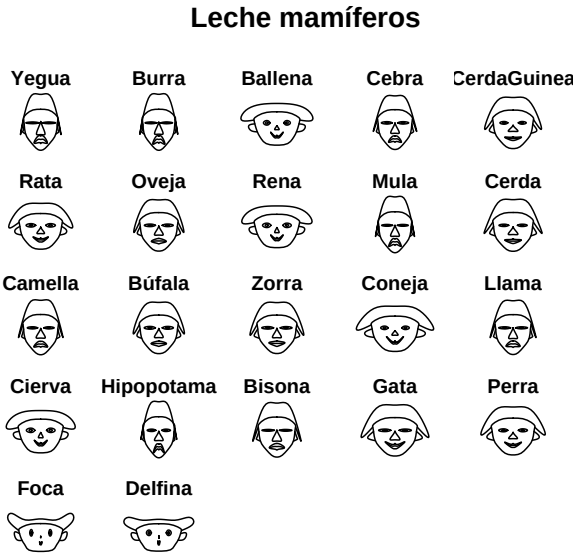


Figura 3: Caras de Chernov de la composición de la leche de mamíferos

La longitud de cada rayo se puede establecer siguiendo distintos criterios:

I Se tipifica cada variable, y sobre el cero común se le da al eje correspondiente la longitud del valor tipificado.

II Se transforma la variable X_i de forma que tome valor 1 cuando el valor de X_i sea máximo y 0 cuando sea mínimo.

III Se transforma la variable X_i de forma que esté acotada entre 0 y 1, para ello

$$\frac{x_i - min(x_i)}{max(x_i)}$$

4.1. Gráfico de estrellas con R

La función que permite obtener este gráfico en R es `stars`, que posee, entre sus distintas opciones, la posibilidad de representarlo en el semicírculo superior (opción full=FALSE), en segmentos (draw.segments=TRUE), en colores (elegiendo una paleta de colores), en forma de radar (location=c(0,0), key.loc=c(0,0)), etc.

Ejemplo 4 *Se considera el fichero de datos* `iris` *del paquete* `datasets` *de* ***R****. Representaremos mediante estrellas los 20 primeros individuos considerando las cuatro variables métricas. Figura 4*

```
library(datasets)
data(iris, package="datasets")
stars(iris[1:20,1:4],draw.segments=TRUE,col.segments=c(2:5))
```

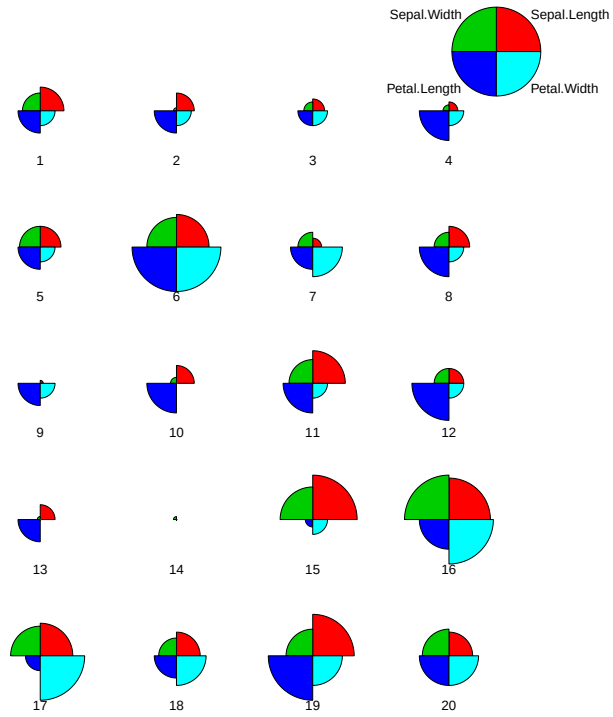


Figura 4: Grafico de estrellas para Iris

5. Matrices de dispersión

Como generalización del diagrama de dispersión entre dos variables continuas, se pueden dibujar en un único gráfico todas las combinaciones de diagramas de dispersión de un conjunto de variables.

5.1. Diagramas de dispersión con R

Puede accederse a este tipo de gráfico desde el submenu `Gráficas` de **Rcmdr**, aunque si se desea explotar todas las posibilidades habrá que editar las instrucciones añdiendo algunos parámetros.

Por defecto, **Rcmdr** usa las instrucciones que se dan a continuación para generar la matriz de dispersión en función de la especie. Figura 5

```
library(datasets)
data(iris, package="datasets")
scatterplot.matrix(~Petal.Length+Petal.Width+Sepal.Length+Sepal.Width |
Species, reg.line=lm, smooth=TRUE, span=0.5, diagonal= 'density',
by.groups=TRUE, data=iris)
```

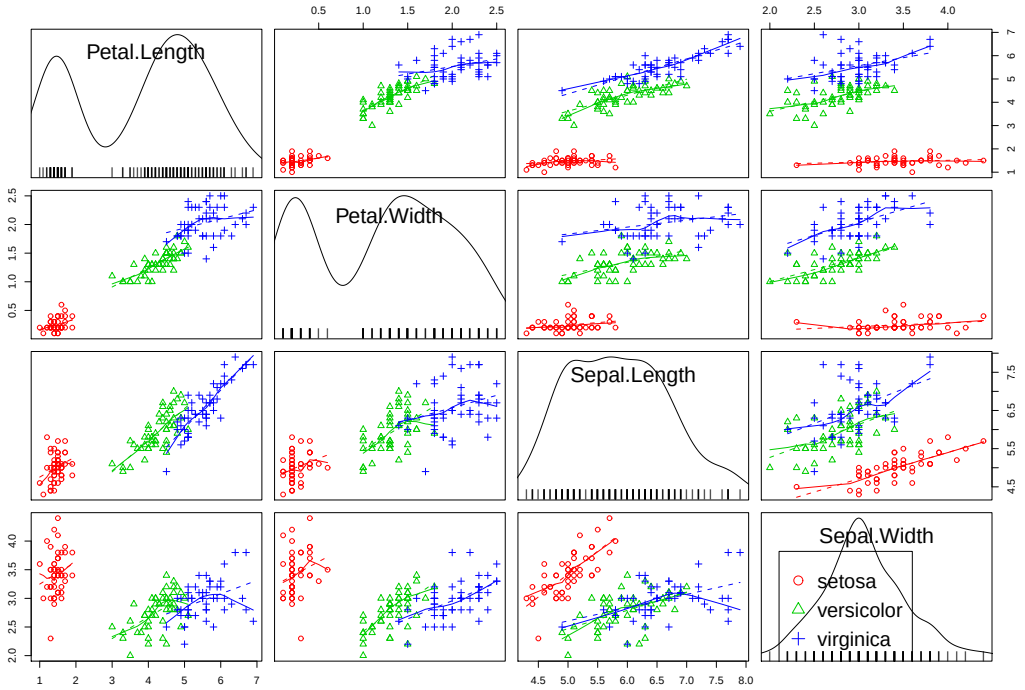


Figura 5: Matriz de dispersión de las flores de iris en función de la especie

Otra posibilidad es construir una matriz que incluya las correlaciones y los gráficos de dispersión, para ello se considera el código que se incluye a continuación, resultando la figura 6

```
panel.cor <- function(x, y, digits=2, prefix=, cex.cor)
usr <- par(üsr"); on.exit(par(usr))
par(usr = c(0, 1, 0, 1))
r <- cor(x, y)
txt <- format(c(r, 0.123456789), digits=digits)[1]
txt <- paste(prefix, txt, sep=)
if(missing(cex.cor))
cex <- 0.8/strwidth(txt)
text(0.5, 0.5, txt, cex = cex)
pairs(iris[1:4], lower.panel=panel.smooth,
upper.panel=panel.cor)
```

6. Ejercicios

- Carga los datos del fichero “UScereal” contenido en el paquete “MASS” y construye las curvas de Andrews para los distintos cereales según la variable “mfr” (empresa fabricante) utilizando las distintas variables cuantitativas. Modifica el orden de las variables y compara los resultados obtenidos.
  - Con los datos anteriores obtén las caras de Chernov con la función “faces2”
  - Construye los gráficos de estrella

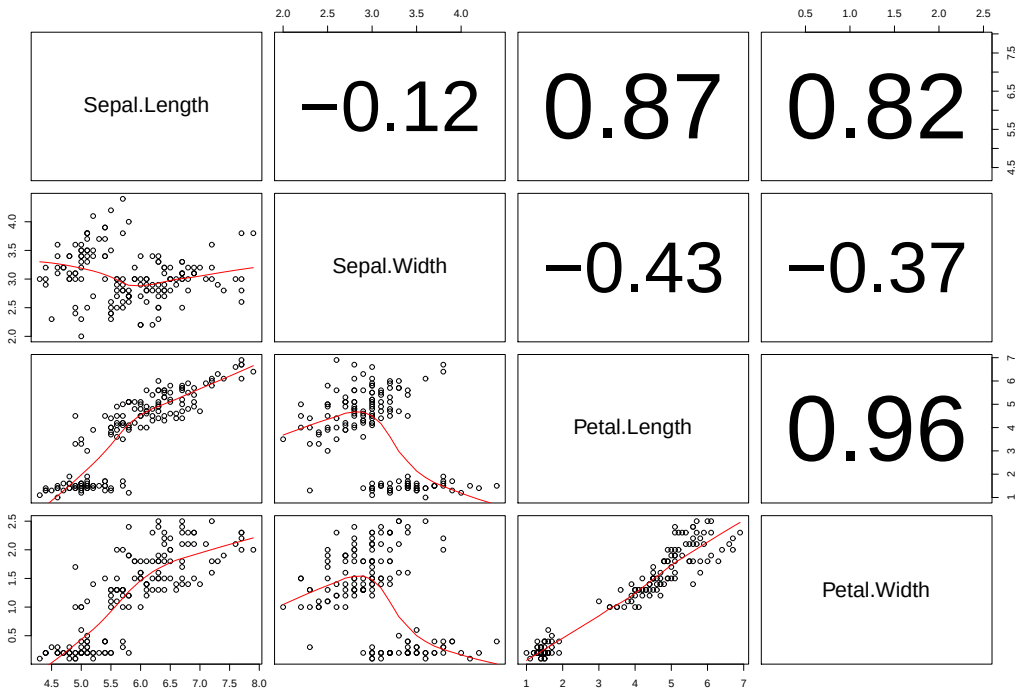


Figura 6: Matriz de dispersión de las flores de iris con correlaciones

- Construye la matriz de dispersión de las variables “calories”, “carbo”, “fat” y “fibre”

- Utilizando los datos de la leche de las hembras de los mamíferos, `Leche_mamiferos.dat`, genera las curvas de Andrews, los gráficos de estrella y las caras de Chernov para los individuos c(1,2,3,8,20,22)

  - Cambia el orden las variables para las curvas de Andrews y observa el efecto que produce.

- Utilizando el fichero Chile, aplica las representaciones gráficas expuestas a los individuos 2, 46, 49 y 129, utilizando las variables “population”, “age”, “income” y “statusquo”.

UNIVERSIDAD DE CADIZ
Departamento de Estadística e I.O.
Grupo de investigación TeLoyDisRen

Análisis de Componentes Principales con R

1. Introducción

Muchos estudios estadísticos presentan un modelo de datos en que el que, para un conjunto de individuos, se han medido una serie de características que los investigadores han considerado importantes. Dado que los responsables de la investigación no quieren dejar de considerar información que podría ser relevante, dentro de las restricciones establecidas de economía de recursos y tiempo, la cantidad de variables que incorporan en primera instancia a la base de datos suele ser muy grande.

En general, para alcanzar los objetivos marcados en un estudio se necesitaría información de distinta naturaleza que suele clasificarse en varios bloques homogéneos. Por ejemplo, un estudio epidemiológico requerirá de información sobre parámetros fisiológicos, bioquímicos, medioambientales, etc. Y dentro de cada uno de esos bloques se medirán varias características, que a menudo presentarán rasgos comunes. Siguiendo con el ejemplo, será habitual que si se analizan parámetros fisiológicos, individuos que tengan valores altos de alguna característica–talla alta–, conserve la tendencia en el resto–peso alto, brazos largos–, etc.

Las reflexiones anteriores plantean dos cuestiones que según el caso pueden llegar a ser problemáticas, una evidente de dimensión, es decir de tamaño del problema a resolver, y otra más sutil derivada de la existencia de información redundante entre variables, que se puede calibrar, "grosso modo", a través de la matriz de correlaciones. En realidad ambos problemas están muy relacionados, dado que si se restringe el estudio a unas pocas variables y el investigador tiene un cierto grado de conocimiento de la situación, es muy factible que pueda elegir grupos de variables de forma que las correlaciones entre variables de grupos distintos se aproximen a cero.

El Análisis de Componentes Principales(ACP) viene a dar respuesta a las dos cuestiones planteadas, clasificándose, dentro del conjunto de técnicas multivariantes de Reducción de la dimensión. El ACP puede ser una técnica finalista que de respuesta a alguno de los objetivos del estudio, aunque lo más habitual es que sea un algoritmo previo a la aplicación de otras técnicas más sofisticada, como puede ser el Análisis Factorial, o incluso una herramienta que construya variables incorreladas para aplicar, por ejemplo, una Regresión Múltiple.

Formalmente, el **ACP** es una técnica estadística multivariante de simplificación o reducción de la dimensión, que permite transformar un conjunto de variables correlacionadas en otro conjunto de variables ortogonales denominadas *Componentes* o *Ejes principales*. Para aplicar el **ACP** se requiere que todas las variables de la matriz de datos sean cuantitativas o asimilables a éstas. La consecuencia inmediata de esta restricción es que, en general, los datos van a tener una distribución Normal multidimensional, o bien podrán ser transformados o divididos para que se de este supuesto. El objetivo concreto del **ACP** es encontrar un subespacio k-dimensional, $k < p$, desde el que "ver" de forma óptima la configuración geométrica de la nube de puntos p–dimensional.

2. Enfoque geométrico del ACP

El punto de partida para la aplicación de un ACP es una matriz *nxp* de individuos–variables.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Desde un punto de vista geométrico, la matriz de datos se puede ver como la representación de los individuos(filas) mediante puntos en el espacio de p dimensiones definido por las variables(columnas), conformando lo que se conoce como una *nube de puntos*. Las coordenadas del individuo i-ésimo serán $(x_{i1}, x_{i2}, \cdots, x_{in})$.

La representación del conjunto de los n puntos en el espacio p–dimensional tiene, en general, una forma de *hiperelipsoide*. En tres dimensiones, el elipsoide se parecería a un balón de rugby aplastado contra el suelo. Cuando p aumenta la nube de puntos se hace menos densa –siempre que el número de individuos, n, permanezca constante–. Los puntos tienen un *centro de gravedad* cuyas coordenadas coinciden con las medias de las variables. Si sustraemos de cada punto su centro-media el centro de gravedad se convierte en el orígen de coordenadas. *(Figura 1)*.

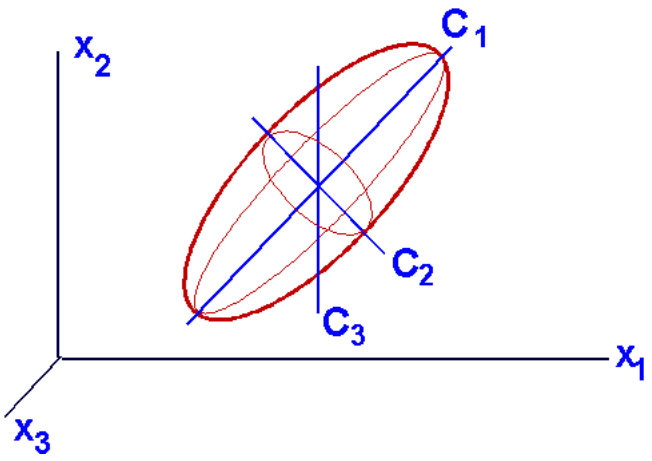


Figura 1: Elipsoide de dispersión

La correlación entre dos variables es una medida de la desviación de la circularidad en la proyección de la nube de puntos, depués de la estandarización, en el plano formado por los ejes coordenados de las variables. Es decir es una medida de la excentricidad de la elipse.

En el *espacio de variables*, cada variable viene representada en un espacio de dimensión $n-1$ –si se han estandarizado previamente– por un punto. Aunque es más pertinente

considerar cada variable como un vector desde el orígen hasta dicho punto. Los *espacios Objeto y Variable* son claramente entidades no independientes, debido a que ambos se derivan de la misma matriz de datos. Sin embargo, hay una cierta falta de simetría en la manera en que se dan las relaciones en los dos espacios. Las técnicas multivariables que se basan en el análisis de la matriz *nxn* de distancias entre individuos se conocen como *Q-técnicas*, mientras que aquellas cuyo punto de partida es la matriz *pxp* de relaciones entre variables se conocen como *R-técnicas*.

3. Ejemplo de resolución del ACP en dos dimensiones

Supongamos la distribución de la *altura* y el *peso* de un grupo de personas representada en la *figura 2*. Es posible que la representación en las variables originales no sea demasiado relevante y que interese una *rotación de ejes* sin cambiar la configuración de los puntos.

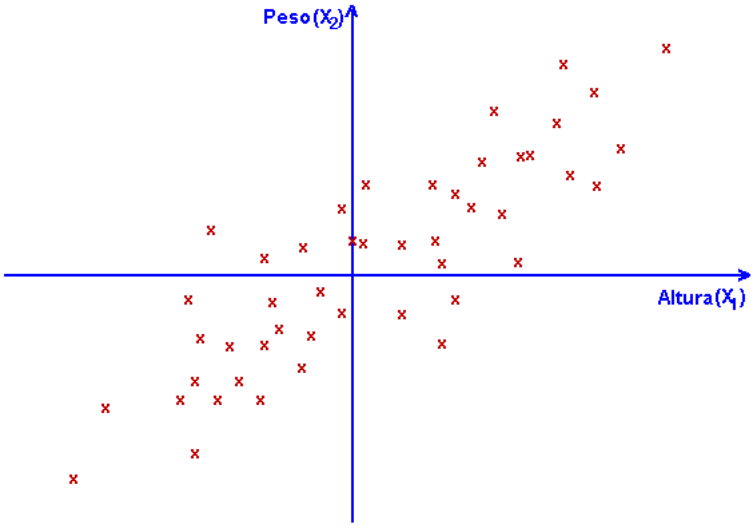


Figura 2: Dispersion_pesoaltura

Si consideramos un eje en la dirección que marca la tendencia de los puntos y un segundo eje perpendicular a éste, los individuos pueden explicarse de forma más intuitiva que con las mediciones originales. *(Figura 3)*. De hecho, podríamos relacionar el primer eje, Y_1, con la tendencia central del conjunto de los individuos, mientras que el segundo, Y_2, mide la separación de los mismos de la tendencia central.

Las coordenadas (y_1, y_2) en función de las (x_1, x_2) para cualquier punto de la muestra vienen dadas por:

$$\begin{aligned} y_1 &= x_1 cos\alpha + x_2 sen\alpha \\ y_2 &= -x_1 sen\alpha + x_2 cos\alpha \end{aligned}$$

Vemos que Y_1 e Y_2 son combinaciones lineales de Peso y Altura. Por otra parte, si proyectamos los puntos sobre los nuevos ejes, vemos que mientras que el rango de valores

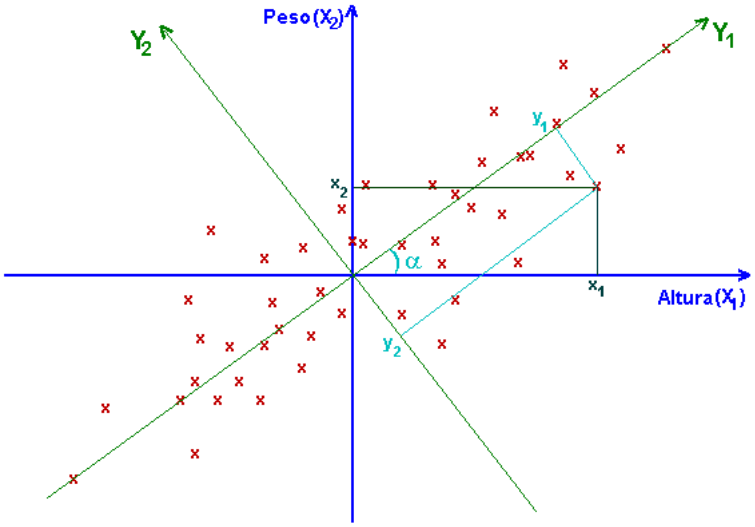


Figura 3: Rotacion_ pesoaltura

de Y_1 es grande, el de Y_2 es considerablemente más pequeño. Podríamos concluir que la aproximación unidimensional en Y_1 de la muestra bidimensional original es razonablemente buena, ya que los desplazamientos de los puntos en la proyección es pequeña. En otras palabras, si en lugar de dar *peso* y *altura* damos el Y_1 caracterizamos a los individuos bastante bien.

Lo que acabamos de hacer no deja de ser una aproximación gráfica a la solución del problema. La determinación de la solución exacta pasa por la obtención del valor de α, para lo que habrá que establecer un criterio. A la vista de la solución gráfica, el criterio que emplearemos consistirá en minimizar los desplazamientos de los puntos en su proyección sobre el eje Y_1. *(Figura 4)*. Si llamamos $P_i^{'} \quad i = 1, \ldots, n$ a la proyección del punto P_i sobre Y_1.

El criterio de obtención de α nos lleva a $Min \sum_{i=1}^{n} (P_i P_i^{'})^2$; aplicando el Teorema de Pitágoras se tiene:

$$\begin{array}{rcl} (OP_i)^2 & = & (OP_i^{'})^2 + (P_i P_i^{'})^2 \\ \sum (OP_i)^2 & = & \sum (OP_i^{'})^2 + \sum (P_i P_i^{'})^2 \\ \frac{1}{n-1} \sum (OP_i)^2 & = & \frac{1}{n-1} \sum (OP_i^{'})^2 + \frac{1}{n-1} \sum (P_i P_i^{'})^2 \end{array}$$

Por tanto, $Min \frac{1}{n-1} \sum (P_i P_i^{'})^2$ equivale a $Max \frac{1}{n-1} \sum (OP_i^{'})^2$. Puesto que O es el centroide de los puntos, $\frac{1}{n-1} \sum (OP_i^{'})^2$ es la varianza muestral cuando los individuos vienen dados por sus coordenadas Y_1 (Hotelling, 1933).

4. Caso p-dimensional

Para el caso p-dimensional, se trata de encontrar un eje Y_1 tal que la proyección de los puntos sobre ese eje tenga máxima dispersión.

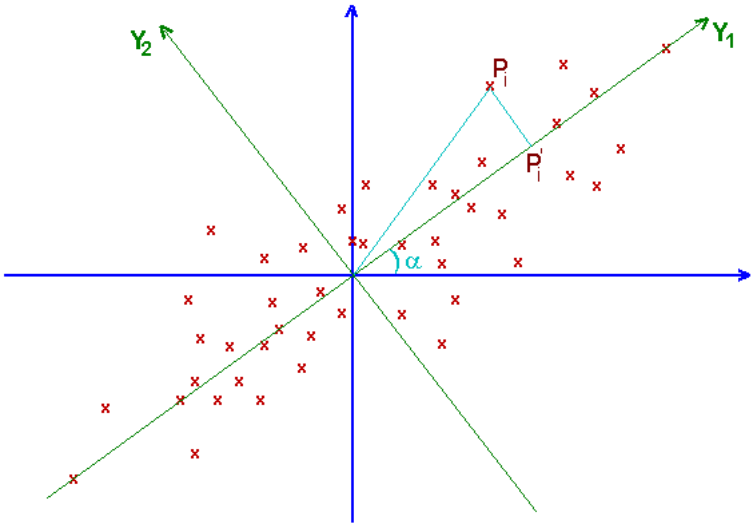


Figura 4: Proyección_ pesoaltura

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$s.a. \quad \sum_{j=1}^{p} a_{1j}^2 = 1$$

$$V[Y_1] \quad \text{máxima}$$

Para el resto de componentes $Y_2, \ldots, Y_p$ se tendría que:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

$$s.a. \quad Y_i \perp Y_1, Y_2, \ldots, Y_{i-1}$$

$$\sum_{j=1}^{p} a_{ij}^2 = 1$$

$$V[Y_i] \quad \text{máxima}$$

En definitiva, cada eje se obtiene maximizando la dispersión de los puntos proyectados sobre él y exigiendo que sea ortogonal a los anteriores. A los coeficientes a_{ij} se les conoce como pesos y representan la importancia de la variable X_j para explicar el comportamiento de la componente Y_i.

Si la nube de puntos se encontrara contenida en un subespacio de dimensión $t < p$, entonces el número de componentes que se pueden obtener sería precisamente t. Lo que esto significa es que $p - t$ de las variables originales se explican de forma exacta mediante combinaciones lineales del resto.

4.1. Interpretación matemática de las componentes principales

Las componentes principales desde un punto de vista matemático son los autovalores de la *matriz de covarianzas* $\mathcal{S}$ o de la *matriz de correlaciones* $\mathcal{R}$. Las varianzas de las componentes coinciden con los autovalores asociados e indican el peso de la componente en cuestión. En realidad, podemos asociar la varianza de una componente a la cantidad de información que retiene de la matriz de datos original. Las componentes son ortogonales y no comparten información, con lo cual el ACP puede ser una técnica previa a la realización de otras técnicas multivariantes, como puede ser un *Análisis de Regresión*, evitándose en este caso problemas de *multicolinealidad*.

4.2. Reducción de la dimensión

Si deseamos obtener la mejor representación k-dimensional, $k < p$, del modelo geométrico p-dimensional, simplemente necesitamos proyectar los puntos en el subespacio k-dimensional definido por las k primeras componentes, $Y_1, Y_2, \ldots, Y_k$. Para un individuo I_s con coordenadas originales $(x_{s1}, \ldots, x_{sp})$, las nuevas coordenadas en el espacio de proyección vendrían dadas por:

$$\begin{aligned} y_{s1} &= a_{11}x_{s1} + a_{12}x_{s2} + \cdots + a_{1p}x_{sp} \\ y_{s2} &= a_{21}x_{s1} + a_{22}x_{s2} + \cdots + a_{2p}x_{sp} \\ \cdots &\quad \cdots \qquad\qquad \cdots \\ y_{sk} &= a_{k1}x_{s1} + a_{k2}x_{s2} + \cdots + a_{kp}x_{sp} \end{aligned}$$

Repitiendo esta operación para todos los individuos, obtenemos la matriz nxk de las puntuaciones en las k primeras componentes. La reducción de la dimensión conlleva una pérdida de información que deberí ser evaluada. Como contrapartida, las componentes retenidas son ortogonales y no comparten información (lineal) entre ellas.

A menudo se consideran solo las dos primeras componentes y se obtiene una aproximación bidimensional de la configuración original de los puntos en p dimensiones. El plano definido por las dos primeras componentes es el plano de mejor ajuste para los n puntos, en el sentido introducido por **Pearson (1901)** de minimizar los desplazamientos perpendiculares de los puntos al plano. La belleza de este método es que si, a partir del plano, quisiéramos encontrar la mejor proyección tridimensional, solo necesitaríamos añadir las puntuaciones de la tercera componente.

4.3. Propiedades y consideraciones prácticas del PCA

Para la realización de un ACP se trabaja con la *Matriz de Covarianzas*, $\mathcal{S}$, de las variables originales, siendo discutible su aplicación cuando:

- la naturaleza de las variables originales es muy heterogénea y habría problemas para interpretar una combinación lineal de las mismas.

- las variables originales vienen dadas en unidades muy distintas.

El segundo de los problemas puede subsanarse mediante la estandarización previa de las variables originales. Así en vez de trabajar con la Matriz de Covarianzas lo haríamos con la *Matriz de Correlaciones*, $\mathcal{R}$. Las componentes obtenidas a partir de $\mathcal{R}$ son distintas de las obtenidas a partir de $\mathcal{S}$, es más, el conocimiento de uno de los conjuntos de componentes no posibilita que el otro pueda ser derivado.

Por otra parte, no debemos perder de vista que la proyección en k componentes no deja de ser una aproximación a la realidad muestral y que pueden producirse distorsiones más o menos graves. Tales distorsiones para un punto I_s pueden deberse a:

- una puntuación alta en una componente $t > k$
- una alta suma de cuadrados residual $y_{sk+1}^2 + y_{sk+2}^2 + \ldots + y_{sp}^2$

Estos valores altos indican que el individuo se encuentra lejos de su proyección en el espacio determinado por las k primeras componentes.

4.4. Criterios para el número de componentes a retener

Una cuestión muy importante en el ACP es el número de componentes a retener, o, lo que es lo mismo, la determinación de la *dimensionalidad esencial* del modelo geométrico. La medida de la adecuación de la proyección de los n puntos en un subespacio definido por las k primeras componentes está basada en alguna función de la varianza de dichos puntos proyectados. Si llamamos λ_i a la varianza de la proyección de los puntos sobre la componente i-ésima, la ortogonalidad de las componentes hace que la varianza total de las k componentes retenidas sea

$$S_k^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_k$$

La varianza total S^2 de los datos originales es la suma de las varianzas de cada variable X_i y coincide con la suma de los elementos de la diagonal de la matriz $\mathcal{S}$.

$$traza(\mathcal{S}) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = S_p^2$$

Lo anterior supone que en realidad el ACP solo supone una rotación del sistema de referencia y que la varianza total de las X_i coincide con la varianza total de las Y_i. La proporción de varianza explicada por las k primeras componentes viene dada por

$$P_k = \frac{S_k^2}{S_p^2} = \frac{\sum_1^k \lambda_i}{\sum_1^p \lambda_i} = \frac{\sum_1^k \lambda_i}{traza(\mathcal{S})}$$

Por tanto, P_k puede ser una medida razonable de la bondad de ajuste de la proyección en el subespacio de las k primeras componentes. Desde una óptica descriptiva, el ACP puede considerarse bueno cuando el conjunto de componentes retenidas explica al menos el 75 % de la varianza total. No obstante, basar la elección de k solo en el valor P_k puede ser arbitrario, por lo que analizaremos otros procedimientos más objetivos.

4.4.1. Gráfico de sedimentación

La idea de este procedimiento es representar de forma indexada los valores $\lambda_1, \lambda_2, \ldots, \lambda_p$ y unir los puntos por una poligonal. *(Figura 5).*

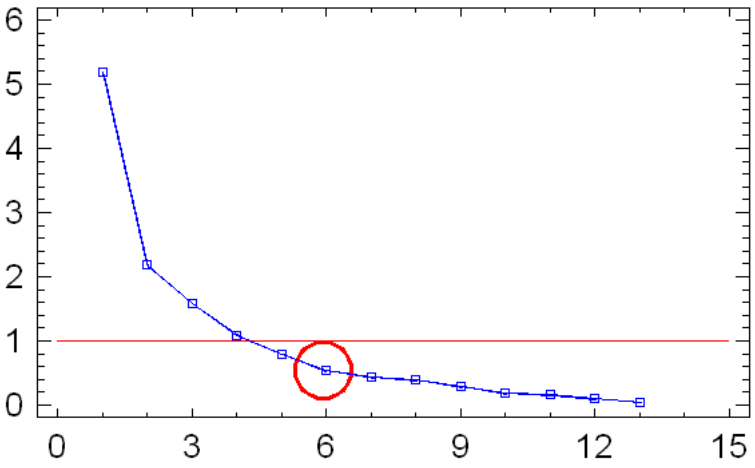


Figura 5: Gráfico de sedimentación

Tipicamente, este gráfico muestra bruscos descensos en las primeras componentes que se van suavizando hasta *sedimentarse* a partir de un determinado índice. Si argumentamos que las componentes que se corresponden con la porción plana del gráfico representan componentes de ruido no diferenciables del sistema, deberíamos elegir k como el primer índice donde la pendiente se suaviza.

4.4.2. λ medio

Otra alternativa para la elección de k es calcular la media de los λ_i,

$$\bar{\lambda} = \frac{\sum_1^p \lambda_i}{p}.$$

Si los datos originales tienen una estructura esférica, es decir, si no hay direcciones multivariables que den una mejor representación que los datos originales, entonces todos los λ_i serían aproximadamente igual a $\bar{\lambda}$. Por tanto, los $\lambda_i > \bar{\lambda}$ serán importantes, mientras que los $\lambda_i < \bar{\lambda}$ serán despreciables. Cuando el ACP se hace a partir de $\mathcal{R}$, entonces $\bar{\lambda} = 1$, de forma que, con este criterio, retendremos aquellas componentes que verifiquen que $\lambda_i \geq 1$. Lo que hacemos es retener aquellas componentes que nos dan más información (varianza) que las variables originales estandarizadas y despreciar el resto.

4.5. Interpretación de las componentes

Dado que las componentes son combinaciones lineales de las variables originales, es posible que exista una interpretación física de los coeficientes de la combinación. La aportación de una variable original X_i a la interpretación de una componente Y_j depende del

tamaño del coeficiente a_{ij}. Si a_{ij} es grande la variable interviene de forma importante, mientras que si es pequeño podríamos eliminar la variable X_i de la combinación lineal. Este plantamiento tiene, en cualquier caso, dos matices importantes: por una parte, la idea de grande o pequeño es subjetiva, y, por otra, está relacionado con la componente en cuestión, dado que un valor de coeficiente insignificante para una componente podría ser significativo para otra.

El coeficiente de correlación r_{ij} entre una componente Y_i y una variable original X_j con desviación típica σ_j vale

$$r_{ij} = \frac{\sqrt{\lambda_i} a_{ij}}{\sigma_j}; \qquad i = 1, \ldots, k; \qquad j = 1, \ldots, p,$$

estos coeficientes representan la parte de la varianza de la variable explicada por la componente. A partir de los r_{ij} se obtiene la matriz de *cargas factoriales* (pxk)

$$\begin{aligned} X_1 &= r_{11}Y_1 + \cdots + r_{k1}Y_k \\ X_1 &= r_{12}Y_1 + \cdots + r_{k2}Y_k \\ &\vdots \\ X_p &= r_{1p}Y_1 + \cdots + r_{kp}Y_k. \end{aligned}$$

$C_j = \sum_{i=1}^{k} r_{ij}^2 \quad j = 1, \ldots, p$ se denomina *comunalidad* y representa el tanto por uno de la variable X_j explicada por el conjunto de las k componentes retenidas. Si alguna de las variables tiene una comunalidad muy baja, ello supone que la información que contiene se encuentra concentrada en las $p - k$ componentes descartadas. Si esta variable nos resulta de especial interés, habrá que plantearse recuperar alguna componente más. Por otra parte, $\lambda_i = \sum_{j=1}^{p} r_{ij}^2 \quad , j = 1, \ldots, k$, es la cantidad de información (varianza) del total que contiene la componente Y_i .

5. Resolución de un caso práctico con R

Para ilustrar la obtención de las componentes principales de una matriz de datos con R utilizaremos el fichero de datos que nos da la composición de la leche de un grupo de mamíferos(tabla 1):

En **Rcmdr** elegimos `Estadísticos`→`Análisis dimensional`→Análisis de componentes principales.... En la ventana de diálogo que se recoge en la *figura 6* se seleccionan las variables numéricas para el análisis, en este caso las cuatro, obteniéndose la siguiente secuencia de órdenes de **R**:

```
.PC <- princomp(~Agua+Grasa+Lactosa+Proteina, cor=TRUE, data=leche)
unclass(loadings(.PC)) # component loadings
.PC$sd^2 # component variances
remove(.PC)
```

Al objeto de aprovechar mejor el contenido del objeto creado, $.PC$, ejecutaremos de nuevo desde la ventana de **Rcmdr** el conjunto de funciones salvo `remove(.PC)`, evitando así la eliminación del mismo.

	Animal	Agua	Proteina	Grasa	Lactosa
1	Yegua	90.1	2.6	1.0	6.9
2	Burra	90.3	1.7	1.4	6.2
3	Ballena	64.8	11.1	21.2	1.6
4	Cebra	86.2	3.0	4.8	5.3
5	Cerda	81.9	7.4	7.2	2.7
6	Rata	72.5	9.2	12.6	3.3
7	Oveja	82.0	5.6	6.4	4.7
8	Rena	64.8	10.7	20.3	2.5
9	Mula	90.0	2.0	1.8	5.5
10	Cerda	82.8	7.1	5.1	3.7
11	Camella	87.7	3.5	3.4	4.8
12	Búfala	82.1	5.9	7.9	4.7
13	Zorra	81.6	6.6	5.9	4.9
14	Coneja	71.3	12.3	13.1	1.9
15	Llama	86.5	3.9	3.2	5.6
16	Cierva	65.9	10.4	19.7	2.6
17	Hipopótama	90.4	0.6	4.5	4.4
18	Bisona	86.9	4.8	1.7	5.7
19	Gata	81.6	10.1	6.3	4.4
20	Perra	76.3	9.3	9.5	3.0
21	Foca	46.4	9.7	2.0	0.0
22	Delfina	44.9	10.6	34.9	0.9

Cuadro 1: Componentes de la leche de mamíferos

- Se crea el objeto `.PC` mediante la instrucción `princomp`, cuyos argumentos son:
  - $\sim$ Agua+Grasa+Lactosa+Proteina (variables seleccionadas)
  - cor=TRUE (trabaja con la matriz de correlaciones)
  - data=leche (fichero de datos)

  A continuación, se imprimen los pesos mediante la instrucción `unclass (loadings(.PC))`, en realidad `unclass` lo único que hace es desproveer de sus atributos al objeto `loadings(.PC)`.

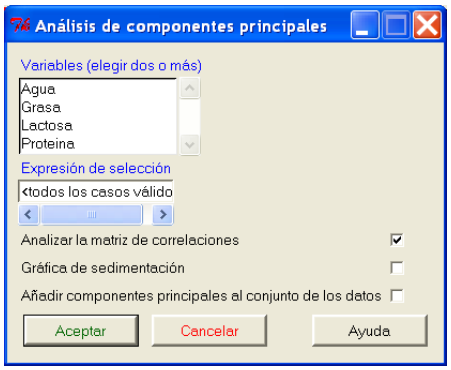


Figura 6: Ventana de diálogo ACP

```
> .PC <- princomp(~ Agua+Grasa+Lactosa+Proteina, cor=TRUE, data=leche)
> unclass(loadings(.PC)) # component loadings
              Comp.1       Comp.2       Comp.3      Comp.4
 Agua        0.5198985    0.2684160 -0.3938995   0.7088735
 Grasa      -0.5056630   -0.5002503  0.2028317   0.6729887
 Lactosa     0.5116260   -0.0262597  0.8519732   0.1081247
 Proteina   -0.4607053    0.8228079  0.2790057   0.1813658
```

- **Autovalores**. Se le indica al programa que imprima los autovalores, puede verse como solo uno de ellos es superior a uno.

```
> .PC$sd^2 # component variances
    Comp.1       Comp.2       Comp.3       Comp.4
3.532491829  0.357399346  0.103333242  0.006775583
```

  Dividiendo el primer autovalor por 4 y la suma de los dos primeros autovalores por 4, se obtiene la cantidad de información que retiene la primera componente y las dos primeras componentes, respectivamente. En este caso, estos valores son 0,8831 y 0,9724; es decir, la primera componente contiene el 88,31 % de la información total y las dos primeras componentes el 97,24 %, por lo que si despreciamos las componentes restantes se pierde únicamente el 2,76 % de la información contenida en las variables originales.

- **Inclusión de las componentes en la matriz de datos**. Para obtener las puntuaciones de los individuos en las componentes obtenidas imprimimos los *scores*.

```
>.PC$scores
          Comp.1       Comp.2       Comp.3       Comp.4
  1    2.38374741 -0.30636434  0.56619910  0.048868536
  2    2.28833659 -0.52266865  0.15736658 -0.004579167
  3   -2.25431223  0.30112008 -0.16047770  0.052848134
  4    1.52899713 -0.44780856  0.01594885 -0.002488356
  5   -0.09354439  0.42577897 -0.71257629 -0.017837689
  6   -0.79170338  0.39201512  0.11092119 -0.062296176
  7    0.76665037  0.01142319  0.09222869 -0.033691134
  8   -1.89749966  0.23571426  0.22553752  0.030419375
  9    2.01444760 -0.46656042 -0.14031796 -0.023137227
  10   0.37231526  0.45752277 -0.32052966 -0.055359672
  11   1.44471466 -0.22558378 -0.25824589 -0.013318559
  12   0.65987031  0.01374638  0.14171991  0.082168271
  13   0.70039560  0.25928536  0.27151474 -0.023025306
  14  -1.67928539  1.09502949 -0.27207529 -0.021437691
  15   1.58538076 -0.15878529  0.19328869 -0.021965041
  16  -1.75610549  0.21456985  0.20492981  0.043440387
  17   1.76773191 -0.89861627 -0.74479173  0.029215489
  18   1.58297616  0.13059975  0.27278912 -0.041904716
  19   0.07438972  1.07279694  0.31709664  0.153387926
  20  -0.59182309  0.64464949 -0.20107904 -0.062800704
  21  -4.26246527 -1.36443684 -0.08851314  0.186256609
  22  -3.84321458 -0.86342751  0.32906585 -0.242763287
```

  Mediante la instrucción:

```
leche<-data.frame(leche,.PC$scores)
```

  incorporamos las nuevas puntuaciones a la matriz de datos.

- **Gráfico de sedimentación**. Para obtener el gráfico de sedimentación basta con plotear los autovalores de forma indexada, obteniéndose la *figura 7*:

```
$>$ plot(.PC$sd^2,type=''l'',col= ``red'',main=''Gráfico de sedimentación'', xlab=''nº.
autovalor'',ylab=''valor'')
```

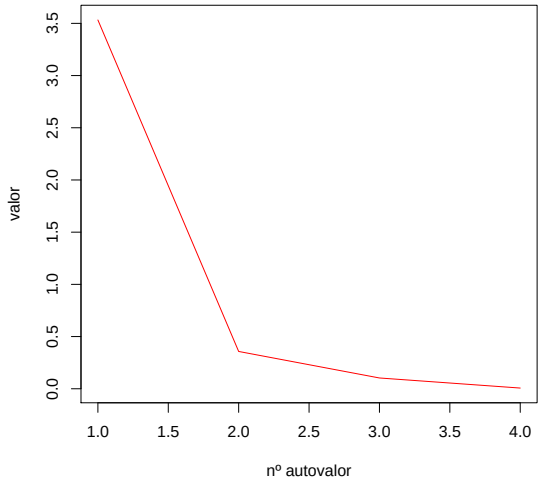


Figura 7: Gráfico de sedimentación

- **Gráfico de las puntuaciones de los individuos en las dos primeras componentes**. Para ver como se situan los individuos en el plano conformado por las dos primeras componentes, dibujamos el gráfico de dispersión, identificando los puntos con el nombre del animal mediante la serie de instrucciones siguiente.

```
prin1<-.PC$scores[,1]
prin2<-.PC$scores[,2]
plot(prin1,prin2,pch=19,xlab=''1ª componente'',ylab=''2ª componente'',col=''red'',
cex=0.5,ylim=c(min(prin2)-0.5, max(prin2)+0.5), xlim=c(min(prin1)-0.5,max(prin1)+1),
col.lab=''blue'',col.axis=''blue'',col.main=''red'', main=''Puntuaciones en las
dos primeras components principales de la leche de mamíferos'',cex.main=0.8,
text(prin1,prin2,labels=Leche$Animal, adj=c(0,1),cex=0.7))
abline(h=0,lty=3)
abline(v=0,lty=3)
```

En la *figura 8* se puede apreciar el comportamiento claramente diferenciado del resto de la leche de la Delfina y la Foca en las dos primeras componentes. Recordemos que esas dos primeras componentes suponen el 97,24 % del total, por lo que el mapa de posiciones de los individuos refleja de forma casi exacta la realidad.

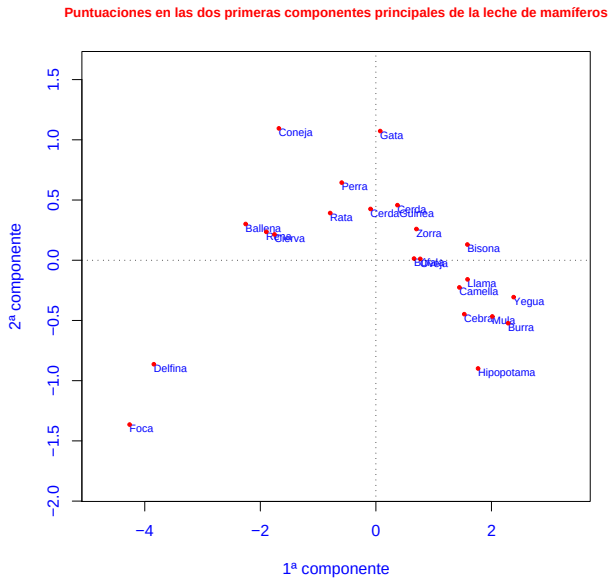


Figura 8: Nube de puntos en las dos primeras componentes

- **Gráficos biplot**. Una vez obtenidas las componentes principales, el *gráfico biplot* representa en las dos primeras componentes, o *plano principal*, tanto a los individuos como a las variables originales, de forma que se obtiene una visión real de la disposición de los individuos en las variables originales y el grado de relación lineal entre éstas. Para obtener el biplot ejecutamos las instrucciones siguientes:

```
#gráfico biplot
biplot(princomp(leche[,2:5]), xlabs=leche[,1])
```

  Puede observarse en la *figura 9* como están dispuestas a las variables y los individuos en el plano principal, que como hemos comprobado recoge el 97.24 % de la inercia global.

6. Análisis de tamaño y forma

Una de las aplicaciones más interesantes del ACP a los estudios morfométricos es el *Análisis de Tamaño y Forma*. Esta aplicación tiene su origen en los trabajos de Jolicoeur y Hosimann (1960) en los que intentaban clasificar tortugas atendiendo a la altura, longitud y anchura de sus caparazones. Para que un ACP pueda verse desde la perspectiva de un Análisis de Tamaño y Forma deben verificarse las condiciones de Rao (1971):

- Todos los coeficientes de la primera componente principal deben ser positivos, es decir, si $Y_1 = a_{11}X_1 + \ldots + a_{1p}X_p$, entonces $a_{1i} > 0$ para todo i. Esto implica que

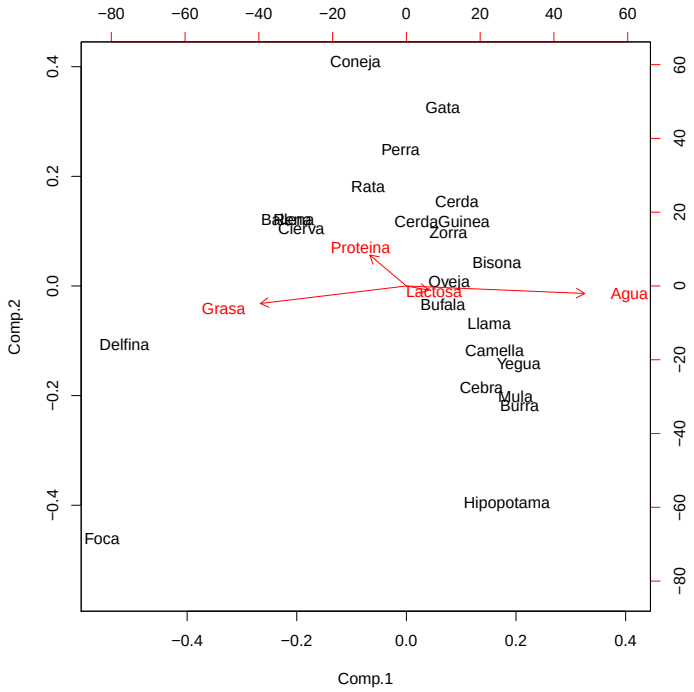


Figura 9: Gráfico biplot

si crece cualquiera de las variables originales entonces crece la primera componente. Y_1 es la componente de tamaño.

- Los coeficientes del resto de componentes no deben tener todos el mismo signo. En ese caso $Y_2, \ldots, Y_p$ son componentes de forma.

7. Ejercicios propuestos

1. Realiza un análisis de Componentes principales a las variables del archivo "obesidad.dat". Las variables de dicho archivo, de izquierda a derecha, son Peso en Kgs. a los dos años, Altura en cms. a los dos años, Peso en Kgs. a los nueve años, Altura en cms. a los nueve años, Circunferencia de la pierna a los nueve años, Medida de fuerza a los nueve años, Peso en kgs. a los dieciocho años, Altura en cms. a los dieciocho años, Circunferencia de la pierna a los dieciocho años, Medida de fuerza a los dieciocho años, Medida de obesidad a los dieciocho años en una escala de 7 puntos (1= delgado; 7=obeso).

   - Obtén las componentes principales
   - Decide cuantas componentes retener
   - Interpreta las variables retenidas en función de los coeficientes que las ligan con las variables originales.

- Representa en un gráfico biplot los individuos y las variables.

2. Utilizando el fichero de datos "UScereal" incluido en el paquete "MASS", responde a las siguientes cuestiones:
   - ¿Qué variables están más correlacionadas?
   - ¿Tiene sentido aplicar la técnica de componentes principales?
   - ¿Cuántas componentes habría que retener?
   - Comenta los resultados obtenidos.
   - Representa en un gráfico biplot los individuos y las variables.
3. Realiza un análisis de Tamaño y Forma al fichero de las tortugas de Jolicoeur y Hosimann (1960). Intenta representar los individuos usando representaciones multivariantes y establecer estereotipos a partir del análisis realizado.

UNIVERSIDAD DE CADIZ
Departamento de Estadística e I.O.
Grupo de investigación TeLoyDisRen

Análisis de correspodencia con R

1. Introducción

Pretendemos en este tema estudiar el comportamiento de un conjunto de `variables categóricas` organizadas en `tablas de frecuencias`, para lo cual usaremos la técnica multivariante conocida como **Análisis de Correspondencias (AC)**. Antes de entrar en materia, conviene recordar algunos conceptos y recursos estadísticos básicos relacionados con las `tablas de frecuencias`.

2. Análisis de tablas de frecuencias.

Las técnicas desarrolladas para variables cuantitativas son aproximadamente válidas cuando nos encontramos con variables categóricas ordenadas, sobre todo si el número de clases es alto, pero, en general, no son aceptables para el caso de variables no ordenadas.

El tratamiento de las variables cualitativas se basa necesariamente en la presencia o ausencia de un cierto matiz o cruce de matices –si hablamos de más de una variable–, renunciándose al carácter numérico, con lo que ello conlleva. Este tipo de análisis será casi obligatorio, como acabamos de indicar, para variables categóricas no ordenadas, recomendable para variables categóricas ordenadas, tanto más, cuantas menos sean las categorías, y de utilidad para variables continuas, cuando optemos por hacer clases agrupando valores en intervalos.

La información en este tipo de situaciones se organiza mediante las denominadas `tablas de frecuencias`, dedicándose los distintos análisis existentes a la detección de concentraciones por encima o por debajo de la proporción que le correspondería en un reparto lineal a los cruces de clases de cada una de las variables. Cuando hay sólo dos variables tendremos una tabla de frecuencias de `doble entrada` y cuando existan más de dos variables la tabla será de `múltiple entrada`. Este tipo de tablas se denominan en la literatura *Tablas de contingencia*. Para el caso de dos variables la tabla será:

A \ B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}

Donde la variable A posee I clases, categorías o matices distintos y la variable B J. Siendo n_{ij} el número de individuos que presentan el carácter A_i de A y el carácter B_j de B.

Para representar tablas múltiples no tendremos más remedio que hacerlo a través de varias tablas de doble entrada. Así, si tenemos tres variables, necesitaremos una tabla de doble entrada de las dos primeras variables para cada matiz de la tercera. La mejor manera de proceder en estos casos es considerar las dos primeras variables a aquellas que tengan un mayor número de clases, puesto que de esta manera reduciremos el número de tablas, siendo éste igual al producto del número de clases del resto de las variables.

2.1. Probabilidades marginales y condicionadas.

Dentro del estudio de las tablas de frecuencia tienen especial importancia los conceptos de distribuciones marginales y condicionadas, el primero de ellos nos indicará como se comporta cada variable independientemente del resto, mientras que el segundo nos permitirá analizar las relaciones de influencia entre dichas variables. Para el caso de dos variables, las marginales se obtienen sumando las filas y las columnas, así, en la tabla:

A \ B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	$Total$
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
$Total$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	$n_{..}$

se recoge la distribución marginal de A, que viene dada por sus valores y la columna de la derecha y la marginal de B, determinada de igual forma por sus clases y la última fila. Los valores $n_{i.}$ y $n_{.j}$, se obtienen sumando la fila i-ésima y la columna j-ésima, respectivamente, y nos dan el número de individuos que presentan el carácter A_i de A y B_j de B, en ese orden; mientras que $n_{..}$, representa al total de individuos de la población en estudio. La anterior tabla está formada por las frecuencias absolutas, también podríamos considerar las frecuencias relativas sin más que dividir cada una de las frecuencias absolutas por $n_{..}$, con lo que la tabla se transforma en:

B / A	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	$Total$
A_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1J}	$f_{1.}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2J}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{iJ}	$f_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	f_{I1}	f_{I2}	$\cdots$	f_{Ij}	$\cdots$	f_{IJ}	$f_{I.}$
$Total$	$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.j}$	$\cdots$	$f_{.J}$	1

Observemos que ahora la casilla inferior derecha vale 1, es decir, la tabla se ha transformado en una distribución de frecuencias relativas o de probabilidades, aunque el uso de probabilidad implicaría que estamos hablando de la distribución de la población bajo estudio.

Si $f_{.j} > 0$, entonces la distribución condicionada de A_i para A, conocido que B toma el valor B_j, viene dado por:

$$fr(A_i/B_j) = \frac{fr(A_i \cap B_j)}{fr(B_j)} = \frac{f_{ij}}{f_{.j}}$$

Obviamente, la condición de que $fr(B_j)$ debe ser estrictamente mayor que cero se impone para que esté definido el cociente.

En formato de probabilidades, la definición de probabilidad condicionada de A_i para A, conocido que B toma el valor B_j, viene dado por:

$$Pr(A_i/B_j) = \frac{Pr(A_i \cap B_j)}{Pr(B_j)}$$

Ejemplo 1. *Si consideramos la distribución de probabilidades que relaciona el color del pelo y de los ojos de un colectivo de personas:*

		Color de ojos		
		Azul(A)	*Marrón(M)*	*Verde(V)*
Color de pelo	*Rubio(R)*	*.12*	*.15*	*.03*
	Castaño(C)	*.22*	*.34*	*.04*
	Pelirrojo(P)	*.06*	*.01*	*.03*

La distribución marginal del Color del pelo(P) viene dada por la suma de las filas:

$$\begin{aligned} Pr(PR) &= 0.3 \\ Pr(PC) &= 0.6 \\ Pr(PP) &= 0.1 \end{aligned}$$

De igual forma la distribución del Color de Ojos(O) está determinada por la suma de columnas:

$$\begin{aligned} Pr(OA) &= 0.4 \\ Pr(OM) &= 0.5 \\ Pr(OV) &= 0.1 \end{aligned}$$

La distribución condicional de que un individuo tenga pelo Rubio dado que tiene los ojos Azules es:

$$Pr(PR/OA) = \frac{Pr(PR \cap OA)}{Pr(OA)} = \frac{0{,}12}{0{,}4} = 0{,}3$$

que coincide con $Pr(PR)$, lo que implica que PR y OA son independientes, es decir, el conocimiento de que una persona tiene los ojos azules no nos proporciona información sobre si tiene o no el pelo rubio.

Por otro lado,

$$Pr(PC/OA) = \frac{Pr(PC \cap OA)}{Pr(OA)} = \frac{0{,}22}{0{,}4} = 0{,}55$$

cuando $Pr(PC) = 0{,}6$. Es decir, cuando conocemos que una persona tiene los ojos azules la probabilidad de que posea el pelo castaño, disminuye respecto al porcentaje global de la población.

Ejemplo 2. *Consideremos la tabla de tres variables donde se relaciona el estatus económico (Alto, Bajo), la residencia (Madrid, Cádiz) y las preferencias sobre bebidas(Cerveza, Otras):*

	Cerveza		*Otras*		*Total*
	Madrid	*Cádiz*	*Madrid*	*Cádiz*	
Alta	*.021*	*.009*	*.049*	*.021*	*.1*
Baja	*.189*	*.081*	*.441*	*.189*	*.9*
Total	*.210*	*.090*	*.490*	*.210*	*1.0*

Podemos comprobar que las tres variables son completamente independientes. Por ejemplo:

$$Pr(Alta/Madrid, Cerveza) = \frac{Pr(Alta, Madrid, Cerveza)}{Pr(Madrid, Cerveza)} =$$

$$= \frac{0{,}021}{0{,}210} = 0{,}1 = Pr(Alta)$$

Lo mismo ocurriría con cualquier otra combinación. Alternativamente, la condición de independencia es equivalente a que la probabilidad conjunta de cualquier cruce de categorías coincide con el producto de las marginales de dichas categorías, por ejemplo:

$$Pr(Baja, Madrid, Cerveza) = 0{,}189 = (0{,}9) \cdot (0{,}7) \cdot (0{,}3) =$$

$$= Pr(Baja) \cdot Pr(Madrid) \cdot Pr(Cerveza)$$

En definitiva, el conocimiento de que alguna(s) de las variables toma determinados valores no nos da información sobre los valores que tomarán el resto.

Ejemplo 3. *La siguiente tabla contiene las probabilidades asociadas a las ocho combinaciones del estatus socioeconómico (Alto, Bajo), la opción política (Liberal, Conservador) y la afiliación (Demócrata, Republicano), de un determinado colectivo:*

	Demócrata		*Republicano*		*Total*
	Liberal	*Conservador*	*Liberal*	*Conservador*	
Alta	*.12*	*.12*	*.04*	*.12*	*.4*
Baja	*.18*	*.18*	*.06*	*.18*	*.6*
Total	*.30*	*.30*	*.10*	*.30*	*1.0*

Para cualquier combinación de categorías el factor socieconómico es independiente de los otros dos. Por ejemplo:

$$Pr(Alta, Liberal, Republicano) = 0{,}04 = (0{,}4) \cdot (0{,}1) =$$

$$= Pr(Alta) \cdot Pr(Liberal, Republicano)$$

Sin embargo esto no ocurre con cualquier otra división de las tres variables en dos grupos. Así,

$$Pr(Alta, Liberal, Republicano) = 0{,}04 \neq (0{,}4) \cdot (0{,}16) =$$

$$= Pr(Liberal) \cdot Pr(Alta, Republicano)$$

O también:

$$Pr(Alta, Liberal, Republicano) = 0{,}04 \neq (0{,}4) \cdot (0{,}16) =$$
$$= Pr(Republicano) \cdot Pr(Alta, Liberal)$$

3. Distribuciones asociadas a variables categóricas

En esta sección repasaremos las principales características de las distribuciones asociadas a variables categóricas, como son fundamentalmente la Binomial y la Multinomial.

3.1. La distribución Binomial.

Si consideramos un evento en el que se dan dos posibles resultados, éxito y fracaso, con probabilidades respectivas p y 1-p, constantes a lo largo del tiempo, y repetimos dicho experimento un número n de veces, la variable binomial X viene dada por el número de veces que ocurre uno de los posibles resultados, el éxito por ejemplo, en las n pruebas. Notamos que $X \sim B(n; p)$. La distribución de X viene dada por

$$Pr(X = r) = \binom{n}{r} p^r (1-p)^{n-r},$$

para $r = 0, 1, \cdots, n$. Puede demostrarse fácilmente que:

$$\mathrm{E(X)} = \mathrm{np}$$
$$\mathrm{V(X)} = \mathrm{np(1\text{-}p)}$$

Si consideramos los dos sucesos, éxito y fracaso, a la vez, tendríamos que:

$$X_1 \sim B(n; p)$$
$$X_2 \sim B(n; 1-p)$$

De tal forma que ahora

$$\mathrm{E}(X_2) = \mathrm{n(1\text{-}p)}$$
$$\mathrm{V}(X_2) = \mathrm{n(1\text{-}p)p}$$

Además, puede demostrarse que

$$Cov(X_1, X_2) = -np(1-p),$$

y de aquí que

$$Corr(X_1, X_2) = -1,$$

existiendo una relación lineal exacta e inversa entre X_1 y X_2.

3.2. La distribución multinomial.

La distribución multinomial es una generalización de la binomial, que nos ofrece la estructura probabilística para situaciones en las que, para una cierta prueba, existen más de dos resultados posibles con probabilidades constantes. Si se dan q resultados y X_i expresa el número de veces que aparece el resultado i-ésimo en n realizaciones de la prueba, tendremos que, generalizando la expresión **??**, podemos escribir

$$(X_1, X_2, \cdots, X_q) \sim M(n; p_1; p_2, \cdots; p_q)$$

La distribución de tal variable es

$$Pr(X_1 = r_1, X_2 = r_2, \cdots, X_q = r_q) = \frac{n!}{r_1! \cdots r_q!} p_1^{r_1} \cdots p_q^{r_q} = \frac{n!}{\prod_{i=1}^{q} r_i!} \prod_{i=1}^{q} p_i^{r_i}$$

para $r_i \geq 0$ y $r_1 + \cdots + r_q = n$. Observemos que para $q = 2$ se tiene justamente la distribución binomial. En general, se tiene que para cada componente

$$X_i \sim B(n; p_i),$$

y además

$$\begin{aligned} \mathrm{E}(X_2) &= np_i \\ \mathrm{V}(X_2) &= np_i(1-p_i) \end{aligned}$$

Tambien puede demostrarse que

$$Cov(X_i, X_j) = -np_i p_j$$

.

Ejemplo 4 *Supongamos una muestra de 50 individuos extraídos de una población cuyas probabilidades asociadas vienen dadas por la tabla del Ejemplo 3:*

	Demócrata		*Republicano*		*Total*
	Liberal	*Conservador*	*Liberal*	*Conservador*	
Alta	*.12*	*.12*	*.04*	*.12*	*.4*
Baja	*.18*	*.18*	*.06*	*.18*	*.6*
Total	*.30*	*.30*	*.10*	*.30*	*1.0*

El número de individuos pertenecientes a cada una de las ocho categorías tiene una distribución multinomial con n=50 y los p_i de la tabla. El número esperado de observaciones en cada categoría viene dado por

	Demócrata		*Republicano*	
	Liberal	*Conservador*	*Liberal*	*Conservador*
Alta	*6*	*6*	*2*	*6*
Baja	*9*	*9*	*3*	*9*

De forma trivial se podrían calcular las varianzas y las covarianzas. Consideremos ahora la siguiente distribución entre las distintas categorías

	Demócrata		*Republicano*	
	Liberal	*Conservador*	*Liberal*	*Conservador*
Alta	*5*	*7*	*4*	*6*
Baja	*8*	*7*	*3*	*10*

La probabilidad de esta tabla viene dada por:

$$\frac{50!}{5!7!4!6!8!7!3!10!}(0{,}12)^5(0{,}12)^7(0{,}04)^4(0{,}12)^6(0{,}18)^8(0{,}18)^7(0{,}06)^3(0{,}18)^{10} = 0{,}000007$$

4. Tablas de doble entrada.

Antes de entrar a fondo en el **Análisis de correspondencia** conviene analizar el tratamiento clásico de las tablas de frecuencias. Comenzaremos con el análisis de las tablas de doble entrada y dentro de éstas las denominadas tablas 2×2.

4.1. Tablas 2×2.

Consideremos dos variables binomiales medidas sobre un conjunto de individuos y organizadas en tabla de doble entrada. Supongamos que hemos extraído una muestra de tamaño $n_{..}$, la tabla de contingencia viene dada por:

		Columnas		
		1	2	*Totales*
Filas	1	n_{11}	n_{12}	$n_{1.}$
	2	n_{21}	n_{22}	$n_{2.}$
	Totales	$n_{.1}$	$n_{.2}$	$n_{..}$

Donde las notaciones se corresponden con las expuestas en la sección anterior. Para el conjunto del colectivo las probabilidades vienen dadas por:

		Columnas		
		1	2	*Totales*
Filas	1	p_{11}	p_{12}	$p_{1.}$
	2	p_{21}	p_{22}	$p_{2.}$
	Totales	$p_{.1}$	$p_{.2}$	$p_{..}$

Nosotros estamos interesados en estimar los p_{ij}, desarrollar modelos para ellos y construir contrastes. Los valores esperados, basados en algún modelo estadístico, pueden expresarse como:

		Columnas		
		1	2	*Totales*
Filas	1	m_{11}	m_{12}	$m_{1.}$
	2	m_{21}	m_{22}	$m_{2.}$
	Totales	$m_{.1}$	$m_{.2}$	$m_{..}$

donde obviamente $m_{ij} = n_{..} \times p_{ij}$.

4.1.1. Producto de binomiales. Homogeneidad de poblaciones

Si en el esquema de la tabla 2×2 las filas representan dos variables binomiales, podemos estar interesados en contrastar que éstas son iguales, es decir

$$H_0 : p_{11} = p_{21} \quad y \quad p_{12} = p_{22}$$

Una aplicación de este problema sería la de comprobar que las proporciones de una determinada característica coinciden en dos poblaciones, se dice que las poblaciones son homogéneas; para ello se tomarían sendas muestras

en las dos poblaciones y se aplicaría el procedimiento que se detalla. Observe que puesto que $p_{i1} + p_{i2} = 1$, basta testar

$$H_0 : p_{11} = p_{21}$$

o equivalentemente

$$H_0 : p_{11} - p_{21} = 0$$

Si H_0 es verdadera entonces $p = p_{11} = p_{21}$ y $\hat{p} = n_{\cdot 1}/n_{\cdot\cdot}$, y de igual forma $(1-p) = p_{12} = p_{22}$ y $(1-\hat{p}) = n_{\cdot 2}/n_{\cdot\cdot}$. Los valores esperados son $m_{ij} = n_{i\cdot}p_{ij}$, de tal forma que

$$\hat{m}_{ij} = n_{i\cdot}(n_{\cdot j}/n_{\cdot\cdot})$$

Estando definido el estadístico Chi-cuadrado de Pearson como

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

Si H_0 es verdadera entonces n_{ij} y $\hat{m}_{ij}$ deben estar próximos y sus diferencias al cuadrado deben ser razonablemente pequeñas. Lo contrario ocurrirá si H_0 es falsa.

Si H_0 es verdadera, χ^2 se distribuye como una Chi-cuadrado con un grado de libertad.

4.1.2. Dos binomiales independientes.

Si suponemos que nos encontramos ante un modelo de independencia entre dos variables binomiales, se deben preservar las proporciones, es decir, $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ para todo i y todo j; y puesto que los $p_{i\cdot}$ y los $p_{\cdot j}$ pueden estimarse por:

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \quad y \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}},$$

los m_{ij} pueden estimarse por:

$$\hat{m}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

El estadístico Chi-cuadrado de Pearson viene definido por:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

Debiendo realizarse el mismo análisis que hacíamos en el apartado anterior.

4.2. Tablas I × J. Producto de multinomiales

Podemos generalizar lo visto en el punto anterior, de tal forma que podemos encontrarnos ante una de las dos situaciones siguientes: un producto de multinomiales o una multinomial. En el caso de que tengamos una distribución producto de multinomiales, la hipótesis de que todas ellas son idénticas, o que las poblaciones son homogéneas, nos llevaría a contrastar la hipótesis

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{Ij} \quad \forall \quad j = 1, \cdots, J,$$

donde cada muestra tiene una distribución multinomial y por tanto

$$m_{ij} = n_{i.} p_{ij}$$

Si H_0 es verdadera p_{ij} es la misma para todos los valores de i, estimándose dichas p_{ij} por

$$\hat{p}_{ij} = n_{.j}/n_{..}$$

de donde

$$\hat{m}_{ij} = n_{i.}(n_{.j}/n)$$

Y el estadístico Chi-cuadrado es ahora:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

De tal forma que cuando H_0 es verdadera y la muestra es grande, X^2 se distribuye aproximadamente como una Chi-cuadrado con $(I-1)(J-1)$ grados de libertad.

4.2.1. Dos multinomiales independientes

En el caso de una distribución multinomial, la hipótesis de independencia entre las dos variables es:

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j} \quad i = 1, \cdots, I, \quad j = 1, \cdots, J$$

Las probabilidades marginales se pueden estimar como

$$\hat{p}_{i.} = n_{i.}/n_{..}$$

y

$$\hat{p}_{.j} = n_{.j}/n_{..}$$

Puesto que $m_{ij} = n_{..}p_{ij}$, si H_0 es cierta podemos estimar los m_{ij} como

$$\hat{m}_{ij} = n_{..}\,\hat{p}_{i.}\hat{p}_{.j} = n_{..}(n_{i.}/n_{..})(n_{.j}/n_{..}) = n_{i.}n_{.j}/n_{..}$$

Y el estadístico Chi-cuadrado es ahora:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

De tal forma que cuando H_0 es verdadera y la muestra es grande, X^2 se distribuye aproximadamente como una Chi-cuadrado con $(I-1)(J-1)$ grados de libertad.

5. Análisis de correspondencia

Englobada dentro del análisis factorial, el **Análisis de correspondencia (CA)** es una técnica multivariante de reducción de la dimensión en la que se estudia las relaciones de dependencia entre variables categóricas o variables cuantitativas categorizadas. Creada por Jean-Paul Benzecri (1930), se basa en métodos geométricos y prescinde de las distribuciones de probabilidad y los métodos de inferencia. El **CA** es en realidad un **Análisis de Componentes Principales** aplicado a variables cualitativas; utiliza la distancia no euclídea chi-cuadrado para convertir las frecuencias en coordenadas métricas. La salida de un **CA** es una representación gráfica en un espacio dimensional de escasas variables sintéticas o factores, idealmente dos, que pueden ser interpretados o nombrados y que además deben condensar el máximo posible de información. Sobre el plano o planos principales se situan las categorías de los caracteres, interpretándose la proximidad entre dos categorías en términos de relación entre las mismas. Dependiendo de que se consideren dos variables cualitativas o más de dos, el **CA** recibe el nombre de `Análisis de correspondencia simple` o `Análisis de correspondencia múltiple` Supongamos que los datos corresponden a dos criterios de clasificación o a un único criterio medido en varias poblaciones, los cuales se disponen en una tabla de contingencia o una de homogeneidad, respectivamente.

			Niveles						
		A	A_1	A_2	$\cdots$	A_i	$\cdots$	A_I	*Total*
	B								
	B_1		n_{11}	n_{12}	$\cdots$	n_{1i}	$\cdots$	n_{1I}	$n_{1.}$
	B_2		n_{21}	n_{22}	$\cdots$	n_{2i}	$\cdots$	n_{2I}	$n_{2.}$
Niveles	$\vdots$		$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
o	B_j		n_{j1}	n_{j2}	$\cdots$	n_{ji}	$\cdots$	n_{jI}	$n_{j.}$
Poblaciones	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
	B_J		n_{J1}	n_{J2}	$\cdots$	n_{Ji}	$\cdots$	n_{JI}	$n_{J.}$
	Total		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot i}$	$\cdots$	$n_{\cdot I}$	n

La distribución de frecuencias de los niveles de A en la población o nivel B_j viene dado por el vector de coordenadas

$$B_j : (\frac{n_{j1}}{n_{j.}}, \frac{n_{j2}}{n_{j.}}, \cdots, \frac{n_{jI}}{n_{j.}}) \qquad \forall j = 1, 2, \cdots, J$$

que se conoce como perfil de la fila j. Las coordenadas del perfil, $\frac{n_{ji}}{n_{j.}}$, pueden interpretarse como la probabilidad condicionada del suceso A_i a la población o nivel B_j. De igual forma se pueden obtener los perfiles por columna,

$$A_i : (\frac{n_{1i}}{n_{.i}}, \frac{n_{2i}}{n_{.i}}, \cdots, \frac{n_{Ji}}{n_{.i}}) \qquad \forall i = 1, 2, \cdots, I$$

A partir de aquí, podemos obtener las distancias entre elementos de A o de B usando la distancia chi-cuadrado, también conocida como distancia de Benzecri. Entre dos poblaciones o niveles B_j y $B_{j'}$, la distancia de Benzecri es igual a:

$$d^2(B_j, B_{j'}) = \sum_{i=1}^{I} \frac{1}{n_{.i}} (\frac{n_{ji}}{n_{j.}} - \frac{n_{j'i}}{n_{j'.}})^2 == \sum_{i=1}^{I} (\frac{n_{ji}}{\sqrt{n_{.i}} n_{j.}} - \frac{n_{j'i}}{\sqrt{n_{.i}} n_{j'.}})^2$$

Lo que es equivalente a que las poblaciones o niveles B_j están representados por una configuración de J puntos en un espacio euclídeo $\mathbf{R}^I$ de coordenadas:

$$B_j : (\frac{n_{j1}}{\sqrt{n_{.1}} n_{j.}}, \cdots, \frac{n_{jI}}{\sqrt{n_{.I}} n_{j.}})$$

Una vez construidos los perfiles y obtenidas las nuevas coordenadas, el **CA** se convierte en un problema de representación de individuos mediante un

Análisis de Componentes Principales. La matriz sobre la que habrá que trabajar es:

$$\mathbf{X} = (\frac{n_{ij}}{\sqrt{n_{.i}n_{j.}}})$$

cuyas filas son las coordenadas de las poblaciones/caracter B en el espacio euclídeo del caracter A. Las casillas de (X) son las frecuencias condicionadas por filas y estandarizadas por la variabilidad de su columna, representada por la raiz cuadrada de la marginal de la misma. Como acabamos de comentar esta matriz (X) puede tratarse como una matriz de datos estandar, con la única restricción de que las coordenadas de cada individuo fila suman lo mismo, por lo que el espacio de representación será en la práctica $I - 1$. No obstante, dado el proceso de estandarización seguido, debemos corregir el procedimiento para que aquellas filas con mayores frecuencias absolutas tengan más peso en la representación. En definitiva, la matriz $\mathbf{X}$ se sustituye por:

$$\mathbf{Z} = (\frac{n_{ij}}{\sqrt{n_{.i}n_{j.}}})$$

Las componentes principales se obtienen buscando los autovalores y autovectores de la matriz $\mathbf{Z}'\mathbf{Z}$. La restricción de que cada fila(columna) está estandarizada, supone que el primero de los autovalores es igual a 1, que se asocia a un autovector que no dice nada sobre la estructura de los datos. El resto de autovalores son menores a 1, siendo el máximo número que se puede calcular igual al mínimo entre el número de filas y columnas menos 1.

$$1 > \mu_2 = \lambda_2 \geq \mu_3 = \lambda_3 \geq \cdots$$

La suma de todos estos autovalores válidos nos da la inercia o cantidad de información que contiene el conjunto de datos. Una representación en las dos primeras dimensiones es ideal, siempre que éstas retengan un buen porcentaje de la inercia total. Dada la simetría del problema, los autovalores obtenido a partir de filas y columnas coinciden, mientras que los correpondientes autovectores están relacionados, permitiendo una representación conjunta de ambos elementos. Una forma rápida de obtener los autovetores de ambas representaciones, filas o columnas, es obtener los autovectores de la matriz de dimensión más pequeña, $\mathbf{Z}'\mathbf{Z}$ o $\mathbf{Z}\mathbf{Z}'$ y obtener los otros autovectores multiplicando los obtenidos por los autovalores.

6. Resolución de casos prácticos con R

Empezaremos trabajando algunos ejemplos de variables categóricas, sobre las que se establecen hipótesis sobre las probabilidades, independencia u homogeneidad.

6.1. Aplicaciones del test chi-cuadrado de bondad de ajuste

- Para contrastar si un dado no está trucado se lanza 60 veces, obteniéndose los siguientes resultados:

x_i	1	2	3	4	5	6
n_i	7	12	10	11	8	12

La hipótesis a contrastar es que $p_i = 1/6, \forall i$, con lo que se tiene que $E_i = 60(1/6) = 10, \forall i$. Para resolver con **R** ejecutamos las instrucciones siguientes:

```
> n< −c(7,12,10,11,8,12)
>chisq.test(n)
Chi-squared test for given probabilities
data: n
X-squared = 2.2, df = 5, p-value = 0.8208
```

Lo que, dado que el p-valor es muy alto nos llevaría a admitir la honradez del dado.

- Se desea estudiar la relación entre la orientación hacia las ciencias de un grupo de alumnos de un instituto y el nivel de estudios de la madre:

	Ninguno	Básico	Medio	Superior
Orientado	23	12	34	32
No orientado	18	42	16	27

En este caso es posible usar Rcmdr para introducir datos y hacer el análisis, usaremos la secuencia de instrucciones `Estadísticos`→ `Tablas de contingencia`→`Introducir y analizaruna tabla de doble entrada...`1. Lo que nos ofrece la salida de **R**

```
 Pearson's Chi-squared test
data: .Table
X-squared = 24.1629, df = 3, p-value = 2.31e-05
```

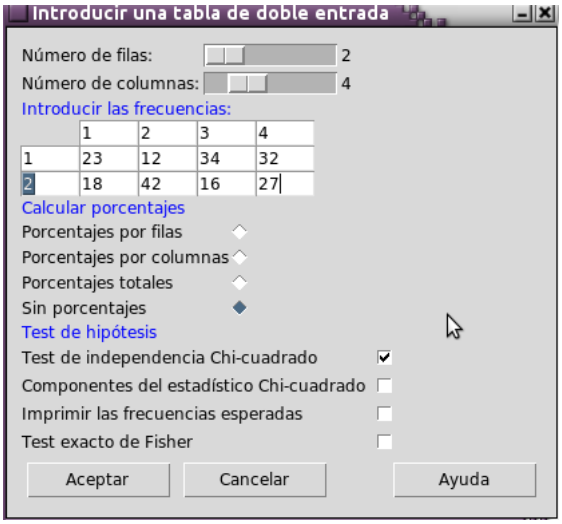


Figura 1: Ventana de entrada de datos de tablas

En este caso, el p-valor tan pequeño nos hace rechazar la hipótesis de independencia de caracteres y admitir una fuerte relación entre el nivel de estudios de la madre y la orientación a las ciencias de sus hijos.

La limitación que tiene el test chi-cuadrado es que solo ofrece resultados globales y no entra a detallar el mapa de relaciones entre las distintas categorías de los dos caracteres. Este handicap se puede superar usando **CA**

6.2. Análisis de correspondencia con R

La función `ca` que se encuentra dentro del paquete del mismo nombre realiza **Análisis de correspondencias simple**, para ilustrar su uso usaremos el fichero de datos `smoke` que se carga con `ca`. `smoke` contiene frecuencias de hábitos de fumar (ninguno, ligero, medio y alto) para grupos de personal (dirigentes de alto rango, gerentes menores, empleados mayores, empleados menores y secretarios) en una empresa ficticia.

En primer lugar analizaremos la independencia de los caracteres mediante el test chi-cuadrado.

```
library(ca) data(smoke) .Test <- chisq.test(smoke, correct=T) .Test
Pearson's Chi-squared test
data: smoke
X-squared = 16.4416, df = 12, p-value = 0.1718
```

Lo que nos indica que, desde un punto de vista inferencial, los caracteres son independientes. No obstante, aplicaremos un análisis de corresponden-

cias. Para ver el comportamiento en las distintas casillas y las aportaciones al estadístico chi-cuadrado ejecutamos:

```
.Test$expected # Valores esperados
smoke # Valores observados
round(.Test$residuals^2, 2) # componentes de la chi-cuadrado
```

Para ver la inercia total de la tabla de datos, se ejecuta:

```
n<-sum(smoke)
#Inercia Total
.Test$statistic/n
X-squared 0.08518986
```

Para obtener los perfiles:

```
#Perfiles fila, MASA
library(abind)
rowPercents(smoke)
#Perfiles columna, MASA
colPercents(smoke)
```

Para aplicar el **CA** usamos la función `ca`.

```
#Analisis de correspondencia simple
ca<-ca(smoke)
ca
Principal inertias (eigenvalues):
             1          2          3
 Value       0.074759   0.010017   0.000414
 Percentage  87.76%     11.76%     0.49%
Rows:
                SM         JM         SE         JE         SC
 Mass       0.056995   0.093264   0.264249   0.455959   0.129534
 ChiDist    0.216559   0.356921   0.380779   0.240025   0.216169
 Inertia    0.002673   0.011881   0.038314   0.026269   0.006053
 Dim. 1    -0.240539   0.947105  -1.391973   0.851989  -0.735456
 Dim. 2    -1.935708  -2.430958  -0.106508   0.576944   0.788435
Columns:
              none      light     medium      heavy
 Mass       0.316062  0.233161  0.321244    0.129534
 ChiDist    0.394490  0.173996  0.198127    0.355109
 Inertia    0.049186  0.007059  0.012610    0.016335
 Dim. 1    -1.438471  0.363746  0.718017    1.074445
 Dim. 2    -0.304659  1.409433  0.073528   -1.975960
```

La salida contiene los valores propios y los porcentajes de inercia explicada para todas las dimensiones posibles. Además proporciona distintos cálculos por filas y columnas: masas, distancias chi-cuadrado a su centroide, inercia y coordenadas estándar. Por defecto las coordenadas se restringen a dos dimensiones. Se puede comprobar que la suma de las inercias de las filas (columnas) coincide con la suma de los autovalores.

Podemos buscar información adicional usando la instrucción `summary`

```
summary(ca)
Principal inertias (eigenvalues):
 dim         value       %    cum%   scree plot
 1        0.074759    87.8    87.8   ************************
 2        0.010017    11.8    99.5   ***
 3        0.000414     0.5   100.0                              Rows:
          --------   -----
 Total:   0.085190   100.0
       name   mass    qlt     inr    k=1   cor     ctr    k=2   cor     ctr
 1 |   SM |     57    893    31 |    -66    92     3 |   -194   800   214 |
 2 |   JM |     93    991   139 |    259   526    84 |   -243   465   551 |
 3 |   SE |    264   1000   450 |   -381   999   512 |    -11     1     3 |
 4 |   JE |    456   1000   308 |    233   942   331 |     58    58   152 |
 5 |   SC |    130    999    71 |   -201   865    70 |     79   133    81 |
Columns:
         name   mass    qlt     inr    k=1   cor     ctr    k=2   cor     ctr
 1 |   none |    316   1000   577 |   -393   994   654 |    -30     6    29 |
 2 |   lght |    233    984    83 |     99   327    31 |    141   657   463 |
 3 |   medm |    321    983   148 |    196   982   166 |      7     1     2 |
 4 |   hevy |    130    995   192 |    294   684   150 |   -198   310   506 |
```

Esta nueva salida, da de nuevo los valores propios y los porcentajes relativos de inercia explicada para todas las dimensiones disponibles. Además, muestra los porcentajes acumulados y un gráfico scree. Los items Filas y Columnas incluyen las coordenadas principales para las dos primeras dimensiones (k = 1 y k = 2). También ofrece las correlaciones al cuadrado (cor) y las contribuciones (ctr) por puntos junto con las coordenadas. Calcula el porcentaje de inercia explicado por cada fila y columna.

En ocasiones, se pueden añadir filas o columnas adicionales en la tabla de datos, de forma que una vez realizado el **CA** y construidas las direcciones principales, estos elemntos adicionales pueden representarse en el(los) plano(s) principal(es). Para ello, se ejecuta la instrucción.

```
summary(ca(smoke, supcol = 1))
```

En este caso se considera la primera columna como columna suplementaria.

7. Ejercicios propuestos

I Aplique correspondencias simples a la tabla que se obtiene a partir de la ejecución del siguiente código de **R**:

```
Tabla <- matrix(c(8,5,0,0,0,12,26,13,0,3,0,9,21,6,7), 3, 5, byrow=TRUE)
rownames(.Table) <- c('Alta', 'Media', 'Baja') colnames(.Table) <-
c('Excelente', 'Bueno', 'Malo', 'Nulo', 'NS/NC') Tabla # Counts
```

II Aplique análisis de correspondencias simples a las variables `ingresos` y `categoria` del fichero `hotel.rda`, para ello:

- Construya la tabla de contingencia.

- Estudie la relación entre dichas variables.
- Aplique la función `ca` a la tabla de contigencia obtenida anteriormente

Proyecto de Innovación

Usted se ha autentificado como Antonio Sánchez Navas (Salir)

knuth.uca.es ▶ Innovación

Cambiar rol a... Activar edición

Personas
- Participantes

Actividades
- Foros
- Recursos

Buscar en los foros

Ir

Búsqueda avanzada

Administración
- Activar edición
- Configuración
- Asignar roles
- Calificaciones
- Grupos
- Copia de seguridad
- Restaurar
- Importar
- Reiniciar
- Informes
- Preguntas
- Archivos
- Desmatricular en Innovación
- Perfil

Mis cursos
- Análisis exploratorio de los datos con R y R-Commander
- Análisis Multivariante y Series Temporales
- datos
- Estadística - Grado Ciencias del Mar
- Estadística CC del Mar
- Foro de discusión y soporte para usuarios de R
- HUPR: Estadística básica con R y R-Commander en Ciencias de la Salud (2011)
- Introducción a la

Diagrama de temas

- Foro de Noticias
- Proyecto
- Aceptación del proyecto
- Web del CIDUI (Barcelona)
- Practica de cluster (propuesta)
  - Practica Anova
- Resolucion proyecto
- Resolucion proyecto 2º semestre

1 **Documentos a presentar antes del 15 de Marzo**
- Información memoria final
- Memoria Final Artículo
- Memoria Final Compromisos y Resultados
- Memoria Final Gestión Económica

2

3

4

5

6

7

8

9

10

Novedades

Agregar un nuevo tema...

19 de dic, 15:33
Manuel Muñoz Márquez
Proyecto subversion más...

Temas antiguos ...

Eventos próximos

No hay eventos próximos

Ir al calendario...
Nuevo evento...

Actividad reciente

Actividad desde miércoles, 25 de julio de 2012, 20:36
Informe completo de la actividad reciente...

Sin novedades desde el último acceso

Estadística y a R
Libro Libre: Estadística Básica con R y R-Commander
Libro Libre: Estadística Descriptiva y Probabilidad
Libro Libre: Inferencia Estadística (2ª Edición Revisada)
RcmdrPlugin.UCA
Técnicas de Análisis Multivariante
TeLoYDisRen
Usuarios de Servidores
CEP: Aprender Estadística con R
Curso base Estadística y R
Curso Moodle (con Antonio Saorín)
Proyecto de Innovación
R básico para la estadística en la docencia y la investigación
Semana de Formación: Introducción al paquete estadístico R

Todos los cursos ...

Moodle Docs para esta página

Usted se ha autentificado como Antonio Sánchez Navas (Salir)

Página Principal