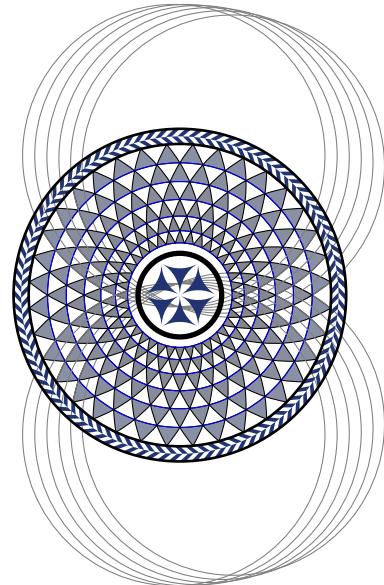


# Análisis inteligente de datos con lenguaje R

Curso introductorio

✿ Débora Chan ✿ Cristina Badano ✿ Andrea Rey ✿





# Índice de contenidos

<b>1</b>	<b>Introducción a la minería de datos</b>	<b>1</b>
1.1	Orígenes . . . . .	1
1.2	Objetivo . . . . .	3
1.3	Aspectos a considerar en la preparación de los datos . . . . .	4
1.4	Dominios de aplicación . . . . .	4
1.5	Software disponible . . . . .	5
1.6	Estadística vs data mining . . . . .	6
1.7	Nueva terminología . . . . .	6
<b>2</b>	<b>Introducción al análisis de datos</b>	<b>9</b>
2.1	Variables: niveles de medición . . . . .	9
2.1.1	Presentación de los datos . . . . .	10
2.2	Medidas descriptivas univariadas . . . . .	13
2.2.1	Medidas de tendencia central . . . . .	13
2.2.2	Medidas de posición o estadísticos de orden . . . . .	15
2.2.3	Medidas de dispersión . . . . .	16
2.2.4	Otras medidas para caracterizar la distribución . . . . .	19
2.2.5	Representación gráfica . . . . .	21
2.2.5.1	Diagrama circular . . . . .	21
2.2.5.2	Gráfico de barras . . . . .	22
2.2.5.3	Gráfico de bastones . . . . .	25
2.2.5.4	Histograma y polígono de frecuencias . . . . .	26
2.2.5.5	<i>Boxplot</i> o diagrama de caja . . . . .	31
2.2.5.6	<i>Boxplots</i> comparativos . . . . .	33
2.3	Información multivariada . . . . .	35
2.3.1	Objetivos del análisis exploratorio . . . . .	39
2.3.1.1	Tabla de clasificación cruzada . . . . .	39
2.3.1.2	Gráfico de mosaicos . . . . .	40
2.3.1.3	Diagrama de dispersión . . . . .	41
2.3.1.4	Dispersograma . . . . .	42

2.3.1.5	Gráfico de coordenadas paralelas . . . . .	42
2.3.1.6	Gráfico de perfiles multivariados . . . . .	44
2.3.1.7	Curvas de nivel . . . . .	45
2.3.1.8	Gráficos de estrellas . . . . .	47
2.3.1.9	Gráficos de caras de Chernoff . . . . .	48
2.4	Medidas de posición y dispersión en datos multivariados . . . . .	49
2.4.1	Propiedades del vector de medias . . . . .	50
2.4.2	Propiedades de la matriz de varianzas y covarianzas . . . . .	50
2.5	Transformación del conjunto de datos . . . . .	51
2.5.1	Transformaciones por variables . . . . .	51
2.5.1.1	Variables aleatorias estandarizadas . . . . .	51
2.5.2	Transformaciones por individuo . . . . .	51
2.6	Análisis multivariado . . . . .	52
2.6.1	Covarianza y Correlación . . . . .	54
2.7	Alternativas robustas para posición y escala . . . . .	60
2.8	Ejercitación . . . . .	64
<b>3</b>	<b>Análisis de componentes principales</b>	<b>69</b>
3.1	Nociones Previas . . . . .	69
3.2	Transformaciones . . . . .	73
3.2.1	Autovalores y Autovectores . . . . .	78
3.2.1.1	Relación entre autovalores, traza y determinante . . . . .	80
3.3	Motivación del problema de reducción de la dimensión . . . . .	82
3.4	Análisis de componentes principales . . . . .	88
3.4.1	Definición de las componentes . . . . .	90
3.4.2	Variabilidad explicada por las componentes principales . . . . .	91
3.4.3	Variabilidad de las componentes principales . . . . .	93
3.4.4	Cantidad de componentes principales . . . . .	93
3.4.4.1	Criterio 1: Porcentaje de variabilidad explicada . . . . .	94
3.4.4.2	Criterio 2: Criterio de Kaiser . . . . .	94
3.4.4.3	Criterio 3: Criterio del bastón roto . . . . .	95
3.4.4.4	Criterio 4: Prueba de esfericidad . . . . .	95
3.4.5	Estimación de las componentes principales . . . . .	95
3.4.6	Escalas de medida . . . . .	99
3.4.7	Cargas o <i>loadings</i> . . . . .	99
3.4.8	Interpretación de las componentes principales . . . . .	104
3.4.9	<i>Biplot</i> . . . . .	104
3.5	Componentes principales robustas . . . . .	117
3.6	Ejercitación . . . . .	128

<b>4 Contrastes de independencia y homogeneidad</b>	<b>133</b>
4.1 Contraste de Hipótesis . . . . .	133
4.1.1 Nivel de significación . . . . .	138
4.1.2 Relaciones entre los errores de tipo I y II . . . . .	139
4.1.3 Potencia de un contraste . . . . .	139
4.1.4 Concepto de $p$ -valor . . . . .	140
4.2 Contrastes de homogeneidad e independencia . . . . .	140
4.2.1 Contraste de independencia . . . . .	140
4.2.2 Test Chi cuadrado de independencia . . . . .	143
4.2.2.1 Hipótesis de interés . . . . .	143
4.2.3 Test Chi cuadrado de homogeneidad . . . . .	144
4.2.3.1 Hipótesis de interés . . . . .	146
4.2.4 Estadístico de contraste . . . . .	147
4.2.5 Región crítica . . . . .	147
4.2.6 Limitaciones . . . . .	151
4.2.7 Aplicación en R . . . . .	152
4.2.8 Test Exacto de Fisher . . . . .	152
4.3 Ejercitación . . . . .	157
<b>5 Análisis de correspondencias</b>	<b>161</b>
5.1 Introducción . . . . .	161
5.1.0.1 Perfiles medios . . . . .	168
5.1.1 Inercia total . . . . .	169
5.1.1.1 Guía para la interpretación gráfica del <i>biplot</i> simétrico . . . . .	180
5.1.1.2 Otra representación gráfica . . . . .	185
5.1.1.3 Estadístico de Pearson y la inercia . . . . .	194
5.1.1.4 Interpretación geométrica de la inercia . . . . .	195
5.1.2 Análisis de correspondencias múltiples . . . . .	197
5.1.2.1 Matriz de Burt . . . . .	200
5.1.2.2 Examen de los puntos . . . . .	208
5.2 Ejercitación . . . . .	215
<b>Referencias</b>	<b>218</b>



# Índice de figuras

2.1	Escala visual . . . . .	10
2.3	Variabilidad y rango . . . . .	16
2.4	Asimetría negativa o a izquierda . . . . .	19
2.5	Simetría . . . . .	19
2.6	Asimetría positiva o a derecha . . . . .	20
2.7	Distintos tipos de curtosis . . . . .	21
2.9	Diagrama circular con etiquetas . . . . .	22
2.10	Diagrama de barras . . . . .	23
2.11	Diagrama de barras superpuestas . . . . .	24
2.12	Diagrama de barras adyacentes . . . . .	25
2.13	Diagrama de bastones . . . . .	26
2.14	Histograma . . . . .	27
2.15	Polígono de frecuencias . . . . .	28
2.17	Histogramas con distintos intervalos . . . . .	29
2.18	Comparación de métodos para el cómputo de intervalos . . . . .	30
2.19	Simetría en <i>boxplots</i> . . . . .	33
2.21	<i>Boxplots</i> comparativos . . . . .	35
2.23	Diagrama de mosaicos . . . . .	40
2.24	Diagrama de dispersión para tres poblaciones . . . . .	41
2.25	Dispersograma . . . . .	43
2.26	Gráfico de coordenadas paralelas . . . . .	44
2.27	Gráfico de perfiles . . . . .	45
2.28	Gráfico de la distribución Normal Bivariada . . . . .	46
2.29	Gráfico de las curvas de nivel de la distribución Normal Bivariada . . . . .	47
2.30	Gráfico de estrellas . . . . .	48
2.31	Gráfico de caras de Chernoff para galletitas saladas . . . . .	49
2.33	Crontrol univariado . . . . .	53
2.34	Control multivariado . . . . .	55
2.35	Signo de la covarianza . . . . .	57
2.36	Correlograma . . . . .	59

2.37 Detección multivariada de <i>outliers</i> . . . . .	62
3.1 Vectores en coordenadas . . . . .	70
3.2 Dependencia lineal entre vectores . . . . .	71
3.4 Bases para $\mathbb{R}^2$ . . . . .	74
3.5 Modelo de datos a proyectar . . . . .	74
3.6 Simetría respecto del eje de abscisas . . . . .	75
3.7 Rotación de ángulo $\pi$ . . . . .	76
3.8 Proyección ortogonal de un punto sobre el plano $xy$ . . . . .	77
3.9 Simetría respecto de la recta $y = x$ . . . . .	79
3.11 Dispersograma entre dos variables . . . . .	84
3.12 Dispersograma en tres dimensiones desde distintos puntos de vista . . . . .	86
3.13 Dispersograma en 3D clasificado por grupos . . . . .	87
3.14 Ejes principales . . . . .	88
3.15 Direcciones principales en el espacio tridimensional . . . . .	89
3.17 Gráfico de sedimentación . . . . .	97
3.18 Cargas de la primera componente principal . . . . .	100
3.19 Cargas de la segunda componente principal . . . . .	101
3.20 <i>Biplot</i> para nadadores . . . . .	105
3.21 Caras de Chernoff para nadadores . . . . .	107
3.23 Gráfico de sedimentación para aspirantes . . . . .	112
3.24 Cargas para los aspirantes . . . . .	113
3.25 <i>Biplots</i> para los aspirantes . . . . .	114
3.26 Comparación de <i>boxplots</i> para nadadores . . . . .	118
3.27 Diagramas de dispersión para nadadores . . . . .	119
3.28 Análisis clásico de cargas para los nadadores con los datos agregados . . . . .	120
3.29 Análisis clásico de <i>screeplot</i> para los nadadores con los datos agregados . . . . .	121
3.30 Análisis clásico de <i>biplot</i> para los nadadores con los datos agregados . . . . .	121
3.31 Análisis robusto (MCD) de <i>screeplot</i> para los nadadores con los datos agregados . . . . .	126
3.32 Análisis robusto (MCD) de <i>biplot</i> para los nadadores con los datos agregados . . . . .	126
4.1 Ejemplo de regiones en un contraste bilateral . . . . .	137
4.2 Representación de los errores de un test . . . . .	138
4.3 Poblaciones según variable de color . . . . .	146
4.4 Ejemplos de la distribución $\chi^2$ según sus grados de libertad . . . . .	148
4.5 Distribución $\chi^2$ y zona crítica . . . . .	149
5.2 Perfiles de nivel cultural según atención . . . . .	169
5.3 Contribución de filas a la dimensión 1 . . . . .	177
5.4 Contribución de columnas a la dimensión 1 . . . . .	177

5.5	Puntos fila - AC . . . . .	178
5.6	Puntos columna - AC . . . . .	178
5.7	<i>Biplot</i> simétrico - AC . . . . .	179
5.8	Perfiles fila de las actividades universitarias . . . . .	183
5.9	Caras de Chernoff para fila de las actividades universitarias . . . . .	186
5.10	Representación en 3D de las actividades universitarias . . . . .	186
5.11	Plano de representación de las actividades universitarias . . . . .	188
5.12	<i>Biplot</i> simétrico de las actividades universitarias . . . . .	189
5.13	Contribución de las filas de las actividades universitarias . . . . .	190
5.14	Contribución de las columnas de las actividades universitarias . . . . .	190
5.16	Ejemplos de inercias . . . . .	196
5.18	Contribución a la inercia de las variables para la dimensión 1 . . . . .	202
5.19	Contribución a la inercia de los individuos para la dimensión 1 . . . . .	202
5.20	Categorías variables - ACM . . . . .	203
5.21	Individuos agrupados por género - ACM . . . . .	203
5.22	Individuos agrupados por estado civil - ACM . . . . .	204
5.24	Contribución a la inercia de las variables . . . . .	211
5.25	Contribución a la inercia de los individuos . . . . .	212
5.26	<i>Biplot</i> simétrico para la empresa . . . . .	212
5.27	Empleados agrupados por género . . . . .	213



# Índice de tablas

1.1	Comparación de características . . . . .	6
2.1	Ejemplo de distribución de frecuencias . . . . .	11
2.2	Ejemplo de variable discreta . . . . .	12
2.3	Ejemplo de frecuencias absolutas . . . . .	12
2.4	Ejemplo de frecuencias porcentuales . . . . .	13
2.5	Distribución de frecuencias: caso 1 . . . . .	15
2.6	Distribución de frecuencias: caso 2 . . . . .	15
2.7	Modelo de base de datos . . . . .	36
2.8	Base de datos para las galletitas . . . . .	37
2.9	Cantidad de parámetros en función de las variables . . . . .	38
2.10	Consideraciones para la compra de un auto . . . . .	39
2.11	Distancias entre <i>outliers</i> . . . . .	63
2.12	Distancias de Mahalanobis . . . . .	63
2.13	Datos candidatas a recepcionistas . . . . .	64
3.1	Tiempos por tramos en competencia de natación . . . . .	72
3.2	Tiempos por tramos en competencia de natación ampliada . . . . .	73
3.3	Análisis sobre riesgo cardíaco . . . . .	83
3.4	Variabilidad de las componentes principales usando las variables estandarizadas . . . . .	96
3.5	Variabilidad de las componentes principales usando las variables originales . . . . .	97
3.6	Cargas para los nadadores . . . . .	100
3.7	Datos de los nadadores estandarizados por columna . . . . .	102
3.8	Puntajes ( <i>scores</i> ) de los nadadores . . . . .	103
3.9	Estadística descriptiva univariada para los nadadores . . . . .	103
3.10	Autovalores y autovectores . . . . .	110
3.11	Esfericidad de Bartlett . . . . .	110
3.12	Criterios de evaluación . . . . .	111
3.13	Variabilidad explicada . . . . .	112
3.14	Cargas de los datos de los aspirantes . . . . .	113

3.15	Nuevos nadadores . . . . .	118
3.16	PCA clásico con nuevos datos . . . . .	119
3.17	Análisis de componentes principales usando MCD . . . . .	125
4.1	Errores en un test . . . . .	138
4.2	Nivel de violencia según la edad . . . . .	141
4.3	Frecuencias relativas del nivel de violencia según la edad . . . . .	141
4.4	Formato teórico del nivel de violencia según la edad . . . . .	142
4.5	Cálculos para el análisis de independencia . . . . .	143
4.6	Frecuencias observadas y esperadas del nivel de violencia según la edad . . . . .	144
4.7	Frecuencias teóricas de homogeneidad . . . . .	145
4.8	Datos enfermedad según tabaquismo . . . . .	148
4.9	Frecuencias observadas y esperadas de enfermedad según tabaquismo . . . . .	150
4.10	Similitudes y diferencias entre ambas pruebas . . . . .	151
4.11	Tabla de contingencia de $2 \times 2$ . . . . .	153
4.12	Depresión por Sexo . . . . .	154
4.13	Combinaciones para Fisher . . . . .	155
4.14	Probabilidades asociadas a las combinaciones . . . . .	156
4.15	Ingreso a <i>Twitter</i> según sexo . . . . .	159
4.16	Especialidad según zona . . . . .	159
4.17	Presencia de angioma según tipo de embarazo . . . . .	160
5.1	Tabla de contingencia . . . . .	163
5.2	Tabla de probabilidades estimadas . . . . .	164
5.3	Nivel cultural según atención . . . . .	165
5.4	Distribución marginal del nivel de atención . . . . .	165
5.5	Distribución marginal del nivel de cultura . . . . .	165
5.6	Distribución conjunta de niveles culturales y de atención . . . . .	165
5.7	Frecuencias esperadas bajo independencia . . . . .	167
5.8	Probabilidades condicionales dado el nivel ‘Atento’ . . . . .	168
5.9	Representación de niveles como simulaciones ( <i>dummies</i> ) . . . . .	170
5.10	Representación de niveles para un ejemplo sencillo . . . . .	171
5.11	Tabla de contingencia y matriz $F$ para un ejemplo sencillo . . . . .	171
5.12	Matriz $F_r$ para un ejemplo sencillo . . . . .	171
5.13	Inercias principales (autovalores) . . . . .	181
5.14	Perfiles de las filas . . . . .	181
5.15	Perfiles de las columnas . . . . .	182
5.16	Registro viajes de intercambio . . . . .	182
5.17	Perfiles fila de los viajes de intercambio . . . . .	183
5.18	Perfiles columna de los viajes de intercambio . . . . .	185

5.19	Datos para analizar ausencia de depresión según práctica deportiva . . . . .	191
5.20	Frecuencias esperadas bajo independencia para ausencia de depresión según práctica deportiva . . . . .	192
5.21	Residuos para analizar ausencia de depresión según práctica deportiva . . . . .	193
5.22	Distancias entre perfiles fila . . . . .	197
5.23	Distancias entre perfiles columna . . . . .	197
5.24	Características observadas . . . . .	198
5.25	Matriz disyuntiva para las características observadas . . . . .	199
5.26	Matriz $G^t$ . . . . .	199
5.27	Matriz de Burt para el Ejemplo 5.10 . . . . .	201
5.28	Situación de los empleados de una empresa . . . . .	206
5.29	Matriz disyuntiva para la situación de los empleados . . . . .	207
5.30	Agregado de categoría para los empleados . . . . .	209
5.31	Matriz disyuntiva para la empresa . . . . .	213
5.32	Matriz de Burt para la empresa . . . . .	214
5.33	Inercias para la empresa . . . . .	214
5.34	Coordenadas de representación para la empresa . . . . .	214
5.35	Hábito de fumar según puesto de trabajo . . . . .	215



# Capítulo 1

## Introducción a la minería de datos

*Somewhere, something incredible  
is waiting to be known.*

— Carl Sagan

### 1.1 Orígenes

La minería de datos, *data mining* en inglés, surge con el análisis de los datos sociales de Quetelet, los datos biológicos de Galton y los datos agronómicos de Fisher.

Adolphe Quetelet (1796-1874) hizo un gran aporte en el área de la Física Social. Entre sus notables conclusiones podemos mencionar las siguientes:

- ✿ El delito es un fenómeno social que puede conocerse y determinarse estadísticamente.
- ✿ Los delitos se cometen año a año con absoluta regularidad y precisión.
- ✿ Posibles causas de la actividad delictiva pueden ser la pobreza, el clima, la miseria, el analfabetismo, entre otras.

Francis Galton (1822-1911) fue el primero en aplicar métodos estadísticos en el estudio de las Ciencias Humanísticas enfocando en la herencia de la inteligencia. Algunos de sus resultados son:

- ✿ Creación del concepto estadístico de correlación y regresión hacia la media, altamente promovido.
- ✿ Introducción del uso de cuestionarios y encuestas con el objetivo de obtener datos sobre las comunidades humanas.
- ✿ Desarrollo de estudios genealógicos, biográficos y antropométricos aplicando estadística.

Ronald Fisher (1890-1962) trabajó desde 1919 como estadístico en la estación agrícola de *Rothamsted Research*, donde desarrolló el análisis de la varianza que aplicó a datos que provenían de cultivos. Realizó aportes fundamentales en el área de la genética de poblaciones, entre los cuales podemos mencionar:

- ✿ El principio de Fisher.
- ✿ El modelo de selección sexual denominado *runaway*.
- ✿ La hipótesis del hijo sexy.

Hasta hace pocos años la única estrategia para extraer información de utilidad de una base de datos, era la Estadística clásica. Sin embargo, los tamaños y la disponibilidad de las bases han crecido notablemente gracias a la tecnología informática. La minería de datos brinda una respuesta al análisis de gigantescas bases de datos, que suponen cierta complejidad y donde la Estadística clásica resulta un recurso limitado.

La edad de oro de la Estadística clásica puede ubicarse después de la Segunda Guerra Mundial. Su metodología ocupó un lugar de relevancia en la evaluación de ciertos resultados. Sin embargo, el escenario actual tiene características diferentes al de aquella época.

Se evidencia un aumento considerable en la cantidad de datos

- ✿ colectados,
- ✿ almacenados,
- ✿ accesibles,
- ✿ distribuidos.

El origen de estos datos puede ser a partir de

- ✿ transacciones bancarias,
- ✿ reservas de aerolíneas,
- ✿ llamadas y mensajes por celulares,
- ✿ registros de atención de pacientes,
- ✿ datos obtenidos por sensores remotos,
- ✿ operaciones con tarjetas de crédito,
- ✿ búsquedas en *internet*,
- ✿ compras en supermercados.

Estos datos son huellas o rastros que dejamos en nuestro cotidiano accionar.



Estas gigantescas bases de datos, plantean un nuevo escenario que nos conduce a preguntarnos más que el qué, el **por qué** de las cosas.

El valor de la información no reside en los datos concretos, sino en la forma de correlacionarlos para descubrir patrones y estructuras ocultas.

El desafío es tolerar la imprecisión, la confusión, “aceptar el desorden natural del mundo”, a cambio de “un sentido más completo de la realidad”. La herramienta para ello es la **minería de los datos**.

## 1.2 Objetivo

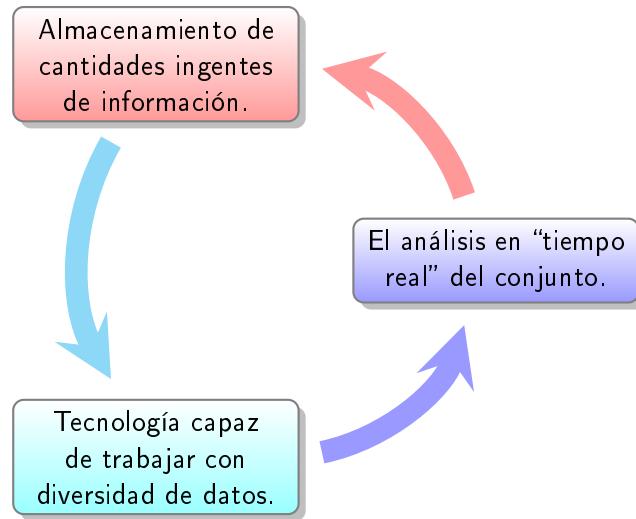
El *data mining* forma parte de un proceso conocido como “descubrimiento de conocimiento a partir de los datos” (*KDD: Knowledge Discovery in Databases*).

El objetivo es extraer información de una gran base de datos, sin disponer de conocimiento previo para construir patrones y/o relaciones sistemáticas de valor, así como anomalías.

Las soluciones que aporta la minería de datos se basan en la implementación, mediante programación, de interfaces de uso general y algoritmos propios. Estos posibilitan una exploración y organización eficiente de la información. Dichos algoritmos apoyan la identificación de regularidades para quienes deben tomar decisiones.

En esta disciplina confluyen técnicas provenientes de diferentes áreas como:

- ✿ bases de datos y Computación,
- ✿ aprendizaje automático,
- ✿ visualización,
- ✿ inteligencia artificial,
- ✿ Estadística,
- ✿ aprendizaje de máquina incluyendo redes neuronales,
- ✿ procesamiento de imágenes.



## 1.3 Aspectos a considerar en la preparación de los datos

Al analizarse una gran base de datos es importante considerar cuestiones de diversa índole, tales como:

- ✿ objetivos del análisis,
- ✿ disponibilidad de medios para resolver el problema,
- ✿ estructura y preparación de los datos,
- ✿ costos insumidos por el estudio
- ✿ necesidad de interpretación de resultados,
- ✿ redacción de un informe que incluya los alcances de las conclusiones y sea comprensible para todos los interesados.

## 1.4 Dominios de aplicación

En *data mining* han surgido diversos dominios de aplicación, entre los cuales cabe mencionar:

- ✿ análisis y procesamiento de imágenes y señales,
- ✿ análisis multidimensional de procesos,
- ✿ análisis de datos textuales,

- ✿ *web mining*,
- ✿ detección de fraudes,
- ✿ bioinformática.

Se generaron nuevos desafíos que demandan la creación de herramientas específicas para este contexto. En el marco de las respuestas a estos desafíos han surgido nuevos productos de software para el manejo de grandes cantidades de datos.

## 1.5 Software disponible

Entre las múltiples ofertas de *software* para *data mining* podemos citar las siguientes:

- ✿ *SAS Enterprise Miner* desarrollado por la empresa multinacional *SAS Corporation* con sede en Cary, Carolina del Norte, Estados Unidos, permite crear modelos predictivos y descriptivos para grandes volúmenes de datos.
- ✿ *R* es un entorno y lenguaje de programación libre con un enfoque al análisis estadístico, nacido como una reimplementación de *software* libre del lenguaje *S*, adicionado con soporte para alcance estático.
- ✿ *Phyton* es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible, siendo un lenguaje multiparadigma, debido a que soporta orientación a objetos, programación imperativa y programación funcional.
- ✿ *Statistica Data Miner*, desarrollado por Dell, provee un completo y exhaustivo conjunto de paquetes para la manipulación y análisis de datos.
- ✿ *SPSS Clementine* es una aplicación de *software* de análisis de texto y minería de datos de *IBM*, que se utiliza para construir modelos predictivos y realizar otras tareas analíticas, con una interfaz visual que permite aprovechar estos algoritmos sin programación.
- ✿ *T (Textual)* aplicado al análisis de datos simbólicos.
- ✿ *ISL Decision Systems*, es un producto que convierte datos en decisiones de negocios, cuyas últimas aplicaciones incluyen detección de fraude, fidelidad de la clientela, análisis de ventas, segmentación directa por correo y predicción de audiencia televisiva.
- ✿ *Salford Systems* se especializa en el estado del arte de la tecnología de aprendizaje por máquina diseñado para asistir a científicos en todos los aspectos del desarrollo de modelos predictivos.

- ✿ *MineSet*, desarrollado por *Silicon Graphics*, ofrece herramientas para analizar, minar y visualizar datos.
- ✿ *WEKA* es un *software* libre desarrollado en *Java* que consiste de una colección de algoritmos de aprendizaje de máquina para tareas de *data mining*.
- ✿ *SODAS (Symbolic Official Data Analysis System)* es un software modular software en el cual cada método estadístico es manipulado como un ícono y los íconos son enlazados en una cadena.
- ✿ *IBM Intelligent Miner* es un conjunto que consta de productos para el modelado, evaluación y visualización de minería inteligente.
- ✿ *SPAD (Système Portable pour l'Analyse de Données)*, permite implementar una estrategia de análisis adecuada al tratamiento exploratorio multivariado de grandes tablas de datos.

## 1.6 Estadística vs data mining

En la Tabla 1.1 se muestran algunas diferencias entre los campos de la Estadística (clásica) y de la Minería de datos.

Análisis estadístico	Data mining
Procedimiento hipotético deductivo	Procedimiento inductivo
Técnicas confirmatorias	Técnicas exploratorias
Supuestos iniciales	Sin supuestos iniciales
No se vale de herramientas informáticas	Amplia difusión entre especialistas en Computación

Tabla 1.1: Comparación de características

## 1.7 Nueva terminología

Algunos de los términos que mencionaremos a continuación, forman parte del lenguaje de trabajo de esta disciplina.

### M2M: *Machine to Machine*

**M2M** o **máquina a máquina** es un concepto genérico que se refiere al intercambio de información o comunicación en formato de datos entre dos máquinas remotas [11].

Los sistemas M2M permiten a las empresas disponer de infraestructuras y servicios más inteligentes, ágiles y eficientes, dado que estos sistemas facilitan el control de fraudes, la reducción de

costos, el ahorro de tiempo y el monitoreo en tiempo real del negocio. Actualmente se utiliza, entre otras muchas aplicaciones, en:

- ✿ gestión de flotas,
- ✿ alarmas domésticas,
- ✿ contadores de agua, gas o electricidad,
- ✿ telemantenimiento de ascensores,
- ✿ estaciones meteorológicas,
- ✿ terminal punto de venta,
- ✿ máquinas *vending*.

### **IoT: Internet of Things**

El término **IoT** o *internet de las cosas*, fue acuñado en 1999 por el investigador británico Kevin Ashton [2] y se refiere a la interconexión digital de objetos cotidianos mediante *internet*. Ashton por aquellos años trabajaba en el *Massachusetts Institute of Technology* (MIT) como cofundador y director ejecutivo del Centro de Auto-ID desarrollando un sistema de sensores e identificadores de radio frecuencia (RFID).

El primer dispositivo ‘conectado’ fue una máquina de *Coca-Cola* en la Universidad Carnegie a principios de 1980. Los programadores podían conectarse a la máquina a través de *internet*, comprobando el estado de la máquina y determinando si había o no había una bebida fría antes de decidirse a hacer el viaje a la máquina.

Inicialmente, el término *internet* de las cosas se usaba denotando una conexión avanzada de dispositivos, sistemas y servicios que trasciende el tradicional M2M y abarca una amplia variedad de dominios y aplicaciones. Actualmente, todos los aparatos domésticos comunes pueden ser modificados para trabajar en un sistema IoT. Por lo tanto, no debemos preocuparnos si tenemos adaptadores de redes Wi-Fi, sensores de movimiento, cámaras, micrófonos u otros instrumentos como básculas inalámbricas y monitores de presión arterial inalámbricos o los nuevos dispositivos usables (*wearables* en inglés) como gafas, relojes inteligentes ya que todos se pueden conectar a la *internet* de las cosas.

### **WoT: Web of Things**

El término **WoT** o *red de las cosas* se refiere a los enfoques, estilos arquitectónicos de *software* y patrones de programación que permiten que objetos del mundo real formen parte de la *World Wide Web*. Su principal objetivo es el modo de conectar objetos en red [9].

Cosas es un término de sentido amplio que alude a los objetos físicos, pero también a objetos etiquetados como códigos de barra, redes de sensores inalámbricos, máquinas o productos electrónicos de consumo.

La *Web of Things* proporciona una capa de aplicación que simplifica la creación de aplicaciones de IoT.

### IoE: Internet of Everything

El término **IoE** o **internet del todo** tiene un sentido amplio y alude a la conexión inteligente entre la gente, los dispositivos, los datos en proceso y las cosas [15]. Es una filosofía en la que el futuro de la tecnología se compone de muchos tipos diferentes de dispositivos y elementos conectados a *internet* global.

IoE describe un mundo de millones de objetos con sensores que detectan y evalúan su estado. Todos están conectados a través de redes públicas o privadas utilizando diversos protocolos.

Los expertos sostienen que *Internet of Everything* reinventará las industrias en tres niveles: proceso de negocio, modelo de negocio y momento de negocio.

# Capítulo 2

## Introducción al análisis de datos

*La Estadística es una ciencia que demuestra que si mi vecino tiene dos autos y yo ninguno, los dos tenemos uno.*

— George Bernad Shaw

### 2.1 Variables: niveles de medición

El análisis descriptivo es el paso inicial generalmente recomendado para comprender la estructura de los datos disponibles y la extracción de la información relevante para el análisis.

Describir cualquier situación real, por ejemplo, las características físicas de una raza de vacas, la situación financiera de una empresa, las particularidades de la producción de una planta, requiere tener en cuenta simultáneamente el comportamiento y la interacción entre las variables.

Las variables pueden ser, según su nivel de medición:

- ✿ **Categóricas o cualitativas:** Las distintas modalidades que adoptan estas variables sólo se distinguen por ser diferentes, no se puede establecer un ordenamiento entre ellos. Son ejemplos de estas variables: color de cabello, tipo de auto, sexo.
- ✿ **Cuasicuantitativas u ordinales:** En estas variables, si bien se puede ordenar las modalidades que adopta, no se puede establecer una distancia entre ellas. Por ejemplo: calificación de examen (A, B, C, D y E), estadío de una enfermedad (I, II, III o IV).
- ✿ **Cuantitativas discretas:** Estas variables toman valores numéricos siendo que entre dos valores consecutivos de las mismas no existen valores intermedios. Pueden tomar un conjunto a lo sumo numerable de valores, vinculándose generalmente al proceso de contar. Son ejemplos de estas variables: cantidad de hijos, cantidad de materias aprobadas, dinero en una billetera.

- ✿ **Cuantitativas continuas:** Estas variables también toman valores numéricos, pero entre dos valores de la variable existen infinitos valores intermedios, asociándose generalmente al proceso de medir. Son ejemplos de estas variables: peso, edad, duración de un llamado.

Existen otras formas de medición, asociadas generalmente a la subjetividad del individuo.

Por ejemplo las **escalas analógicas o visuales** que se utilizan en muchas ocasiones para que el paciente indique el grado de alguna variable “de nivel subjetivo” como dolor, bienestar, agrado, acuerdo-desacuerdo o sensaciones en general.

Un ejemplo de ello es el tratamiento del dolor, ver Figura 2.1. A los pacientes se les suele pedir que indiquen en una línea entre 0 y 10 que une los extremos sin dolor y dolor intolerable, cuál es su posición.

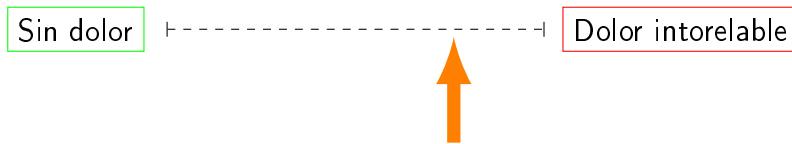


Figura 2.1: Escala visual

Estas escalas son útiles para evaluar la progresión de un mismo individuo pero debe tenerse en cuenta el carácter subjetivo de esta escala a la hora de intentar comparar entre individuos.

Otro ejemplo podría ser el estudio de la satisfacción de los clientes con algún servicio en particular previa y posterior a alguna mejora o actualización de dicho servicio.

Es usual que el método de análisis de este tipo de variables esté basado en rangos de *scores*.

### 2.1.1 Presentación de los datos

Una vez definida la base de datos con toda la información disponible, es necesario ordenarla y organizarla, a fin de facilitar su comprensión e interpretación. El análisis sobre los datos crudos, puede resultar inabordable.

Surge naturalmente la siguiente pregunta:

*¿Cómo convendría entonces organizar la información?*

Si se desea analizar una sola variable, el paso inicial más sencillo es confeccionar una tabla denominada distribución de frecuencias; que tiene un aspecto particular para cada tipo de variable de las consideradas.

#### 1. Para **datos cualitativos**:

Las clases se definen según el interés de la investigación.

Se cuenta la cantidad de observaciones de cada clase. A dicha cantidad se la conoce como frecuencia absoluta observada.

**Ejemplo 2.1.** Estudiamos los tipos de autos vendidos en una concesionaria de Capital Federal durante el mes pasado.



<https://flic.kr/p/bx4uHH>

Para ello, se construye una distribución de frecuencias donde a cada categoría o modalidad de la variable se le asigna su frecuencia absoluta; es decir, el número de veces que se ha registrado dicha categoría en la muestra de observaciones.

Modelo	Frecuencia
Utilitario	6
Familiar	10
Cupé	7
Camioneta	12
Sedán	17

Tabla 2.1: Ejemplo de distribución de frecuencias



## 2. Para datos cuantitativos:

- ✿ En el caso de variables **discretas**, las modalidades quedan definidas por los valores del recorrido de la variable.
- ✿ En el caso de variables **continuas**, es necesario definir intervalos que cubran el recorrido de la variable en estudio, denominados “intervalos de clase”.

- \* En **ambos casos**, se registra la frecuencia absoluta de cada modalidad (cantidad de observaciones en ella) o de cada intervalo (cantidad de observaciones dentro del rango del intervalo definido).

**Ejemplo 2.2.** Estudiamos ahora la evolución de las ventas de vehículos de alta gama, en la misma sucursal durante los últimos 24 meses.

Alta gama	Meses
1	2
2	3
3	7
4	4
5	8

Tabla 2.2: Ejemplo de variable discreta



La Tabla 2.2 indica que en 7 de los meses observados se han vendido 3 vehículos de alta gama, 8 meses en los que se han vendido 5 vehículos de alta gama, etc.

**Ejemplo 2.3.** Estamos interesados en investigar la cantidad de proteínas en gramos consumidas por día per cápita para una muestra de habitantes de distintos partidos del Gran Buenos Aires.

Intervalo de clase	$f_i$ (frec. absoluta)
[7, 9)	6
[9, 11)	10
[11, 13)	4
[13, 15)	7
[15, 17)	5

Tabla 2.3: Ejemplo de frecuencias absolutas

La Tabla 2.3 informa, por ejemplo, que 4 individuos consumieron entre 11 y 13 gramos de proteínas por día. Pero no nos da una idea de la concentración de nuestra población de interés en dicha categoría. Por este motivo, es usual incorporar las frecuencias porcentuales en estas tablas.

Para calcular las frecuencias porcentuales, es necesario recordar que la suma de las frecuencias observadas en las  $m$  modalidades de la variable  $f_i$ , con  $1 \leq i \leq m$ , es igual a la cantidad

total de observaciones  $n$ , registradas en las mismas de la variable; es decir, se tiene que  $f_1 + f_2 + \dots + f_m = n$ .

La frecuencia relativa se calcula dividiendo la frecuencia absoluta por la cantidad total de observaciones  $f_i/n$  y la frecuencia porcentual  $f_{r_i}$  se obtiene multiplicando estos resultados por 100. Así, por ejemplo, la frecuencia relativa de la clase  $[11, 13)$  resulta  $4/32 = 0.125$  y su frecuencia porcentual es 12.5%. Repitiendo este procedimiento para todos los intervalos de clase obtenemos la distribución de frecuencias porcentuales o relativas dadas en la Tabla 2.4.

Intervalo de clase	$f\%$ (frec. porcentual)
[7, 9)	18.75
[9, 11)	31.25
[11, 13)	12.5
[13, 15)	21.88
[15, 17)	15.62

Tabla 2.4: Ejemplo de frecuencias porcentuales

Ahora tenemos una idea de la magnitud de la frecuencia y podemos apreciar que la mayoría de los individuos observados consumen entre 9 y 11 gramos de proteínas por día.



## 2.2 Medidas descriptivas univariadas

### 2.2.1 Medidas de tendencia central

Las medidas de tendencia central son resúmenes estadísticos que pretenden representar a un conjunto de valores con un solo valor. Definen, de alguna manera, el punto en torno al cual se encuentra ubicado el conjunto de los datos. A continuación presentamos los ejemplos más difundidos de medidas de tendencia central.

**Media aritmética o promedio muestral:** es el promedio de las observaciones registradas y se calcula a partir de un conjunto de datos dado  $\{x_1, x_2, \dots, x_n\}$ , como

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

#### Propiedades

- \* Es de cálculo sencillo.

- ✿ Se puede calcular sólo para escalas de medición cuantitativas.
- ✿ Preserva la dependencia lineal; es decir, si  $y = ax + b$  entonces  $\bar{y} = a\bar{x} + b$ .
- ✿ No puede aplicarse a datos censurados.
- ✿ Es muy sensible a la presencia de valores extremos (muy alejados del conjunto de datos), vale decir que no es una medida robusta.

**Mediana:** se define como un valor que divide a la distribución ordenada en dos partes iguales, cada una de las cuales contiene el 50% de las observaciones. Si la muestra ordenada es:  $x^{(1)} \leq x^{(2)} \leq \dots \leq x^{(n)}$ , entonces la mediana es

$$\tilde{x} = \begin{cases} x^{(\frac{n+1}{2})} & \text{si } n \text{ es impar} \\ \frac{x^{(\frac{n}{2})} + x^{(\frac{n}{2}+1)}}{2} & \text{si } n \text{ es par} \end{cases}$$

### Propiedades

- ✿ Es de cálculo sencillo.
- ✿ Se puede calcular para escalas de medición al menos ordinales.
- ✿ Preserva la dependencia lineal; es decir, si  $y = ax + b$  entonces  $\tilde{y} = a\tilde{x} + b$ .
- ✿ No es sensible a la presencia de valores extremos, por lo que es una medida robusta.

**Moda:** es la observación de mayor frecuencia y suele notarse por  $Mo$ . No es una medida muy estable, dado que una sola observación puede cambiar el valor de la moda. Además, puede no ser única, de hecho existen distribuciones bimodales o multimodales, en cuyo caso no resulta una medida de tendencia central muy informativa.

**Ejemplo 2.4.** Presentamos varios casos para el cálculo de las medidas previamente definidas.

- ✿ Para los datos  $\{12, 12, 15, 18, 23\}$ , se tiene que  $\tilde{x} = 15$ ,  $\bar{x} = 16$  y  $Mo = 12$ .
- ✿ Si los datos son  $\{12, 12, 15, 17, 25, 25\}$ , entonces  $\tilde{x} = \frac{15 + 17}{2} = 16$ ,  $\bar{x} = 17.67$  y existen dos modas  $Mo = 12$  y  $Mo = 25$ .
- ✿ Las medidas para los datos de la Tabla 2.5 son  $\tilde{x} = \frac{x^{(25)} + x^{(26)}}{2} = \frac{3 + 7}{2}$ ,  $\bar{x} = 4.9$  y  $Mo = 7$ .
- ✿ Las medidas correspondientes a los datos que se presentan en la Tabla 2.6 son  $\tilde{x} = x^{(25)} = 6$ ,  $\bar{x} \cong 4.86$  y  $Mo = 6$ .

$x_i$	$f_i$	$F_i$
2	10	10
3	15	25
7	20	45
8	5	50

Tabla 2.5: Distribución de frecuencias: caso 1

$x_i$	$f_i$	$F_i$
1	10	10
5	14	24
6	21	45
8	4	49

Tabla 2.6: Distribución de frecuencias: caso 2



Las tercera columnas de las Tablas 2.5 y 2.6 contienen las frecuencias absolutas acumuladas  $F_i$ , que resultan de la suma de todas las frecuencias absolutas de las categorías menores de la variable, simbólicamente  $F_k = \sum_{i=1}^k f_i$ .

**Media  $\alpha$ -podada:** se define como el promedio de los datos centrales recortando el  $\alpha\%$  de los valores más grandes y el  $\alpha\%$  de los valores más chicos. Se denota como  $\bar{x}_\alpha$ . Esta medida tiene como posiciones extremas a la media aritmética y a la mediana que se corresponden con  $\alpha\% = 0$  y  $\alpha\% = 50$  respectivamente.

**Ejemplo 2.5.** Calculemos la media podada al 10% para los siguientes datos:

$$2 - 4 - 5 - 6 - 7 - 7 - 8 - 8 - 8 - 9 - 9 - 10 - 13 - 14 - 14 - 14 - 15 - 15 - \mathbf{15} - 25$$

Sin considerar los números en negrita,

$$\bar{x}_{0.10} = \frac{5 + 6 + 7 \cdot 2 + 8 \cdot 3 + 9 \cdot 2 + 10 + 13 + 14 + 14 \cdot 3 + 15 \cdot 2}{16} = 10.125$$



## 2.2.2 Medidas de posición o estadísticos de orden

Si bien hemos visto que la mediana es una medida de tendencia central, también puede pensarse como un estadístico de orden, dado que se calcula en función de los datos ordenados.

Recordemos que los datos ordenados de menor a mayor se denotan como  $x^{(1)} \leq x^{(2)} \leq \dots \leq x^{(n)}$ . Entonces  $x^{(1)}$  es el **valor mínimo** observado y  $x^{(n)}$  es el **valor máximo** observado. Estos son dos casos particulares de estadísticos de orden.

**Cuantiles:** son ciertos valores del recorrido de la variable que permiten subdividir el conjunto de datos en partes iguales, todas formadas por la misma cantidad de observaciones. Los cuantiles pueden o no corresponder a valores observados. Los más usados son los **cuartiles**  $Q$  que dividen las observaciones en cuatro partes iguales, los **deciles**  $D$  que lo hacen en diez partes iguales y los **percentiles**  $P$  que lo hacen en 100 partes iguales.

**Cuartiles:** cada una de las cuatro partes iguales en que dividen las observaciones contiene un cuarto o 25% de la información. Se denotan  $Q_1$ ,  $Q_2$  y  $Q_3$  y se denominan primer, segundo y tercer cuartil. Observemos que el segundo cuartil coincide con la mediana.

### 2.2.3 Medidas de dispersión

Las medidas de dispersión indican la variabilidad de los datos. La mayoría cuantifica el grado de concentración de los datos alrededor de una medida de posición. Presentaremos a continuación las medidas de dispersión más difundidas.

**Rango muestral:** se define como la diferencia entre el valor máximo y el valor mínimo de la muestra, es decir,

$$rg(x) = x^{(n)} - x^{(1)}$$

Si bien es una medida de cálculo sencillo, no resulta en general muy informativa.

En la Figura 2.3 se pueden apreciar tres conjuntos de datos con el mismo rango pero diferente grado de concentración alrededor del centro.

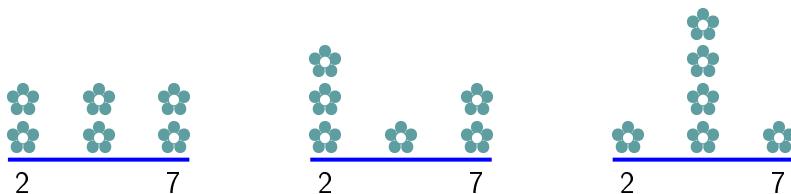


Figura 2.3: Variabilidad y rango

**Varianza Muestral:** se define como el promedio de los cuadrados de las distancias de las observaciones a la media muestral; es decir,

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

**Observación:** Algunos autores definen la varianza muestral usando como denominador  $n - 1$  en lugar de  $n$ . El fundamento teórico para esta expresión es que la varianza muestral calculada de esta forma es un estimación mas precisa de la varianza poblacional, especialmente cuando  $n$  es pequeño.

## Propiedades

- ✿ Es de cálculo sencillo.
- ✿ Se puede calcular sólo para variables cuantitativas.
- ✿ Si  $y = ax + b$ , entonces  $s_y^2 = a^2 s_x^2$ .
- ✿ Las unidades de medición de la varianza son el cuadrado de las unidades de los datos originales.
- ✿ Es muy sensible a la presencia de valores extremos. No es una medida *robusta*.
- ✿ En los casos en que la media no resulta adecuada como medida de tendencia central, tampoco la varianza lo es como medida de dispersión.

**Desviación estándar muestral:** Se define como la raíz cuadrada de la varianza y permite retornar a las unidades de medición originales. En símbolos:

$$s_x = \sqrt{s_x^2}$$

**Coeficiente de variación (CV):** es una medida de dispersión relativa porque mide la proporción que representa el desvío estándar de la media aritmética. Se define como el cociente entre el desvío estándar y el promedio muestral. Es usual que se exprese en porcentaje, dado que es una medida de dispersión relativa, mientras que las anteriores son medidas de dispersión absolutas.

Cuando se quiere comparar la dispersión de dos conjuntos de datos, si estos tienen valores de media similares y comparten la unidad de medición, basta con comparar las desviaciones estándar respectivas. Sin embargo, si las unidades de medición de ambos conjuntos no son las mismas o los valores de la media son diferentes, no corresponde utilizar las desviaciones estándar para comparar las dispersiones de ambos conjuntos. Se usa entonces, el coeficiente de variación.

Cuando por alguna de las causas que hemos mencionado, no resulta adecuada la media como representación de la tendencia central de nuestros datos, tampoco será adecuado informar la variabilidad utilizando varianza, desvío estándar o coeficiente de variación.

Analicemos algunas alternativas para estos casos.

**Rango intercuartílico (RI):** es un valor numérico que informa el rango del 50% de los valores centrales del conjunto de datos. Se define como la diferencia entre el tercer cuartil y el primero. Simbólicamente:

$$RI = Q_3 - Q_1$$

**MAD:** es la mediana de los desvíos absolutos respecto de la mediana. La sigla proviene del inglés *Median Absolute Deviation*.

**Ejemplo 2.6.** En el siguiente conjunto de observaciones  $\{2, 3, 5, 8, 13, 27\}$ , es clara la presencia de un valor muy alejado del conjunto de datos.

$$\text{La mediana es } \tilde{x} = \frac{5+8}{2} = 6.5.$$

Los desvíos respecto de la mediana resultan:  $-4.5, -3.5, -1.5, 1.5, 6.5, 20.5$ .

Los valores absolutos de los desvíos ordenados de menor a mayor son  $1.5, 1.5, 3.5, 4.5, 6.5, 20.5$ .

$$\text{La mediana de los valores absolutos de los desvíos es } MAD = \frac{3.5 + 4.5}{2} = 4.$$

Para hacer la MAD comparable con la desviación estándar, se propone la normalización de la misma

$$MADN(X) = \frac{MAD(X)}{0.6745}$$

La justificación de esta normalización es que en caso de normalidad coinciden el desvío estándar y la MADN [14].

Para comprender el sentido de esta constante, consideremos  $Z \sim N(0, 1)$  y notemos por  $med(X) = \widetilde{X}$ .

Por definición,

$$MAD(Z) = med(|Z - med(Z)|)$$

y puesto que  $Z$  es una variable simétrica con media nula,  $med(Z) = 0$ . Luego,  $MAD(Z) = med(|Z|)$ .

Si llamamos  $W = |Z|$ , entonces  $MAD(Z) = med(W)$ .

Por otro lado,  $F_W(w) = 2\phi(w) - 1$  y buscamos  $\widetilde{w}$  tal que  $F(\widetilde{w}) = 0.5$ . En efecto,

$$F_W(w) = P(W \leq w) = P(|Z| \leq w) = \Phi(w) - \Phi(-w) = \Phi(w) - [1 - \Phi(w)] = 2\phi(w) - 1$$

Entonces  $F(\widetilde{w}) = 2\phi(\widetilde{w}) - 1 = 0.5$ , por lo que  $\phi(\widetilde{w}) = 0.75$  y  $\widetilde{w} = 0.6745$ .

Dado que  $\sigma(Z) = 1$  y  $MAD(Z) \cong 0.6745$ , se desprende que

$$\frac{MAD(Z)}{\sigma(Z)} \cong 0.6745.$$

Generalizando para cualquier distribución gaussiana, si  $X \sim N(\mu; \sigma)$ ,

$$MAD\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}MAD(X - \mu) = \frac{1}{\sigma}MAD(X) \cong 0.6745$$

y por lo tanto  $\frac{MAD(X)}{\sigma} \cong 0.6745$ .



## 2.2.4 Otras medidas para caracterizar la distribución

En esta sección introducimos medidas de análisis estadístico.

**Coeficiente de asimetría muestral de Fisher:** es una medida que describe la asimetría de la distribución de los datos con respecto a la media muestral. Su expresión analítica es

$$sk_F(x) = \frac{\sqrt{n} \sum_{j=1}^n (x_j - \bar{x})^3}{\left[ \sum_{j=1}^n (x_j - \bar{x})^2 \right]^{\frac{3}{2}}}$$

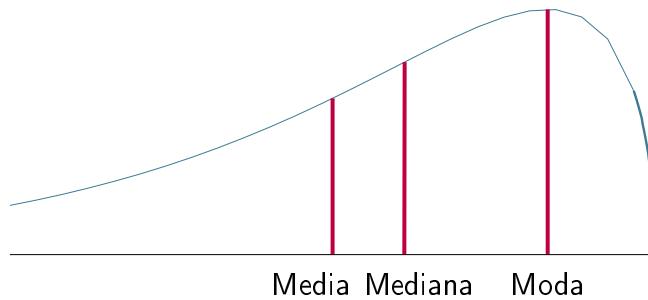


Figura 2.4: Asimetría negativa o a izquierda

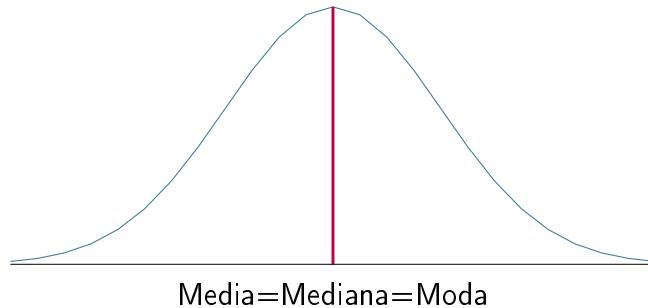


Figura 2.5: Simetría

Cuando los datos proceden de una distribución simétrica (Figura 2.5), como la distribución normal,  $sk(x) \approx 0$ , la mediana coincide con la moda y el promedio muestral. Sin embargo, como puede observarse en las Figuras 2.4 y 2.6), la media es ‘arrastrada’ ante la presencia de valores extremos (muy grandes o muy chicos).

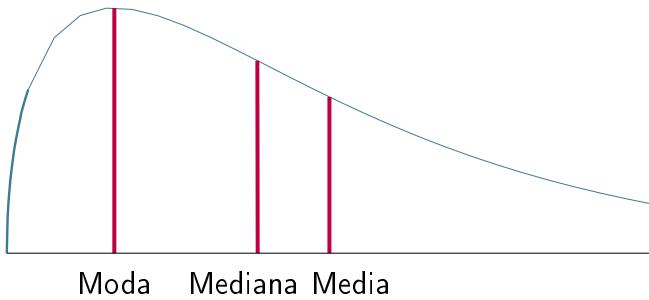


Figura 2.6: Asimetría positiva o a derecha

**Coeficiente de asimetría de Pearson:** mide la asimetría cuantificando la separación entre la moda respecto de la desviación estándar, siendo

$$sk_P(x) = \frac{\bar{x} - Mo(x)}{s_x}$$

Este coeficiente es menos usual dado que requiere que la distribución sea unimodal.

**Coeficiente de asimetría de Bowley:** toma como referencia los cuartiles para determinar si la distribución es simétrica o no, focalizando en el 50% de los valores centrales de la distribución. Su expresión es

$$sk_B(x) = \frac{(q_3 - q_2) + (q_1 - q_2)}{q_3 - q_1} = \frac{q_3 + q_1 - 2\tilde{x}}{q_3 - q_1}$$

Se utiliza en general cuando la media y el desvío estándar no son representativos del conjunto de observaciones.

**Coeficiente de curtosis muestral:** es una medida que describe el grado de apuntamiento de una distribución. También puede entenderse como una descripción del comportamiento de las colas de la distribución de las observaciones. Una mayor curtosis no implica una mayor varianza, ni viceversa. La expresión analítica para su cálculo es:

$$k(x) = \frac{n \sum_{j=1}^n (x_j - \bar{x})^4}{\left[ \sum_{j=1}^n (x_j - \bar{x})^2 \right]^2}$$

Cuando los datos proceden de una distribución simétrica, como la distribución normal,  $k(x_i) \cong 3$ . Las distribuciones leptocúrticas tienen coeficientes superiores a 3 y las platicúrticas coeficientes menores a 3.

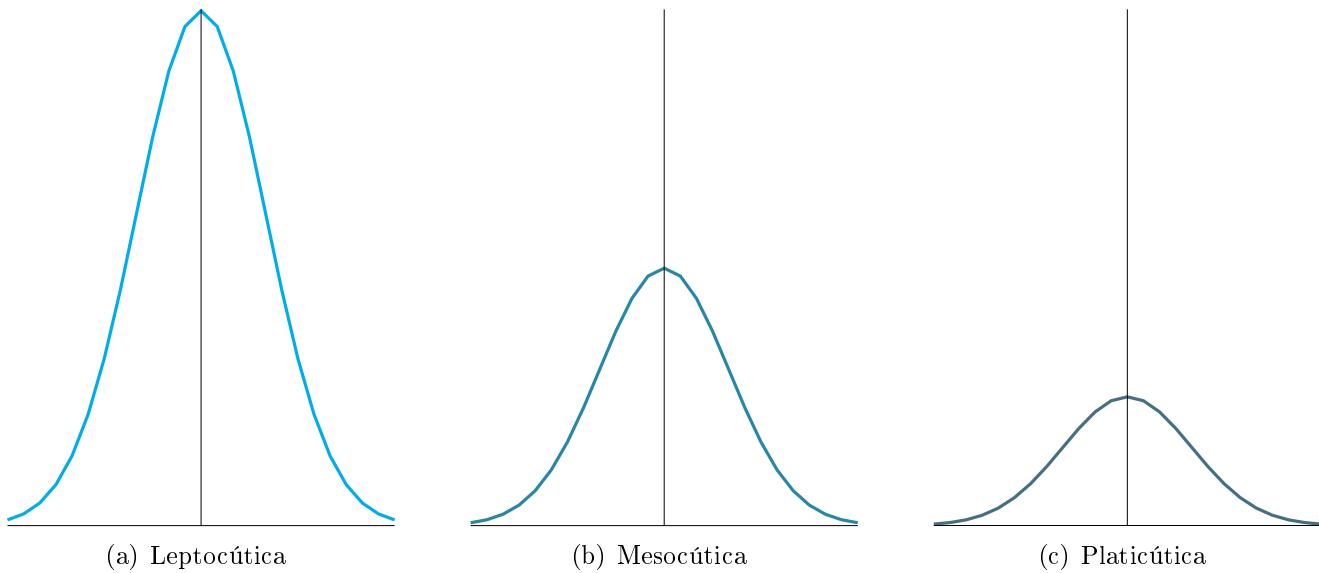
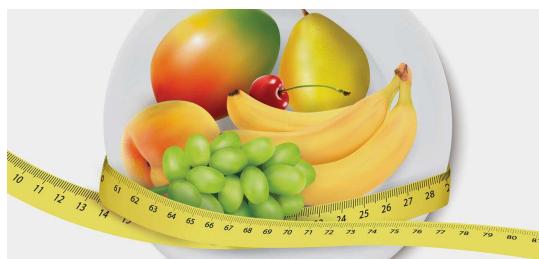


Figura 2.7: Distintos tipos de curtosis

## 2.2.5 Representación gráfica

Sobre el eje de las abscisas (eje horizontal) se representan las distintas categorías, valores o intervalos de la variable en estudio. Sobre el eje las ordenadas (eje vertical) se representan las frecuencias absolutas, las frecuencias relativas o las porcentuales.

En varios de los ejemplos que siguen utilizaremos una base de datos sobre índice de masa corporal (IMC) infantil.



<https://flic.kr/p/FsKKYp>

### 2.2.5.1 Diagrama circular

Es adecuado para representar la distribución de variables cualitativas y cuasicuantitativas. Permite visualizar la proporción captada por cada categoría de la variable.

El Código 2.1 produce la Figura 2.9. Los datos son extraídos de <https://goo.gl/Dpxn9Z>.

```
library(plotrix) # Paquete para manipular dibujos
library(readxl) # Permite leer archivos xlsx

IMCinfantil=read_excel("C:/.../IMCinfantil.xlsx")
# Importa la base con la cual se va a trabajar
attach(IMCinfantil) # Se pone la base en la memoria

frec.catpeso=table(CatPeso) # Calcula las frecuencias de las categorías de peso
etiquetas=c("Deficiente", "Normal", "Obeso", "Con sobrepeso") # Define etiquetas

pie3D(frec.catpeso, labels=etiquetas, explode=0.5, labelcex=0.8, radius=2,
height=0.1, shade=0.7,
col=c("palegreen1", "paleturquoise", "plum2", "lightpink1"))
# Produce un diagrama circular
```

Código 2.1: Generación de un diagrama circular

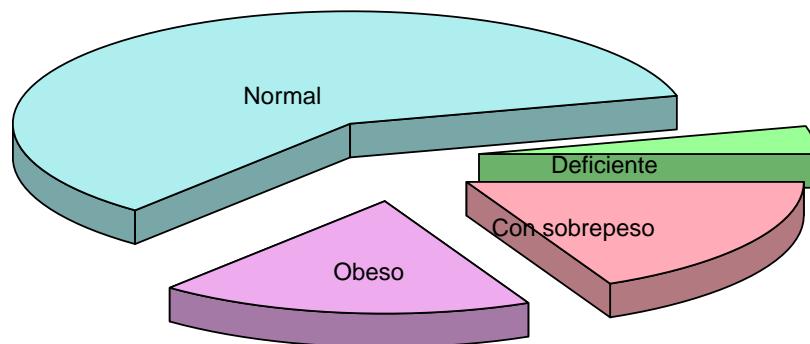


Figura 2.9: Diagrama circular con etiquetas

### 2.2.5.2 Gráfico de barras

Es adecuado para representar variables cualitativas y aventaja al diagrama circular pues que permite apreciar la distribución conjunta de más de una variable.

A modo de ejemplo, exhibimos la Figura 2.10 producida por el Código 2.2. Los datos son extraídos de <https://goo.gl/Dpxn9Z>.

```
library(readxl) # Permite leer archivos xlsx

IMCinfantil=read_excel("C:/.../IMCinfantil.xlsx")
# Importa la base con la cual se va a trabajar
attach(IMCinfantil) # Se pone la base en la memoria

barplot(table(CatPeso), ylab="Cantidad"),
names.arg=c("Deficiente", "Normal", "Obeso", "Con_sobrepeso"),
col=c("palegreen1", "paleturquoise", "plum2", "lightpink1"))
# Produce un diagrama de barras
```

Código 2.2: Generación de un diagrama de barras

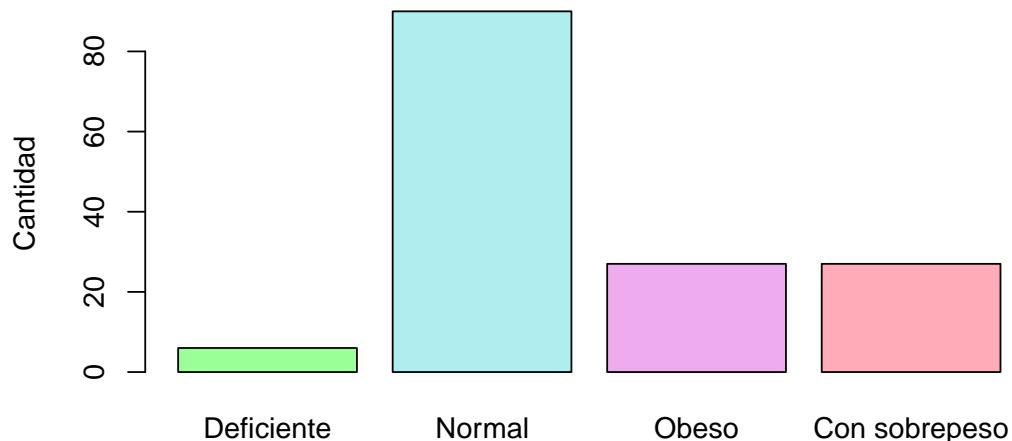


Figura 2.10: Diagrama de barras

### Barras superpuestas

Este tipo de gráfico es útil cuando queremos apreciar la distribución en dos subconjuntos de individuos. A modo de ejemplo, la Figura 2.11 producida por el Código 2.3. Los datos son extraídos de <https://goo.gl/Dpxn9Z>.

```
library(readxl) # Permite leer archivos xlsx
library(ggplot2) # Paquete para confeccionar dibujos
```

```

IMCinfantil=read_excel("C:/.../IMCinfantil.xlsx")
# Importa la base con la cual se va a trabajar
attach(IMCinfantil) # Se pone la base en la memoria

datos=data.frame(table(SEXO, CatPeso)) # Arregla los datos

ggplot(data=datos, aes(x=CatPeso, y=Freq, fill=SEXO)) +
  geom_bar(stat="identity", colour="blue") +
  scale_fill_brewer(palette="Paired") +
  xlab("Categoría de peso") +
  ylab("")
# Produce un diagrama de barras superpuestas

```

Código 2.3: Generación de un diagrama de barras superpuestas

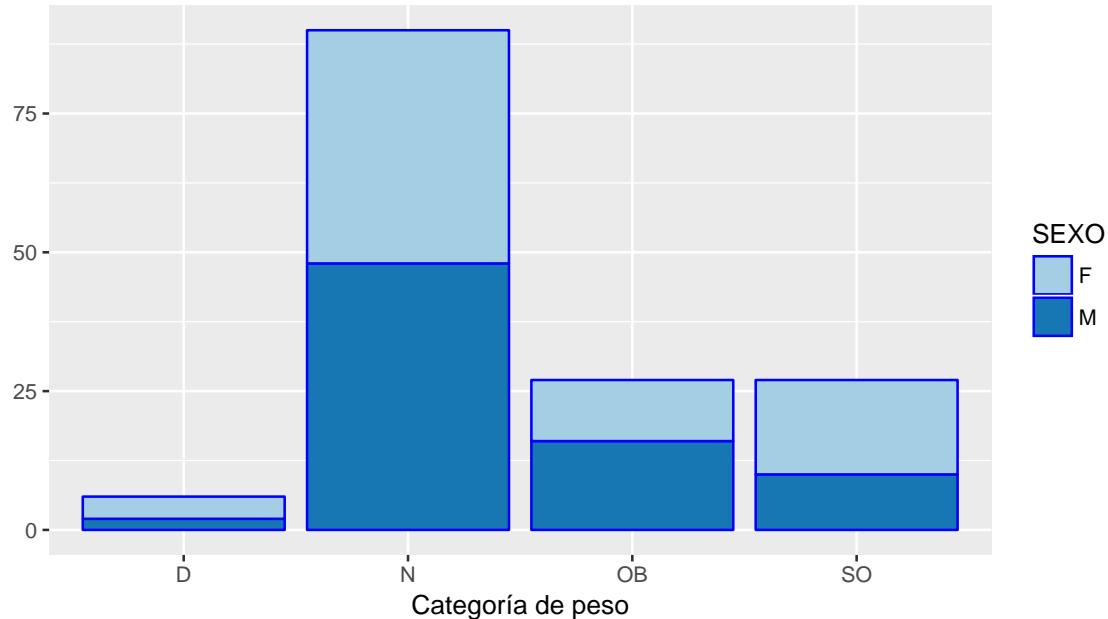


Figura 2.11: Diagrama de barras superpuestas

### Barras adyacentes

En este tipo de esquemas, las barras pueden estar en posición vertical u horizontal. En la Figura 2.12, generada por el Código 2.4, se muestra un ejemplo. Los datos son extraídos de <https://goo.gl/Dpxn9Z>.

```

library(readxl) # Permite leer archivos xlsx
library(ggplot2) # Paquete para confeccionar dibujos

```

```

IMCinfantil=read_excel("C:/.../IMCinfantil.xlsx")
# Importa la base con la cual se va a trabajar
attach(IMCinfantil) # Se pone la base en la memoria

datos=data.frame(table(SEXO, CatPeso)) # Arregla los datos

ggplot(data=datos, aes(x=CatPeso, y=Freq, fill=SEXO)) +
  geom_bar(stat="identity", colour="blue", position="dodge") +
  coord_flip() +
  scale_fill_brewer(palette="Paired") +
  xlab("Categoría de peso") +
  ylab("") +
  # Produce un diagrama de barras adyacentes

```

Código 2.4: Generación de un diagrama de barras adyacentes

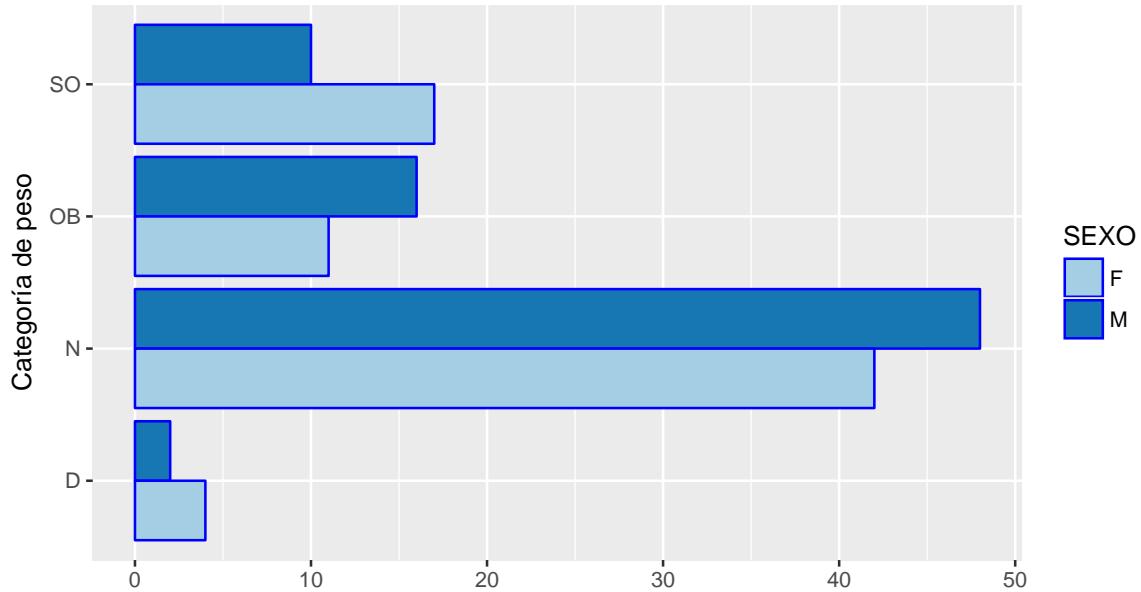


Figura 2.12: Diagrama de barras adyacentes

### 2.2.5.3 Gráfico de bastones

Es adecuado para representar la distribución de frecuencias de una variable discreta. Mostramos como el Código 2.5 genera la Figura 2.13.

```

Modelo=2010:2016 # Ingresá datos
Ventas=c(2,3,7,4,9,0,5) # Ingresá datos

```

```

plot(Modelo, Ventas, type="h", lty="solid", lwd=4,
col=c("palegreen1", "paleturquoise", "plum2", "lightpink1", "deepskyblue3",
"darkorchid2", "indianred1"))
# Produce un diagrama de bastones

```

Código 2.5: Generación de un diagrama de bastones

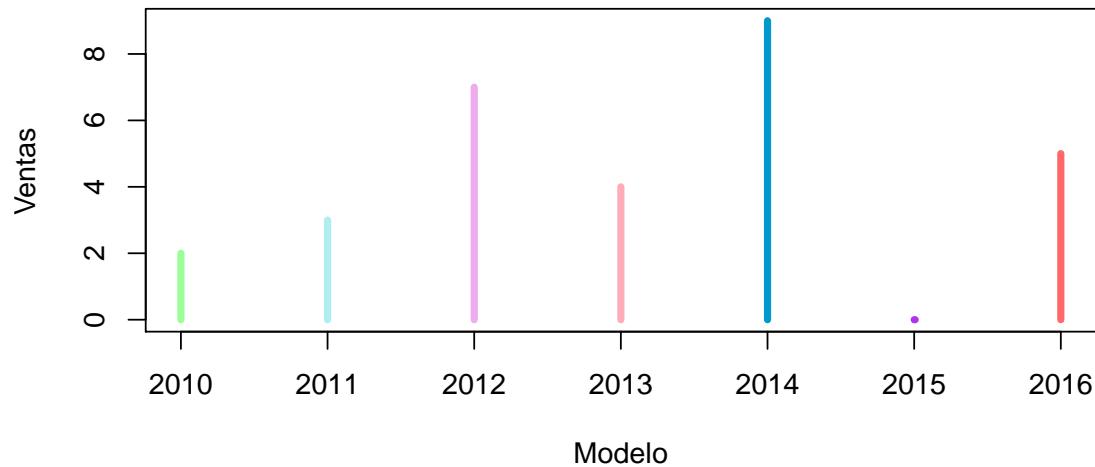


Figura 2.13: Diagrama de bastones

#### 2.2.5.4 Histograma y polígono de frecuencias

Se utiliza para representar distribuciones de frecuencias correspondientes a variables continuas.

El histograma es un método muy utilizado para presentar los datos. Muestra la forma de la distribución de los datos de la misma manera que la función de densidad muestra las probabilidades. El rango de los valores de los datos es dividido en intervalos y se grafica la cantidad o proporción de observaciones que caen dentro de cada intervalo.

Uniendo los puntos medios de las bases superiores de los rectángulos del histograma se construye un polígono de frecuencias. Si la longitud de las bases de los rectángulos se redujera indefinidamente, el polígono de frecuencias tendería a la curva de densidad de la distribución.

Las Figuras 2.14 y 2.15 se obtienen mediante el Código 2.6. Los datos son extraídos de <https://goo.gl/Dpxn9Z>.

```

library(readxl) # Permite leer archivos xlsx

IMCinfantil=read_excel("C:/.../IMCinfantil.xlsx")
# Importa la base con la cual se va a trabajar
attach(IMCinfantil) # Se pone la base en la memoria

hist(PESO, col="paleturquoise3", border="royalblue", breaks=seq(0,85,5),
density=20, angle=70, ylab="", main="")
# Produce un histograma

pto.medio=seq(2.5,82.5,5) # Toma los puntos medios de las barras
alt.dens=hist(PESO, breaks=seq(0,85,5), plot=F)$counts
# Busca la altura de las barras
points(pto.medio, alt.dens, type="l", lwd=2, col="mediumslateblue")
# Agrega el polígono de frecuencias

```

Código 2.6: Generación de un histograma y su polígono de frecuencias

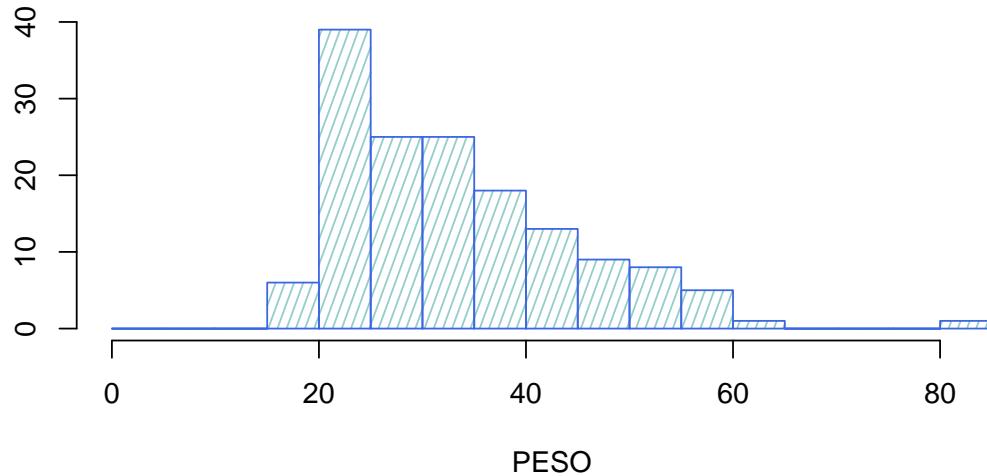


Figura 2.14: Histograma

**Ejemplo 2.7.** Vamos a utilizar el data set `iris` en R.

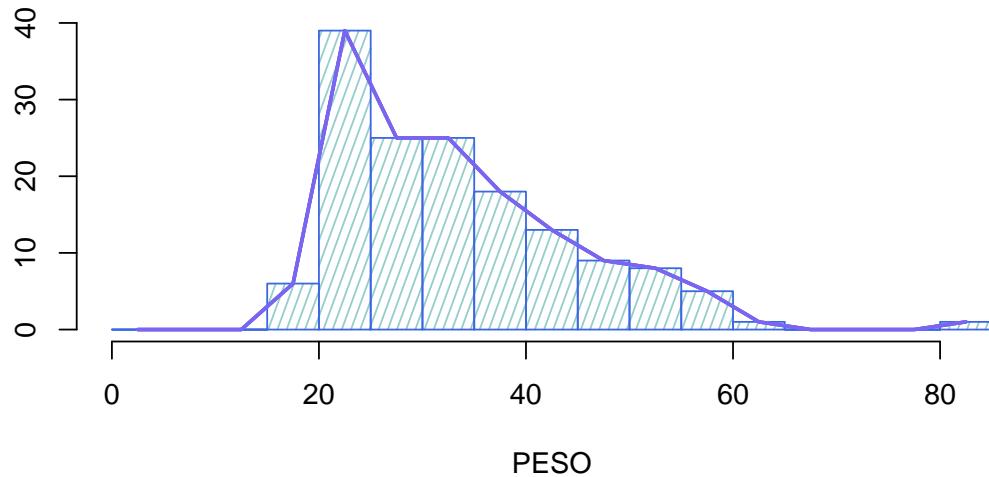


Figura 2.15: Polígono de frecuencias



<https://flic.kr/p/9vfUeF>

Hay que tener en cuenta que el comando `hist` de R dibuja las frecuencias absolutas. Si le agregamos el parámetro opcional `prob=TRUE`, grafica las frecuencias relativas. La Figura 2.17 (ver Código 2.7) muestra tres histogramas del mismo conjunto de datos utilizando diferente cantidad de intervalos.

```
par(mfrow=c(1,3)) # Permite realizar diagramas conjuntos
hist(iris$Sepal.Length, nclass=4, prob=TRUE, ylab="Densidad",
```

```

col="lightsteelblue", border="lightsteelblue4",
xlab="Longitud del sépalo", main="4 clases")

hist(iris$Sepal.Length, nclass=30, prob=TRUE, ylab="Densidad",
col="lightsteelblue", border="lightsteelblue4",
xlab="Longitud del sépalo", main="30 clases")

hist(iris$Sepal.Length, breaks='FD', prob=TRUE, ylab="Densidad",
col="lightsteelblue", border="lightsteelblue4",
xlab="Longitud del sépalo", main="Freedman-Diaconis")

```

Código 2.7: Generación de un histogramas variando la cantidad de clases

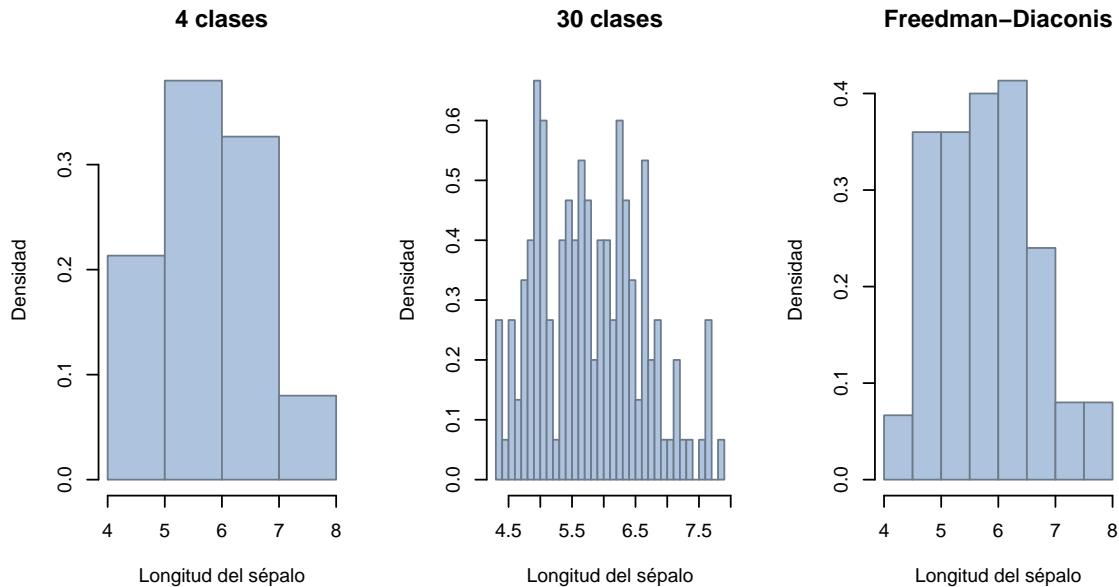


Figura 2.17: Histogramas con distintos intervalos

El parámetro `nclass` da una cantidad sugerida de clases para la función `hist`. Si la cantidad de clases es excesiva el histograma resultante es muy irregular, mientras que si la cantidad es escasa la forma del histograma está sobreavizada.

Entonces para mostrar la distribución subyacente, la pregunta es:

*¿Cómo elegir la cantidad de los intervalos de clase para el histograma o bien el ancho de los mismos?*

Varios autores propusieron respuestas alternativas a esta pregunta.

El **número de intervalos**,  $k$ , sugerido por las siguientes tres reglas depende de la cantidad  $n$  de datos. Las reglas proponen tomar la parte entera y son

- \*  $k = \lfloor 10 \log(n) \rfloor$ , Dixon y Kronmal (1965)

- \*  $k = \lfloor 2\sqrt{n} \rfloor$ , Velleman (1976)

- \*  $k = \lfloor 1 + \log_2(n) \rfloor$ , Sturges (1926)

En la Figura 2.18 se muestra la comparación entre las tres opciones.

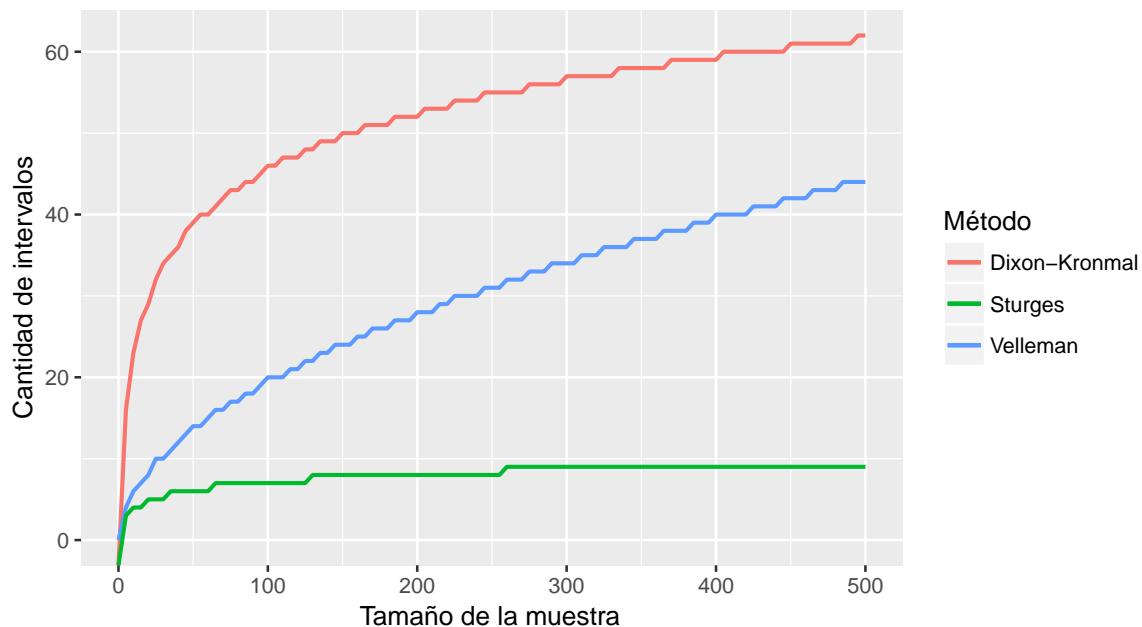


Figura 2.18: Comparación de métodos para el cómputo de intervalos

Entre otras reglas que estiman el ancho de los intervalos de clase podemos mencionar:

- \*  $h_n = 3.49sn^{-1/3}$ , Scott (1979)

- \*  $h_n = 2Rn^{-1/3}$ , Freedman y Diaconis (1981)

donde  $s$  es la desviación estándar de los datos y  $R$  es el rango intercuartil.

## 2.2.5.5 *Boxplot* o diagrama de caja

John Wilder Tukey (1915-2000) propuso este gráfico para presentar datos numéricos, apreciar características importantes de la distribución y comparar distintas distribuciones. Está basado en las medidas de posición. Es un gráfico de fácil lectura.

- ✿ Se dibuja un rectángulo o caja (*box*) cuyos extremos son los cuartiles primero y tercero. Dentro de ella, se dibuja un segmento que corresponde a la mediana o segundo cuartil.
- ✿ A partir de cada extremo, se dibuja un segmento o bigote (*whisker*), hasta el dato más alejado que está, a lo sumo, a 1.5 veces RI del extremo de la caja.
- ✿ Se denominan *outliers* moderados a los datos cuya distancia a uno de los extremos de la caja es mayor que 1.5 veces el RI y menor que 3 veces el RI. Mientras que los *outliers* severos son los datos que están a una distancia mayor a 3 veces el RI de uno de los extremos de la caja.

A partir de un *boxplot* se pueden apreciar los siguientes aspectos de la distribución de un conjunto de datos:

- ✿ posición
- ✿ dispersión
- ✿ asimetría
- ✿ puntos anómalos o *outliers*

Los *boxplots* son especialmente útiles para comparar varios conjuntos de datos, pues nos dan una rápida impresión visual de sus características.

### Datos atípicos, salvajes o *outliers*

Los datos recolectados poseen con frecuencia una o más observaciones atípicas; es decir datos alejados de alguna forma del patrón general del conjunto. La media y la varianza muestrales son buenos resúmenes estadísticos cuando no existen observaciones atípicas o outliers. Sin embargo, en presencia de estos datos salvajes, es conveniente recurrir a medidas más robustas.

La detección de observaciones atípicas es importante, pues su presencia puede determinar o influenciar fuertemente los resultados de un análisis estadístico clásico. Esto ocurre porque muchas de las técnicas habitualmente usadas son muy sensibles a la presencia de este tipo de observaciones, especialmente en el caso de datos multivariados.

Los *outliers* deben ser **cuidadosamente inspeccionados**. Si no hay evidencia de error y su valor es posible **no deben ser eliminados**. Pueden estar alertando de anomalías de un tratamiento o patología, conjuntos especiales de clientes, etc.

La presencia de *outliers* puede indicar que la escala elegida no es la más adecuada, podemos tener una idea de cuán influyentes son los datos, en función de su alejamiento del conjunto general.

**Ejemplo 2.8.** Para la siguiente muestra con  $n = 13$ , tenemos los siguientes datos:

$$\{14, 18, 24, 26, 35, 39, 43, 45, 56, 62, 68, 92, 198\}$$

Para observar el tipo de distribución en el boxplot, es decir, para ver si es simétrica o asimétrica, deben observarse: las distancias entre cuartiles, la posición de la mediana dentro de la caja y el tamaño de los bigotes.

Se observa claramente que el valor 198 está alejado del grupo de valores restantes, por lo que 198 aparenta ser un valor atípico (*outlier*). Inspeccionaremos los datos para confirmar esta hipótesis o no.

¿Se trata de un *outlier salvaje*?

- ✿  $\tilde{x} = 43$
- ✿  $Q_1 = 25$
- ✿  $Q_3 = 65$
- ✿  $R.I. = 65 - 25 = 40$
- ✿  $Q_3 + 1.5 \cdot R.I. = 65 + 60 = 125$
- ✿  $Q_1 - 1.5 \cdot R.I. = 25 - 60 = -35$
- ✿  $VAS = 92$  (valor adyacente superior: mayor valor observado inferior a 125, es el extremo superior del segundo bigote.)
- ✿  $VAI = 14$  (valor adyacente inferior: menor valor observado superior a -35, es el extremo inferior del primer bigote.)
- ✿  $Q_3 + 3 \cdot R.I. = 65 + 120 = 185$
- ✿  $Q_1 - 3 \cdot R.I. = 25 - 120 = -95$
- ✿  $198 > Q_3 + 3 \cdot R.I.$  por lo tanto es un *outlier severo*.



En la Figura 2.19 podemos apreciar el aspecto del *boxplot* para distribuciones simétricas y asimétricas.

### Observaciones

- ✿ Si la distribución es simétrica, vemos que la Mediana está ubicada en el centro de la caja y que los bigotes tienen longitudes similares.

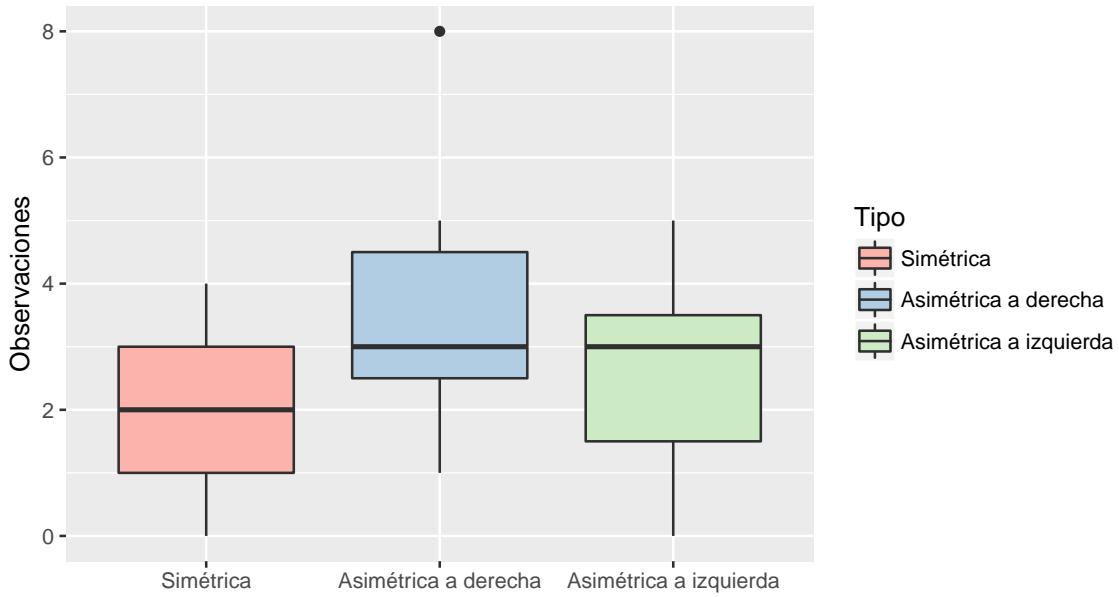


Figura 2.19: Simetría en *boxplots*

- ✿ Si la distribución presenta asimetría positiva (o hacia la derecha), la Mediana se ubica más cerca del Q1, y/o el bigote inferior es de menor tamaño que el bigote superior. Es probable que aparezcan valores atípicos altos.
- ✿ Si la distribución presenta asimetría negativa (o hacia la izquierda), se da la situación inversa de la anterior.
- ✿ Otra forma usual para detectar datos atípicos es la Regla de los tres desvíos. Se define para una observación  $x_i$ , su transformación:

$$t_i = \frac{x_i - \bar{x}}{s}$$

Puesto que en una distribución normal es muy baja la probabilidad  $P(|Z| > 3)$ , entonces se señala como *outlier* a los valores que superan a 3 en valor absoluto. Es decir  $|t_i| > 3$ .

- ✿ Cuando hay varios *outliers* puede que la influencia de ellos se enmascare, es decir que para ciertas medidas se compense el efecto de unos con el efecto de otros.

#### 2.2.5.6 Boxplots comparativos

La representación gráfica conjunta de los *boxplots* correspondientes a las distribuciones de una misma variable en distintos subconjuntos, permite comparar el comportamiento de esta variable en

cada uno de ellos.

**Ejemplo 2.9.** Se desea comparar las mediciones de varios laboratorios respecto del contenido calórico, en kcal, de cierto alimento balanceado. Se sabe que el verdadero valor central del contenido calórico es de 4 kcal para las muestras seleccionadas. Los resultados de las mediciones arrojadas por cada uno de los laboratorios se han representado en el *boxplot* comparativo de la Figura 2.21 generado por el Código 2.8 con datos extraídos de <https://goo.gl/SRd9SR>.



<https://flic.kr/p/nTVq75>

```
library(ggplot2) # Paquete para confeccionar dibujos
library(readxl) # Permite leer archivos xlsx

kcalab=read_excel("C:/.../kcalab.xlsx")
# Importa la base con la cual se va a trabajar
datos=data.frame(kcalab) # Arregla los datos

ggplot(data=datos, aes(y=kcal), colour=factor(Laboratorio)) +
  geom_boxplot(aes(x=Laboratorio, fill=factor(Laboratorio))) +
  xlab("") +
  ylab("Calorías") +
  theme(axis.text.x=element_blank(), axis.ticks=element_blank(),
        axis.line=element_line(colour="royalblue", size=0.5, linetype="solid")) +
  labs(fill='Laboratorio') +
  scale_fill_brewer(palette="BuPu")
# Produce un diagrama comparativo de boxplots
```

Código 2.8: Generación de un *boxplot* comparativo

## Observaciones

- \* Los laboratorios 1 y 3 son los de mayor precisión en sus mediciones.
- \* Los laboratorios 3 y 6 presentan datos atípicos altos.

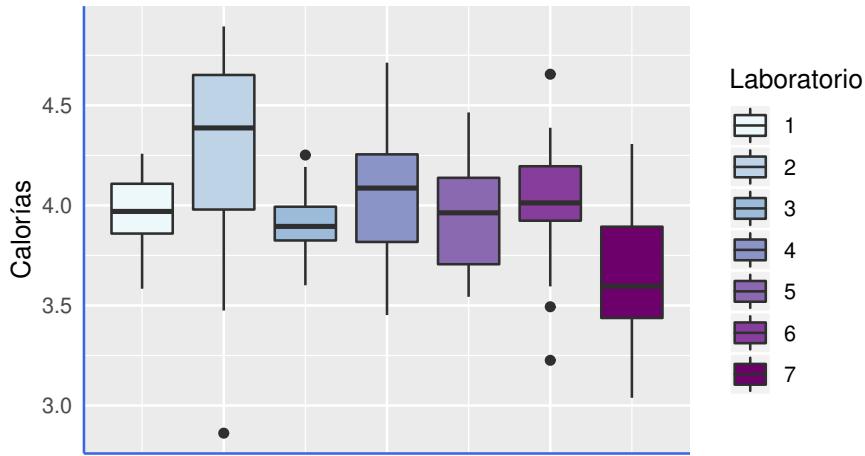


Figura 2.21: *Boxplots* comparativos

- ✿ Todos los laboratorios, excepto el 1 y el 3, presentan asimetría en la distribución de sus mediciones.
- ✿ El laboratorio 2 presenta una asimetría negativa en los valores centrales.
- ✿ El laboratorio 7 presenta una asimetría positiva en los valores centrales.
- ✿ Si se sabe que el verdadero contenido es de 4,00, el laboratorio que deberíamos elegir es el 1, pues entre todos los laboratorios que tienen la mediana próxima al verdadero valor, es el más preciso (menor amplitud del diagrama).



Hasta acá hemos analizado como recopilar, organizar, resumir y representar información de un conjunto de datos respecto de una única variable de interés. Aunque, en rigor de verdad, nuestro objetivo nunca se centra en una sola variable, interesándonos en general el comportamiento de un conjunto de variables.

## 2.3 Información multivariada

La forma más usual en la que se presenta un conjunto de datos multivariados es una tabla donde se listan los valores de  $p$  variables observadas sobre  $n$  elementos.

- ✿ Las **variables** aparecen en las columnas y son características o atributos que toman modalidades diferentes en los individuos de la población. Interesa estudiar el comportamiento de este conjunto de variables en este conjunto de observaciones.

	Variable <sub>1</sub>	...	Variable <sub>j</sub>	...	Variable <sub>p</sub>
Individuo <sub>1</sub>	$X_{1,1}$	...	$X_{1,j}$	...	$X_{1,p}$
⋮	⋮		⋮		⋮
Individuo <sub>i</sub>	$X_{i,1}$	...	$X_{i,j}$	...	$X_{i,p}$
⋮	⋮		⋮		⋮
Individuo <sub>n</sub>	$X_{n,1}$	...	$X_{n,j}$	...	$X_{n,p}$

Tabla 2.7: Modelo de base de datos

- ✿ Los **individuos** aparecen en las filas. Son los ejemplares o elementos sobre los cuales se miden los atributos.
- ✿ Las tablas tendrán entonces  $n$  **filas** y  $p$  **columnas**; siendo  $n$  el número de individuos observados o unidades de análisis y  $p$  la cantidad de variables de interés sobre las cuales basaremos nuestro análisis.

Los datos pueden ser acomodados en una matriz de la siguiente manera

$$X = \begin{pmatrix} & \text{Variables en columnas} \\ x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad \begin{matrix} & \text{Individuos en filas} \\ & \end{matrix}$$

Denotaremos a cada elemento genérico de esta matriz como  $x_{ij}$ , que representa el valor de la variable  $j$  observado sobre el individuo  $i$  (fila  $i$ , columna  $j$ ).

**Ejemplo 2.10.** Los datos de las galletitas se exhiben en la Tabla 2.8.



<https://flic.kr/p/a6Hs83>

Marca	Valor energético cal/100g	Carbohidratos g/100g	Proteinas g/100g	Grasas g/100g	Sodio mg/100g
Marca 1	439	65.0	11.0	15	574.00
Marca 2	466	57.0	10.0	22	828.00
Marca 3	445	69.0	11.0	14	12.00
Marca 4	478	67.0	5.6	21	363.00
Marca 5	464	70.0	6.3	18	263.00
Marca 6	463	66.0	7.1	19	136.00
Marca 7	438	69.0	11.0	13	431.00
Marca 8	418	69.0	6.3	13	201.00
Marca 9	423	70.0	6.8	13	241.00
Marca 10	444	73.0	9.0	13	375.00
Marca 11	407	70.0	6.0	12	106.70
Marca 12	437	60.0	6.7	18	76.67
Marca 13	410	56.7	6.3	18	66.70
Marca 14	493	60.0	7.6	24	1066.00
Marca 15	424	65.0	11.0	13	892.00
Marca 16	462	55.0	11.0	22	931.00
Marca 17	421	63.0	11.0	14	624.00

Tabla 2.8: Base de datos para las galletitas

En este ejemplo, con respecto a la matriz de datos,  $p = 5$  y  $n = 17$ . El valor  $x_{23} = 10$  representa la cantidad en gramos de proteínas cada 100 g de la segunda de las marcas elegidas; es decir, para las galletitas de la Marca 2 (segunda fila).

El análisis de datos multivariantes tiene por objeto el **estudio estadístico de varias variables** medidas en un subconjunto de elementos de una población.

La descripción de los datos multivariantes **comprende el estudio de cada variable aisladamente y también de las relaciones que quedan definidas entre ellas.**

Para entender la complejidad del problema con el cual nos vamos a enfrentar, pensemos que, en casos univariados, basta con estimar dos parámetros para la variable:

- ✿ uno de centralidad (por ejemplo la media),
- ✿ uno de dispersión (por ejemplo la varianza).

En el caso de una población  $p$ -variada; donde se han observado o medido  $p$  características sobre cada individuo, se dispondrá de  $p$  medias,  $p$  varianzas y  $\frac{p(p - 1)}{2}$  covarianzas (concepto que trataremos en detalle más adelante).

Vale decir que, en lugar de estimar dos parámetros debemos aproximar el valor de:

$$2p + \frac{p(p - 1)}{2} = \frac{p^2 + 3p}{2}$$

parámetros.

En la Tabla 2.9 se puede apreciar cómo crece la cantidad de parámetros a medida que aumenta la cantidad de variables observadas sobre cada individuo.

Variables	Parámetros a estimar
2	5
3	9
4	14
5	20
6	27
7	35
8	44
9	54
10	65

Tabla 2.9: Cantidad de parámetros en función de las variables

**Ejemplo 2.11.** En el Ejemplo 2.10 se tiene que  $p = 5$ , lo que implica estimar 20 parámetros. ■

### 2.3.1 Objetivos del análisis exploratorio

Algunos de los objetivos que se fijan en el análisis exploratorio son los siguientes.

- ✿ Conocer los datos.
- ✿ Descubrir regularidades.
- ✿ Verificar la existencia de estructuras ocultas.
- ✿ Entender los patrones descubiertos.
- ✿ Resumir información.
- ✿ Hallar asociaciones de variables.
- ✿ Detectar anomalías.

Con estos propósitos resultará de utilidad disponer de los datos de forma tal que podamos observar y describir estos patrones.

Veremos a continuación algunas otras formas de presentar y representar conjuntos de datos multivariados.

#### 2.3.1.1 Tabla de clasificación cruzada

Se han tabulado las consideraciones respecto del consumo y de la garantía, que tienen 1441 clientes en el momento de decidir la compra de un auto 0 km. y en la Tabla 2.10 se presenta la distribución conjunta de estas dos variables.

		Se tuvo en cuenta el consumo		
		NO	SI	TOTAL
Se tuvo en cuenta la garantía	NO	258	280	538
	SI	184	719	903
	TOTAL	442	999	1441

Tabla 2.10: Consideraciones para la compra de un auto

Cada una de ellas tiene dos niveles, por lo cual la tabla tiene dos filas y dos columnas, sin considerar la fila y la columna de totales.

Cuando las dos variables consideradas son categóricas, una representación adecuada es el gráfico de mosaicos.

### 2.3.1.2 Gráfico de mosaicos

Se utiliza para representar **distribuciones conjuntas multivariadas**.

En la Figura 2.23 de mosaicos, producida mediante el Código 2.9, se representan los datos de la Tabla 2.10 que indica las consideraciones tomadas antes de comprar un auto.

```
gar.no=c(258,    280) # Carga de datos
gar.si=c(184,    719)

mat=rbind(gar.no, gar.si) # Combina datos
colnames(mat)=c("No\_considera\_consumo", "Considera\_consumo")
# Pone nombre a las columnas
rownames(mat)=c("No\_considera\_garantía", "Considera\_garantía")
# Pone nombre a las filas

mosaicplot(mat, col=c("skyblue", "royalblue"), cex.axis=0.8, main="")
# Produce un diagrama de mosaicos
```

Código 2.9: Generación de un diagrama de mosaicos

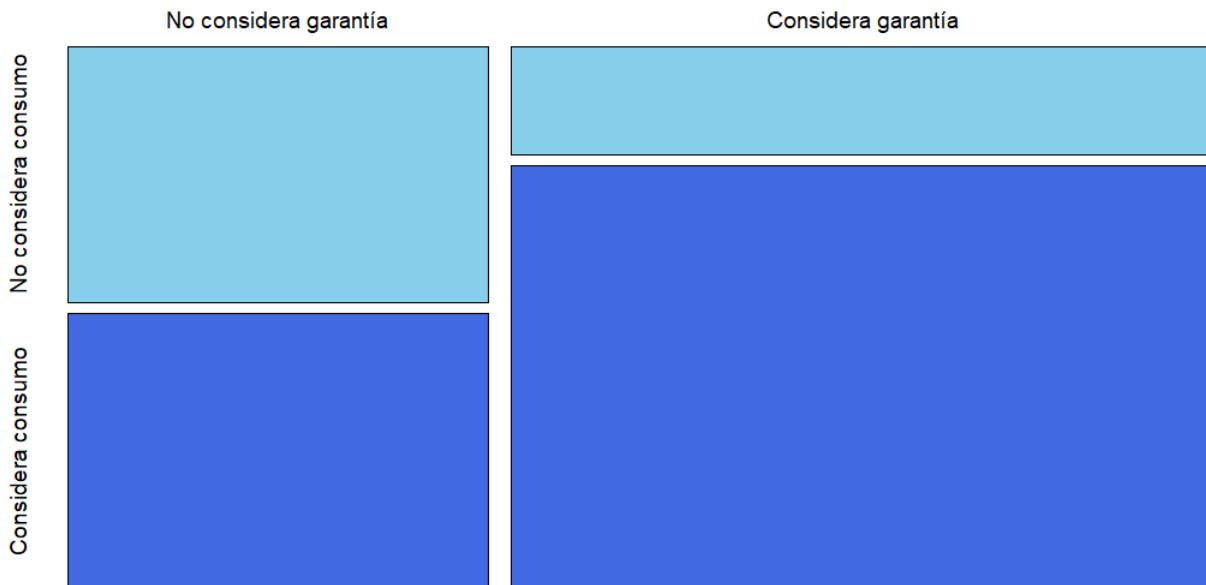


Figura 2.23: Diagrama de mosaicos

En la Figura 2.23 se aprecia que es menor la proporción de compradores que han tenido en cuenta el consumo entre los que consideraron la garantía que entre los que no han tenido en cuenta la garantía, en el momento de decidir la compra.

### 2.3.1.3 Diagrama de dispersión

Vamos a utilizar el conjunto de datos `mtcars` en R, donde se han medido características de consumo, cilindradas, peso, número de carburadores y trasmisión en diferentes modelos de autos. Con el Código 2.10 generamos el diagrama de dispersión de la Figura 2.24.

```
library(ggplot2) # Paquete para confeccionar dibujos

mtcars$cilind=factor(mtcars$cyl) # Declara las cilindradas como factor

ggplot(mtcars, aes(wt, mpg)) +
  geom_point(aes(colour=cilind)) +
  xlab("Peso") +
  ylab("Millas por galón") +
  labs(colour='Cilindrada')
# Produce un diagrama de dispersión
```

Código 2.10: Generación de un diagrama de dispersión

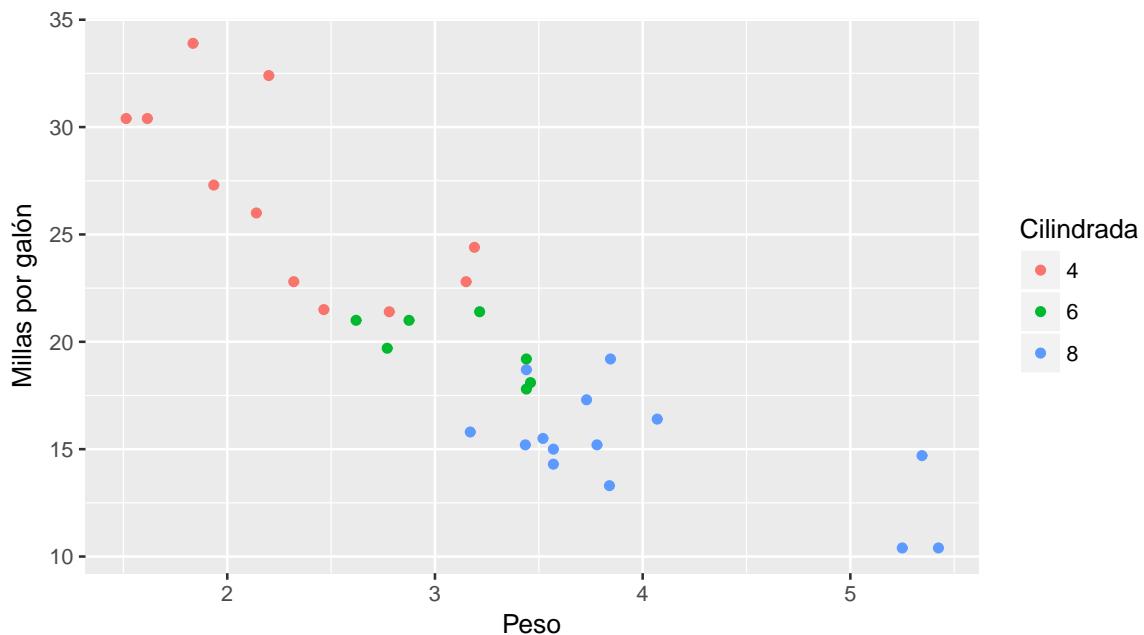


Figura 2.24: Diagrama de dispersión para tres poblaciones

En la Figura 2.24 se han representado tres variables y podemos apreciar simultáneamente:

- ✿ Características individuales de la variable ‘Peso’.
- ✿ Características individuales de la variable ‘Millas por galón’.

- ✿ Posicionamiento de los grupos definidos por las cilindradas respecto de ambas.
- ✿ Posicionamiento relativo de los grupos.
- ✿ Relación entre variables cuantitativas por grupo definido por las cilindradas y en general.

#### 2.3.1.4 Dispersograma

Cuando sobre un conjunto de individuos se han medido varias variables cuantitativas, puede resultar de interés visualizar si existe vinculación entre pares de estas variables. Para esta visualización es muy útil el dispersograma.

Utilizamos nuevamente el conjunto de datos disponibles en <https://goo.gl/Dpnx9Z> y, mediante el Código 2.11, generamos el dispersograma de la Figura 2.25.

```
library(readxl) # Permite leer archivos xlsx
IMCinfantil=read_excel("C:/.../IMCinfantil.xlsx")
# Importa la base con la cual se va a trabajar
attach(IMCinfantil) # Se pone la base en la memoria
SEX=4*(SEXO=="F")+5*(SEXO=="M")
#define una variable cuantitativa para el factor SEXO
base.niños=data.frame(EDAD, PESO, TALLA, IMC, CC)
# Arma una sub-base con variables numéricas
pairs(base.niños, pch=19, cex=0.8, col=SEX)
# Produce un diagrama de dispersión de a pares
```

Código 2.11: Generación de un dispersograma

En el dispersograma de la Figura 2.25 se aprecia la variación conjunta de cada par de variables de la base, en general y por sexo (azul corresponde a mujeres y celeste a varones).

#### 2.3.1.5 Gráfico de coordenadas paralelas

Los gráficos de coordenadas paralelas son una alternativa para la visualización datos multidimensionales.

- ✿ En lugar de usar ejes perpendiculares ( $x, y, z$ ) se utilizan ejes paralelos.
- ✿ Cada atributo es representado en uno de estos ejes paralelos con sus respectivos valores.
- ✿ Se escalan los valores de los distintos atributos para que la representación de los mismos tenga la misma altura.
- ✿ Cada individuo se representa mediante una línea que une los puntos que le corresponden en los distintos ejes.

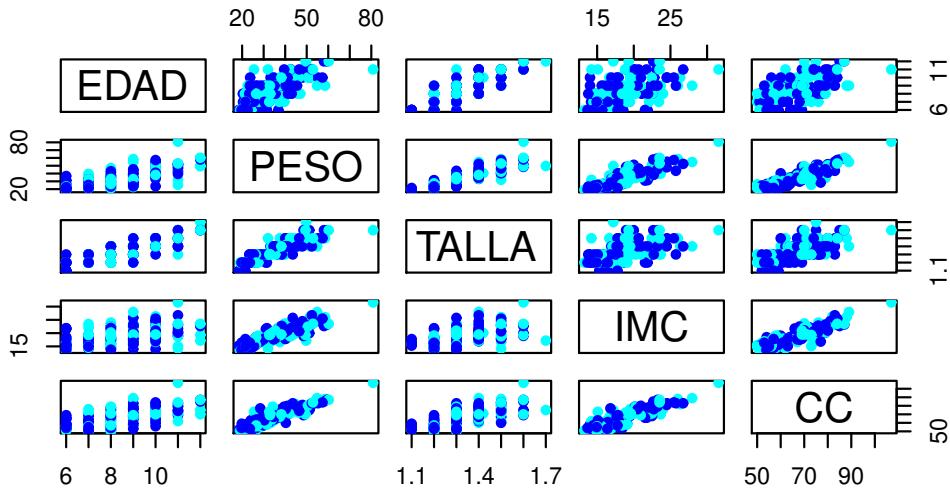


Figura 2.25: Dispersograma

- ✿ De esta forma, se puede apreciar la similitud de las observaciones.
- ✿ También puede compararse la forma de distintos subgrupos o definir patrones, realizando el gráfico con diferentes colores para cada subgrupo.

Con los datos Iris de R, construimos la Figura 2.26 mediante el Código 2.12.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(GGally) # Paquete que extiende funciones de ggplot2

ggparcoord(data=iris, columns=1:4, mapping=aes(color=as.factor(Species))) +
  scale_color_discrete("Especies", labels=levels(iris$Species)) +
  xlab("") +
  ylab("") +
  scale_x_discrete(limit=c("Sepal.Length", "Sepal.Width", "Petal.Length",
  "Petal.Width"),
  labels=c("Longitud del sépalo", "Ancho del sépalo",
  "Longitud del pétalo", "Ancho del pétalo"))
# Produce diagrama de coordenadas paralelas
```

Código 2.12: Generación de un gráfico de coordenadas paralelas

En la Figura 2.26 se puede apreciar que hay representadas tres especies, cada una de ellas con un color distinto. La relación entre longitud y ancho del sépalo es claramente distinta en el grupo

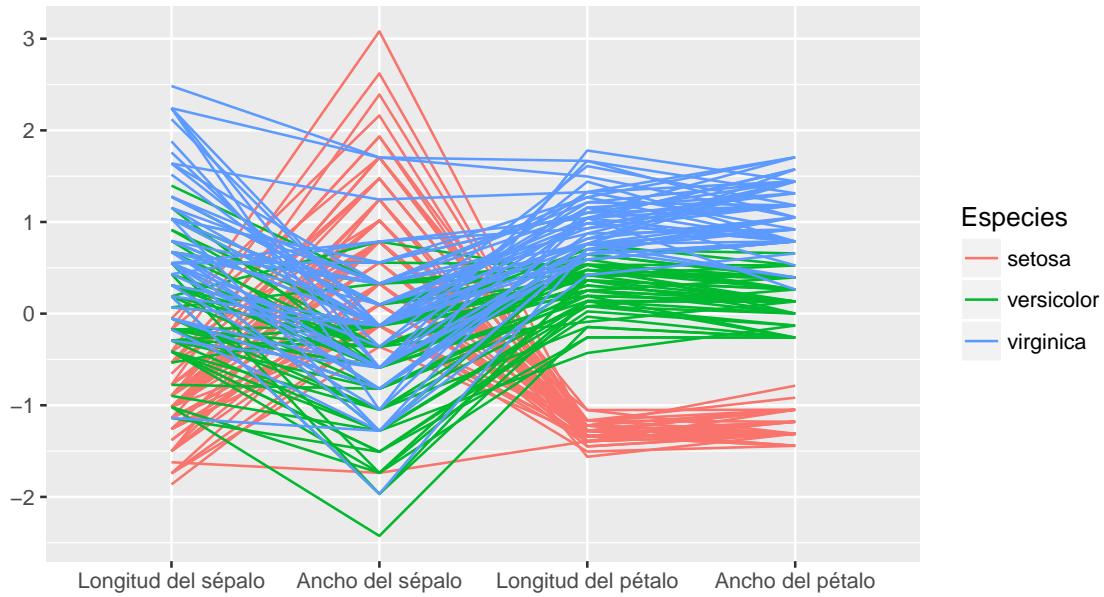


Figura 2.26: Gráfico de coordenadas paralelas

de ‘virginica’ y ‘versicolor’ respecto del grupo ‘setosa’. Una apreciación similar puede realizarse con respecto a los anchos del pétalo y el sépalo.

### 2.3.1.6 Gráfico de perfiles multivariados

Se representan los valores medios o medianos de cada una de las variables observadas en distintos individuos en las diferentes categorías en las que se clasifica a los grupos o a los individuos. Esto permite comparar la posición central de estas variables en los distintos individuos o grupos definidos.

Con los datos de disponibles en <https://goo.gl/yDmQE2> sobre ciertas características de diferentes tipos de galletitas, se construye la Figura 2.27 mediante el Código 2.13. Se aprecia en la misma que la composición nutricional media de las galletitas dulces y saladas es similar en todas las variables estudiadas, excepto en el contenido de sodio.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(readxl) # Permite leer archivos xlsx
library(reshape) # Paquete para reestructurar datos

galle=read_excel("C:/.../galletitas.xlsx")
# Importa la base con la cual se va a trabajar

dulces=split(galle, galle$Tipo)$dulce # Agrupa las dulces
saladas=split(galle, galle$Tipo)$salada # Agrupa las saladas
med.dul=apply(dulces[,2:6], 2, mean) # Calcula las medias de las dulces
```

```

med.sal=apply(saladas[,2:6], 2, mean) # Calcula las medias de las saladas

data.plot=data.frame(group=c(1,2,3,4,5), value1=med.dul+7, value2=med.sal)
melteddata = melt(data.plot, id = 'group')
# Arregla datos para gráfico

ggplot(melteddata, aes(x = group, y = value, colour = variable)) +
  geom_line() +
  xlab("Variables") +
  ylab("Medias") +
  scale_x_discrete(limit=c("1", "2", "3", "4", "5"),
  labels=c("Calorías", "Carbohidratos", "Proteinas", "Grasas", "Sodio")) +
  labs(colour='Tipo') +
  scale_colour_manual(labels = c("Dulces", "Saladas"),
  values=c("royalblue", "green4"))

```

Código 2.13: Generación de un gráfico comparativo de perfiles

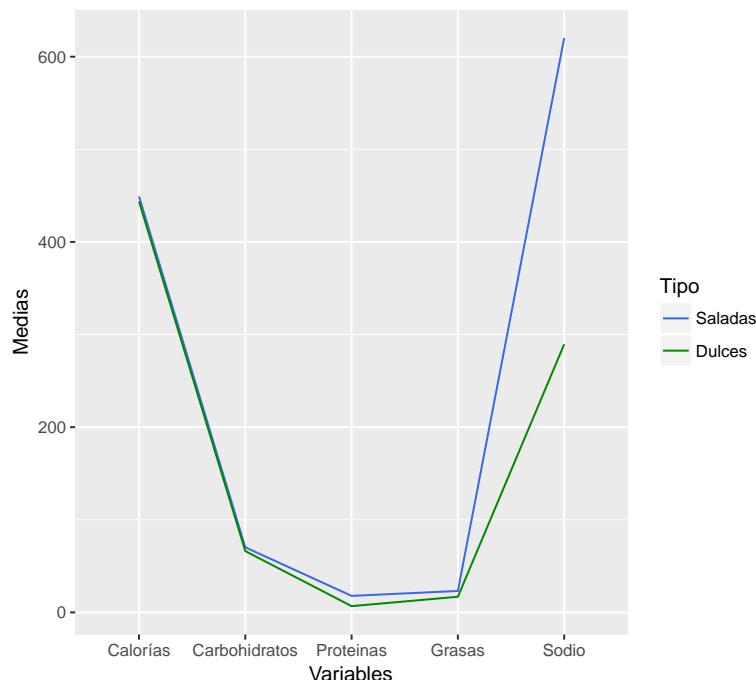


Figura 2.27: Gráfico de perfiles

### 2.3.1.7 Curvas de nivel

Las curvas de nivel unen puntos de igual cantidad de observaciones. De este modo, los distintos colores ayudan a identificar regiones de mayor o menor densidad de observaciones.

Mostramos el caso de la distribución Normal Bivariada en las Figuras 2.28 y 2.29, ambas fueron generadas mediante el Código 2.14.

```
fun=function(x,y) exp(-x^2-y^2)
# Define la función de distribución Normal Bivariada con ro=0

x=seq(-3,3,0.1)
y=x
# Asigna valores a las variables

persp(x, y, outer(x,y,fun), theta = -15, phi = 30, r = sqrt(3), d = 3,
col="deepskyblue1", xlab = "x", ylab = "y",
zlab ="z")
# Produce un dibujo de la Normal Bivariada

filled.contour(outer(x,y,fun), axes=TRUE, frame.plot=FALSE,
color.palette = topo.colors, plot.axes=FALSE)
# Grafica las curvas de nivel de la Normal Bivariada
```

Código 2.14: Generación de las curvas de nivel de la Normal Bivariada

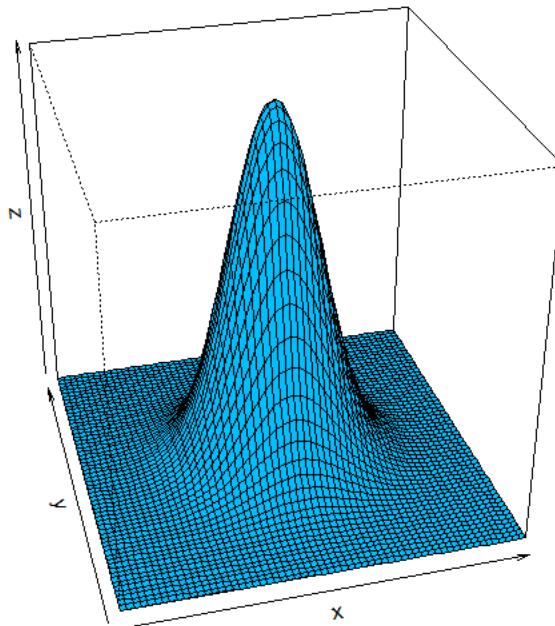
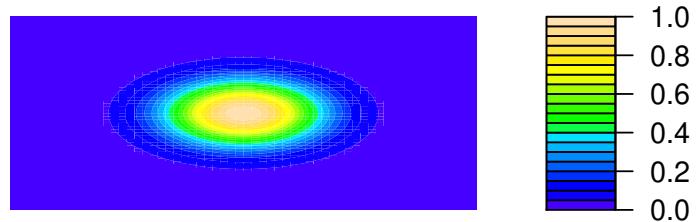


Figura 2.28: Gráfico de la distribución Normal Bivariada



**Figura 2.29:** Gráfico de las curvas de nivel de la distribución Normal Bivariada

#### 2.3.1.8 Gráficos de estrellas

Cuando todas las variables consideradas son cuantitativas para poder detectar estructuras similares, es adecuado el gráfico de estrellas.

Queremos encontrar similitudes entre individuos o grupos del conjunto de datos considerado. Con los datos del archivo `mtcars` de R, seleccionamos los primeros nueve modelos de autos. Cada variable es representada con un radio de una estrella, la longitud del radio está dada por el valor de la variable en un individuo o bien por el promedio de observaciones de esa variable en el grupo. Por ejemplo podríamos representar en una estrella los autos familiares y en otra los utilitarios.

Mostramos un ejemplo de ello en la Figura 2.30, generada con el Código 2.15.

```
autos=mtcars[1:9,] # Toma las primeras nueve marcas de la base
row.names(autos)=c("Mazda", "Mazda_Wag", "Datsun", "Hornet_D", "Hornet_S",
"Valiant", "Duster", "Merc_D", "Merc")
# Coloca etiquetas

stars(autos, full=F, cex=0.8, flip.labels=T, len=0.9, col.stars=cm.colors(9))
# Produce un diagrama de estrellas
```

**Código 2.15:** Generación de un gráfico de estrellas

En la Figura 2.30 se aprecia similitud en la estructura de los modelos *Mazda* y *Mazda Wag*, así como también son similares los modelos *Merc D* y *Merc*.

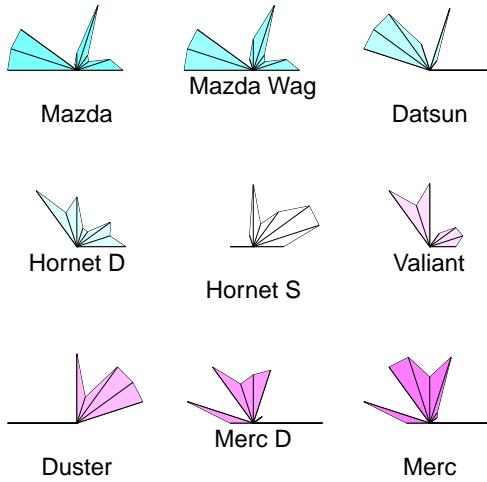


Figura 2.30: Gráfico de estrellas

### 2.3.1.9 Gráficos de caras de Chernoff

Las caritas de Chernoff [5] son un método gráfico mediante el cual ciertas características cuantitativas de un grupo de observaciones se asocian con datos físicos de la cara de una persona. Esto permite realizar un dibujo que representa dichas características, y visualizar fácilmente similitudes y diferencias entre individuos, dado que estamos habituados a hacerlos con personas.

En la Figura 2.31, generada con el Código 2.16 con datos extraídos de <https://goo.gl/yDmQE2>, se muestran caras de Chernoff para ciertas marcas de galletitas saladas. En la misma, se aprecian similitudes entre las marcas 8, 9 y 11 por un lado y entre las marcas 5 y 6 por otro.

```
library(tcltk2) # Paquete que permite hacer caras de Chernoff
library(aplpack) # Paquete que permite hacer caras de Chernoff
library(readxl) # Permite leer archivos xlsx

galle=read_excel("C:/.../galletitas.xlsx")
# Importa la base con la cual se va a trabajar

saladas=split(galle, galle$Tipo)$salada # Agrupa las saladas

faces(saladas[,2:6], nrow.plot=2, ncol.plot=5, face.type=1,
labels=saladas$Marca)
# Produce un diagrama de caras de Chernoff
```

Código 2.16: Generación de caras de Chernoff

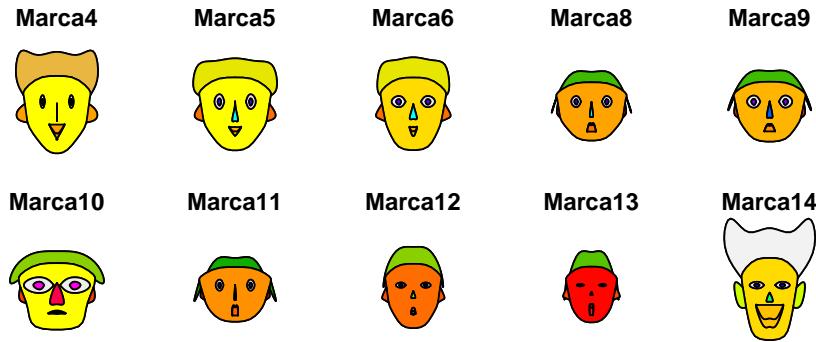


Figura 2.31: Gráfico de caras de Chernoff para galletitas saladas

## 2.4 Medidas de posición y dispersión en datos multivariados

Un conjunto de  $p$  variables observadas sobre  $n$  individuos puede representarse mediante una matriz  $X \in \mathbb{R}^{n \times p}$ . Definimos el **vector de medias muestral** como:

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p) \in \mathbb{R}^p$$

donde  $\bar{x}_i$  se refiere al promedio de la  $i$ -ésima variable (columna) observada; es decir, es un vector formado por la media de cada una de las variables observadas.

Además, se define la **matriz de varianzas y covarianzas muestral** como:

$$\widehat{\Sigma} = \frac{1}{n}(X - \bar{X})^t(X - \bar{X})$$

donde  $\bar{X}$  es una matriz que en cada una de sus columnas tiene el promedio muestral de la variable respectiva repetido tantas veces como individuos tiene el conjunto de observaciones.

La matriz  $\widehat{\Sigma}$ , de tamaño  $p \times p$ , resulta ser simétrica y su diagonal principal está formada por las varianzas muestrales de cada una de las variables observadas; mientras que fuera de su diagonal, se encuentran las covarianzas muestrales de cada par de variables.

### 2.4.1 Propiedades del vector de medias

Sean  $X, Y \in \mathbb{R}^{n \times p}$  matrices que guardan los datos observados, y sean  $A, B \in \mathbb{R}^{p \times k}$  y  $C \in \mathbb{R}^{n \times k}$  matrices escalares, entonces:

- \*  $\overline{XA + C} = \bar{X}A + C.$
- \*  $\overline{XA + YB} = \bar{X}A + \bar{Y}B.$

### 2.4.2 Propiedades de la matriz de varianzas y covarianzas

- \* La matriz de covarianzas muestral es simétrica:  $\hat{\Sigma}^t = \hat{\Sigma}$ , es decir que para todo  $i, j$  se cumple que  $\hat{\Sigma}_{ij} = \hat{\Sigma}_{ji}$ .
- \*  $\hat{\Sigma} = \frac{1}{n}(X - \mathbb{1}_n\bar{x})^t(X - \mathbb{1}_n\bar{x})$ , siendo  $\mathbb{1}_n$  el vector columna de  $n$  unos.
- \*  $\hat{\Sigma}$  estima a la matriz de varianzas y covarianzas poblacional  $\Sigma = E[(X - \mathbb{1}_n\mu)^t(X - \mathbb{1}_n\mu)]$  que también es simétrica.
- \* La matriz de covarianzas (poblacional o muestral) es semidefinida positiva; es decir, que todos sus autovalores son mayores o iguales a cero.
- \* Si  $Y = XA + B$ ,  $\Sigma_Y = A^t\Sigma_X A$ , siendo  $A \in \mathbb{R}^{p \times k}$  y  $B \in \mathbb{R}^{n \times k}$  matrices de escalares.

**Ejemplo 2.12.** Vamos a buscar la matriz de covarianza muestral correspondiente al conjunto de observaciones dado por  $X = \begin{pmatrix} 10 & 4 \\ 15 & 1 \\ 20 & 7 \end{pmatrix}$ .

Tenemos que

$$\bar{x} = (15 \ 4), \quad \bar{X} = \mathbb{1}_3\bar{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (15 \ 4) = \begin{pmatrix} 15 & 4 \\ 15 & 4 \\ 15 & 4 \end{pmatrix} \quad \text{y}$$

$$\hat{\Sigma} = \frac{1}{3} \left[ \begin{pmatrix} 10 & 4 \\ 15 & 1 \\ 20 & 7 \end{pmatrix} - \begin{pmatrix} 15 & 4 \\ 15 & 4 \\ 15 & 4 \end{pmatrix} \right]^t \left[ \begin{pmatrix} 10 & 4 \\ 15 & 1 \\ 20 & 7 \end{pmatrix} - \begin{pmatrix} 15 & 4 \\ 15 & 4 \\ 15 & 4 \end{pmatrix} \right] = \frac{1}{3} \begin{pmatrix} 50 & 15 \\ 15 & 18 \end{pmatrix} = \begin{pmatrix} 16.6 & 5 \\ 5 & 6 \end{pmatrix}$$



## 2.5 Transformación del conjunto de datos

En algunas ocasiones, para optimizar el análisis de la información disponible, es conveniente realizar transformaciones a los datos. Las transformaciones pueden ser por filas o por columnas, o sea por individuos o por variables, dependiendo de los objetivos de las mismas.

Los objetivos más usuales de estas transformaciones son:

- ✿ Hacer comparables las magnitudes.
- ✿ Modificar la escala de medición.
- ✿ Satisfacer alguna propiedad estadística.

### 2.5.1 Transformaciones por variables

Las transformaciones por variables se aplican con el objeto de hacer comparables los valores asignados a los distintos individuos u objetos de análisis. Por ejemplo, cuando un grupo de jueces deben evaluar un conjunto de individuos o productos, suele ocurrir que algunos de ellos tengan tendencia a poner puntuaciones muy altas o muy bajas de manera subjetiva, lo cual sesga el estudio. Para neutralizar estas diferencias se utilizan transformaciones por filas tales como las que veremos a continuación.

#### 2.5.1.1 Variables aleatorias estandarizadas

Suele denominarse a la transformación de estandarizado como *z-scores* o puntuaciones  $Z$ , ya que tienen la característica de tener media 0 y varianza 1. Las mismas se realizan restando a las observaciones el valor medio muestral y dividiendo esta diferencia por el desvío estándar muestral. Simbólicamente,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j^2}} \quad (2.1)$$

Estas transformaciones tienen sentido en el caso en que la media y el desvío resulten una buena representación de la centralidad y la dispersión respectivamente. En caso contrario, pueden considerarse en forma alternativa la mediana y la desviación intercuartil o la mediana y el MAD.

### 2.5.2 Transformaciones por individuo

Se aplican con el objeto de hacer comparables los valores de los distintos individuos. En el caso de varios jueces que evalúan un conjunto de individuos o productos. Se sabe que un juez podría tener

una tendencia a puntuaciones muy altas o muy bajas lo cual sesgaría el estudio. Para neutralizar la influencia de esta tendencia, se realizan transformaciones por fila. Por ejemplo, la siguiente

$$T(x) = \begin{cases} \frac{x - \bar{x}}{x_{\max} - \bar{x}} & \text{si } x > \bar{x} \\ \frac{x - \bar{x}}{\bar{x} - x_{\min}} & \text{si } x < \bar{x} \end{cases}$$

La transformación de las puntuaciones superiores a la media de cada juez resultarán positivas, mientras que las que resulten inferiores a la media resultarán negativas. A las puntuaciones superiores se las normaliza por la distancia entre la media y el máximo, mientras que a las inferiores por la distancia entre la media y el mínimo.

## 2.6 Análisis multivariado

*¿En qué nos beneficia realizar el análisis conjunto de todas las variables?*

**Ejemplo 2.13.** Consideremos un conjunto de cajas producidas por una máquina o un operador. Si observamos el comportamiento de una sola variable, podemos detectar si alguna observación está alejada de la mayor parte de los datos. Con los datos extraídos de <https://goo.gl/uWiUtv>) mediante el Código 2.17 generamos la Figura 2.33.



<https://flic.kr/p/9qBNAs>

```
library(ggplot2) # Paquete para confeccionar dibujos
library(dplyr) # Paquete para manipular datos
library(readxl) # Permite leer archivos xlsx
```

```

datos=read_excel("C:/.../controlunivariado.xlsx")
# Importa la base con la cual se va a trabajar
attach(datos) # Se pone la base en la memoria

dat=datos %>% group_by(Obs, Clase) # Reagrupa la base
exp_names <- c('A'="Bajo_control", 'B'="Fuera_de_control",
'C'="Fuera_de_control") # Cambia etiquetas

ggplot(dat, aes(x=Obs, y= Valor, group=Clase, colour=Clase)) +
facet_wrap(~Experimento, labeller=as_labeller(exp_names)) +
geom_point() +
geom_hline(yintercept=1, linetype="dashed") +
geom_hline(yintercept= 3, linetype="dashed") +
xlab("Observaciones") +
ylab("") +
theme(legend.position="none") +
scale_color_manual(values=c("royalblue", "indianred3"))
# Produce un diagrama

```

Código 2.17: Generación de un gráfico de control univariado

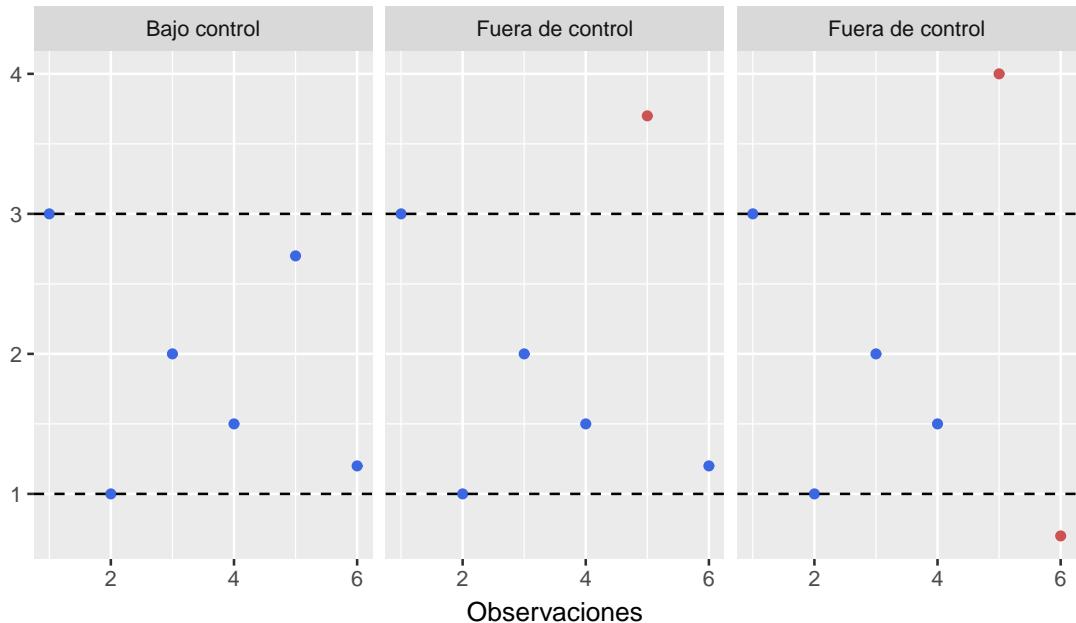


Figura 2.33: Control univariado

En la Figura 2.33 podemos apreciar si el dato excede o está por debajo de las especificaciones, pero no podremos apreciar si la forma es la adecuada o no.

El *scatter plot* o gráfico de dispersión (ver Figura 2.34 y Código 2.18), nos permite identificar variables que siguen el patrón general de interacción pero se alejan del centro de las variables. Asimismo permite identificar puntos que están dentro del rango de ambas variables pero la forma de su interacción no es la forma general del grupo.

```
library(MASS)
# Paquete con funciones y bases de datos para la librería de Venables y Ripley

dat=mvrnorm(n=60,c(10,5), cbind(c(0.7,0.5),c(0.5,0.4)), tol=1e-6,
empirical=FALSE, EISPACK=FALSE)
# Genera los datos
datos=data.frame(dat)
# Arregla los datos

ggplot(datos, aes(x=X1,y=X2)) +
geom_point(colour="royalblue") +
geom_point(aes(x=11.6,y=3.3), colour="indianred3") +
stat_ellipse(aes(x=X1, y=X2), colour="orchid3", type="norm") +
geom_hline(yintercept=3, linetype="dashed", colour="forestgreen") +
geom_hline(yintercept=7, linetype="dashed", colour="forestgreen") +
geom_vline(xintercept=8, linetype="dashed", colour="forestgreen") +
geom_vline(xintercept=12, linetype="dashed", colour="forestgreen") +
xlab("") +
ylab("")
# Produce un diagrama
```

Código 2.18: Generación de un gráfico de control multivariado

Nos preguntamos ahora, ¿qué podemos observar en el dispersograma 2.25?

- ✿ Cuáles variables parecen asociadas.
- ✿ Cuáles variables no parecen asociadas.
- ✿ Qué sentido se le encuentra a dichas asociaciones.
- ✿ Qué fuerza se le encuentra a dichas asociaciones.

Sin embargo, deberíamos encontrar un modo de cuantificar estas apreciaciones, siendo la covarianza muestral una forma posible.

## 2.6.1 Covarianza y Correlación

### Covarianza muestral

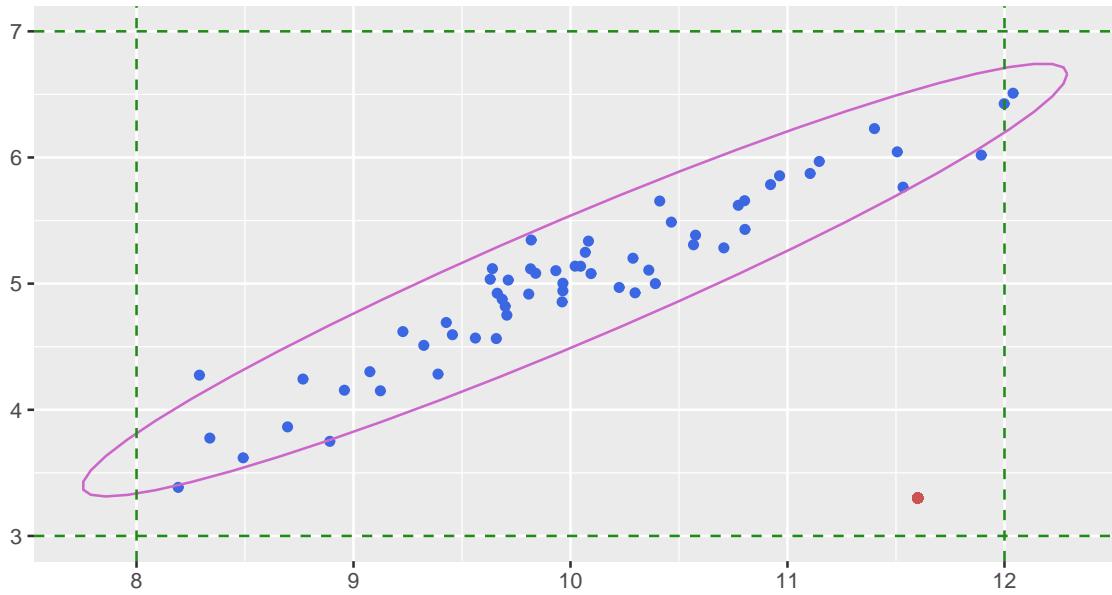


Figura 2.34: Control multivariado

Es una medida de asociación lineal entre dos variables. Se calcula sobre el conjunto de observaciones  $x_{ij}$ , mediante la siguiente fórmula:

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

la matriz de varianzas y covarianzas es de la forma

$$\Sigma = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

donde

- ✿  $s_{ik} > 0$  indica una asociación lineal positiva entre los datos de las variables.
- ✿  $s_{ik} < 0$  indica una asociación lineal negativa entre los datos de las variables.
- ✿  $s_{ik} = 0$  indica que no hay una asociación lineal entre los datos de las variables.

### Propiedades destacables de la covarianza

- ✿  $Cov(X, X) = Var(X)$ .
- ✿  $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$ .
- ✿ Dados vectores aleatorios  $X$  e  $Y$  y matrices de constantes  $A$  y  $B$ , vale que  $Cov(AX, BY) = ACov(X, Y)B^t$ .
- ✿ La covarianza sólo detecta asociación lineal, mientras que otros tipos de asociación no son captadas por esta medida.

**Ejemplo 2.14.** Sean  $X$  e  $Y$  dos variables aleatorias tales que  $\mu_X = 4$ ,  $\sigma_X^2 = 2$ , siendo  $Y = -2X + 3$ . Utilizando propiedades de la varianza y de la esperanza matemática, tenemos que:

$$\mu_Y = -2 \cdot 4 + 3 = -5, \quad \sigma_Y^2 = 4 \cdot 2 = 8 \quad \text{y}$$

$$Cov(X, Y) = Cov(X, -2X + 3) = -2Cov(X, X) = -2Var(X) = -2 \cdot 2 = -4$$

Luego, si consideramos el vector aleatorio  $(X, Y)$ , por lo visto, la matriz de covarianzas está dada por

$$\Sigma = \begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}$$

Es inmediato observar que el determinante de esta matriz es nulo; es decir, esta matriz es **singular**.

¿Por qué sucede esto?

Debido a que una de las variables es función lineal de la otra, el conjunto formado por ambas resulta linealmente dependiente y, por lo tanto, el determinante es nulo.



El valor (magnitud) de la covarianza depende las unidades en que se miden las variables. Este es un defecto que puede salvarse realizando una estandarización. De este modo se obtiene una medida de la fuerza de la relación que no depende de las unidades de medición.

$$Cov(aX + b, cY + d) = acCov(X, Y) \quad \forall a, b, c, d \in \mathbb{R}$$

**Observación:** La varianza muestral es la covarianza muestral entre los datos de la  $i$ -ésima variable con ella misma, algunas veces se denota como  $s_{ii}$

## Correlación muestral

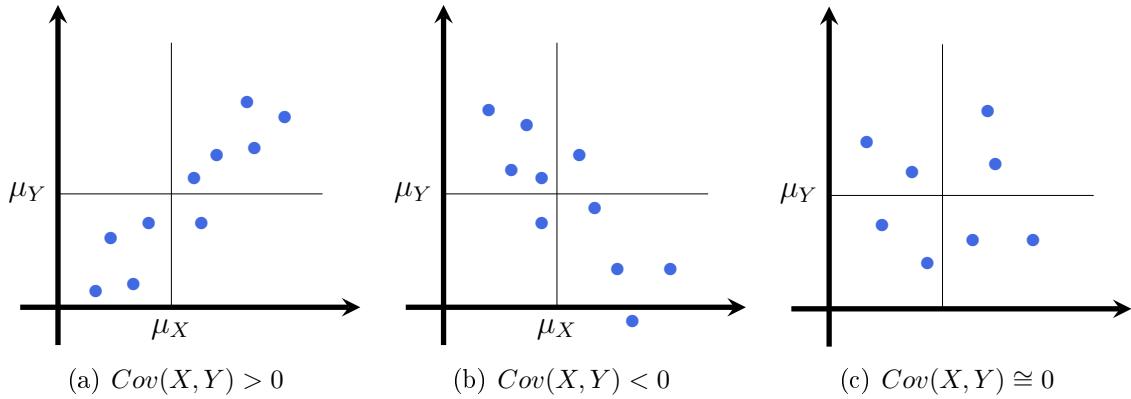


Figura 2.35: Signo de la covarianza

Considerando las variables estandarizadas con la ecuación 2.1, el coeficiente de correlación lineal es una medida de asociación lineal para las variables, definida como la covarianza de los datos estandarizados. Para los datos de la  $i$ -ésima y  $k$ -ésima variable se define como

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

La matriz de correlación muestral es de la forma

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{pmatrix}$$

Entonces,  $r_{jk}$  es la correlación muestral entre  $Z_j$  y  $Z_k$ , columnas  $j$  y  $k$  de las variables estandarizadas.

Tanto  $s_{ik}$  como  $r_{ik}$  son muy sensibles a la presencia de datos atípicos (*outliers*). En presencia de datos atípicos será recomendable utilizar otras medidas de asociación.

### Propiedades de la correlación muestral

- ✿  $|r_{ik}| \leq 1$ .
- ✿ Si  $r_{ik} = 1$  significa que los datos yacen sobre una línea recta de pendiente positiva.
- ✿ Si  $r_{ik} = -1$  significa que los datos yacen sobre una línea recta de pendiente negativa.
- ✿ Si  $0 < r_{ik} < 1$  significa que los datos se ubican alrededor de una línea recta de pendiente positiva.

- Si  $-1 < r_{ik} < 0$  significa que los datos se ubican alrededor de una línea recta de pendiente negativa.
- Si  $r_{ik} = 0$  indica que no hay una asociación lineal entre las dos variables.

### Traza de una matriz

Llamamos **traza** de una matriz cuadrada a la suma de los elementos de la diagonal principal.

Simbólicamente, si  $A \in \mathbb{R}^{n \times n}$ ,  $tr(A) = \sum_{i=1}^n a_{ii}$ .

Siempre es posible calcular la traza de una matriz cuadrada. La traza es un número real, puede ser positivo, negativo o nulo. En el caso de las matrices de varianzas y covarianzas, como en el caso de las matrices de correlación, la traza es positiva.

**Ejemplo 2.15.** Si  $A = \begin{pmatrix} 2 & 3 \\ -4 & 8 \end{pmatrix}$ , entonces  $tr(A) = 2 + 8 = 10$ . ■

### Traza de la matriz de varianzas y covarianzas

Debido a que en una matriz de covarianzas, la diagonal principal está constituida por las varianzas de las variables, que son valores mayores o iguales a cero, la traza de la misma es no negativa. En este caso, la traza es la suma de las varianzas de las variables consideradas en el conjunto de datos por lo cual indica de alguna forma la magnitud del problema.

### Traza de la matriz de correlaciones

En el caso de la matriz de correlaciones, la diagonal principal está constituida por unos, que representan las correlaciones de cada variable consigo misma. En este caso, la traza es igual a la cantidad de variables involucradas en el problema. En el Ejemplo 2.15,  $tr(Corr(A)) = 1 + 1 = 2$  variables.

Retomando el Ejemplo 2.14, la matriz de covarianzas es

$$\Sigma = \begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}$$

y la matriz de correlación es

$$Corr = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Luego,  $tr(\Sigma) = 2 + 8 = 10$  y  $tr(Corr) = 1 + 1 = 2$ .

### Correlogramas

Nos permiten visualizar la fuerza y el sentido de la correlación entre un conjunto de variables.

Con los datos disponibles en <https://goo.gl/Dpnx9Z>, y mediante el código 2.19, se genera la Figura 2.36.

- ✿ El color azul indica correlación positiva.
- ✿ El color rojo indica correlación negativa.
- ✿ Cuanto mayor es la intensidad del color más cercano a 1 en el caso positivo y a  $-1$  en el caso negativo se encuentra el coeficiente de correlación.

```
library(corrplot) # Paquete para representaciones gráficas de matrices
library(readxl) # Permite leer archivos xlsx

IMCinfantil=read_excel("C:/.../IMCinfantil.xlsx")
# Importa la base con la cual se va a trabajar
attach(IMCinfantil) # Se pone la base en la memoria

base.niños=data.frame(EDAD,PESO,TALLA,IMC,CC)
# Arma una sub-base con las variables numéricas de IMCinfantil
base.niños$CC=max(base.niños$CC)-base.niños$CC
# Cambia la variable para que correlacione en forma negativa con las restantes
M=cor(base.niños) # Calcula la matriz de correlación
corrplot.mixed(M, lower="number", upper="shade", addshade="all")
# Produce un correlograma
```

Código 2.19: Generación de un correlograma

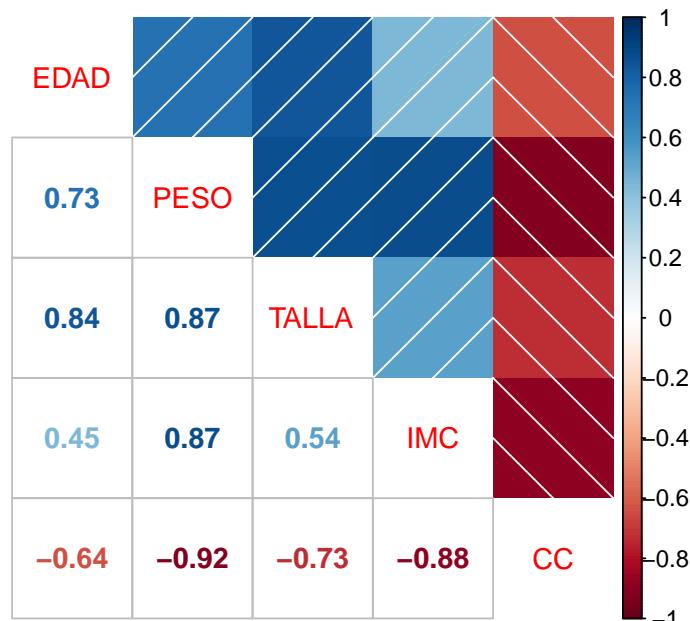


Figura 2.36: Correlograma

En la Figura 2.36 se puede apreciar lo siguiente:

- ✿ Las variables EDAD, PESO, TALLA e IMC correlacionan positivamente entre sí.
- ✿ Todas las variables correlacionan negativamente con CC (que es una modificación de la variable original para lograr correlación negativa).
- ✿ Es más intensa la correlación entre PESO y TALLA que entre EDAD y PESO.
- ✿ Es más intensa la correlación entre IMC y CC que entre EDAD y CC.

## 2.7 Alternativas robustas para posición y escala

Las estadísticas robustas proponen métodos similares a los de la estadística clásica, pero que no se vean afectados por la presencia de observaciones atípicas (*outliers* en inglés) u otras desviaciones de los supuestos de un modelo.

Por lo general las observaciones atípicas en bases grandes de datos no pueden ser eficientemente detectadas analizando por separado cada variable. La detección resulta más eficiente estudiando el conjunto general de todas las variables.

Los *outliers*, en casos multivariados, pueden provocar dos tipos de efectos:

- ✿ El **efecto de enmascaramiento** se produce cuando un grupo de *outliers* esconden a otro/s. Es decir, los *outliers* enmascarados se harán visibles cuando se elimine/n el o los *outliers* que los esconden.
- ✿ El **efecto de inundación** ocurre cuando una observación sólo es *outlier* en presencia de otra/s observación/es. Si se quitara/n la/s última/s, la primera dejaría de ser *outlier*.

### Distancia de Mahalanobis

Este concepto fue introducido por Mahalanobis [13] y se diferencia de la distancia euclídea pues considera la correlación entre las variables. Esta distancia es muy usada en Estadística Multivariada.

Precisamente, sean  $X$  e  $Y$  dos variables aleatorias pensadas como vectores columna y con la misma distribución de probabilidad. Si  $\Sigma$  es la matriz de covarianzas, se define la **distancia de Mahalanobis** como

$$d_m(X, Y) = \sqrt{(X - Y)^t \Sigma^{-1} (X - Y)}$$

### Vector de medianas

En [20] los autores proponen sustituir el vector de medias por un vector de medianas y calcular la matriz de covarianza para el conjunto de las  $k$  observaciones con menor distancia de Mahalanobis al vector de medianas.

Realizar una estimación robusta de la matriz de covarianzas puede entenderse como estimar la covarianza de una buena parte de los datos.

### MVE (Minimum Volume Ellipsoid) (Elipsoide de volumen mínimo)

Este estimador se basa en la idea de buscar el elipsoide de menor volumen que cubra  $m$  de las  $n$  observaciones. Puede ser calculado mediante un algoritmo de remuestreo [22].

Se ha demostrado que este estimador es eficiente, equivariante por transformaciones afines y tiene un alto punto de ruptura. Esto lo convierte en un estimador **robusto** de posición y escala para datos multivariados.

Dado que se trata de un estimador de bajo sesgo; es decir, que la diferencia entre la estimación y el valor real del parámetro de interés es pequeña, resulta una buena estrategia para la detección de valores atípicos multivariados [3].

### MCD (Minimum Covariance Determinant) (Determinante de mínima covarianza)

El objetivo a minimizar en este caso es el determinante de la matriz de covarianzas de  $m$  observaciones de las  $n$  disponibles. Este estimador de posición y dispersión multivariado robusto puede calcularse de manera eficiente con el algoritmo de FAST-MCD propuesto por Rousseeuw y Van Driessen [19].

Puesto que la estimación de la matriz de covarianza es la base de muchos métodos estadísticos multivariados, esta propuesta fue utilizada para desarrollar técnicas robustas multivariadas.

**Ejemplo 2.16.** En el Código 2.20 se muestra cómo calcular los valores de los conceptos previamente definidos utilizando el archivo `stack.x` de R.

```
library(MASS)
# Paquete con funciones y bases de datos para la librería de Venables y Ripley
library(lattice) # Paquete para visualizar datos
library(grid) # Paquete con un sistema para gráficos
library(DMwR) # Paquete con funciones para data mining

cov1=cov.rob(stack.x, method="mcd", nsamp="exact") # Calcula MCD
cov2=cov.rob(stack.x, method="mve", nsamp="best") # Calcula MVE
cov3=cov.rob(stack.x, method="classical", nsamp="best")
# Calcula la matriz de covarianzas clásica
center1=apply(stack.x, 2, mean) # Calcula el vector de medias
center2=apply(stack.x, 2, median) # Calcula el vector de medianas

dcov1=0 ; dcov2=0 ; dcov3=0 # Inicializaciones

for(i in 1:21){
dcov1[i]=mahalanobis(stack.x[i,], cov1$center, cov1$cov, inverted = FALSE)
dcov2[i]=mahalanobis(stack.x[i,], cov2$center, cov2$cov, inverted = FALSE)
dcov3[i]=mahalanobis(stack.x[i,], cov3$center, cov3$cov, inverted = FALSE)
}
# Calcula distancias de Mahalanobis utilizando las distintas estimaciones
# de la matriz de covarianzas
round(cbind(dcov1,dcov2,dcov3),2)
# Combina las tres distancias para observar el resultado
```

```

distancias.outliers=lofactor(stack.x, k=5)
# Calcula las distancias teniendo en cuenta cinco vecinos

plot(density(distancias.outliers), col="royalblue", main="",
xlab="n=21, ancho de banda = 0.06518", ylab="Densidad")
# Dibuja la densidad estimada de las distancias de Mahalanobis de las
# observaciones

outliers=order(distancias.outliers, decreasing=T)[1:5]
# Arroja las observaciones correspondientes a las cinco distancias mayores
print(outliers)

```

Código 2.20: Cálculo en estadística robusta

La Figura 2.37 muestra la densidad estimada de las distancias de Mahalanobis de las observaciones realizadas.

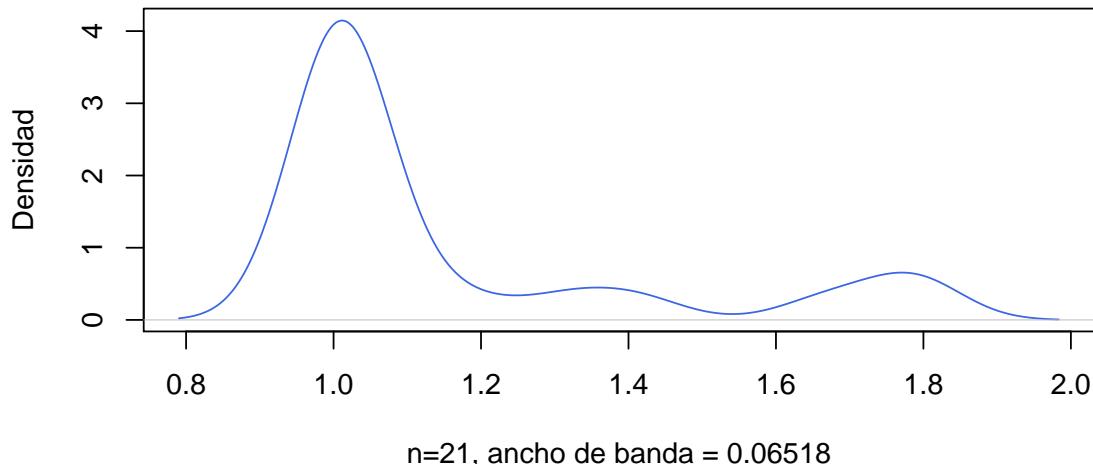


Figura 2.37: Detección multivariada de *outliers*

En la Tabla 2.11 se exhiben las distancias calculadas teniendo en cuenta cinco vecinos, mientras que en la Tabla 2.12 se muestran las distancias de Mahalanobis utilizando las distintas estimaciones propuestas para la matriz de covarianzas. Se marcaron en negrita los *outliers* encontrados teniendo en cuenta cinco vecinos.

1.785212	1.788115	1.670663	0.988912	0.986892	0.985768	1.006635
1.006635	0.986567	0.990893	1.019990	1.021364	1.028066	1.025351
1.169776	1.032410	1.314515	1.052058	1.048615	0.991374	1.410668

Tabla 2.11: Distancias entre *outliers*

Observación	MVE	MCD	MCov
1	<b>30.56</b>	<b>30.56</b>	<b>5.08</b>
2	<b>31.78</b>	<b>31.78</b>	<b>5.4</b>
3	<b>17.62</b>	<b>17.62</b>	<b>2.54</b>
4	2.52	2.52	1.62
5	1.41	1.41	0.09
6	1.71	1.71	0.6
7	2.94	2.94	3.43
8	2.94	2.94	3.43
9	1.5	1.5	1.85
10	3.75	3.75	3.05
11	2.23	2.23	2.15
12	3.66	3.66	3.39
13	2.76	2.76	2.2
14	2.85	2.85	3.16
15	4.97	4.97	2.86
16	3.12	3.12	1.67
17	<b>5.91</b>	<b>5.91</b>	<b>7.29</b>
18	2.32	2.32	2.26
19	2.92	2.92	2.54
20	0.46	0.46	0.65
21	<b>13.38</b>	<b>13.38</b>	<b>4.74</b>

Tabla 2.12: Distancias de Mahalanobis

## 2.8 Ejercitación

### Ejercicio 1. Transformaciones de datos

Seis candidatas son evaluadas para el puesto de recepcionista en una empresa, para lo cual se las somete a dos entrevistas. En la primera de ellas, son evaluadas por el responsable del Departamento de Recursos Humanos de la empresa, al cual denominaremos Juez 1, mientras que en la segunda son evaluadas por el responsable del área de la cual van a depender, que llamaremos Juez 2. La asignación de puntajes se basa en los siguientes tópicos: cordialidad, presencia y manejo de idiomas. Los puntajes asignados independientemente por estos jueces se encuentran en la Tabla 2.13.

Candidatas	Juez 1			Juez 2		
	Cordialidad	Presencia	Idioma	Cordialidad	Presencia	Idioma
Mariana	80	90	70	60	78	80
Maia	80	90	60	65	90	65
Sabrina	90	60	50	70	60	50
Daniela	80	50	50	70	58	40
Alejandra	70	60	50	55	70	65
Carla	90	85	60	80	90	40

Tabla 2.13: Datos candidatas a recepcionistas

1. Calcular el promedio por juez de cada una de las aspirantes. ¿Cuál de ellas seleccionaría cada uno de los jueces? ¿Existe coincidencia?
2. Calcular el promedio de cada una de las aspirantes tomando en cuenta todos los aspectos evaluados y ambos jueces.
3. Transformar las puntuaciones observadas de modo tal que cada una de las seis variables tenga media 0 y dispersión 1. ¿Cuál es el objetivo de esta transformación?
4. Transformar las puntuaciones de modo tal que cada candidata tenga para cada juez media 0 y dispersión 1. ¿Cuál es el objetivo de esta transformación?
5. Graficar los perfiles multivariados de cada una de las candidatas para ambas transformaciones. ¿Qué puede observarse?

### Ejercicio 2. Tipos de variables resúmenes

Se han registrado sobre 1500 individuos (ver <https://goo.gl/ZcakZq>), las siguientes variables:

**ID:** número de identificación del registro de datos

**Nac.:** indica la nacionalidad que puede ser Argentina, Brasilera, Canadiense, Uruguaya

**Edad:** cumplida en años

**Sexo:** Masculino (1) y Femenino (2)

**Estatura:** en metros

**Interés:** de conexión, siendo chat (1), correo electrónico (2), buscadores (3), software (4), música (5), deportes (6) y otros (7)

**Tiempo:** tiempo promedio de uso promedio por día en minutos

**Temp.:** temperatura media anual de la zona de residencia

**Autos:** cantidad de autos en la manzana de residencia

**Cig.:** cantidad de cigarrillos consumida mientras se utiliza *Internet*

1. Clasificar las variables de la base de datos y, para las que sean numéricas, construir un gráfico de coordenadas paralelas.
2. Construir la tabla de frecuencias de la variable Sexo. ¿Hay algún valor que pueda llamar la atención? ¿Qué tipo de error podría ser?
3. Ordenar los datos por la variable Edad. ¿Se encuentra algún valor extraño? ¿Qué tipo de error podría ser?
4. Construir la tabla de frecuencias de la variable Interés. ¿Se encuentra algún valor que pueda llamar la atención? ¿Qué tipo de error podría ser?
5. Proceder de forma similar para las variables Temperatura, Autos y Cigarrillos.
6. Eliminar de la base de datos aquellos valores que no son posibles y que probablemente corresponden a un error de tipeo. Detallar valores o registros que llamen la atención pero que no deban ser eliminados necesariamente.
7. ¿Para cuáles de las variables tiene sentido calcular la media? ¿Y la mediana?
8. ¿Cuáles de las variables parecerían simétricas a partir de estos resúmenes? Confirmar estas observaciones mediante un *boxplot*.
9. Calcular la desviación intercuartil y detectar presencia de valores salvajes moderados y severos.

### Ejercicio 3. Gráficos univariados y multivariados

En la base de datos que se puede encontrar en <https://goo.gl/FVqX22>, se han registrado para 49 gorriones las siguientes variables zoo métricas:

**Largo:** medida del largo total del ave

**Alas:** extensión alar del ave

**Cabeza** medida del largo del pico y la cabeza del ave

**Pata:** medida del largo del húmero del ave

**Cuerpo:** medida del largo de la quilla del esternón del ave

**Sobrevida:** indicando por 1 si el ave está viva y por -1 si no lo está

1. Indicar en cada caso de qué tipo de variable se trata.
2. Confeccionar un informe univariado para cada variable.
3. Realizar un histograma, en el caso en que corresponda, ensayando el número de intervalos que conviene utilizar en cada variable e indicando si se basa en algún criterio.
4. Realizar un *boxplot* comparativo para cada una de estas variables, particionando por el grupo definido por la supervivencia del ave. ¿Podría ser que alguna de estas variables estuviera relacionada con la supervivencia; es decir, que tomara valores muy distintos en ambos grupos? Analizar en todos los casos la presencia de *outliers*.
5. Construir gráficos bivariados para todas las variables en cuestión, particionando por el grupo de supervivencia y considerando un color para cada grupo. ¿Se observa alguna regularidad que pueda explicar la supervivencia?
6. Construir la matriz de diagramas de dispersión. ¿Podría considerarse que algún par de estas medidas están relacionadas? Estudiar si la asociación de algunas de estas medidas es diferente en alguno de los grupos.

### Ejercicio 4.

Se han registrado, respecto de 26 razas de perros, las siguientes características sobre base de datos que se encuentra disponible en <https://goo.gl/eNJ8GU>:

**Raza:** nombre de la raza del perro

**Tamaño:** con los niveles pequeño (1), mediano (2) y grande (3)

**Peso:** con los niveles liviano (1), medio (2) y pesado (3)

**Velocidad:** con los niveles lento (1), mediano (2) y rápido (3)

**Inteligencia:** con los niveles alta (1), media (2) y baja (3)

**Afectividad:** con los niveles alta (1), media (2) y baja (3)

**Agresividad:** con los niveles alta (1), media (2) y baja (3)

**Función:** con las categorías caza, utilitario y compañía.

1. Realizar un gráfico de estrellas por raza y utilizando las variables tamaño, peso, velocidad, inteligencia y afectividad.
2. Idem al inciso anterior por función.
3. Idem al primer inciso por agresividad.
4. En el primer gráfico se observan estrellas similares. ¿Podría decirse que las razas en cuestión son parecidas?

#### Ejercicio 5. Matriz de covarianzas

Para la base de datos disponible en <https://goo.gl/FVqX22>, se piden los siguientes puntos.

1. Calcular la dimensión de la base de datos notando por  $n$  al número de observaciones y por  $p$  a la cantidad de variables observadas sobre cada individuo.
2. Hallar el vector de medias, la matriz de varianzas y covarianzas y la matriz de correlaciones. ¿Qué características tienen estas matrices?
3. Explicar qué representan los elementos  $m_{11}$  y  $m_{31}$  de la matriz de varianzas y covarianzas.
4. Explicar qué representa los elementos  $m_{22}$  y  $m_{13}$  de la matriz de correlaciones.
5. Relacionar los elementos  $m_{21}$ ,  $m_{11}$  y  $m_{22}$  de la matriz de varianzas y covarianzas con el elemento  $m_{12}$  de la matriz de correlaciones.
6. Hallar una nueva variable e incorporarla en la base de datos, llamada **Diferencia** y que mida la diferencia entre el largo total y el largo del húmero.
7. Calcular el vector de medias y las matrices de varianzas y covarianzas y la matriz de correlaciones de la nueva base de datos, relacionando el nuevo vector de medias con el anterior.

8. Hallar la traza de las cuatro matrices calculadas, explicando el significado de cada uno de los resultados obtenidos. ¿Qué trazas no aumentan al agregar una variable? Explicar.

**Ejercicio 6.** Propiedades de la matriz de covarianzas

Para los datos de la Tabla 2.13 se pide:

1. Calcular el vector de medias e interpretar los valores.
2. Hallar las matrices de varianzas y covarianzas y de correlaciones para la submatriz de puntuaciones del primer juez y del segundo juez por separado. Repetir para el conjunto total.
3. ¿Se puede decir que la suma de las dos primeras submatrices da como resultado la matriz del grupo total? De no ser así, explicar el motivo.
4. ¿Se cumple esta relación para las trazas, para el vector de medias y para los vectores de medianas?

**Ejercicio 7.** Medidas de posición y escala robustas

Con los datos disponibles en <https://goo.gl/ZcakZq>,

1. Seleccionar las variables numéricas y agregar 5 observaciones que no sean atípicas en forma univariada pero que sí lo sean en forma multivariada. Utilizar las medidas robustas para detectar estos valores.
2. Agregar cuatro observaciones que sean *outliers* pero que aparezcan enmascaradas. Utilizar estrategias robustas para detectar su presencia.

# Capítulo 3

## Análisis de componentes principales

*Life is the art of drawing  
sufficient conclusions from  
insufficient premises.*

— Samuel Butler

### 3.1 Nociones Previas

En el Apéndice ?? se presentan conceptos elementales de Álgebra Lineal y que serán utilizados a lo largo de este capítulo. Para mayores detalles de álgebra lineal ver entre otros [10, 4]).

Para introducirnos en el concepto de vector en  $\mathbb{R}^n$ , consideramos un punto de coordenadas  $P = (v_1, v_2, \dots, v_n)$ , el vector  $v = (v_1, v_2, \dots, v_n)$  es el segmento rectilíneo orientado cuyo punto inicial es el origen de coordenadas de  $\mathbb{R}^n$  y cuyo punto final es el punto  $P$ . A modo de ejemplo, mostramos en la Figura 3.1 los vectores  $u = (3, 0)$ ,  $v = (1, 3)$  y  $w = (-2, 1)$  en  $\mathbb{R}^2$ .

Consideremos el espacio vectorial  $\mathbb{R}^n$ . Si  $\alpha \in \mathbb{R}$ ,  $v = (v_1, v_2, \dots, v_n)$  y  $w = (w_1, w_2, \dots, w_n)$  son dos vectores en  $\mathbb{R}^n$ , las operaciones del espacio vectorial son la suma y el producto por un escalar definidas como

- \*  $v + w = (v_1 + w_1, v_2 + w_2, \dots, v_n + w_n)$
- \*  $\alpha v = (\alpha v_1, \alpha v_2, \dots, \alpha v_n)$

Para elementos de un espacio vectorial  $\mathbb{V}$ , decimos que el vector  $u$  es **combinación lineal** de los vectores  $v$  y  $w$  cuando existen escalares  $\alpha$  y  $\beta$  tales que  $u = \alpha v + \beta w$ .

Notemos que el vector nulo siempre se puede obtener como combinación lineal de cualquier conjunto de vectores tomando todos los escalares iguales a cero.

Analicemos geométricamente qué significa una combinación lineal. Para ello consideremos como espacio vectorial de referencia  $\mathbb{V}$  a  $\mathbb{R}^2$  o a  $\mathbb{R}^3$ .

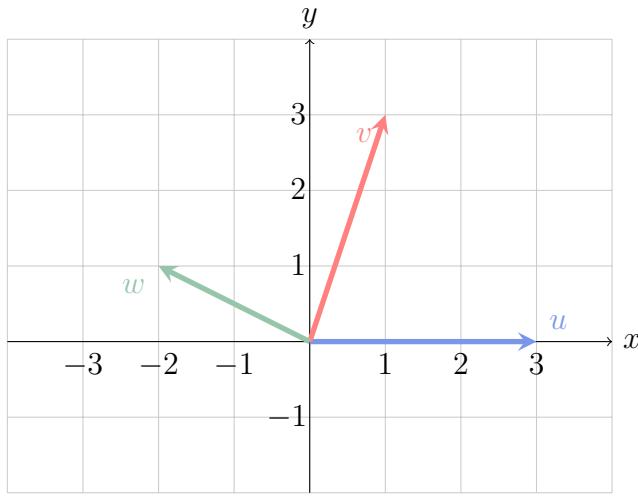


Figura 3.1: Vectores en coordenadas

- ✿ Las posibles combinaciones lineales de un único vector son los múltiplos de este vector; es decir, comparten la misma dirección pues pertenecen a una misma recta.
- ✿ Si dos vectores no nulos en el plano  $\mathbb{R}^2$  no tienen la misma dirección, cualquier otro vector del plano puede escribirse como combinación lineal de ellos.

Un concepto fundamental es el de **dependencia lineal**.

Sea un conjunto de vectores  $v_1, v_2, \dots, v_n$  en un espacio vectorial  $\mathbb{V}$ . Se dice que los vectores  $v_1, v_2, \dots, v_n$  son **linealmente dependientes (l.d.)** si el vector nulo puede escribirse como una combinación lineal de elementos de este conjunto con al menos un escalar distinto de cero. Simbólicamente: existe  $\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n = 0$  con algún  $\alpha_i \neq 0$ . En caso contrario, se dice que los vectores  $v_1, v_2, \dots, v_n$  son **linealmente independientes (l.i.)**; en este caso la única combinación lineal que da el vector nulo tiene todos los escalares iguales a cero. Simbólicamente,  $\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n = 0$  implica que  $\alpha_1 = \alpha_2 = \dots = \alpha_n = 0$ .

Pensando este concepto desde un enfoque estadístico, dos vectores (variables) son l.d. cuando la información que brinda uno de ellos es redundante con la información que brinda el otro. En este caso, se puede ver que uno de ellos es múltiplo del otro. Por ejemplo, la estatura medida en centímetros es l.d. con la medida en metros.

El hecho de que tres vectores (variables) sean l.d. estadísticamente significa que la información de una de las variables es una combinación lineal de la información de las otras dos.

En la Figura 3.2 se exhiben estos conceptos desde un punto de vista gráfico. En el primer caso se representan los vectores  $v = (a, a)$  y  $-v = (-a, -a)$  que son l.d., pues su suma es una combinación lineal que da por resultado el vector nulo siendo ambos escalares iguales a uno.

Sea  $T = \{v_1, v_2, \dots, v_n\} \subseteq \mathbb{V}$  un conjunto de vectores de un espacio vectorial  $\mathbb{V}$ .

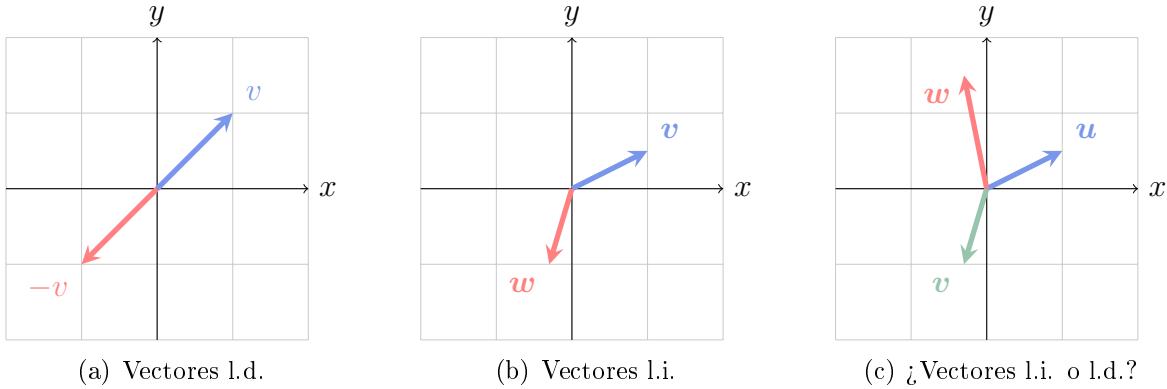


Figura 3.2: Dependencia lineal entre vectores

Al conjunto de todos los vectores que pueden expresarse como combinación lineal de elementos de  $T$  se lo denomina **espacio generado** por  $T$  y se lo denota como  $\text{gen}(T) = \langle T \rangle$ .

El espacio generado por un único vector; es decir, todos los vectores que son múltiplos del mismo, es una recta. Cuando dos vectores no pertenecen a una misma recta, el espacio generado por ellos es un plano. Recíprocamente, todo vector que pertenece a ese plano resulta combinación lineal de los dos vectores que podemos denominar generadores.

**Ejemplo 3.1.** En la Tabla 3.1 se muestra la base de datos de los tiempos empleados por catorce nadadores en cada uno de los cuatro tramos de una competencia.



<https://flic.kr/p/cpmtS5>

Podríamos estar interesados, por ejemplo, en:

- \* el tiempo medio empleado por cada nadador en los primeros dos tramos,
- \* el tiempo medio empleado por cada nadador en los últimos dos tramos,

\* la diferencia entre los dos promedios anteriores.

:

Designemos con  $v_i$  al vector de tiempos empleados por los nadadores durante el tramo  $i$ -ésimo para  $i = 1, 2, 3, 4$ . De este modo las combinaciones lineales de interés son:

$$* w_1 = \frac{1}{2}v_1 + \frac{1}{2}v_2$$

$$* w_1 = \frac{1}{2}v_3 + \frac{1}{2}v_4$$

$$* w_3 = w_1 - w_2$$

Agregamos a la Tabla 3.1 la información correspondiente a cada una de las combinaciones lineales definidas sobre las variables originales, lo presentamos en la Tabla 3.2.

Nadador	Tramo 1	Tramo 2	Tramo 3	Tramo 4
1	10	10	13	12
2	12	12	14	15
3	11	10	14	13
4	9	9	11	11
5	8	8	9	8
6	8	9	10	9
7	10	10	8	9
8	11	12	10	9
9	14	13	11	11
10	12	12	12	10
11	13	13	11	11
12	14	15	14	13
13	10	10	12	13
14	15	14	13	14

Tabla 3.1: Tiempos por tramos en competencia de natación

En la Tabla 3.2 podemos apreciar cuáles nadadores tardaron más en promedio en los dos primeros tramos que en los dos segundos, de igual manera podemos determinar cuán grande es esta diferencia a favor o en contra.

Notemos que, las nuevas variables  $w_1$ ,  $w_2$  y  $w_3$  pertenecen al espacio generado por las primeras cuatro variables.



Nadador	Tramo 1	Tramo 2	Tramo 3	Tramo 4	$w_1$	$w_2$	$w_3$
1	10	10	13	12	10.0	12.5	-2.5
2	12	12	14	15	12.0	14.5	-2.5
3	11	10	14	13	10.5	13.5	-3.0
4	9	9	11	11	9.0	11.0	-2.0
5	8	8	9	8	8.0	8.5	-0.5
6	8	9	10	9	8.5	9.5	-1.0
7	10	10	8	9	10.0	8.5	1.5
8	11	12	10	9	11.5	9.5	2.0
9	14	13	11	11	13.5	11.0	2.5
10	12	12	12	10	12.0	11.0	1.0
11	13	13	11	11	13.0	11.0	2.0
12	14	15	14	13	14.5	13.5	1.0
13	10	10	12	13	10.0	12.5	-2.5
14	15	14	13	14	14.5	13.5	1.0

Tabla 3.2: Tiempos por tramos en competencia de natación ampliada

Sea  $B = \{v_1, v_2, \dots, v_n\}$  un conjunto de vectores de un espacio vectorial  $\mathbb{W}$ . Se dice que  $B$  es una **base** de  $\mathbb{W}$  si los vectores de  $B$  son linealmente independientes y además generan a  $\mathbb{W}$ ,  $gen(B) = \mathbb{W}$ .

Todo espacio vectorial admite infinitas bases pero se puede probar que todas esas bases poseen la misma cantidad de elementos. A dicha cantidad de elementos se la denomina **dimensión** del espacio vectorial. Con la notación anterior,  $dim(\mathbb{W}) = n$ . De esta manera, una recta se genera con un único vector no nulo y por ende es un espacio unidimensional; es decir, de dimensión 1. El espacio generado por dos vectores linealmente independientes (un plano en  $\mathbb{R}^3$  o el plano coordenado  $\mathbb{R}^2$ ) es un espacio bidimensional o de dimensión 2.

Para el espacio  $\mathbb{R}^n$  se conoce como **base canónica** al conjunto de  $n$  vectores donde cada vector tiene un 1 en la coordenada  $i$ -ésima coordena y 0 en las restantes. En la Figura 3.4 se muestran dos ejemplos de bases distintas para el espacio  $\mathbb{R}^2$ . La primera es la base canónica y sus vectores se simbolizan  $E = \{e_1, e_2\}$ , donde  $e_1 = (1, 0)$  y  $e_2 = (0, 1)$ . se nota  $E = \{e_1, e_2, \dots, e_n\}$ , donde  $e_i$  es el  $i$ -ésimo vector canónico.

## 3.2 Transformaciones

Frecuentemente, en el análisis multivariado es conveniente transformar un espacio vectorial dado en otro de distinta dimensión. Para ello se aplican diversos tipos de transformaciones que, en general, son transformaciones lineales. Dentro de este conjunto de transformaciones lineales, las más usuales

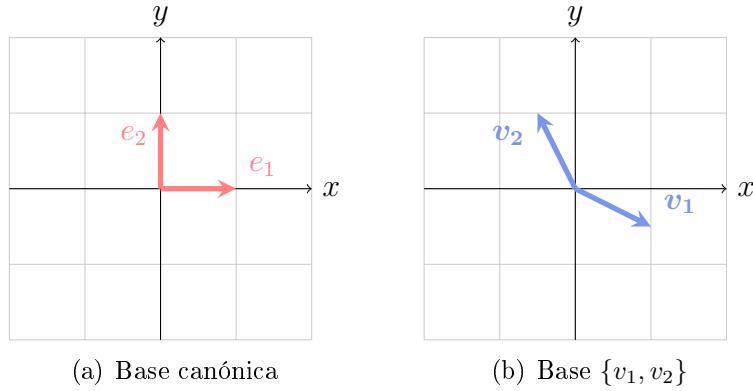


Figura 3.4: Bases para  $\mathbb{R}^2$

son las **proyecciones** y las **rotaciones**.

### Ejemplo 3.2. Proyecciones

Dado un conjunto de datos como los representados en la Figura 3.5, podría interesarnos encontrar una proyección que maximice la sombra o una proyección que discrimine mejor los colores proyectados, claramente no se trata de la misma proyección.

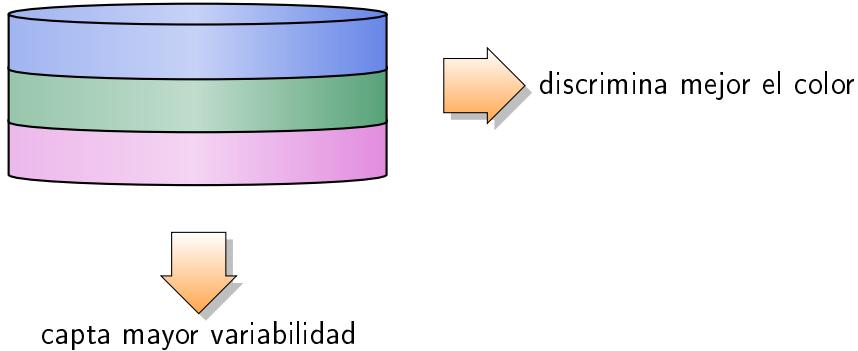


Figura 3.5: Modelo de datos a proyectar

### Ejemplo 3.3. Simetrías

Consideremos la siguiente transformación del plano en sí mismo,  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  definida por

$$T(x, y) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ -y \end{pmatrix}$$

En la Figura 3.6 se exhibe el efecto que tiene esta trasformación sobre un triángulo en el plano.

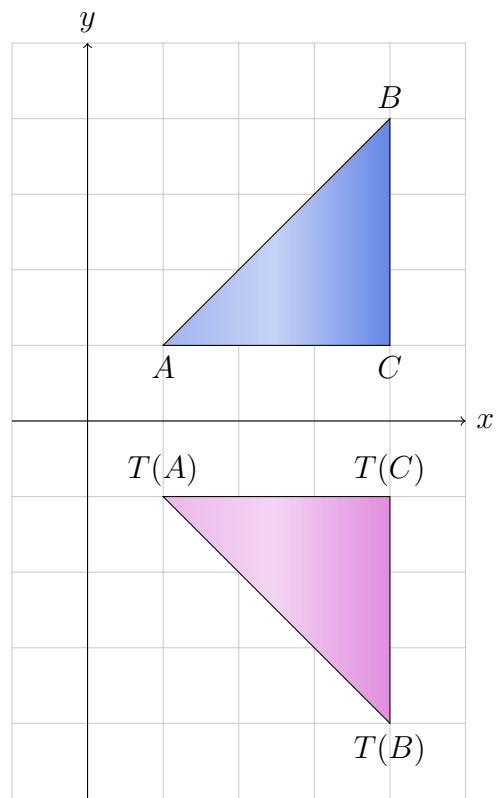


Figura 3.6: Simetría respecto del eje de abscisas

Una pregunta interesante es qué sucede si componemos esta transformación  $T$  consigo misma; es decir, la aplicamos sobre los transformados de la figura original  $B$ . Simbólicamente  $T \circ T(B)$ .

La matriz asociada a esta transformación en las bases canónicas es  $M_E(T) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ . Es fácil ver que  $M_E(T)^2 = I$  por lo que  $T \circ T = id$ ; es decir, aplicar dos veces seguidas esta transformación es equivalente a aplicar la transformación identidad (la cual transforma a cada vector en sí mismo).



### Ejemplo 3.4. Rotaciones

Consideremos la siguiente transformación del plano en sí mismo,  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  definida por

$$T(x, y) = \begin{pmatrix} \cos(\pi) & -\sin(\pi) \\ \sin(\pi) & \cos(\pi) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ -y \end{pmatrix}$$

En la Figura 3.7 se exhibe el efecto que tiene esta trasformación sobre un triángulo en el plano.

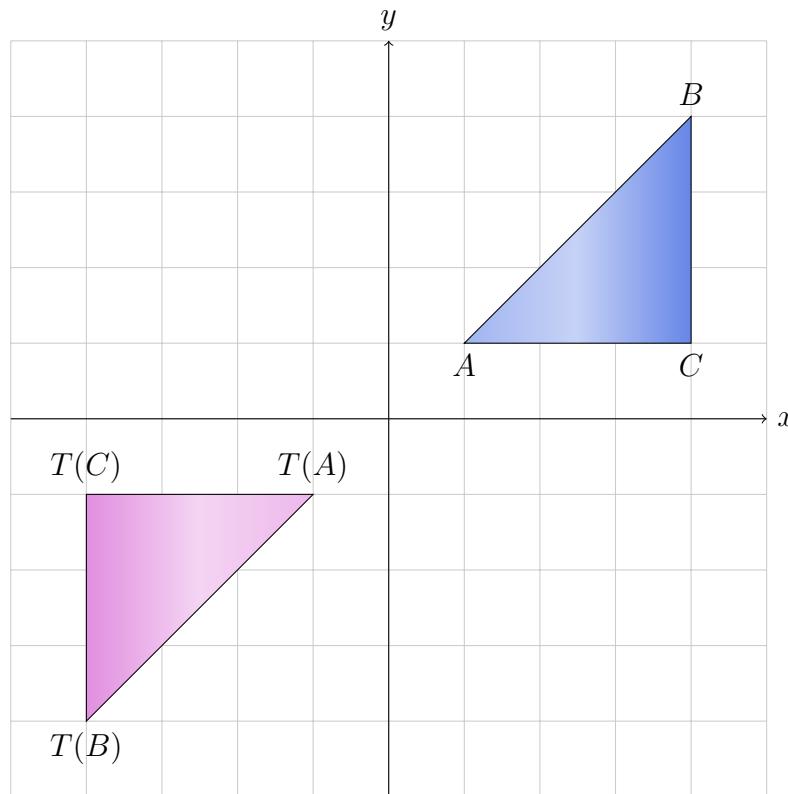


Figura 3.7: Rotación de ángulo  $\pi$

En general, la matriz de rotación de ángulo  $\theta$  en sentido antihorario está dada por

$$\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

■

### Ejemplo 3.5. Proyecciones ortogonales

La proyección ortogonal sobre el plano  $xy$ , es decir el plano de ecuación  $z = 0$ , es la transformación lineal que asigna a un punto  $P = (x, y, z)$  del espacio tridimensional, el punto  $P' = (x, y, 0)$  (ver Figura 3.8). La matriz asociada a esta transformación en las bases canónicas es

$$M_E(T) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

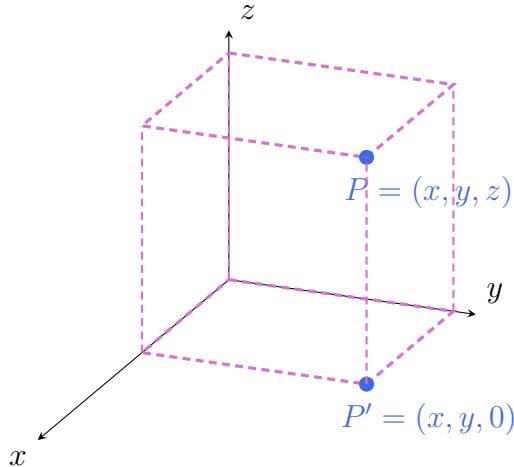


Figura 3.8: Proyección ortogonal de un punto sobre el plano  $xy$

■

### Ejemplo 3.6. Proyectores

Una transformación lineal  $P : \mathbb{V} \rightarrow \mathbb{V}$  es un **proyector** si al aplicarla por segunda vez no se altera el resultado obtenido en la primera. Simbólicamente, satisface  $(P \circ P)(v) = P(P(v)) = P(v)$  para todo  $v \in \mathbb{V}$ . Un ejemplo de proyector es la transformación definida en el Ejemplo 3.5.

Cabe destacar que existen vectores que, al aplicarles una transformación lineal conservan su dirección o permanecen constantes. En el caso de esta transformación, todos los vectores del plano  $xy$  permanecen constantes. En efecto, si  $v = (x, y, 0)$  pertenece al plano  $xy$ , se verifica que  $T(v) = Iv$ .

■

Esta última idea nos conduce a la siguiente sección.

### 3.2.1 Autovalores y Autovectores

Sea  $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$  una transformación lineal de un espacio en sí mismo, con matriz asociada en la base canónica  $A \in \mathbb{R}^{n \times n}$ . Se dice que  $v \in \mathbb{R}^n - \{0\}$  es un **autovector** asociado al **autovalor**  $\lambda \in \mathbb{R}$  si se verifica que  $T(v) = \lambda v$ . La expresión matricial de esta condición es  $Av^t = \lambda v^t$ .

El espacio generado por todos los autovectores asociados a un autovalor  $\lambda$  se denomina **autoespacio** asociado al autovalor  $\lambda$ . Simbólicamente se expresa  $S_\lambda = \{v \in \mathbb{R}^n / Av^t = \lambda v^t\}$ . Observemos que el vector nulo **no** es un autovector (por definición) pero sí pertenece al autoespacio de cualquier autovalor.

Los autovalores y autovectores de una transformación lineal  $T$  se corresponden con los de su matriz asociada en las bases canónicas.

**Ejemplo 3.7.** Para la proyección en plano horizontal vista en el Ejemplo 3.6, recordemos que los vectores del plano  $xy$  se transforman en sí mismos, por lo tanto son autovectores asociados al autovalor 1.

$$S_1 = \{v / v = (x, y, 0), \forall x, y \in \mathbb{R}\} = \langle(1, 0, 0), (0, 1, 0)\rangle$$



**Ejemplo 3.8.** Consideremos la transformación lineal cuya matriz asociada en las bases canónicas es  $M_E(T) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ . Entonces el vector  $(x, y)$  se transforma en el vector  $(y, x)$ . Nos preguntamos qué vectores se transforman en sí mismos o en un múltiplo de sí mismos (ver Figura 3.9).

Se puede observar que los vectores sobre la recta  $y = x$  permanecen fijos y los de la recta  $y = -x$  transforman en sus opuestos.

- \* el vector  $(1, 1)$  es un autovector de autovalor 1, pues se cumple que  $T(x, x) = (x, x) = 1(x, x)$ .
- \* el vector  $(1, -1)$  es un autovector de autovalor  $-1$ , pues se cumple que  $T(x, -x) = (-x, x) = -1(x, -x)$ .

En síntesis, esta transformación tiene dos direcciones principales.



Observemos que, por definición, si  $\lambda$  es un autovalor de  $A \in \mathbb{R}^{n \times n}$ , existe un vector no nulo  $v \in \mathbb{R}^n$  tal que  $Av^t = \lambda v^t$ . O en forma equivalente  $(A - \lambda I)v^t = 0$ . Dado que debe existir  $v$  no nulo, entonces necesariamente

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

Este determinante queda expresado en función de la variable  $\lambda$  y recibe el nombre de **polinomio característico** de  $A$  y se denota  $\chi_A(\lambda) = \det(A - \lambda I)$ . Resulta luego que, los autovalores de  $A$  son las raíces de su polinomio característico.

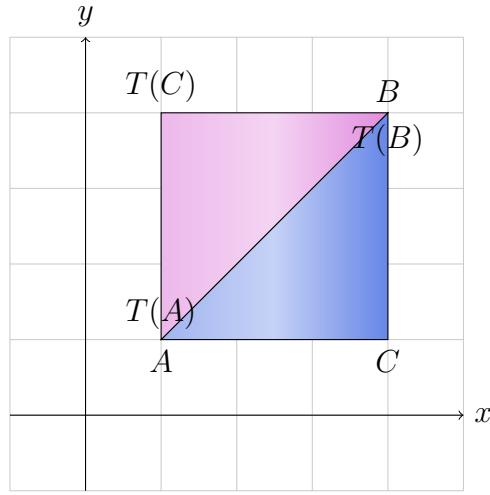


Figura 3.9: Simetría respecto de la recta  $y = x$

**Ejemplo 3.9.** Sea  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  transformación tal que  $M_E(T) = A = \begin{pmatrix} 2 & 3 \\ 3 & -6 \end{pmatrix}$ .

El polinomio característico está dado por

$$\begin{aligned}\chi_A &= \det(A - \lambda I) = \left| \begin{pmatrix} 2 & 3 \\ 3 & -6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = \left| \begin{pmatrix} 2-\lambda & 3 \\ 3 & -6-\lambda \end{pmatrix} \right| \\ &= (2-\lambda)(-6-\lambda) - 9 = \lambda^2 + 4\lambda - 21\end{aligned}$$

Igualando a cero, obtenemos sus raíces que son  $\lambda_1 = 3$  y  $\lambda_2 = -7$ .

Para hallar los autovectores, resolvemos los siguientes sistemas homogéneos:



$$(A - 3I) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 2 & 3 \\ 3 & -6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 3 \begin{pmatrix} x \\ y \end{pmatrix}$$

Al resolver el sistema nos encontramos con que las dos ecuaciones son equivalentes entre sí, por lo que nos quedamos con la primera ecuación  $-x + 3y = 0 \Leftrightarrow x = 3y$ .

Luego los vectores de la forma  $(3y, y) = y(3, 1)$  son autovectores asociados al autovalor 3 y el espacio generado de dimensión 1 es una recta cuyo vector director es  $(3, 1)$ .



$$(A + 7I) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 2 & 3 \\ 3 & -6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -7 \begin{pmatrix} x \\ y \end{pmatrix}$$

Nuevamente las dos ecuaciones son equivalentes, entonces resolvemos  $3x + y = 0 \Leftrightarrow y = -3x$ .

Luego los vectores de la forma  $(x, -3x) = x(1, -3)$  son autovectores asociados al autovalor  $-7$  y el espacio generado de dimensión 1 es una recta cuyo vector director es  $(1, -3)$ .



## Observaciones:

- ✿ Para cuantificar el tamaño de un vector  $v = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ , se puede calcular su **norma** (su longitud) mediante  $\|v\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$ .
- ✿ Si bien las posiciones de dos vectores o sus orientaciones pueden diferir mucho, estas medidas permiten comparar de alguna forma su fuerza.
- ✿ Nos va a interesar cuantificar el tamaño de la variabilidad de un conjunto, lo que equivale a cuantificar el tamaño de la matriz de covarianzas.

### 3.2.1.1 Relación entre autovalores, traza y determinante

Dada una matriz cuadrada  $A$ ; es decir con igual cantidad de filas y de columnas, para cuantificar su tamaño se han utilizado con frecuencia estas dos funciones:

- ✿ la **traza** o suma de los elementos de su diagonal, denotada por  $tr(A)$ .
- ✿ el **determinante** denotado por  $det(A)$ .

Estas dos funciones, nos darán una idea del tamaño de la variabilidad del conjunto y están muy relacionadas con los autovalores de la matriz.

Veamos cómo son estas funciones en el caso de una matriz de  $2 \times 2$  y cómo se vinculan con los autovalores de la matriz. Sea

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- ✿ la traza  $tr(A) = a + d$ .
- ✿ el determinante  $det(A) = ad - bc$ .
- ✿ el polinomio característico es  $\chi_A(\lambda) = (a - \lambda)(d - \lambda) - bc = \lambda^2 - (a + d)\lambda + ad - bc = \lambda^2 - tr(A)\lambda + det(A)$ .
- ✿ las raíces de este polinomio, supongamos  $\lambda_1$  y  $\lambda_2$ , son los autovalores de  $A$ .

Por propiedades de las raíces de un polinomio mónico de grado  $n$ , se sabe que la suma de las mismas coincide con el opuesto del coeficiente del término de grado  $n - 1$  y que su producto es igual al término independiente. En nuestro caso,  $\lambda_1 + \lambda_2 = tr(A)$  y  $\lambda_1\lambda_2 = det(A)$ .

Este resultado se puede generalizar de la siguiente manera. Sea  $A \in \mathbb{R}^{n \times n}$  con autovalores  $\lambda_1, \lambda_2, \dots, \lambda_n$  (complejos -si los tuviera- y con repeticiones), se puede demostrar que

$$\textcircled{*} \ tr(A) = \sum_{i=1}^n \lambda_i$$

$$\textcircled{*} \ \det A = \prod_{i=1}^n \lambda_i$$

**Ejemplo 3.10.** Siguiendo los cálculos del Ejemplo 3.9, vimos que los autovalores de  $A$  son  $3$  y  $-7$ . Aplicando el resultado anterior, tenemos que  $tr(A) = -4$  y  $\det(A) = -21$ .



Muchas aplicaciones de esta materia requieren el cálculo de trazas o determinantes. Para realizar estos cómputos, usaremos R. Mostraremos algunos ejemplos en el Código 3.1.

```
A=matrix(c(1,2,-1,1,0,1,3,1,0,0,2,0,0,0,1,-1), nrow=4, ncol=4, byrow=T)
# Ingresa una matriz de 4x4
A # Muestra la matriz

eigen(A)$values # Calcula los autovalores de A

## Comparar los siguientes cálculos:
sum(diag(A)) # Calcula la traza de A
sum(eigen(A)$values) # Calcula la suma de los autovalores de A

## Comparar los siguientes cálculos:
det(A) # Calcula el determinante de A
prod(eigen(A)$values) # Calcula el producto de los autovalores de A

t(A) # Calcula la traspuesta de A
sum(diag(t(A))) # Observar que las trazas de una matriz y su traspuesta son iguales
det(t(A)) # Observar que los determinantes de una matriz y su traspuesta son iguales
eigen(t(A))$values
# Observar que los autovalores de una matriz y su traspuesta son los mismos

solve(A) # Calcula la inversa de A
A%*%solve(A) # verifica que son inversas
det(solve(A))
# Observar que los determinantes de una matriz y su inversa son inversos
eigen(solve(A))$values
# Observar que los autovalores de una matriz y su inversa son inversos

eigen(A)$vectors # Calcula los autovectores de A
eigen(A)$vectors[,1] # Muestra el primer autovector
## Verifiquemos que es autovector de autovalor 2:
A%*%eigen(A)$vectors[,1]
2*eigen(A)$vectors[,1]

sqrt(sum(eigen(A)$vectors[,1]^2)) # Calcula la norma del primer autovector dado
```

---

### Código 3.1: Cálculos matriciales

En el Código 3.1 hemos observado relaciones entre los autovalores de una matriz con los de su traspuesta e inversa. Ahora estamos en condiciones de analizar el problema de la reducción de dimensión.

## 3.3 Motivación del problema de reducción de la dimensión

Supongamos que deseamos explorar en nuestra población los factores de riesgo de sufrir una enfermedad coronaria.

De estudios anteriores sabemos que se consideran como factores de riesgo para la enfermedad coronaria: la hipertensión arterial, la edad, la obesidad, el tiempo de antigüedad en el diagnóstico de hipertensión, el pulso, y el stress.



<https://flic.kr/p/WxGFa5>

Para la investigación se seleccionan al azar 20 pacientes hipertensos de la población objetivo sobre los cuales se miden las siguientes variables:

- ✿  $X_1$ : presión arterial media en mm/Hg
- ✿  $X_2$ : edad en años
- ✿  $X_3$ : peso en kg
- ✿  $X_4$ : superficie corporal en  $m^2$
- ✿  $X_5$ : tiempo transcurrido desde el diagnóstico de hipertensión en años
- ✿  $X_6$ : pulsaciones por minuto
- ✿  $X_7$ : medida asociada al stress

Si solamente conocemos las herramientas de análisis univariado y queremos estudiar las características de este grupo de pacientes en relación a los factores de riesgo, nos van a interesar las descripciones individuales de cada una de las variables consideradas así como las posibles interrelaciones entre las distintas variables.

También podríamos preguntarnos si es posible definir un índice general (o más de uno) que cuantifique la condición frente al riesgo de cada paciente.

La Tabla 3.3 contiene los datos registrados para este grupo de 20 pacientes.

Caso	Presión	Edad	Peso	Superficie	Tiempo	Pulso	Stress
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	49	99.5	2.25	9.3	71	10
7	121	48	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	50	94.1	1.98	5.6	71	21
13	114	49	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87	1.87	3.6	62	18
18	113	46	94.5	1.90	4.3	70	12
19	110	48	90.5	1.88	9.0	71	99
20	122	56	95.7	2.09	7.0	75	99

Tabla 3.3: Análisis sobre riesgo cardíaco

La *dimensión inicial* del problema planteado, entendida como la cantidad de variables consideradas en el análisis, es 7.

Si consideramos solamente dos variables, por ejemplo la presión y la edad, los resultados se pueden presentar mediante un diagrama de dispersión como el que aparece en la Figura 3.11 y que es generado mediante el Código 3.2 con datos extraídos de <https://goo.gl/E9AhVK>. Sobre la figura se ha representado mediante un punto a cada uno de los 20 pacientes, considerando solamente, las mediciones de su peso y de su superficie corporal. En este gráfico es posible observar el tipo de relación entre las dos variables, así como también las similitudes entre los individuos.

Dos individuos con representaciones próximas en el diagrama de dispersión tendrán características similares en estas dos variables, mientras que dos individuos alejados tendrán características diferentes en las mismas.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(ggrepel) # Paquete que manipula etiquetas para gráficos
library(readxl) # Permite leer archivos xlsx

riesgo=read_excel("C:/.../riesgo.xlsx")
# Importa la base con la cual se va a trabajar

ggplot(riesgo, aes(x=Peso, y=Superficie)) +
  geom_point(colour="royalblue", shape=8) +
  xlab("Peso") +
  ylab("Superficie corporal") +
  geom_text_repel(aes(label=rownames(riesgo), size = 2)) +
  theme(legend.position="none")
# Produce un dispersograma
```

Código 3.2: Análisis de dos variables

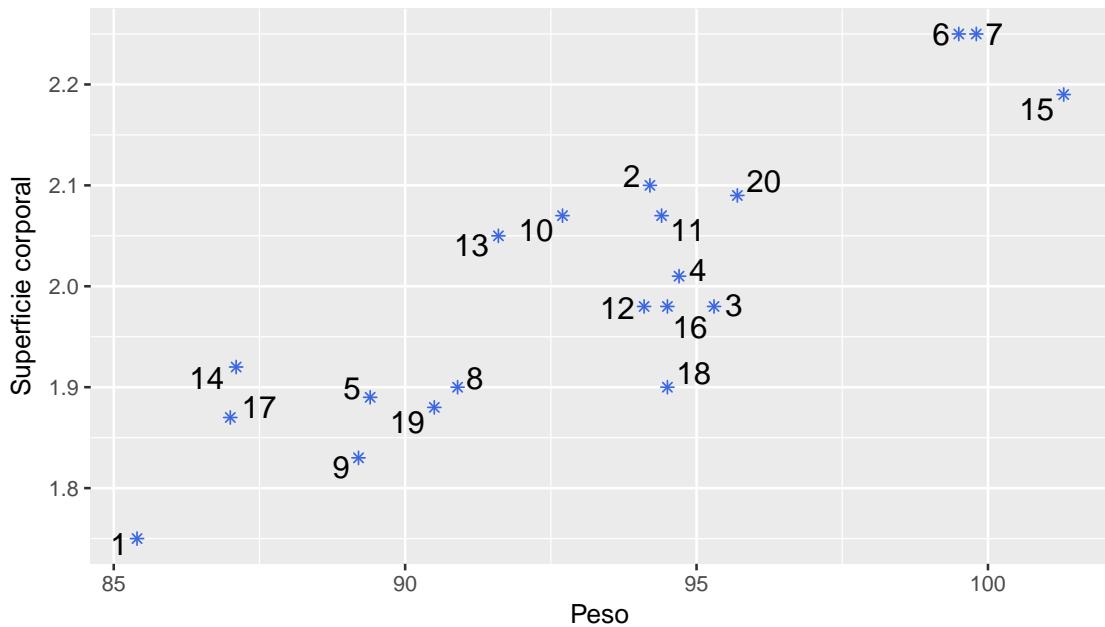


Figura 3.11: Dispersograma entre dos variables

En algunos estudios puede resultar interesante también ver si existen agrupamientos de individuos.

### Representación tridimensional de variables

Considerando las tres primeras variables (presión, edad y peso), aún es posible representarlas en un diagrama de dispersión como se muestra en la Figura 3.12 generada por el Código 3.3 con datos extraídos de <https://goo.gl/E9AhVK>. En algunos utilitarios es posible rotar este gráfico a fin de apreciar la relación entre las variables representadas desde distintos ángulos.

```
library(scatterplot3d) # Paquete para generar gráficos en 3D
library(readxl) # Permite leer archivos xlsx

par(mfrow=c(1,2)) # Permite hacer gráficos simultáneos

riesgo=read_excel("C:/.../riesgo.xlsx")
# Importa la base con la cual se va a trabajar

scatterplot3d(riesgo[,2:4], angle=35, pch=16, color="royalblue", box=FALSE,
grid=TRUE, xlab="Presión", ylab="Edad", zlab="Peso")
scatterplot3d(riesgo[,2:4], angle=225, pch=16, color="royalblue", box=FALSE,
grid=TRUE, xlab="Presión", ylab="Edad", zlab="Peso")
# Producen dispersogramas en 3D con distintos ángulos de visión
```

Código 3.3: Generación de dispersogramas en 3D

Las representaciones tridimensionales sobre el papel son difíciles de interpretar ya que no se tiene una referencia visual clara.

Si lográramos rotar la figura construida para las primeras tres variables de nuestro problema, podríamos apreciar que casi todos los puntos yacen sobre un plano.

Cabe preguntarnos si existe algún sistema de referencia (subespacio), en nuestro ejemplo un plano, cerca de la nube de puntos de forma tal que al proyectar cualquier par de puntos  $A, B$  sobre éste, se minimice la diferencia entre la distancia entre los puntos originales y la distancia entre los puntos proyectados. Es decir que se debe minimizar  $|dist(A, B) - dist(A', B')|$ , siendo  $A'$  y  $B'$  las proyecciones sobre dicho subespacio de los puntos  $A$  y  $B$  respectivamente.

Cuando esto ocurre, se pone de manifiesto que no son necesarias tres dimensiones para describir el conjunto de datos, sino que es posible dar una buena aproximación de la información de estas tres variables utilizando solamente dos.

Cuando el número de variables cuantitativas es superior a tres, el diagrama de dispersión ya no es posible.

Sin embargo, si tuviéramos una variable que indicara el sexo de los pacientes podríamos agregar color a la representación para visualizar los grupos, de modo tal de representar las cuatro variables en un solo gráfico como se muestra en la Figura 3.13 (ver Código 3.4 y datos disponibles en <https://goo.gl/E9AhVK>).

```
library(scatterplot3d) # Paquete para generar gráficos en 3D
library(readxl) # Permite leer archivos xlsx

riesgo=read_excel("C:/.../riesgo.xlsx")
# Importa la base con la cual se va a trabajar
```

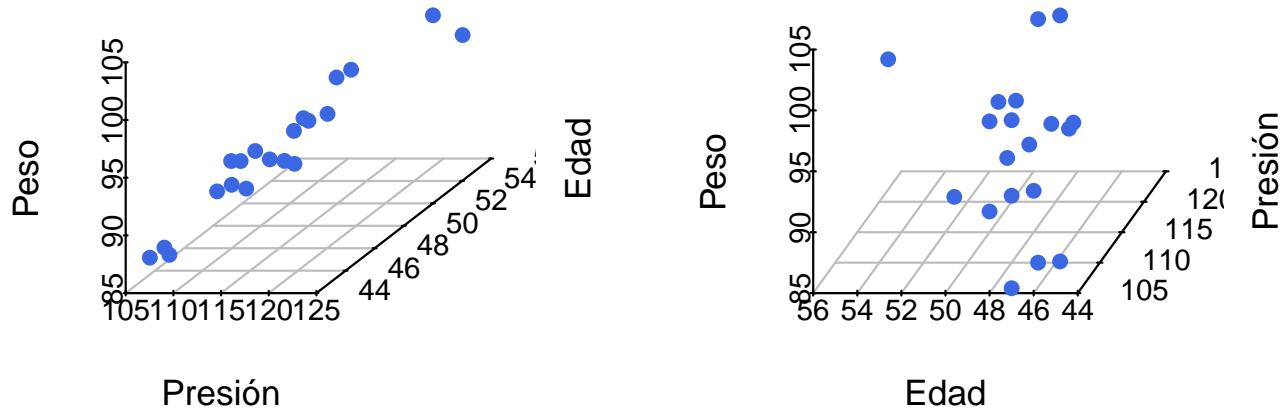


Figura 3.12: Dispersograma en tres dimensiones desde distintos puntos de vista

```

datos=data.frame(x=riesgo$PRESION,
y=riesgo$Peso,
z=riesgo$Edad,
group=riesgo$Sexo)
# Arregla los datos

with(datos, scatterplot3d(x, y, z, box=FALSE, grid=TRUE, pch = 16,
color=ifelse(group=="M", "royalblue", "indianred3"),
xlab="Presión", ylab="Peso", zlab="Edad"))
legend("topright", legend=unique(riesgo$Sexo), title = "Sexo", pch = 16,
col=c("indianred3", "royalblue"))
# Produce un dispersograma en 3D clasificado por grupos

```

Código 3.4: Dispersograma en 3D clasificado por grupos

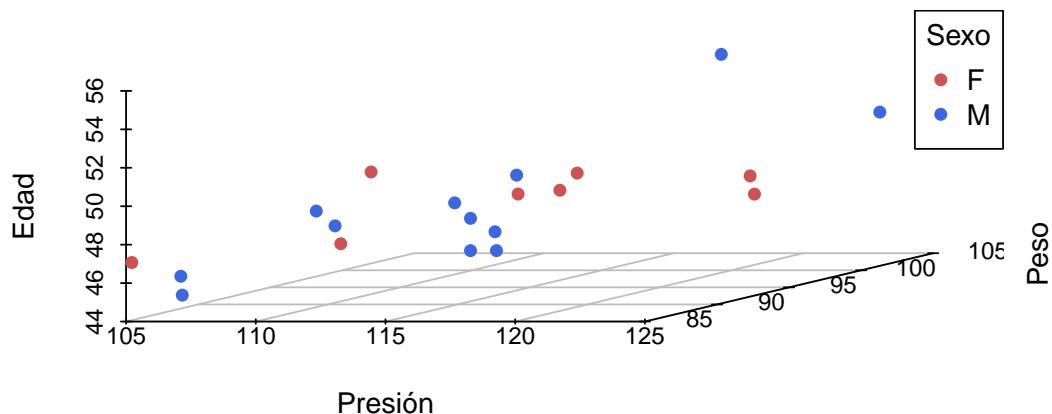


Figura 3.13: Dispersograma en 3D clasificado por grupos

La pérdida de información entre los datos originales y los datos proyectados puede cuantificarse de diversas formas, por ejemplo:

- ✿ variabilidad del conjunto de puntos originales versus variabilidad de las proyecciones.
- ✿ grado de similitud entre las distancias de los puntos originales y las distancias de los puntos proyectados.

Si representamos dos de las variables en un diagrama de dispersión es sencillo distinguir dos direcciones ortogonales, una de las cuales capta la mayor variabilidad del conjunto.

A estas dos direcciones se las suele identificar como **ejes principales** (ver 3.14) y los vectores que las definen resultan ser combinaciones lineales de las variables originales.

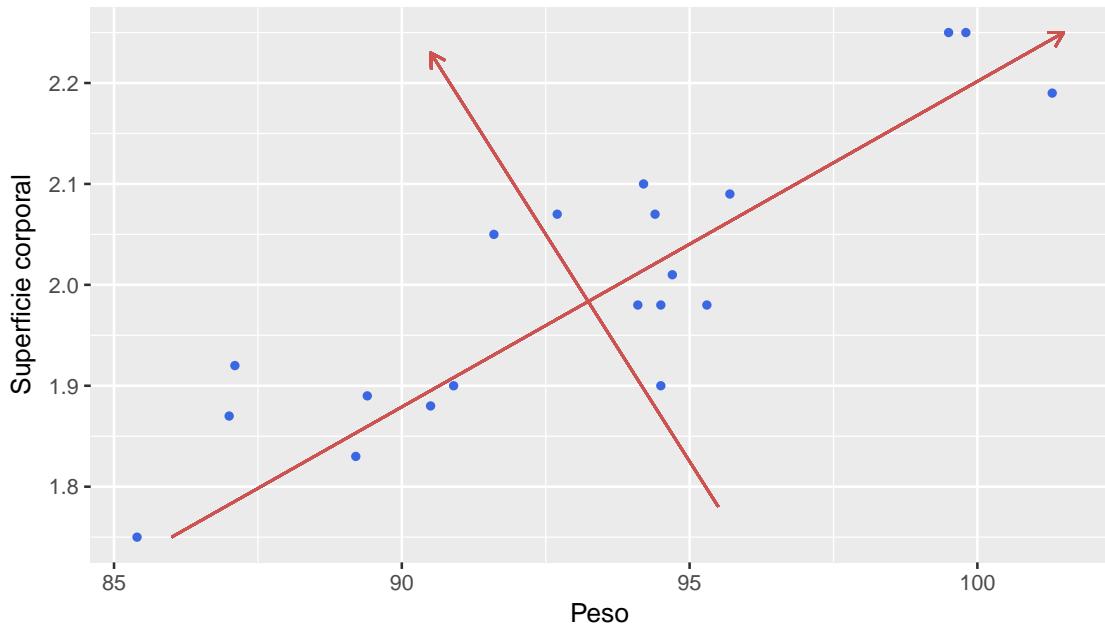


Figura 3.14: Ejes principales

*Estas nuevas variables constituyen un nuevo sistema de referencia.*

Si tuviéramos los puntos graficados en tres dimensiones 3.15, podríamos pensar el problema de cómo hacer para buscar el plano que logra que la proyección de los puntos sobre él tengan la mayor superficie ocupada posible. Una vez reducidos los puntos a dos dimensiones tendríamos el gráfico de la Figura 3.14.

De esta forma, podemos **reducir la dimensión del problema original**, seleccionando los ejes principales sobre el subespacio utilizado para proyectar.

*La reducción de la dimensión es posible cuando las variables están relacionadas entre sí y, por tanto, tienen información común.*

## 3.4 Análisis de componentes principales

El análisis de componentes principales (ACP) es un procedimiento matemático mediante el cual se transforma un conjunto de variables correlacionadas en un conjunto de variables **no correlacionadas**.

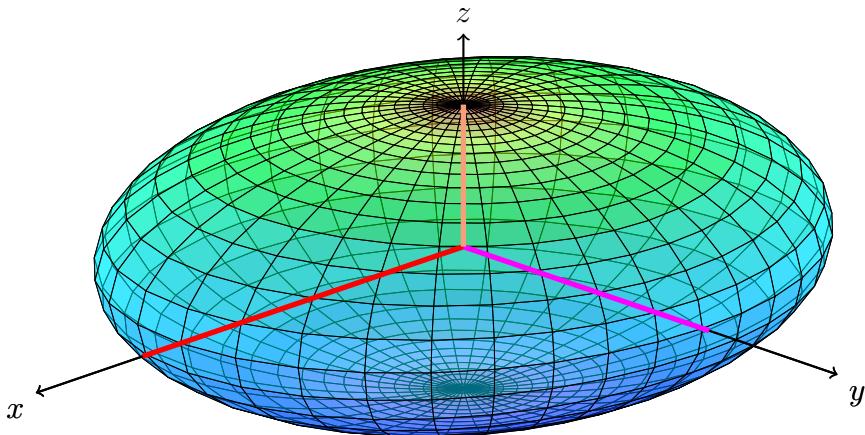


Figura 3.15: Direcciones principales en el espacio tridimensional

cionadas de menor dimensión que se obtienen a partir de combinaciones lineales de las variables originales, de manera tal que se preserve la mayor variabilidad del conjunto original de observaciones. A este nuevo conjunto de variables se lo denomina **componentes principales**.

Este análisis se trata de una técnica descriptiva, libre de distribución y en la cual se trabaja directamente con los datos muestrales.

El ACP es una técnica exploratoria que no establece supuestos por lo tanto **siempre puede aplicarse**. Esta técnica procura hallar aquellas combinaciones lineales de las variables originales que maximizan la varianza. Es decir, minimizan la pérdida de la información inicial. Luego, el propósito fundamental de esta técnica consiste en la **reducción de la dimensión** de los datos con el fin de simplificar la magnitud del problema a estudiar.

Se dispone de una matriz  $X \in \mathbb{R}^{n \times p}$  que contiene las observaciones de  $p$  variables tomadas sobre  $n$  individuos.

Es importante destacar que en el contexto de ACP:

- ✿ todas las variables juegan el mismo papel puesto que no existen variables independientes o dependientes, como en otros modelos estadísticos.
- ✿ el objetivo es reducir la dimensión del problema; es decir, la idea es descartar información redundante.
- ✿ es una alternativa que permite visualizar la información multidimensional.
- ✿ se explora la existencia de variables latentes.
- ✿ sólo tendrá sentido su aplicación en el caso en que las variables originales estén fuertemente correlacionadas.

- ✿ dado que la variabilidad del conjunto está representada por la matriz de covarianzas y ésta se ve influenciada por las unidades de medición, es recomendable realizar el análisis de componentes principales basándose en la matriz de correlaciones.
- ✿ se debe encontrar una condición de ‘finalización’ para determinar el número de componentes principales a seleccionar.
- ✿ los análisis confirmatorios permiten evaluar la estabilidad de las componentes principales y al mismo tiempo, brindan un apoyo para la detección de observaciones atípicas.
- ✿ puede resultar de utilidad para detectar algún tipo de anormalidad en las observaciones.

### 3.4.1 Definición de las componentes

Para resolver los problemas provocados por la multidimensionalidad, los estadísticos han realizado diversos planteos. Todos ellos convergen en la misma solución, que son las componentes principales.

Los distintos planteos alternativos fueron:

- ✿ Buscar aquella combinación lineal de las variables que maximiza la variabilidad (Hotelling [12]).
- ✿ Buscar el subespacio de mejor ajuste por el método de los mínimos cuadrados, minimizando la suma de cuadrados de las distancias de cada punto al subespacio de representación (Pearson [16]).
- ✿ Minimizar la discrepancia entre las distancias euclídeas entre los puntos calculadas en el espacio original y en el subespacio de proyección, que son las coordenadas principales (Gower [7]).



<https://flic.kr/p/7g9KSc>

Como las componentes resultan de la combinación lineal de las variables originales y definen un nuevo espacio de representación de las observaciones, restaría analizar qué variables explican las similitudes o diferencias entre los individuos en este nuevo espacio.

Este análisis se realiza a partir de las correlaciones entre las componentes principales y las variables originales.

### 3.4.2 Variabilidad explicada por las componentes principales

Vamos a deducir la expresión de las componentes principales siguiendo la idea original de Hotelling.

Si designamos a las variables originales por  $X_1, X_2, \dots, X_p$  y la variable  $Y_i$  es una combinación lineal de ellas, entonces

$$Y_i = \sum_{j=1}^p a_{ij} X_j = a_{i1} X_1 + a_{i2} X_2 + \dots + a_{ip} X_p$$

donde  $a_i = (a_{i1}, a_{i2}, \dots, a_{ip}) \in \mathbb{R}^p$ .

Comenzamos buscando  $a_1 \in \mathbb{R}^p$  tal que tenga norma unitaria, simbólicamente  $\|a_1\| = 1$  y tal que la variable  $Y_1$  tenga varianza máxima entre todas las posibles combinaciones lineales de  $X_1, X_2, \dots, X_p$ .

Los coeficientes  $a_{ij}$  se denominan **cargas** (*loadings* en inglés).

*¿Qué formato tiene la solución al problema planteado?*

*¿Cuáles son los valores de las cargas?*

Demostraremos que la variabilidad de la primera componente principal es máxima cuando el vector de cargas es el autovector asociado al mayor autovalor de la matriz de varianzas y covarianzas  $\Sigma$ .

Recordemos que si  $X \in \mathbb{R}^p$  es un vector columna aleatorio y  $a \in \mathbb{R}^p$  es un vector de constantes

$$\text{Var}(aX) = a \text{Var}(X) a^t = a \Sigma a^t$$

Nuestro problema es hallar el vector  $a$  de modo tal que  $a \Sigma a^t$  resulte máximo sujeta a la restricción de norma unitaria; es decir,  $aa^t = 1$ .

Para dar respuesta a este problema utilizaremos el método de multiplicadores de Lagrange. Construimos el multiplicador como la suma de la función a optimizar y una constante  $\lambda$  por la restricción de norma unitaria para el vector buscado,

$$L(a) = a \Sigma a^t + \lambda(aa^t - 1) \tag{3.1}$$

Derivando la expresión (3.1) respecto de  $a$  e igualando a cero, obtenemos

$$\frac{\partial L(a)}{\partial a} = 2\Sigma a^t - 2\lambda I a^t = 0 \tag{3.2}$$

El sistema de ecuaciones que resulta de (3.2) puede expresarse de la forma

$$(\Sigma - \lambda I)a^t = 0 \quad (3.3)$$

El sistema (3.3) admite solución no trivial cuando el determinante de la matriz  $\Sigma - \lambda I$  es nulo.

Como ya hemos visto anteriormente, los valores de  $\lambda$  que son solución de (3.3) son los autovalores de la matriz  $\Sigma$  y el vector  $a$  es el autovector asociado con el autovalor  $\lambda$  y de norma unitaria.

La matriz de covarianzas  $\Sigma$  es real, simétrica y de orden  $p$ , por lo cual tiene  $p$  autovalores reales. Además,  $\Sigma$  es semidefinida positiva lo que dice que sus autovalores son no negativos.

La notación de los autovalores ordenados es la siguiente:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

Entonces, si  $a_1$  es un autovector de norma 1 asociado al autovalor  $\lambda_1$ , se tiene que

$$Var(Y_1) = Var(a_1 X) = a_1 Var(X) a_1^t = a_1 \Sigma a_1^t = a_1 \lambda_1 a_1^t = \lambda_1 \underbrace{a_1 a_1^t}_{=1} = \lambda_1 \quad (3.4)$$

Luego, para maximizar la varianza de  $Y_1$  los coeficientes de la combinación lineal son las componentes del autovector unitario  $a_1$  asociado al mayor autovalor. Es decir, las componentes del autovector  $a_1$  son los coeficientes de la combinación lineal que define la **primera componente principal** dada por

$$Y_1 = a_1 X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

*¿Cómo se define la segunda componente principal?*

Buscamos otra combinación de las variables originales de la forma

$$Y_2 = a_2 X = \sum_{j=1}^p a_{2j} X_j = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

tal que  $Y_2$  tenga varianza máxima entre el conjunto de combinaciones lineales de las variables originales no correlacionadas con  $Y_1$ , sujeto a la condición que  $\|a_2\| = 1$ . Es decir que en el subespacio ortogonal a la primera componente principal, buscamos la segunda componente principal tal que

$$Cov(Y_1, Y_2) = Cov(a_1 X, a_2 X) = a_1 \Sigma a_2^t = a_1 \lambda_2 a_2^t = \lambda_2 \underbrace{a_1 a_2^t}_{=0} = 0$$

El objetivo ahora es maximizar  $Var(Y_2)$  sujeto a las restricciones

\*  $a_1 a_2^t = 0$ ,

\*  $a_2 a_2^t = 1$ .

Repitiendo el procedimiento de multiplicadores de Lagrange, ahora con dos restricciones, se obtiene que la segunda componente principal corresponde al autovector asociado al segundo de los autovalores ordenados en forma decreciente; es decir,  $\lambda_2$ .

En este procedimiento se va construyendo una nueva representación de variables no correlacionadas, que pierde la menor información posible de los datos originales.

### 3.4.3 Variabilidad de las componentes principales

Nos interesa saber ahora, qué parte de la variabilidad total del conjunto logra captar cada componente principal.

La variabilidad de cada variable  $X_k$  está representada por  $Var(X_k) = \Sigma_{kk}$  que es el  $k$ -ésimo elemento de la diagonal principal de la matriz de covarianzas.

La **variabilidad total** del conjunto de datos, es la suma de las varianzas de cada una de las variables; es decir, la traza de la matriz de covarianzas de las variables originales. Simbólicamente, la variabilidad total se calcula como

$$tr(\Sigma) = \Sigma_{11} + \Sigma_{22} + \cdots + \Sigma_{pp}$$

Recordemos que  $tr(\Sigma) = \sum_{i=1}^p \lambda_i$  donde  $\lambda_i$  son los autovalores de la matriz  $\Sigma$ .

Mediante un razonamiento análogo a 3.4 se puede probar que  $\lambda_i = Var(Y_i)$ . Además, los autovalores  $\lambda_i$  son decrecientes por construcción; es decir,  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ .

Entonces cabe preguntarnos lo siguiente

*¿Qué proporción de la variabilidad total logra captar la primera componente principal?*

*¿Qué proporción de esa variabilidad logra captar cada una de las componentes principales consideradas?*

La proporción de la variabilidad que capta la primera componente principal es:

$$\frac{\lambda_1}{tr(\Sigma)} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Como hemos visto, los valores de  $\lambda_i$  van decreciendo, luego la proporción que cada componente logra captar de la variabilidad total, también disminuye.

### 3.4.4 Cantidad de componentes principales

En algún punto, dado que nuestro objetivo es reducir la dimensión del problema original, dejará de tener sentido seguir buscando nuevas componentes principales.

La pregunta es

*¿Cuántas componentes conviene considerar?*

Existen diferentes criterios para decidir el número de componentes principales a elegir [17]. Algunos ya están incorporados en los paquetes estadísticos.

### 3.4.4.1 Criterio 1: Porcentaje de variabilidad explicada

Se define un porcentaje de variabilidad mínimo que se desea explicar y se toman las primeras  $m$  componentes que alcanzan este porcentaje de explicación. Es decir, si se desea explicar el  $q\%$  de la variabilidad, elegimos  $k$  componentes de modo tal que

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\text{tr}(\Sigma)} \geq \frac{q}{100}$$

siendo

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_{k-1}}{\text{tr}(\Sigma)} < \frac{q}{100}$$

En general, no se trabaja con la matriz de covarianza de las variables  $\Sigma$ , sino con su estimación  $\hat{\Sigma}$ , que es la matriz de covarianza muestral. Lo que se puede calcular son los autovalores y autovectores de esta matriz.

Éstos son los estimadores de máxima verosimilitud de los poblacionales, cuando la distribución de los datos es normal multivariada (veremos luego esta distribución con más detalle).

Se dispone de un test para decidir si son suficientes  $q$  componentes principales para explicar el  $p_0\%$  de variabilidad (ver Apéndice ??).

### 3.4.4.2 Criterio 2: Criterio de Kaiser

Obtener las componentes principales a partir de la matriz de correlaciones  $R$  poblacional equivale a suponer que las variables observables tienen varianza 1. Por lo tanto una componente principal con varianza inferior a 1 explica menos variabilidad que una de las variables originales.

El criterio, llamado de **Kaiser**, consiste en retener las  $m$  primeras componentes tales que sus autovalores resulten iguales o mayores que 1. Simbólicamente:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 1 \quad \text{y} \quad \lambda_{m+1} < 1$$

Sin embargo, algunos autores recomiendan utilizar

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0.7$$

basados en estudios de simulación de Montecarlo.

Este criterio puede extenderse a la matriz de covarianzas de la siguiente manera: se eligen las primeras  $m$  componentes tales que

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq \frac{\text{tr}(\Sigma)}{p}$$

y

$$\lambda_{m+1} < \frac{\text{tr}(\Sigma)}{p}$$

Nuevamente, puede considerarse la sugerencia de utilizar como cota inferior a  $\frac{0.7}{p}\text{tr}(\Sigma)$ .

#### 3.4.4.3 Criterio 3: Criterio del bastón roto

Por otra parte, si la proporción de variabilidad explicada por  $Y_1, Y_2, \dots, Y_m$  se estabiliza a partir de un cierto valor de  $m$ , entonces aumentar la dimensión no aportaría cambios significativos.

La representación de la secuencia de valores propios de la matriz de covarianzas, ordenados de mayor a menor, recibe el nombre de **gráfico de sedimentación** o *scree plot*, debido a que se asemeja al perfil de una montaña. En un punto del gráfico la pendiente se suaviza pareciéndose a una meseta, donde se acumularían los sedimentos que caen por la ladera, dando de esta forma el nombre al gráfico. La sugerencia de este criterio es seleccionar las componentes previas a la zona de acumulación de sedimentos.

#### 3.4.4.4 Criterio 4: Prueba de esfericidad

Si las observaciones provienen de una distribución Normal  $p$ -variada y las variables son independientes, entonces no existen direcciones de máxima variabilidad. Es decir, la variabilidad es similar en todas las direcciones. En este caso, la distribución tiene forma similar a una esfera y de ahí el nombre de estos tests.

Este test está basado en un estadístico cuya distribución es Chi Cuadrado,  $\chi^2$ , y se aplica en forma secuencial. En el Apéndice ?? se presenta este test.

La hipótesis nula  $H_0^m$  de este test, plantea que a partir de  $m$ , no hay direcciones de máxima variabilidad; es decir, que a partir de  $m$ , la distribución es esférica.

Si no rechazamos  $H_0^0$ , significa que no hay direcciones principales. Por el contrario, si rechazamos  $H_0^0$ , testeamos  $H_0^1$  y así sucesivamente hasta que no rechacemos  $H_0^m$ . Por ejemplo, si decidimos que hay dos direcciones principales en un conjunto de cinco variables, en este caso  $p = 5$  y  $m = 2$ , significa que rechazamos  $H_0^0$  y  $H_0^1$  pero no rechazamos  $H_0^2$ .

**Importante:** este test suele rechazar la hipótesis nula sólo debido a que el tamaño de la muestra es muy grande. Por tal motivo, existen recomendaciones de aplicarlo sólo cuando  $\frac{n}{p} < 5$ .

### 3.4.5 Estimación de las componentes principales

Hasta acá hemos desarrollado la deducción de las componentes principales y la proporción de variabilidad explicada utilizando la matriz de covarianza poblacional  $\Sigma$ . Sin embargo, en general no se dispone de la matriz  $\Sigma$  y se la estima con la matriz de covarianza muestral  $S$ . Simbólicamente  $\widehat{\Sigma} = S$ .

Análogamente, la matriz de correlación muestral estima a la poblacional. Recordemos que la matriz de correlación equivale a la matriz de covarianza de las variables estandarizadas. De este modo, los autovalores y los autovectores de  $\Sigma$  se estiman respectivamente con los autovalores y los autovectores de  $S$ .

**Ejemplo 3.11.** Retomando el Ejemplo 3.1 de los nadadores, vamos a calcular las componentes principales, a estimar la proporción de variabilidad explicada por cada una de ellas aplicando los

criterios expuestos para decidir la cantidad de componentes que se deberían considerar. Para este estudio utilizamos el Código 3.5 con datos extraídos de <https://goo.gl/MJp9hr>.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(devtools) # Colección de herramientas de desarrollo para paquetes
install_github("vqv/ggbiplot") # Instala paquete desde GitHub
library(ggbiplot) # Paquete para visualización de componentes principales
library(readxl) # Permite leer archivos xlsx

nad=read_excel("C:/.../nadadores.xlsx")
# Importa la base con la cual se va a trabajar

nadadores=data.frame(nad[,2:5])
nad.pca.cov=prcomp(nadadores, center = TRUE, scale. = FALSE)
# Realiza el análisis de componentes principales
nad.pca.cor=prcomp(nadadores, center = TRUE, scale. = TRUE)
# Realiza el análisis de componentes principales para las variables estandarizadas
summary(nad.pca.cor)
summary(nad.pca.cov)
# Realiza un resumen de las variabilidades explicadas por las componentes principales

ggscreeplot(nad.pca.cov, type = c('pev', 'cev')) +
xlab('Número_de_componentes_principales') +
ylab('Proporción_de_la_variabilidad_explícada') +
geom_line(colour='royalblue') +
geom_point(colour='royalblue')
# Produce un gráfico de sedimentación
```

Código 3.5: Análisis de componentes principales de los nadadores

	PC1	PC2	PC3	PC4
<b>Desvío estándar</b>	1.709	0.957	0.348	0.197
<b>Proporción de variabilidad</b>	0.731	0.229	0.030	0.009
<b>Proporción acumulada</b>	0.731	0.960	0.990	1.000

Tabla 3.4: Variabilidad de las componentes principales usando las variables estandarizadas

En las Tablas 3.4 y 3.5 se puede apreciar que:

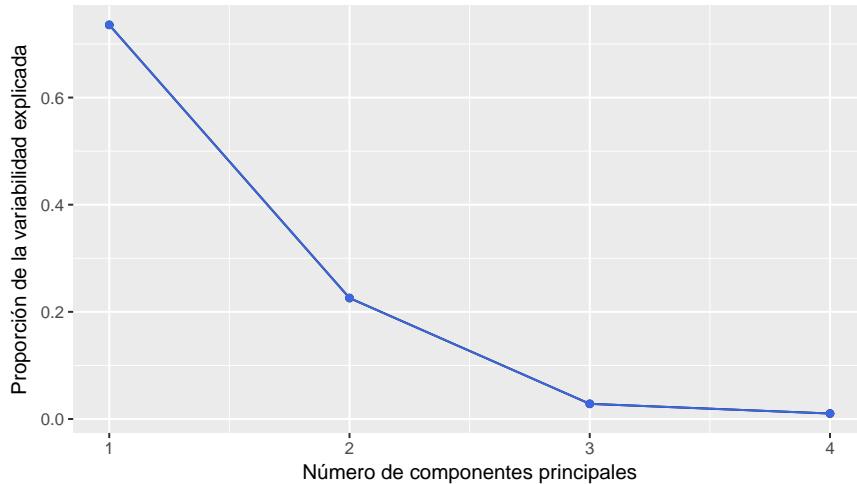
- \* la primera componente principal logra captar el 73% de la variabilidad total.
- \* las primeras dos componentes principales logran captar el 96% de la variabilidad total del conjunto.

	PC1	PC2	PC3	PC4
Desvío estándar	3.584	1.986	0.703	0.422
Proporción de variabilidad	0.736	0.226	0.028	0.010
Proporción acumulada	0.736	0.962	0.989	1.000

**Tabla 3.5:** Variabilidad de las componentes principales usando las variables originales

- ✿ los autovalores disminuyen considerablemente a partir de la tercera componente y alcanzan valores muy por debajo de 1.

En la Figura 3.17 se representan en el eje de abscisas el orden de las componentes y en el eje de ordenadas la proporción de la variabilidad explicada por cada una de ellas.



**Figura 3.17:** Gráfico de sedimentación

### Interpretación del gráfico de sedimentación

En la Figura 3.17 se ve con claridad lo siguiente:

- ✿ las dos últimas componentes principales explican una proporción de variabilidad mucho menor que la que explican las dos primeras.
- ✿ la segunda componente explica algo más del 20% de la variabilidad total.

Analicemos la cantidad de componentes principales a considerar:

- ✿ **Criterio 1:** Si se quiere explicar el 75% de la variabilidad total del conjunto de los nadadores se deben considerar las dos primeras componentes, dado que con una sola no se alcanza ese porcentaje.
- ✿ **Criterio 2:** Si se consideran los autovalores mayores que 1 de las variables estandarizadas, se debe tomar una sola componente. Si en cambio se consideran los mayores que 0.7, se deberían tomar las dos primeras. Recordemos que los autovalores corresponden a la varianza de la componente y por lo tanto debe elevarse al cuadrado el desvío estándar de la salida que se muestra en la Tabla 3.4).
- ✿ **Criterio 3:** En el gráfico de sedimentación 3.17 se aprecia que el quiebre se produce en la segunda componente, lo que coincide con los criterios anteriores.
- ✿ **Criterio 4:** En el Código 3.6 con datos extraídos de <https://goo.gl/MJp9hr> se aplica secuencialmente un test de esfericidad.

```
library(readxl) # Permite leer archivos xlsx

nadadores=read_excel("C:/.../nadadores.xlsx")
# Importa la base con la cual se va a trabajar

pval=0; estad=0; gl=0 # Inicializa las variables
autoval=prcomp(nadadores, center = TRUE, scale. = TRUE)[[1]]
# Guarda los autovalores

p=4; n=14 # Asigna valores a los parámetros
for(m in 1:p){
  r=m+1; u=p-m
  estad[m]=n-(1-(2*p+11)/6)*(u*log(mean(autoval[r:4]))-
    sum(autoval[r:4]))
  # Calcula el estadístico de contraste de cada paso
  gl[m]=(p-1)*p/2 # Calcula los grados de libertad
  pval[m]=1-pchisq(estad[m], gl[m])} # Calcula el p-valor de cada contraste

pval # Muestra los p-valores obtenidos
```

Código 3.6: Test de esfericidad de Bartlet

Este criterio elige las dos primeras componentes principales y rechaza, con nivel 0.05 en el tercer test. La secuencia de  $p$ -valores es: 0.3087 - 0.1935 - 0.0908 - 0.0456.

### 3.4.6 Escalas de medida

Si las escalas de medida de las variables fueran muy diferentes, la variabilidad estaría dominada por las variables con mayores magnitudes de manera que las primeras componentes pueden mostrar simplemente las diferencias de escala de medición. En ese caso conviene tomar las **variables estandarizadas** (matriz estandarizada por columnas), vale decir, centrar las variables y dividirlas por su desvío estándar. En ese caso las componentes estarían calculadas sobre la **matriz de correlaciones**.

Cuando las componentes principales se calculan a partir de las matrices de covarianzas, los factores de carga dependen de la escalas de medida de las variables por lo que son difíciles de interpretar. Mientras que si las componentes principales se calculan a partir de la matriz de correlaciones, las cargas (*loadings* en inglés) son las correlaciones entre las componentes principales y las variables originales. Los factores de carga suelen representarse en un gráfico que permite la interpretación visual de las relaciones. En cualquiera de los casos, podemos calcular la correlación al cuadrado entre las componentes y las variables originales (ver Figuras 3.18 y 3.19).

A estas correlaciones al cuadrado se las denomina usualmente **contribuciones relativas del factor al elemento** y miden la proporción de contribución del elemento a la componente principal.

Las componentes son combinaciones lineales de las variables originales y por ende, se espera que sólo unas pocas (las primeras) recojan la mayor parte de la variabilidad de los datos, obteniéndose así una reducción de la dimensión del problema.

### 3.4.7 Cargas o *loadings*

Estudiaremos ahora el aporte de los *loadings* a este análisis.

- ✿ Si la carga (coeficiente o *loading*) de una de las variables en la componente principal es positiva, significa que la variable y la componente tienen una correlación positiva. En este caso, el coseno del ángulo formado por la componente y la variable es positivo.
- ✿ Si la carga es positiva, un individuo que tenga una puntuación alta en esa variable tendrá valores más altos en esa componente que otro individuo que tiene un menor valor en esa variable y valores similares al primero en las restantes variables.
- ✿ Si por el contrario, la carga es negativa, este hecho indica que dicha variable se correlaciona en forma negativa con la primera componente.
- ✿ Cuando la carga de una variable es negativa para dos individuos con puntuaciones similares en las restantes variables, el que tenga puntuación más alta de los dos en esta variable se ubicará en un valor menor de la componente.

Con el fin de visualizar estas propiedades, se pueden graficar las cargas que tienen las variables originales en las componentes principales.

	PC1	PC2	PC3	PC4
Tramo 1	0.51890	-0.45481	0.18151	0.70061
Tramo 2	0.49743	-0.52759	-0.19481	-0.66049
Tramo 3	0.48743	0.51984	-0.68449	0.15374
Tramo 4	0.49561	0.49452	0.67864	-0.22192

Tabla 3.6: Cargas para los nadadores

En las Figuras 3.18 y 3.19 se representan las primeras dos componentes principales. Las mismas fueron generadas mediante el Código 3.7 y con datos extraídos de <https://goo.gl/MJp9hr>.

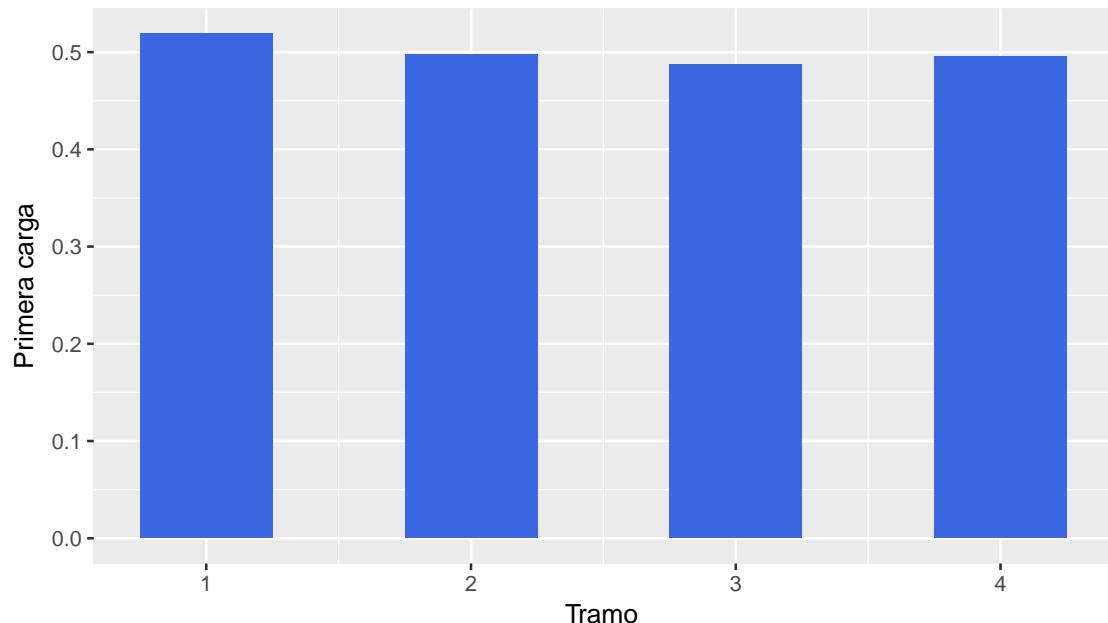


Figura 3.18: Cargas de la primera componente principal

```

library(ggplot2) # Paquete para confeccionar dibujos
library(devtools) # Colección de herramientas de desarrollo para paquetes
install_github("vqv/ggbiplot") # Instala paquete desde GitHub
library(ggbiplot) # Paquete para visualización de componentes principales
library(readxl) # Permite leer archivos xlsx

nad=read_excel("C:/.../nadadores.xlsx")
# Importa la base con la cual se va a trabajar

nadadores=data.frame(nad[,2:5])

```

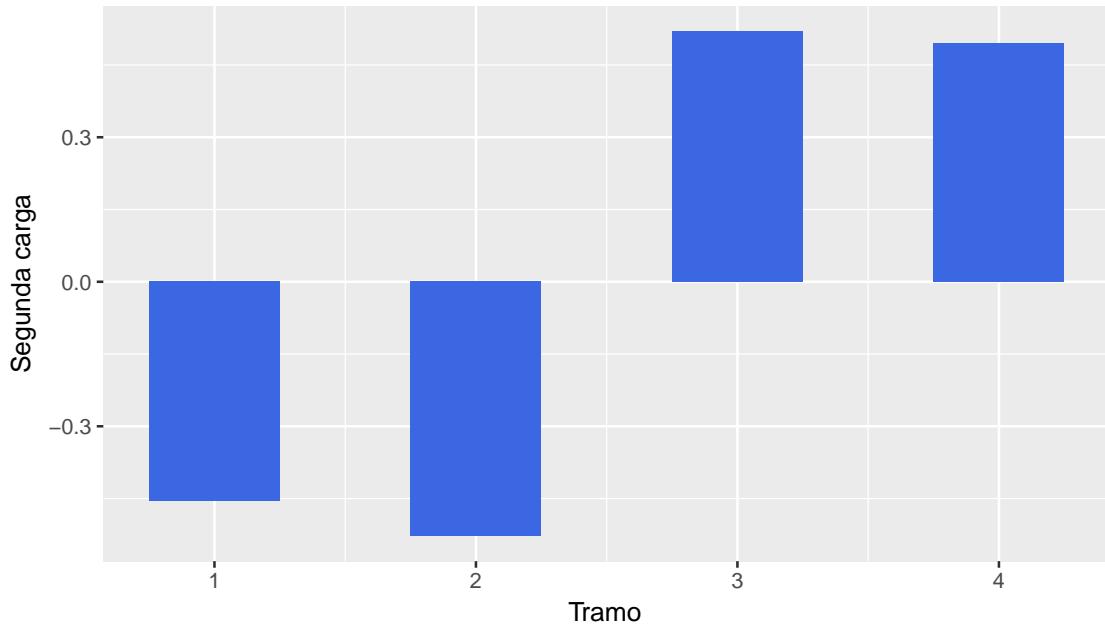


Figura 3.19: Cargas de la segunda componente principal

```

nad.pc=prcomp(nadadores, center=TRUE, scale.=TRUE)

carga1=data.frame(cbind(tramo=1:4,
primeracarga=data.frame(nad.pc$rotation)[,1]))
carga2=data.frame(cbind(tramo=1:4,
segundacarga=data.frame(nad.pc$rotation)[,2]))

ggplot(carga1, aes(tramo, primeracarga), fill=tramo) +
geom_bar(stat="identity", position="dodge", fill="royalblue", width=0.5) +
xlab('Tramo') +
ylab('Primera_carga')

ggplot(carga2, aes(tramo, segundacarga), fill=tramo) +
geom_bar(stat="identity", position="dodge", fill="royalblue", width=0.5) +
xlab('Tramo') +
ylab('Segunda_carga')

```

Código 3.7: Generación de gráficos de cargas

La Tabla 3.6 nos muestra los autovectores asociados a los autovalores presentados en las primeras tablas. Estos autovectores nos dan las cargas de las componentes principales.

Si denotamos a las variables originales estandarizadas con

$$Z_i = \frac{X_i - \bar{X}_i}{s_{X_i}}$$

la expresión para calcular los puntajes o *scores* de la primera componente principal es

$$Y_1 = 0.52Z_1 + 0.50Z_2 + 0.49Z_3 + 0.50Z_4$$

y la expresión para calcular los *scores* de la segunda componente principal es

$$Y_2 = -0.45Z_1 - 0.53Z_2 + 0.52Z_3 + 0.49Z_4$$

En la Tabla 3.7 estandarizamos las variables originales. Luego, con las variables estandarizadas, realizamos el cálculo de los *scores*, que se exhiben en la Tabla 3.8 utilizando las expresiones de las componentes principales.

Nadador	Tramo 1 est.	Tramo 2 est.	Tramo 3 est.	Tramo 4 est.
1	-0.5458	-0.5832	0.7479	0.3357
2	0.3531	0.3774	1.2715	1.7456
3	-0.0963	-0.5832	1.2715	0.8057
4	-0.9952	-1.0635	-0.2992	-0.1343
5	-1.4446	-1.5438	-1.3463	-1.5442
6	-1.4446	-1.0635	-0.8227	-1.0742
7	-0.5458	-0.5832	-1.8698	-1.0742
8	-0.0963	0.3774	-0.8227	-1.0742
9	1.2520	0.8576	-0.2992	-0.1343
10	0.3531	0.3774	0.2244	-0.6042
11	0.8026	0.8576	-0.2992	-0.1343
12	1.2520	1.8182	1.2715	0.8057
13	-0.5458	-0.5832	0.2244	0.8057
14	1.7015	1.3379	0.7479	1.2756

Tabla 3.7: Datos de los nadadores estandarizados por columna

### Estadísticos descriptivos

En la Tabla 3.9 se exhiben la media y la desviación típica de cada una de las variables originales de la base. En todos los tramos, los nadadores han hecho tiempos similares y dispersiones similares. Observemos que estos breves resúmenes no nos permiten distinguir entre los distintos estilos o calidades de nadadores del grupo.

<b>PC1</b>	<b>PC2</b>	<b>PC3</b>	<b>PC4</b>
-0.0424	1.1107	-0.2696	0.0433
1.8559	1.1645	0.3049	-0.1938
0.6790	1.4109	-0.2274	0.3344
-1.2579	0.7918	0.1402	-0.0110
-2.9392	0.0080	-0.0879	0.1432
-2.2122	0.2592	-0.2209	-0.1978
-2.0171	-0.9473	0.5654	-0.0462
-0.7957	-1.1142	-0.2569	-0.2048
0.8640	-1.2438	0.1738	0.2945
0.1809	-0.5419	-0.5731	0.1668
0.6308	-1.0394	0.0923	-0.0204
2.5733	-0.4693	-0.4505	-0.3071
-0.0647	1.0710	0.4077	-0.1415
2.5453	-0.4601	0.4020	0.1403

Tabla 3.8: Puntajes (*scores*) de los nadadores

<b>Tramo</b>	<b>Media</b>	<b>Desvío</b>	<b>n</b>
1	11.21	2.22	14
2	11.21	2.08	14
3	11.57	1.91	14
4	11.29	2.12	14

Tabla 3.9: Estadística descriptiva univariada para los nadadores

### 3.4.8 Interpretación de las componentes principales

- ✿ La primera componente tiene todos las cargas positivas, por lo cual se la considera una componente de tamaño. Es decir que un individuo tendrá puntuación alta en esta componente si ha tardado mucho en todos los tramos o bien si la suma de tiempos que le ha llevado correr la carrera completa es alta. Por el contrario, los individuos que han hecho “buenos tiempos” tendrán valores bajos en esta componente. Esta componente podría denominarse ‘rapidez’.
- ✿ La segunda componente es en cambio, un contraste, se dice que es una componente de forma. Contrasta los tiempos de los primeros dos tramos con los de los últimos dos. Un individuo tendrá alta esta componente si tardó poco al principio y desaceleró en los últimos tramos. Por el contrario, si un individuo reservó su energía en los dos primeros tramos y aceleró en los dos últimos porque está cansado, su segunda componente será baja. Esta componente podría denominarse ‘experiencia en carreras’.

### 3.4.9 Biplot

Para poder interpretar los resultados del análisis, se agrupan las componentes por pares. Se dibuja un gráfico en el que se representan simultáneamente las variables y los valores de cada individuo en el par de componentes principales seleccionada.

A este gráfico se lo conoce como *biplot*.

Graficar los individuos tiene sentido cuando las observaciones son pocas. En contrapartida, cuando disponemos de gran cantidad de observaciones, puede tener sentido identificar grupos de individuos con distintos colores.

La Figura 3.20 muestra un *biplot* para los datos de la base de nadadores y la misma es generada mediante el Código 3.8 y cn datos extraídos de <https://goo.gl/MJp9hr>.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(devtools) # Colección de herramientas de desarrollo para paquetes
install_github("vqv/ggbiplot") # Instala paquete desde GitHub
library(ggbiplot) # Paquete para visualización de componentes principales
library(ggrepel) # Paquete que manipula etiquetas para gráficos
library(readxl) # Permite leer archivos xlsx

nad=read_excel("C:/.../nadadores.xlsx")
# Importa la base con la cual se va a trabajar

nadadores=data.frame(nad[,1:5])
nad.pc=prcomp(nadadores[,2:5], center=TRUE, scale.=TRUE)

ggbiplot(nad.pc, obs.scale=1) +
  geom_point(colour="royalblue") +
  geom_text_repel(aes(label=nadadores[,1])) +
  theme(legend.position="none") +
```

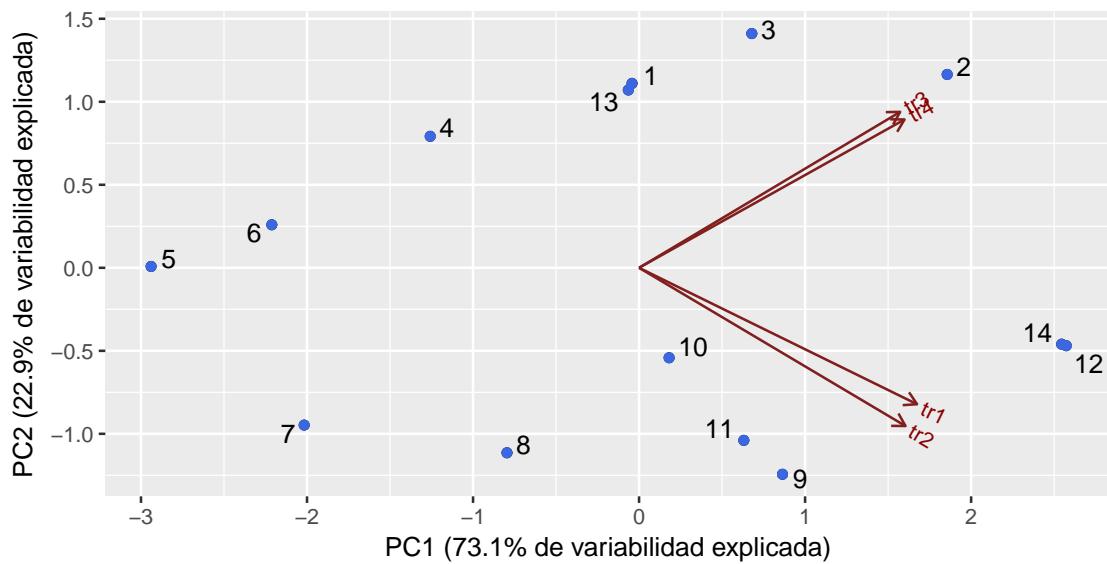


Figura 3.20: *Biplot* para nadadores

```
xlab("PC1_(73.1%_de_variabilidad_explicada)") +
ylab("PC2_(22.9%_de_variabilidad_explicada)")
# Genera un biplot
```

Código 3.8: Generación de un *biplot*

El *biplot* tiene la particularidad de facilitar

- ✿ la interpretación de las distancias entre individuos en términos de similitud en relación a las variables consideradas.
- ✿ la búsqueda de grupos o patrones.
- ✿ la explicación de las componentes principales utilizando las correlaciones con las variables originales.
- ✿ el estudio de las posiciones relativas de los individuos entre sí y respecto de las componentes principales graficadas.

Un *biplot* es una representación gráfica de datos multivariantes. De la misma manera que un diagrama de dispersión muestra la distribución conjunta de dos variables, un *biplot* puede representar tres o más variables.

El *biplot* aproxima la distribución de una muestra multivariante en un espacio de dimensión menor, frecuentemente de dimensión dos. El mismo superpone, sobre la misma representación,

las variables originales de la muestra. El prefijo *bi-* se refiere a la superposición, en la misma representación, de individuos y variables.

En el *biplot* las representaciones de las variables son vectores. Sus proyecciones sobre las componentes principales (ejes del *biplot*) nos dan idea de los *loadings*.

Este tipo de figura resulta útil para describir gráficamente los datos o para mostrar los resultados proporcionados por modelos más formales.

La forma más sencilla del *biplot* es un diagrama de dispersión en el que los puntos representan a los individuos, y los dos ejes a las componentes.

Desde el punto de vista del usuario, los *biplots* serán importantes debido a que su interpretación se basa en conceptos geométricos sencillos como los que se detallan a continuación.

- ✿ La similitud entre individuos es la función inversa de la distancia entre los mismos, sobre la representación *biplot*.
- ✿ Las longitudes y los ángulos de los vectores que representan a las variables, se interpretan en términos de variabilidad y covariabilidad respectivamente.
- ✿ Las relaciones entre individuos y variables se interpretan en términos de producto escalar; es decir, en términos de las proyecciones de los puntos *individuo* sobre los vectores *variable*.

**Ejemplo 3.12.** A continuación citamos conclusiones obtenidas a partir de la interpretación del *biplot* de nadadores dado por la Figura 3.20.

- ✿ En el *biplot* se aprecian las relaciones entre las variables y entre los individuos.
- ✿ Si las variables forman ángulos muy pequeños, significa que están muy correlacionadas.
- ✿ En este gráfico hay dos pares de variables muy correlacionadas *tr1* con *tr2* por un lado y, *tr3* con *tr4* por el otro.
- ✿ Cuando dos variables son ortogonales (perpendiculares) indica que no están correlacionadas.
- ✿ Asimismo, las proyecciones de las cuatro variables sobre el eje de la primera componente principal son todas positivas, mientras que la proyección de las dos primeras variables sobre la segunda componente principal es positiva y la de las dos siguientes sobre la segunda componente principal es negativa.
- ✿ Respecto de los individuos, podemos decir que 5 y 12 son los opuestos respecto de la primera componente principal, el individuo 12 es el más lento del grupo, mientras que el individuo 5 es el más rápido, el que hizo la carrera en menos tiempo; es decir, el ganador.
- ✿ Los individuos 4, 8 y 10 son los más cercanos al origen del nuevo sistema de coordenadas y se los considera nadadores promedio.

- Si pensamos en la segunda componente principal, que explica los estilos de nadar de los participantes, los individuos 2 y 3 se gastan toda la energía en el primer tramo y llegan cansados al segundo tramo, mientras que los individuos 9 y 11 guardan energía para el segundo tramo, donde aceleran y ganan diferencia teniendo estilos similares.

Veamos si con el gráfico de caritas de Chernoff (ver Figura 3.21 generada con el Código 3.9) y con datos extraídos de <https://goo.gl/MJp9hr>, es posible detectar la presencia de los mismos patrones y similitudes que se aprecian en el *biplot*.

```
library(tcltk2) # Paquete que permite hacer caras de Chernoff
library(aplpack) # Paquete que permite hacer caras de Chernoff
library(readxl) # Permite leer archivos xlsx

nad=read_excel("C:/.../nadadores.xlsx")
# Importa la base con la cual se va a trabajar

faces(nadadores, nrow.plot=3, ncol.plot=5, face.type=1,
labels=nadadores$nadador)
# Produce un diagrama de caras de Chernoff
```

Código 3.9: Generación de caras de Chernoff para nadadores

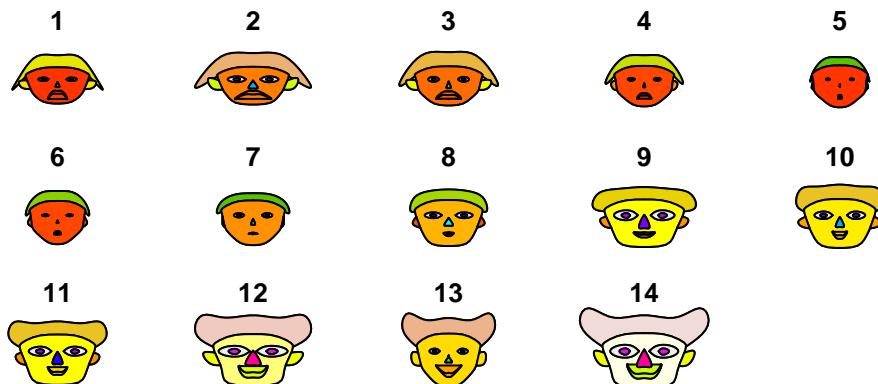


Figura 3.21: Caras de Chernoff para nadadores



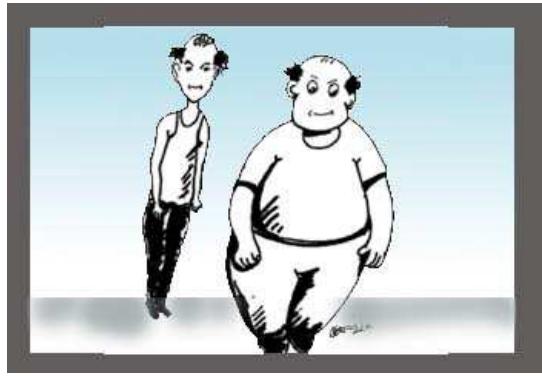
Se visualiza en la Figura 3.21 que hay nadadores similares como 6 y 5 o 9 y 11. Sin embargo en este gráfico no podríamos asegurar cuáles de estos nadadores son más rápidos ni tampoco cuáles son más expertos. Además sólo tiene sentido realizar esta representación si se dispone de una base relativamente chica de datos.

Enunciamos a continuación conceptos clave.

- ✿ La matriz de vectores propios, que denotamos por  $V$ , define un cambio de base del espacio  $\mathbb{R}^p$  en el que se ha representado la matriz de datos originales.
- ✿ Las  $q$  primeras columnas de  $V$  definen la proyección de los puntos en  $\mathbb{R}^p$  sobre el subespacio  $q$ -dimensional de mejor ajuste.
- ✿ Los elementos de  $V$  son los cosenos de los ángulos que forman las variables originales y las componentes principales.
- ✿ Las coordenadas de los individuos en el nuevo sistema de referencia son de la forma  $VX^t = Y^t$ .
- ✿ Estas puntuaciones se denominan *scores* y son representables.
- ✿ El ACP utiliza la información redundante, a través de las correlaciones entre las variables, para reducir la dimensión.
- ✿ Las componentes principales son variables no correlacionadas y, por tanto, cada una de ellas aporta información independiente de la aportada por las restantes.
- ✿ La varianza de la  $i$ -ésima componente principal es el autovalor  $\lambda_i$ .

**Ejemplo 3.13.** Sobre un conjunto de 146 estudiantes de *Data Mining* se midieron las siguientes variables:

- ✿  $X_1$  peso en kilogramos
- ✿  $X_2$  talla en centímetros
- ✿  $X_3$  ancho de hombros en centímetros
- ✿  $X_4$  ancho de caderas en centímetros



<https://flic.kr/p/BPb18>

Se cuenta con la siguiente información:

- ✿ el vector medio muestral del conjunto es

$$\bar{X} = \begin{pmatrix} 54.25 \\ 161.73 \\ 36.53 \\ 30.10 \end{pmatrix}$$

- ✿ La matriz de varianzas y covarianzas muestral es

$$S = \hat{\Sigma} = \begin{pmatrix} 44.70 & 17.79 & 5.99 & 9.19 \\ 17.79 & 26.15 & 4.52 & 4.44 \\ 5.99 & 4.52 & 3.33 & 1.34 \\ 9.19 & 4.44 & 1.34 & 4.56 \end{pmatrix}$$

- ✿ Los autovectores y autovalores de la matriz de varianzas y covarianzas se muestran en la Tabla 3.10.
- ✿ Los  $p$  valores del test de esfericidad de Bartlett se exhiben en la Tabla 3.11.

### Decisión sobre el número de componentes a considerar

- ✿ Si se quisiera explicar el 90% de la variabilidad, alcanzaría con considerar las primeras dos componentes principales.
- ✿ Aplicando el criterio de Kaiser, la media de las varianzas es  $\bar{V} = \frac{tr(S^2)}{p} = \frac{78.74}{4} = 19.685$ . Podemos ver que los dos primeros valores propios son 58.49 y 15.47 siendo ambos mayores que  $0.7\bar{V} = 13.78$ .

	<b>X1</b>	<b>X2</b>	<b>X3</b>	<b>X4</b>
	0.8328	0.5095	0.1882	0.1063
	0.5029	-0.8552	0.2020	0.1232
	0.1363	-0.0588	0.1114	-0.9826
	0.1867	0.0738	-0.9755	-0.0892
<b>Autovalor</b>	58.49	15.47	2.54	2.24
<b>Porcentaje acumulado</b>	74.27	93.92	97.15	100

Tabla 3.10: Autovalores y autovectores

<b>m</b>	<b><math>\chi^2</math></b>	<b>g.l.</b>	<b>p-valor</b>
0	333.90	9	0.439
1	123.80	5	0.013
2	0.39	2	0.009

Tabla 3.11: Esfericidad de Bartlett

- Considerando los  $p$  valores del test de esfericidad de la Tabla 3.11, se deberían considerar las dos primeras componentes.

Las dos primeras componentes principales con las variables estandarizadas tienen la siguiente expresión analítica

$$Y_1 = 0.8328X_1 + 0.5029X_2 + 0.1363X_3 + 0.1867X_4$$

$$Y_2 = 0.5095X_1 - 0.8552X_2 - 0.0588X_3 + 0.0738X_4$$

### Interpretación de las componentes

- La primera componente es la variable con mayor varianza y tiene todos sus coeficientes positivos. Es una componente de tamaño; es decir, ordena a los estudiantes por tamaño, en el sentido de las variables consideradas.
- La segunda componente tiene coeficientes positivos y negativos, por lo que se trata de una componente de forma. Surgen de este modo dos tipologías de estudiante: el atlético y el de formas redondeadas.
- Las componentes de tamaño y de forma son no correlacionadas.
- Las coordenadas de las primeras componentes principales nos permiten interpretar las similitudes entre los individuos con pérdida mínima de información.



Como observaciones generales importantes, destacamos las siguientes:

- ✿ El estudio de componentes principales, como otros métodos multivariados basados en la matriz de varianzas y covarianzas o la matriz de correlaciones, usan una pequeña porción de la información disponible.
- ✿ Se puede obtener la expresión de las componentes, así como el porcentaje de variabilidad explicada, a partir de la matriz de correlaciones o la matriz de covarianzas. Sin embargo, si se dispone sólo de esta información no pueden obtenerse los *scores*.

**Ejemplo 3.14.** Un grupo de 48 individuos se presentó a una selección de personal que convocó una empresa multinacional. Los candidatos fueron entrevistados y evaluados de acuerdo con 15 criterios. En la Tabla 3.12 se exhiben los criterios considerados los cuales constituyen las variables de interés del estudio.

PRO: prolividad	APA: apariencia personal	FAC: formación académica
AMA: amabilidad	SEG: seguridad	LUC: lucidez
HON: honestidad	VEN: arte para vender	EXP: experiencia
CAR: carácter	AMB: ambición	CON: capacidad para conceptualizar
POT: potencial	ADA: capacidad para adaptarse	GRU: entusiasmo para trabajo grupal

Tabla 3.12: Criterios de evaluación

Cada criterio se evaluó con una calificación dentro de la escala del 0 a 10, siendo 0 completamente insatisfactoria y 10 sobresaliente. La evaluación de cada uno de estos 48 individuos, según estos quince criterios se encuentran en el archivo disponible en <https://goo.gl/1TERF3>.

Las Tablas 3.13, 3.14 y las Figuras 3.23, 3.24 se generan con el Código 3.10.

En la Tabla 3.13 se exhiben las proporciones de variabilidad explicadas por cada una de las componentes principales y la variabilidad explicada acumulada.

En la Figura 3.23 se presenta el gráfico de sedimentación o screeplot correspondiente a las componentes halladas.

En la Tabla 3.14 se encuentran los *loadings* o cargas de las primeras cuatro componentes principales.

En la Figura 3.24 se grafican las cargas de las primeras cuatro componentes principales.

En la Figura 3.25 (generada por el Código 3.10) se presentan los *biplots* de las primeras dos componentes principales y de la tercera y cuarta componente principal.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(gridExtra) # Paquete para acomodar gráficos simultáneos
library(devtools) # Colección de herramientas de desarrollo para paquetes
```

Comp.	Desviaci{on} est{andar}	Proporci{on} de variabilidad	Variabilidad acumulada
PC1	2.741	0.501	0.501
PC2	1.434	0.137	0.638
PC3	1.207	0.097	0.735
PC4	1.094	0.080	0.815
PC5	0.860	0.049	0.864
PC6	0.703	0.033	0.897
PC7	0.593	0.023	0.921
PC8	0.557	0.021	0.941
PC9	0.507	0.017	0.958
PC10	0.430	0.012	0.971
PC11	0.391	0.010	0.981
PC12	0.312	0.007	0.987
PC13	0.298	0.006	0.993
PC14	0.254	0.004	0.998
PC15	0.189	0.002	1.000

Tabla 3.13: Variabilidad explicada

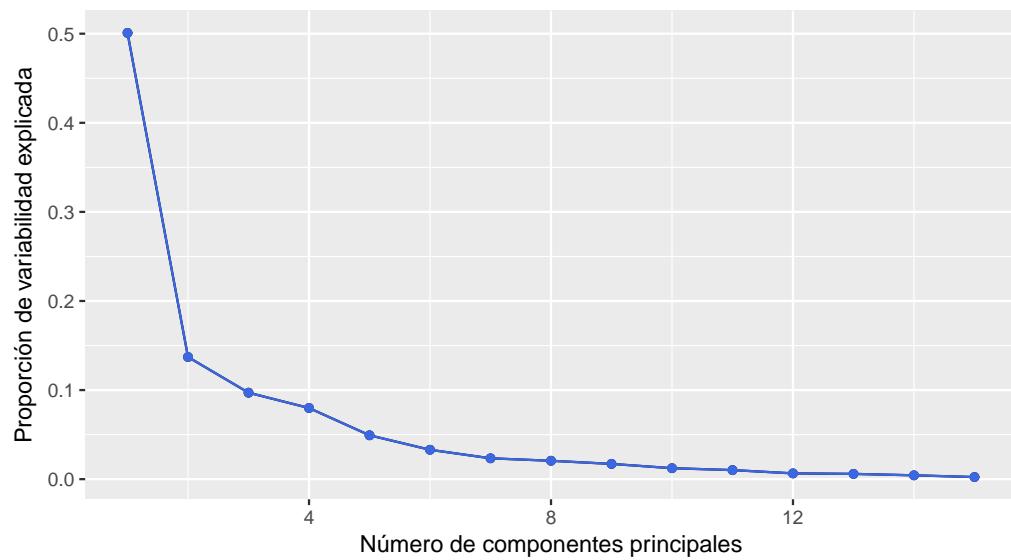


Figura 3.23: Gr{a}fico de sedimentaci{on} para aspirantes

Criterio	PC1	PC2	PC3	PC4
PRO	-0.1624	-0.4288	0.3154	0.0943
APA	-0.2131	0.0353	-0.0229	-0.2622
FAC	-0.0402	-0.2369	-0.4305	-0.6363
AMA	-0.2251	0.1298	0.4658	-0.3454
SEG	-0.2905	0.2489	-0.2410	0.1728
LUC	-0.3149	0.1310	-0.1500	0.0710
HON	-0.1581	0.4054	0.2839	-0.4165
VEN	-0.3243	0.0295	-0.1860	0.1982
EXP	-0.1341	-0.5531	0.0826	-0.0678
CAR	-0.3151	-0.0462	-0.0796	0.1560
AMB	-0.3180	0.0682	-0.2087	0.1993
CON	-0.3315	0.0232	-0.1171	-0.0747
POT	-0.3333	-0.0223	-0.0725	-0.1881
ADA	-0.2592	0.0823	0.4672	0.2014
GRU	-0.2360	-0.4207	0.0892	0.0199

Tabla 3.14: Cargas de los datos de los aspirantes

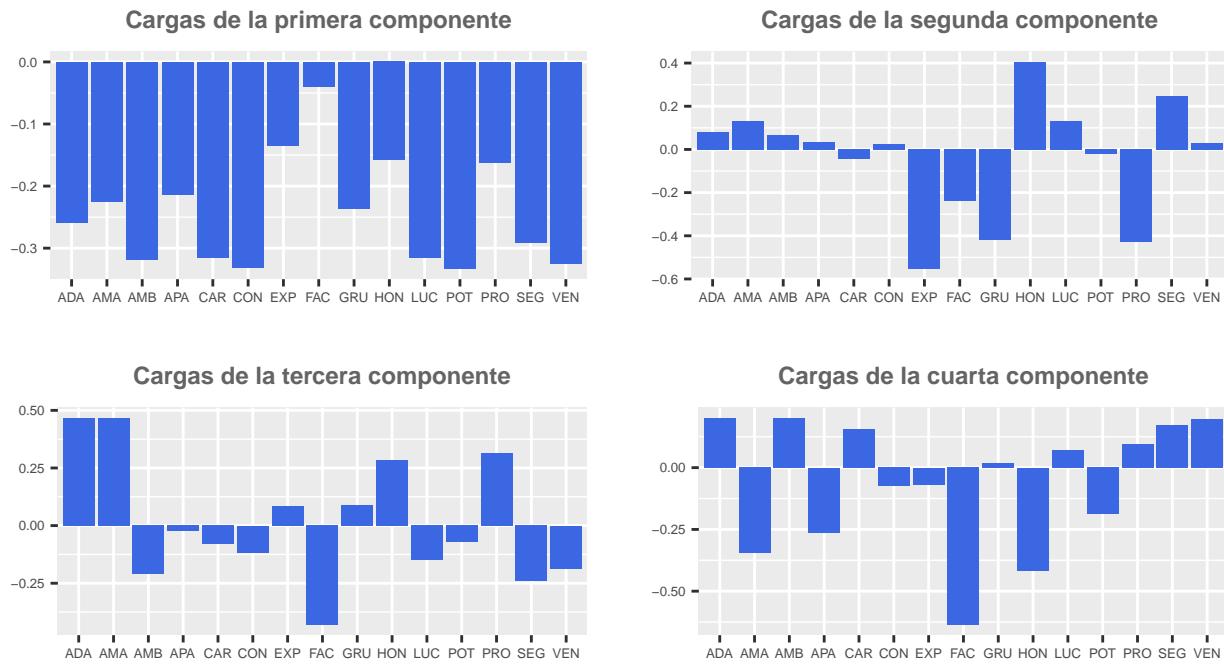


Figura 3.24: Cargas para los aspirantes

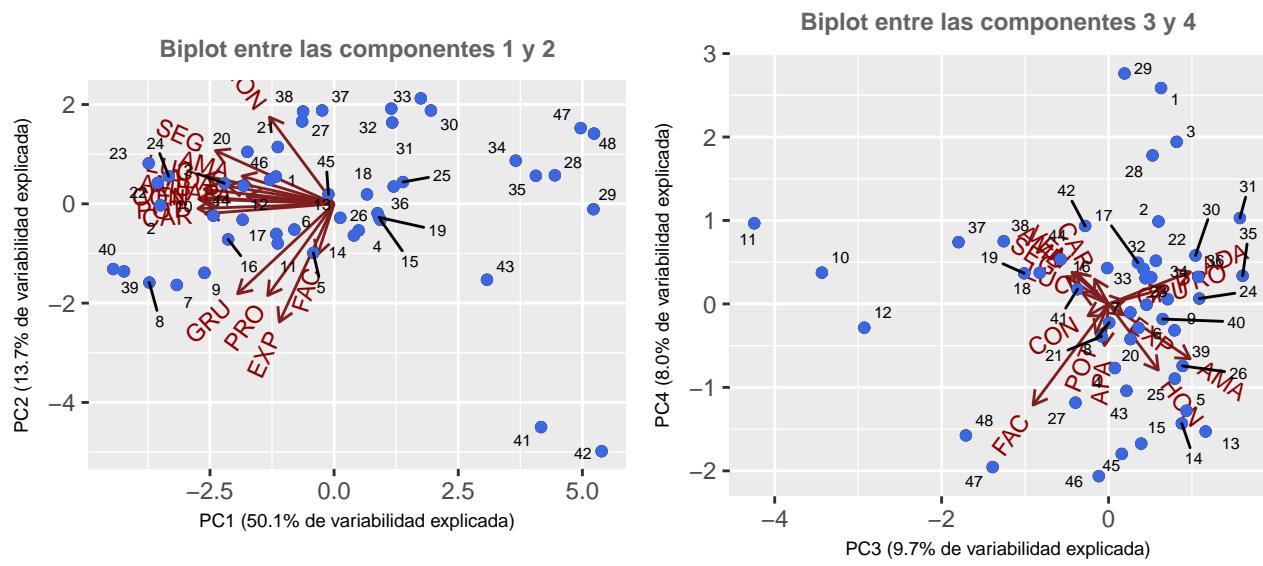


Figura 3.25: *Biplots* para los aspirantes

```

install_github("vqv/ggbiplot") # Instala paquete desde GitHub
library(ggbiplot) # Paquete para visualización de componentes principales
library(ggrepel) # Paquete que manipula etiquetas para gráficos
library(readxl) # Permite leer archivos xlsx

asp=read_excel("C:/.../aspirantes.xlsx")
# Importa la base con la cual se va a trabajar

asp.pca.cor=prcomp(asp[,2:16], center = TRUE, scale. = TRUE)
# Realiza el análisis de componentes principales para las variables estandarizadas

summary(asp.pca.cor) # Muestra la importancia de las componentes principales

ggscreepplot(asp.pca.cor, type = c("pev", "cev")) +
xlab("Número_de_componentes_principales") +
ylab("Proporción_de_variabilidad_explícada") +
geom_line(colour='royalblue') +
geom_point(colour='royalblue')
# Produce un gráfico de sedimentación

# Cálculo de cargas
c1=as.vector(round(asp.pca.cor$rotation[,1],4))
c2=as.vector(round(asp.pca.cor$rotation[,2],4))
c3=as.vector(round(asp.pca.cor$rotation[,3],4))
c4=as.vector(round(asp.pca.cor$rotation[,4],4))

criterio=factor(colnames(asp)[2:16])
datos=data.frame(criterio,c1,c2,c3,c4)
# Acomoda datos para gráfico

load1=ggplot(datos, aes(x=criterio, y=c1))+
geom_bar(stat="identity", position="dodge", fill="royalblue", size=0.5)+ 
ggtitle("Cargas_de_la_primer_a_componente") +
xlab("") +
ylab("") +
theme(axis.text=element_text(size=5),
plot.title=element_text(color="#666666", face="bold", size=9,
hjust=0.5))
# Grafica las cargas de la primera componente principal

load2=ggplot(datos, aes(x=criterio, y=c2))+
geom_bar(stat="identity", position="dodge", fill="royalblue", size=0.5)+ 
ggtitle("Cargas_de_la_segunda_componente") +
xlab("") +
ylab("") +
theme(axis.text=element_text(size=5),
plot.title=element_text(color="#666666", face="bold", size=9,
hjust=0.5))

```

```

# Grafica las segunda de la primera componente principal

load3=ggplot(datos , aes(x=criterio , y=c3))+  

geom_bar(stat="identity" , position="dodge", fill="royalblue" , size=0.5)+  

ggtitle ("Cargas_de_la_tercera_componente") +  

xlab("") +  

ylab("") +  

theme( axis.text=element_text(size=5),  

plot.title=element_text(color="#666666" , face="bold" , size=9,  

hjust =0.5))  

# Grafica las cargas de la tercera componente principal

load4=ggplot(datos , aes(x=criterio , y=c4))+  

geom_bar(stat="identity" , position="dodge", fill="royalblue" , size=0.5)+  

ggtitle ("Cargas_de_la_cuarta_componente") +  

xlab("") +  

ylab("") +  

theme( axis.text=element_text(size=5),  

plot.title=element_text(color="#666666" , face="bold" , size=9,  

hjust =0.5))  

# Grafica las cargas de la cuarta componente principal

grid.arrange(arrangeGrob(load1 , load2 , load3 , load4 , nrow=2))  

# Realiza un gráfico en simultáneo

b12=ggbiplot(asp.pca.cor , obs.scale=1, choices=1:2)+  

geom_point(colour="royalblue") +  

geom_text_repel(aes(label=1:48) , size=2) +  

theme(legend.position="none") +  

xlab("PC1_(50.1%_de_variabilidad_explícada)") +  

ylab("PC2_(13.7%_de_variabilidad_explícada)") +  

ggtitle("Biplot_entre_las_componentes_1_y_2") +  

theme( axis.title=element_text(size=7),  

plot.title=element_text(color="#666666" , face="bold" , size=9,  

hjust =0.5))  

# Genera un biplot entre las componentes 1 y 2

b34=ggbiplot(asp.pca.cor , obs.scale=1, choices=3:4)+  

geom_point(colour="royalblue") +  

geom_text_repel(aes(label=1:48) , size=2) +  

theme(legend.position="none") +  

xlab("PC3_(9.7%_de_variabilidad_explícada)") +  

ylab("PC4_(8.0%_de_variabilidad_explícada)") +  

ggtitle("Biplot_entre_las_componentes_3_y_4") +  

theme( axis.title=element_text(size=7),  

plot.title=element_text(color="#666666" , face="bold" , size=9,  

hjust =0.5))

```

```
# Genera un biplot entre las componentes 3 y 4
grid.arrange(arrangeGrob(b12, b34, nrow=1))
# Realiza un gráfico en simultáneo
```

Código 3.10: Análisis de componentes principales de aspirantes

A partir de las salidas presentadas podríamos decidir:

- ✿ ¿Cuántas componentes principales sería conveniente considerar? basándonos en alguno de los criterios presentados.
- ✿ Si es pertinente la aplicación de esta técnica en este caso.
- ✿ Si las componentes principales son de tamaño o de forma.
- ✿ ¿Qué implica un valor alto en la primera componente o en la segunda?
- ✿ Los nombres que serían adecuados para las componentes principales elegidas.



## 3.5 Componentes principales robustas

La presencia de *outliers* univariados o multivariados puede distorsionar la información de la matriz de covarianza muestral y conducir a resultados erróneos. Luego, se hace necesario contar con técnicas robustas alternativas. Algunas de estas técnicas se basan en métodos de *bootstrap*, las cuales requieren de menos supuestos pero tienen un alto costo computacional.

Otras alternativas propuestas se basan en el reemplazo del vector de medias y de la matriz de covarianzas obtenidas con el método clásico; por el vector de medias y la matriz de covarianzas obtenidos con un método robusto.

Una de las alternativas robustas propuestas es *Minimun Covariance Determinant* (MCD) [21], otra es el estimador de Stahel-Donoho [6] y una tercera propuesta es el *Minimum volume ellipsoid* (MVE) [22].

La idea principal del estimador de Stahel-Donoho es utilizar una ponderación de las observaciones en función de su ‘medida de **alejamiento del conjunto general de datos**’.

La ponderación está basada en proyecciones univariadas sobre la dirección en la cual el alejamiento es máximo. Este estimador para los casos multivariados, tiene dificultades en la medición del grado de alejamiento de las observaciones.

En el caso de conjuntos de datos fuertemente contaminados, se han propuesto correcciones para este estimador que miden esta calidad previamente al cálculo de las ponderaciones [23].

Recordemos que el MCD es un algoritmo para estimar el vector de medias y la matriz de covarianzas a partir de una submuestra cuya principal característica es lograr el determinante mínimo. Se trata de una estimación robusta del vector de medias y de la matriz de covarianzas.

Se espera que en conjuntos de datos que tienen *outliers*, los métodos robustos logren un funcionamiento superior a los métodos clásicos.

**Ejemplo 3.15.** Considerando los datos de nadadores del Ejemplo 3.1, agregamos tres observaciones nuevas que se muestran en la Tabla 3.15

Nadador	Tramo 1	Tramo 2	Tramo 3	Tramo 4
15	18	12	12	10
16	8	15	5	11
17	10	13	12	8

Tabla 3.15: Nuevos nadadores

Observemos los gráficos de las Figuras 3.26 y 3.27 (generados dentro del Código 3.11) para interpretar el objetivo de agregar estos datos.

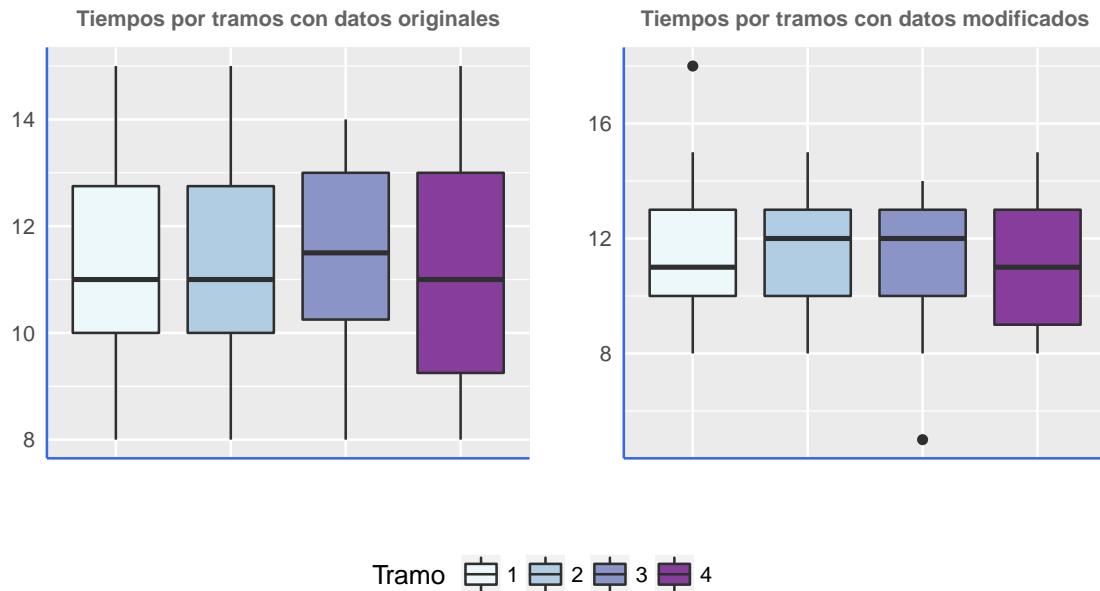


Figura 3.26: Comparación de boxplots para nadadores

Veamos qué efecto tienen estas tres nuevas observaciones sobre el análisis de componentes principales clásico que se muestra en la Tabla 3.16 cuyos datos pueden generarse aplicando el Código 3.11.

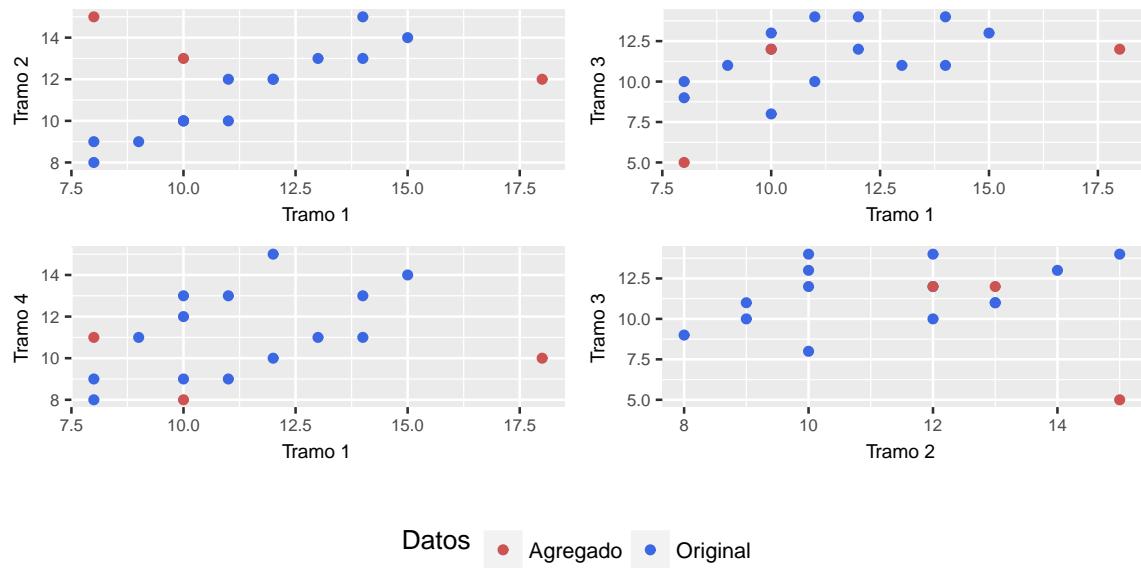


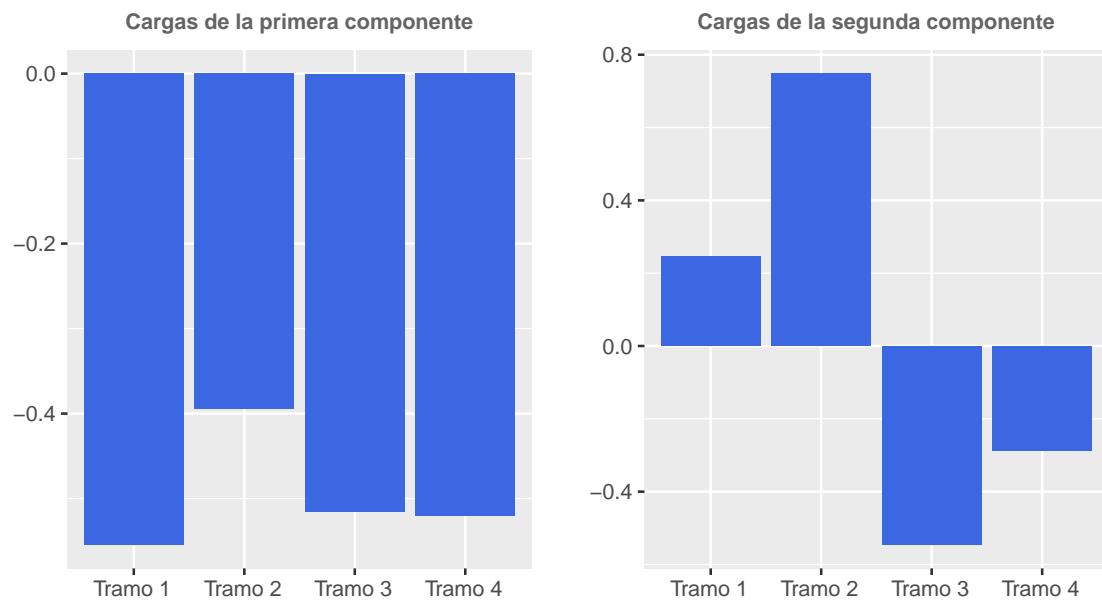
Figura 3.27: Diagramas de dispersión para nadadores

Tabla 3.16: PCA clásico con nuevos datos

Con el objeto de apreciar el impacto de estas tres observaciones nuevas, se sugiere comparar los resultados con los obtenidos en la Tabla 3.4.

Las Figuras 3.28, 3.29 y 3.30 fueron generadas mediante el Código 3.11 con datos extraídos de <https://goo.gl/MJp9hr> y muestran los resultados del análisis clásico aplicado a la base de datos de los nadadores con los datos agregados.

Se sugiere su comparación con las Figuras 3.18, 3.19, 3.17 y 3.20 que muestran los resultados con los datos originales.



**Figura 3.28:** Análisis clásico de cargas para los nadadores con los datos agregados

```
library(ggplot2) # Paquete para confeccionar dibujos
library(gridExtra) # Paquete para acomodar gráficos simultáneos
library(readxl) # Permite leer archivos xlsx
install_github("vqv/ggbiplot") # Instala paquete desde GitHub
library(ggbiplot) # Paquete para visualización de componentes principales
library(ggrepel) # Paquete que manipula etiquetas para gráficos

g_legend<-function(a.gplot){
  tmp <- ggplot_gtable(ggplot_build(a.gplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)}
# Función para obtener leyendas
```

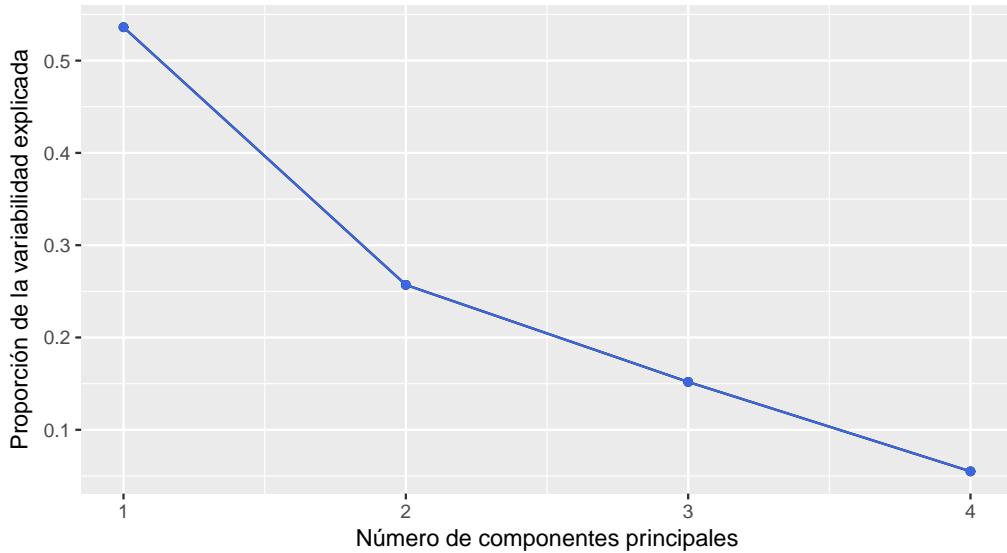


Figura 3.29: Análisis clásico de *screeplot* para los nadadores con los datos agregados

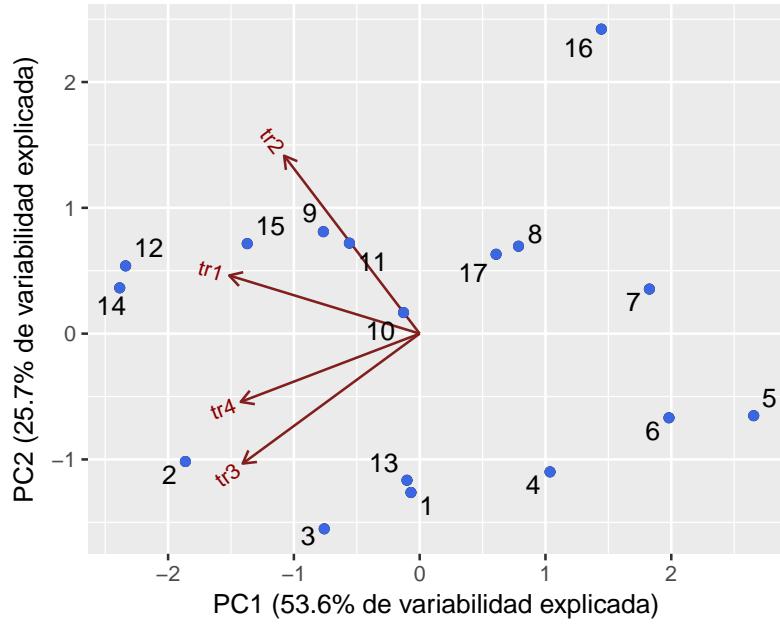


Figura 3.30: Análisis clásico de *biplot* para los nadadores con los datos agregados

```

nad=read_excel("C:/.../nadadores.xlsx")
# Importa la base con la cual se va a trabajar
datos=data.frame(c(nad$tr1, nad$tr2, nad$tr3, nad$tr4),
c(rep("tr1",14), rep("tr2",14), rep("tr3",14), rep("tr4",14)))
# Arregla los datos
colnames(datos)=c("Tiempo", "Tramo")

bp=ggplot(data=datos, aes(y=Tiempo), colour=factor(Tramo)) +
geom_boxplot(aes(x=Tramo, fill=factor(Tramo))) +
ggtitle("Tiempos por tramos con datos originales") +
xlab("") +
ylab("") +
theme(axis.text.x=element_blank(), axis.ticks=element_blank(),
axis.line=element_line(colour="royalblue", size=0.5, linetype="solid")) +
scale_fill_brewer(palette="BuPu", name="Tramo",
breaks=c("tr1", "tr2", "tr3", "tr4"),
labels=c("1", "2", "3", "4")) +
theme(plot.title=element_text(color="#666666", face="bold", size=9,
hjust=0.5)) +
theme(legend.position="bottom")
# Genera un boxplot

nad.cont=rbind(nad, c(15,18,12,12,10), c(16,8,15,5,11), c(17,10,13,12,8))
# Agrega nuevos datos
nad.cont=nad.cont[,-1]
# Quita una columna

datos.cont=data.frame(grupo=c(rep("original",14), rep("nuevo",3)),
c(nad.cont$tr1, nad.cont$tr2, nad.cont$tr3, nad.cont$tr4),
c(rep("tr1",17), rep("tr2",17), rep("tr3",17),
rep("tr4",17)))
colnames(datos.cont)=c("Grupo", "Tiempo", "Tramo")
# Acomoda los datos

bpcont=ggplot(data=datos.cont, aes(y=Tiempo), colour=factor(Tramo)) +
geom_boxplot(aes(x=Tramo, fill=factor(Tramo))) +
ggtitle("Tiempos por tramos con datos modificados") +
xlab("") +
ylab("") +
theme(axis.text.x=element_blank(), axis.ticks=element_blank(),
axis.line=element_line(colour="royalblue", size=0.5,
linetype="solid")) +
scale_fill_brewer(palette="BuPu", name="Tramo",
breaks=c("tr1", "tr2", "tr3", "tr4"),
labels=c("1", "2", "3", "4")) +
theme(plot.title=element_text(color="#666666", face="bold", size=9,
hjust=0.5)) +
theme(legend.position="bottom")

```

```

# Genera un boxplot

mylegend1=g_legend(bpcont)
# Guarda una leyenda

grid.arrange(arrangeGrob(bp + theme(legend.position="none"),
bpcont + theme(legend.position="none"), nrow=1),
mylegend1, nrow=2, heights=c(10, 2.5))
# Realiza un gráfico en simultáneo

datos.tr=split(datos.cont, datos.cont$Tramo)
data=data.frame(datos.tr)
tr1=datos.tr[[1]]
tr2=datos.tr[[2]]
tr3=datos.tr[[3]]
tr4=datos.tr[[4]]
# Acomoda datos para gráfico

p12=ggplot(data, aes(tr1$Tiempo, tr2$Tiempo))+
geom_point(aes(colour=factor(tr1$Grupo))) +
labs(x="Tramo_1", y="Tramo_2", color = "Datos\n") +
scale_color_manual(labels=c("Agregado", "Original"),
values=c("indianred3", "royalblue")) +
theme(axis.title=element_text(size=8),
axis.text=element_text(size=7),
legend.position="bottom")
# Genera un diagrama de dispersión

p13=ggplot(data, aes(tr1$Tiempo, tr3$Tiempo))+
geom_point(aes(colour=factor(tr1$Grupo))) +
labs(x="Tramo_1", y="Tramo_3", color = "Datos\n") +
scale_color_manual(labels=c("Agregado", "Original"),
values=c("indianred3", "royalblue")) +
theme(axis.title=element_text(size=8),
axis.text=element_text(size=7),
legend.position="bottom")
# Genera un diagrama de dispersión

p14=ggplot(data, aes(tr1$Tiempo, tr4$Tiempo))+
geom_point(aes(colour=factor(tr1$Grupo))) +
labs(x="Tramo_1", y="Tramo_4", color = "Datos\n") +
scale_color_manual(labels=c("Agregado", "Original"),
values=c("indianred3", "royalblue")) +
theme(axis.title=element_text(size=8),
axis.text=element_text(size=7),
legend.position="bottom")
# Genera un diagrama de dispersión

```

```

p23=ggplot(data, aes(tr2$Tiempo, tr3$Tiempo))+  

  geom_point(aes(colour=factor(tr2$Grupo)))+  

  labs(x="Tramo_2", y="Tramo_3", color = "Datos\n") +  

  scale_color_manual(labels=c("Agregado", "Original"),  

  values=c("indianred3", "royalblue")) +  

  theme(axis.title=element_text(size=8),  

  axis.text=element_text(size=7),  

  legend.position="bottom")  

# Genera un diagrama de dispersión  

mylegend2=g_legend(p23)  

# Guarda una leyenda  

grid.arrange(arrangeGrob(p12 + theme(legend.position="none"),  

  p13 + theme(legend.position="none"),  

  p14 + theme(legend.position="none"),  

  p23 + theme(legend.position="none"), nrow=2),  

  mylegend2, nrow=2, heights=c(10, 3.5))  

# Realiza un gráfico en simultáneo  

nad.pca=princomp(nad.cont, cor = TRUE, scores = TRUE)  

# Calcula las componentes principales para los nadadores con los datos agregados  

summary(nad.pca) # Muestra la importancia de las componentes principales  

load1=nad.pca$loadings[,1]  

load2=nad.pca$loadings[,2]  

# Calcula las cargas de las componentes principales  

dat=data.frame(cbind(load1,load2))  

x=factor(c("Tramo_1", "Tramo_2", "Tramo_3", "Tramo_4"))  

# acomoda datos para gráfico  

p1=ggplot(dat, aes(x=x, y=load1))+  

  geom_bar(stat="identity", position="dodge", fill="royalblue", size=0.5)+  

  ggtitle("Cargas de la primera componente") +  

  xlab("") +  

  ylab("") +  

  theme(plot.title=element_text(color="#666666", face="bold", size=9,  

  hjust=0.5))  

# Genera un gráfico de barras  

p2=ggplot(dat, aes(x=x, y=load2))+  

  geom_bar(stat="identity", position="dodge", fill="royalblue", size=0.5)+  

  ggtitle("Cargas de la segunda componente") +  

  xlab("") +  

  ylab("") +  

  theme(plot.title=element_text(color="#666666", face="bold", size=9,  

  hjust=0.5))

```

```

# Genera un gráfico de barras

grid.arrange(arrangeGrob(p1, p2, nrow=1))
# Realiza un gráfico en simultáneo

ggscreepplot(nad.pca, type = c('pev', 'cev')) +
  xlab('Número_de_componentes_principales') +
  ylab('Proporción_de_la_variabilidad_explícada') +
  geom_line(colour='royalblue') +
  geom_point(colour='royalblue')
# Produce un gráfico de sedimentación

ggbiplot(nad.pca, obs.scale=1) +
  geom_point(colour="royalblue") +
  geom_text_repel(aes(label=1:17)) +
  theme(legend.position="none") +
  xlab("PC1_(53.6%_de_variabilidad_explícada)") +
  ylab("PC2_(25.7%_de_variabilidad_explícada)")
# Genera un biplot

```

Código 3.11: Análisis de componentes principales de nadadores con datos agregados

A continuación presentaremos las diferentes alternativas robustas para este análisis. Para ello, en la Tabla 3.17 y en las Figuras 3.32 y 3.31 vamos a mostrar los resultados de una de las alternativas robustas y las instrucciones a seguir para aplicar las demás opciones y así poder comparar las salidas obtenidas. Los resultados se obtienen a partir del Código 3.12 con datos extraídos de <https://goo.gl/MJp9hr>.

	PC1	PC2	PC3	PC4
<b>Desviación estándar</b>	1.670	1.031	0.337	0.181
<b>Proporción de variabilidad</b>	0.698	0.266	0.028	0.008
<b>Variabilidad acumulada</b>	0.698	0.963	0.992	1.000

Tabla 3.17: Análisis de componentes principales usando MCD

```

library(readxl) # Permite leer archivos xlsx
install_github("vqv/ggbiplot") # Instala paquete desde GitHub
library(ggbiplot) # Paquete para visualización de componentes principales
library(MASS)
# Paquete con funciones y bases de datos para la librería de Venables y Ripley
library(ggrepel) # Paquete que manipula etiquetas para gráficos

nad=read_excel("C:/.../nadadores.xlsx")
# Importa la base con la cual se va a trabajar

```

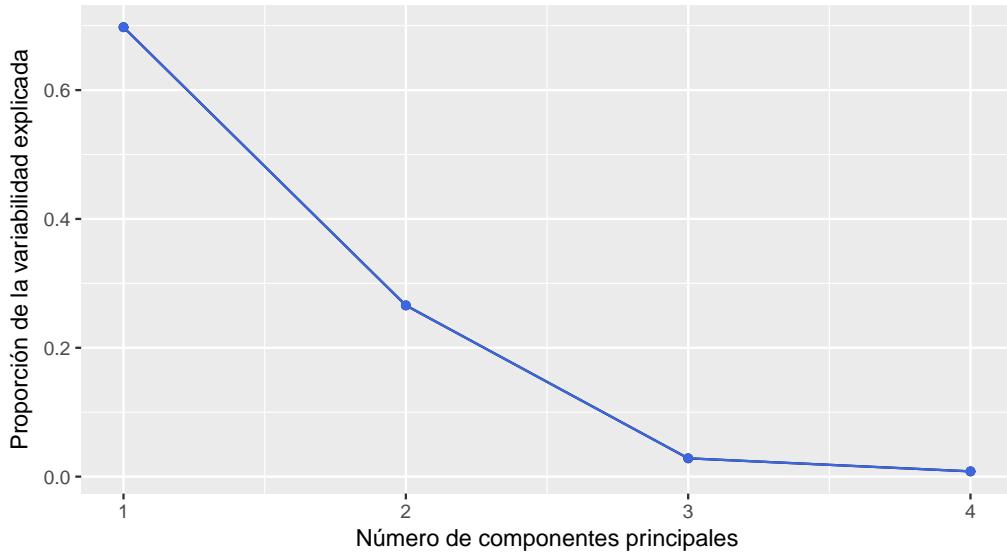


Figura 3.31: Análisis robusto (MCD) de *screeplot* para los nadadores con los datos agregados

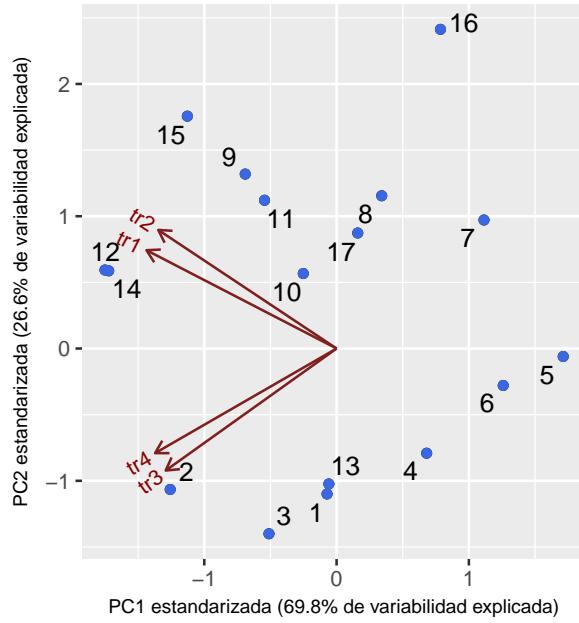


Figura 3.32: Análisis robusto (MCD) de *biplot* para los nadadores con los datos agregados

```

nad.cont=rbind(nad,c(15,18,12,12,10),c(16,8,15,5,11),c(17,10,13,12,8))
# Agrega nuevos datos
nad.cont=nad.cont[,-1]
# Quita una columna

nad.rob.pca1=princomp(nad.cont, cor=TRUE, scores=TRUE,
covmat=MASS::cov.mcd(nad.cont))
summary(nad.rob.pca1)
# Análisis de componentes principales aplicando MCD

ggscreepplot(nad.rob.pca1, type = c('pev', 'cev')) +
xlab('Número_de_componentes_principales') +
ylab('Proporción_de_la_variabilidad_explícada') +
geom_line(colour='royalblue') +
geom_point(colour='royalblue')
# Produce un gráfico de sedimentación

ggbiplot(nad.rob.pca1, choices = 1:2) +
geom_point(colour="royalblue") +
geom_text_repel(aes(label=1:17)) +
theme(legend.position="none") +
xlab("PC1_estandarizada_(69.8%_de_variabilidad_explícada)") +
ylab("PC2_estandarizada_(26.6%_de_variabilidad_explícada)") +
theme(axis.title=element_text(size=8))
# Genera un biplot

#####
# Otras alternativas robustas
#####

nad.rob.pca2=princomp(nad.cont, cor=TRUE, scores=TRUE,
covmat=MASS::cov.rob(nad.cont))
summary(nad.rob.pca2)
# Análisis de componentes principales aplicando el estimador
# resistente de ubicación multivariada y dispersión

nad.rob.pca3=princomp(nad.cont, cor=TRUE, scores=TRUE,
covmat=MASS::cov.mve(nad.cont))
summary(nad.rob.pca3)
# Análisis de componentes principales aplicando MVE

```

Código 3.12: PCA Robusto (nadadores con datos agregados



## 3.6 Ejercitación

**Ejercicio 1.** Consideramos un vector aleatorio  $X = (X_1, X_2, X_3)^t$  de media 0 cuya matriz de varianzas y covarianzas poblacionales está dada por

$$\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

1. Hallar los autovalores y autovectores de la matriz de varianzas y covarianzas.
2. Dar la expresión de las componentes principales  $Y = (Y_1, Y_2, Y_3)^t$  e indicar la proporción de la variabilidad explicada por cada una de ellas.
3. Hallar los *loadings* de la primera componente principal.
4. Hallar los *scores* de las primeras dos componentes principales correspondientes a la observación  $X = (2, 2, 1)^t$ .

**Ejercicio 2.** Considerando los datos de la base disponible en <https://goo.gl/CSZuvH>, se pide:

1. Graficar el *boxplot* de cada una de las variables, indicando si se observa la presencia de valores atípicos.
2. Graficar los diagramas de dispersión de las variables de a pares. Estimar la presencia de correlación entre variables a partir de estos gráficos, indicando si la misma puede considerarse fuerte y el signo de las mismas.
3. Calcular el vector de medias y la matriz de varianzas y covarianzas muestral.
4. Hallar la matriz de correlación muestral. Verificar las estimaciones realizadas visualmente.
5. A partir de estas observaciones, ¿resulta razonable pensar en un análisis de componentes principales para reducir la dimensión del problema?
6. Hallar la primera componente principal y graficar sus coeficientes mediante barras verticales.
7. Indicar qué porcentaje de la variabilidad total logra explicar esta componente. Explicar si se trata de una componente de tamaño o de forma. Es posible ordenar las promotoras en función de esta componente? Si la respuesta es afirmativa, ¿cuál es la mayor y cuál la menor? En caso contrario, explicar por qué no es posible ordenarlos.

**Ejercicio 3.** Consideremos el siguiente conjunto de datos

$$X = \begin{pmatrix} 3 & 6 \\ 5 & 6 \\ 10 & 12 \end{pmatrix}$$

1. Calcular la matriz de covarianza, sus autovalores y autovectores.
2. Hallar las componentes principales y su contribución porcentual a la varianza total.
3. Graficar los datos en  $\mathbb{R}^2$  teniendo en cuenta la base original y luego la base de los dos primeros ejes.
4. Repetir los cálculos con los datos estandarizados e interpretar los resultados obtenidos
5. Verificar que los dos primeros autovectores son ortogonales entre sí. Representar gráficamente estos dos vectores en un gráfico bidimensional y trazar rectas desde el origen hasta la ubicación de cada uno de los vectores en el gráfico.

**Ejercicio 4.** Sea

$$\Sigma = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

la matriz de varianzas y covarianzas poblacionales correspondiente al vector aleatorio  $X = (X_1, X_2, X_3)^t$  siendo:

**X<sub>1</sub>**: puntuación media obtenida en las asignaturas de Econometría

**X<sub>2</sub>**: puntuación media obtenida en las asignaturas de Derecho

**X<sub>3</sub>**: puntuación media obtenida en asignaturas libres

Los datos corresponden a un conjunto de alumnos de la carrera de economía.

1. Calcular los autovalores de la matriz  $\Sigma$ .
2. Interpretar la segunda componente principal sabiendo que el autovector correspondiente es  $w = (0.5744, -0.5744, 0.5744)$ .
3. Cómo se debería interpretar el hecho que un estudiante tuviera segunda una puntuación en la componente principal muy inferior a la de sus compañeros?
4. ¿Cuántas componentes principales serán necesarias para explicar al menos el 80% de la variancia total del conjunto?

**Ejercicio 5.** El conjunto de datos disponible en <https://goo.gl/9Mg4JD>, se refiere a 20 observaciones de suelo, donde se midieron

$x_1$ : contenido de arena,

$x_2$ : contenido de cieno,

$x_3$ : contenido de arcilla,

$x_4$ : contenido de materia orgánica,

$x_5$ : acidez según PH.

1. Comparar los resultados del Análisis en Componentes Principales para la matriz de covarianza y para la matriz de correlación.
2. Los porcentajes de variabilidad que logran explicar cada una de las componentes, ¿son los mismos?
3. ¿Cambia el orden de las componentes?
4. ¿Cambian los *loadings* de las componentes?
5. ¿Cuál de los dos análisis parece más adecuado? ¿Por qué?

**Ejercicio 6.** Los datos disponibles en <https://goo.gl/FVqX22> se refieren a 49 aves, 21 de los cuales sobrevivieron a una tormenta.

1. Estandarizar las variables y calcular la matriz de covarianzas para las variables estandarizadas.
2. Verificar que ésta es la matriz de correlación de las variables originales.
3. ¿Parece adecuado en este caso un análisis de componentes principales? ¿Qué indica el autovalor para una componente principal?
4. ¿Cuántas componentes son necesarias para explicar el 80% de la varianza total? Realizar el grafico de sedimentación, fundamentando la respuesta con este gráfico.
5. ¿Cuál es la expresión de la primera componente principal?
6. ¿Cómo queda expresada la primera componente principal (en función del autovector correspondiente y de las variables)?
7. Encontrar las coordenadas del pájaro 11 en las nuevas componentes.

8. Representar gráficamente en el plano: Eje 1 vs. Eje 2, Eje 1 vs. Eje 3, Eje 2 vs. Eje 3. Interpretar los tres primeros ejes.
9. Realizar un gráfico donde se observen las aves en los nuevos ejes 1 y 2, resaltando con distinto color el grupo de los que sobrevivieron.
10. Utilizar el Análisis en Componentes Principales como método para encontrar *outliers*.

**Ejercicio 7.** Con el objetivo de obtener índices útiles para la gestión hospitalaria basados en técnicas estadísticas multivariantes descriptivas, se recogió información del Hospital de Algeciras correspondiente a los ingresos hospitalarios del período 2007-2008. Se estudiaron las siguientes variables habitualmente monitorizadas por el Servicio Andaluz de Salud, del Sistema Nacional de Salud Español:

**NI:** número de ingresos

**MO:** tasa de mortalidad

**RE:** número de reingresos

**NE:** número de consultas externas

**ICM:** índice cardíaco máximo

**ES:** número de estancias

Las variables se midieron en un total de 22486 ingresos. En el archivo disponible en <https://goo.gl/V2UQ1p> se aprecia la distribución de los valores obtenidos en las variables listadas por los servicios del hospital de Algeciras, Andalucía, España.

1. Calcular las primeras dos componentes principales.
2. Graficar las cargas y explicar la interpretación de las componentes principales.
3. ¿Qué porcentaje de variabilidad logra captar cada una de ellas? Graficar el *scree plot*.
4. ¿Parece adecuado considerar dos componentes principales?
5. Hallar la correlación entre las nuevas variables y las originales.
6. Ordenar los servicios en función de su puntuación en cada una de las dos primeras componentes principales, indicando cuáles son los servicios más demandados y los más complejos.
7. Representar un *biplot* y buscar servicios similares, asociaciones entre las variables. Verificar en este gráfico la representación de las variables originales en las componentes.



# Capítulo 4

## Contrastes de independencia y homogeneidad

*Si tu experimento necesita un estadista, hubiera sido necesario hacer un experimento mejor.*

— Ernest Rutherford

### 4.1 Contraste de Hipótesis

El contraste de hipótesis se propone investigar si una propiedad, que se supone es válida en una cierta población, es compatible con lo observado en una muestra de dicha población.

Se trata de un procedimiento que permite elegir entre dos posibles hipótesis antagónicas o simplemente excluyentes.

Todo contraste de hipótesis estadísticas se basa en la formulación de dos hipótesis mutuamente excluyentes:

- ✿ Hipótesis nula, denotada por  $H_0$
- ✿ Hipótesis alternativa, denotada por  $H_1$

*¿Qué se debe asignar a  $H_0$ ? ¿Y a  $H_1$ ?*

**La hipótesis  $H_0$  es la que se contrasta.** En general, es una afirmación concreta sobre la forma de una distribución de probabilidad, sobre el valor de alguno de los parámetros de una distribución o sobre la vinculación entre distribuciones o parámetros. El nombre de **nula** se refiere a ‘sin valor, efecto o consecuencia’, lo cual sugiere que  $H_0$  debe identificarse con la hipótesis *status quo*; es decir, no habría cambio, diferencia o mejora a partir de la situación actual.

Es importante destacar que la hipótesis nula **nunca se considera probada**, solamente puede ser rechazada o no por la evidencia empírica.

Por ejemplo, la hipótesis de que dos poblaciones tienen la misma media puede ser rechazada fácilmente cuando las mismas difieren notablemente al analizar muestras suficientemente grandes de ambas poblaciones. Sin embargo, no puede ser ‘demostrada’ mediante muestreo puesto que siempre cabe la posibilidad de que las medias difieran en una cantidad lo suficientemente pequeña para que no pueda ser detectada, aún en el caso de que la muestra sea muy grande.

Podemos resumir diciendo que la lógica del contraste de hipótesis se basa en:

- ✿ Se encuentra, o no, evidencia en contra de la hipótesis de nulidad planteada.
- ✿ En caso de no haberse encontrado evidencia, por el momento, no hay motivos para dejar de sostenerla.
- ✿ Si, por el contrario, se encuentra evidencia; se tienen los motivos necesarios para rechazarla.

Dado que descartaremos o no la hipótesis nula en función de la información disponible, o **evidencia empírica**, que surge a partir de las muestras obtenidas, no será posible garantizar que la decisión tomada sea la correcta. Es decir, podría surgir una diferencia como producto de seleccionar la muestra al azar.

La hipótesis alternativa, también llamada **hipótesis del investigador**, es muchas veces la negación de la hipótesis nula, pero puede ser simplemente excluyente sin llegar a incluir todo lo que  $H_0$  excluye.

*¿A qué se refiere una hipótesis estadística?*

Podemos distinguir dos grandes grupos de hipótesis:

- ✿ **Hipótesis paramétricas:** se refieren al valor de algún parámetro, o relaciones entre valores de varios parámetros.

En este caso,  $H_0$  asigna un valor específico o un intervalo de valores al parámetro en cuestión. La opción de igualdad siempre debe formar parte de  $H_0$ .

- ✿ **Hipótesis no paramétricas o de libre distribución:** no se refieren al valor de un parámetro. Las mismas pueden referirse a una forma distribucional, a una estructura relacional de las variables o a la pertenencia de cierta familia.

Recordemos que los **parámetros** son constantes que caracterizan a una distribución teórica o poblacional. Por ejemplo, en la distribución Normal existen dos parámetros:  $\mu$  que indica la media, y  $\sigma$  que indica la dispersión.

Para poder estimar el valor de los parámetros, se utilizan funciones de la muestra denominadas **estadísticos**, cuya distribución se denomina **distribución muestral**.

Algunos ejemplos de estadísticos son los siguientes:

- ✿ la **media muestral**  $\bar{X}$
- ✿ la **mediana muestral**  $\tilde{X}$
- ✿ la **varianza muestral**  $S^2$
- ✿ el **rango muestral**  $X^{(n)} - X^{(1)}$

*¿En qué se basa la regla de decisión?*



El **estadístico de contraste** es un resultado que se obtiene a partir de la muestra y que cumple dos condiciones:

- ✿ proporcionar información empírica relevante sobre la afirmación propuesta en  $H_0$ ,
- ✿ poseer una distribución muestral conocida.

Luego, se debe definir un criterio que permita decidir si se rechaza o no la hipótesis nula  $H_0$ .

Suponiendo cierta la hipótesis de nulidad, conocemos la distribución del estadístico de contraste y partimos el soporte del estadístico en dos regiones o zonas mutuamente excluyentes, que denominaremos **región crítica** o de rechazo y **región de no rechazo** (algunos textos la denominan de aceptación).

- ✿ **región crítica** o **región de rechazo** es el área del soporte de distribución muestral que corresponde a los valores del estadístico de contraste que se encuentran muy alejados de la afirmación establecida. Siendo cierta  $H_0$  es muy poco probable que el estadístico de contraste caiga en esta región.
- ✿ **región de no rechazo** es el área del soporte de la distribución muestral correspondiente a los valores del estadístico de contraste próximos a la afirmación establecida en  $H_0$ . Es decir, los valores del estadístico de contraste que tienen una probabilidad alta de ocurrir siendo  $H_0$  cierta.

El o los **valores críticos** son valores del estadístico de contraste que delimitan la región de rechazo.

El área de la región crítica o de rechazo se denomina **nivel de significación** o **nivel de riesgo** y se representa con la letra  $\alpha$ . De esta manera, el área asignada a la región de no rechazo es  $1 - \alpha$ .

Una vez definidas estas dos zonas, la regla de decisión consiste en:

- \* **rechazar**  $H_0$  si el estadístico de contraste toma un valor perteneciente a la zona de rechazo.
- \* **no rechazar**  $H_0$  si el estadístico de contraste toma un valor perteneciente a la zona de no rechazo.

Entonces, el tamaño de la zona de rechazo o crítica y la de no rechazo, se determina fijando el valor de  $\alpha$ ; es decir, fijando el nivel de significación con el que se desea trabajar. Habitualmente se consideran para el nivel de significación las proporciones 0.10, 0.05 o 0.01.

La forma en que se divide la distribución muestral en zona de rechazo y de no rechazo depende de si el contraste es **bilateral**, situando la región de rechazo en los dos extremos o colas; o **unilateral**, en el cual se sitúa la región de rechazo en uno de los dos extremos o colas.

La zona crítica debe situarse donde puedan aparecer los valores muestrales incompatibles con  $H_0$  y compatibles con  $H_1$ .

Teniendo esto en cuenta, las reglas de decisión se basan en lo siguiente:

- \* Para **contrastos bilaterales** donde la hipótesis alternativa da lugar a una región crítica ‘a ambos lados’ del valor del parámetro, lo más usual es que cada una de las dos regiones de rechazo tengan la misma área ( $\alpha/2$ ). Se rechaza  $H_0$  cuando el estadístico de contraste pertenece a la zona crítica. Esto ocurre cuando el estadístico de contraste toma un valor muy alejado del supuesto en la hipótesis nula. Siendo cierta  $H_0$  la probabilidad de obtener un valor dentro de la región crítica derecha es  $\alpha/2$  y lo mismo en la izquierda.
- \* Para **contrastos unilaterales** en los cuales la región crítica está ‘a un solo lado’, pudiendo ser derecho o izquierdo, el área de la zona crítica o de rechazo es  $\alpha$ . Se rechaza  $H_0$  si el estadístico de contraste pertenece a la zona crítica; es decir, si el estadístico de contraste toma un valor mayor que el valor crítico si la cola es a la derecha y menor que el valor crítico si la cola es a la izquierda.

Una vez planteadas las hipótesis, se debe:

1. Establecer los supuestos.
2. Definir el estadístico de contraste y hallar su distribución muestral.
3. Fijar el nivel de significación.
4. Elegir convenientemente el tamaño muestral  $n$ .



Figura 4.1: Ejemplo de regiones en un contraste bilateral

##### 5. Deducir la región crítica para una muestra de tamaño $n$ .

Recién entonces, se toma una muestra aleatoria de tamaño  $n$  y se aplica el test. Se decide luego

- ✿ rechazar  $H_0$  si el estadístico de contraste pertenece a la zona crítica.
- ✿ no rechazar  $H_0$  si el estadístico pertenece a la zona de no rechazo.

En caso de rechazar  $H_0$ , se está afirmando que la hipótesis nula es falsa; es decir, que hemos conseguido probar que esa hipótesis es falsa con una probabilidad  $\alpha$  de equivocarnos. Por el contrario, si no se rechaza, no significa que estemos afirmando que la hipótesis sea verdadera. Simplemente, decimos que no tenemos evidencia empírica suficiente para rechazarla y que la misma se considera compatible con los datos. Es importante tener en cuenta que, aunque se mantenga y no se rechace  $H_0$ , nunca se puede afirmar que ésta sea verdadera.

La toma de decisión puede implicar dos tipos de error:

- ✿ **Error de tipo I** es el que se comete cuando se decide rechazar la hipótesis nula  $H_0$  siendo en realidad verdadera. La probabilidad de cometer este error resulta

$$P(\text{Rechazar } H_0 / H_0 \text{ es verdadera}) = \alpha$$

- ✿ **Error de tipo II** es el que se comete cuando se decide no rechazar la hipótesis nula  $H_0$  siendo en realidad falsa. La probabilidad de cometer este error resulta

$$P(\text{No rechazar } H_0 / H_0 \text{ es falsa}) = \beta$$

Por lo tanto tenemos que

- ✿  $1 - \alpha$  es la probabilidad de tomar una decisión correcta cuando  $H_0$  es verdadera.
- ✿  $1 - \beta$  es la probabilidad de tomar una decisión correcta cuando  $H_0$  es falsa.

El problema de usar un procedimiento basado en datos muestrales es que, debido a la variabilidad del muestreo, la muestra obtenida puede resultar no representativa, y por ende, conducir a un error.

	No se rechaza $H_0$	Se rechaza $H_0$
$H_0$ es verdadera	Decisión correcta	Error de tipo I ( $P = \alpha$ )
$H_0$ es falsa	Error de tipo II ( $P = \beta$ )	Decisión correcta

Tabla 4.1: Errores en un test

#### 4.1.1 Nivel de significación

El **nivel de significación** de un test se puede definir como la máxima probabilidad de rechazar la hipótesis nula  $H_0$  cuando ésta es cierta. Por lo tanto, el nivel de significación representa el riesgo máximo admisible para rechazar  $H_0$  siendo ella cierta. Este nivel debe ser elegido por el investigador antes de realizar el contraste, para que el mismo no influya sobre su decisión.

Por otro lado, la probabilidad de cometer un error de tipo II,  $\beta$ , es un valor desconocido que depende de los siguientes factores:

- \* la hipótesis  $H_1$  que se considere verdadera,
- \* el valor de  $\alpha$ ,
- \* el tamaño del error típico (desviación típica muestral) utilizado para efectuar el contraste.

Cuanto más se aleje el verdadero valor del valor supuesto en  $H_0$ , más se alejará la distribución del estadístico de contraste bajo  $H_0$  de la del estadístico bajo  $H_1$ . En consecuencia, más pequeña será el área  $\beta$  marcada con rojo en la Figura 4.2. Así, el valor de  $\beta$  depende del valor concreto que se haya supuesto en  $H_1$  para el parámetro de interés. Es más,  $\alpha$  y  $\beta$  se relacionan de forma inversa, siendo menor  $\beta$  cuanto mayor es  $\alpha$ . El solapamiento entre las curvas correspondientes a uno y otro parámetro, establecidos en  $H_0$  y  $H_1$ , será tanto mayor cuanto menor sea la distancia entre ambos parámetros ( $\mu_0$  y  $\mu_1$  de la Figura 4.2).

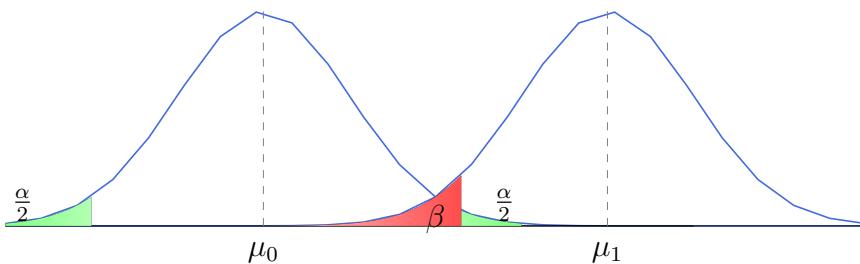


Figura 4.2: Representación de los errores de un test

### 4.1.2 Relaciones entre los errores de tipo I y II

Un buen procedimiento estadístico es aquel para el cual la probabilidad de cometer cualquier tipo de error resulta pequeña. La elección de un valor particular para efectuar el corte de la región de rechazo determina las probabilidades de errores de tipo I y de tipo II.

El valor del nivel de significación,  $\alpha$ , se establece ‘a priori’ por lo tanto es único. Sin embargo, hay un valor diferente de  $\beta$  por cada valor del parámetro escogido en  $H_1$ .

Como  $\alpha$  es fijado por el investigador, trataremos de elegir un procedimiento tal que la probabilidad de cometer el error de tipo II sea la menor posible.

Usualmente, se diseñan los contrastes de tal manera que el nivel de significación sea  $\alpha = 0.05$ . Aunque en ocasiones se usan los valores 0.10 o 0.01 para adoptar condiciones más relajadas o más estrictas, respectivamente.

### 4.1.3 Potencia de un contraste

Se llama **potencia de la prueba** y se denota por  $\pi$ , a la probabilidad de decidir por  $H_1$  cuando ésta es cierta; es decir,

$$\pi = P(\text{Rechazar } H_0 / H_0 \text{ es falsa}) = 1 - \beta$$

El concepto de potencia se utiliza para medir la *bondad* de un contraste de hipótesis. Cuanto más lejana se encuentra la hipótesis  $H_1$  de  $H_0$ , menor es la probabilidad de incurrir en un error tipo II y, por consiguiente, la potencia tomará valores más próximos a 1.

Si la potencia en un contraste es siempre próxima a 1, entonces se dice que la prueba de hipótesis es muy potente para contrastar  $H_0$  ya que en ese caso las muestras serán, con alta probabilidad, incompatibles con  $H_0$  cuando  $H_0$  sea falsa.

Por tanto puede interpretarse la potencia de un contraste como su sensibilidad o capacidad para detectar la falsedad de la hipótesis nula. Dicho con otras palabras, la potencia cuantifica la capacidad del criterio utilizado para rechazar  $H_0$  cuando ésta es falsa.

Es deseable en un contraste de hipótesis que las probabilidades de ambos tipos de error sean lo más pequeñas posibles. Sin embargo, con una muestra de tamaño establecido, disminuir la probabilidad del error de tipo I ( $\alpha$ ), conduce a incrementar la probabilidad del error de tipo II ( $\beta$ ).

Una estrategia válida para aumentar la potencia del contraste; esto es, disminuir la probabilidad de error de tipo II, es aumentar el tamaño muestral, lo que en la práctica conlleva a un incremento de los costos del estudio que se quiere realizar.

El concepto de potencia nos permite valorar cuál, entre dos contrastes con la misma probabilidad de error de tipo I, es preferible.

El objetivo consiste en tratar de escoger entre todos los contrastes posibles con un valor de  $\alpha$  establecido, aquel que tenga mayor potencia; esto es, menor probabilidad de incurrir en el error de tipo II ( $\beta$ ).

#### 4.1.4 Concepto de *p*-valor

Cuando se realiza un contraste de hipótesis se sabe que, a partir del nivel de significación, se genera una partición del soporte de la distribución muestral en la zona de aceptación y la región crítica o de rechazo.

El ***p*-valor** es la probabilidad, suponiendo cierta  $H_0$ , de obtener una muestra como la obtenida o más alejada aún que la hipótesis de nulidad, en el sentido de la hipótesis alternativa.

Cuanto menor resulte el *p*-valor, mayor es la seguridad con la que rechazamos  $H_0$ . El *p*-valor resulta de esta forma, una manera de cuantificar la seguridad del rechazo de  $H_0$ .

## 4.2 Contrastes de homogeneidad e independencia

Presentaremos fundamentalmente el **test Chi cuadrado**, que es uno de los más conocidos para estudiar datos categóricos [18]. Mostraremos que las hipótesis que se pueden testear con el estadístico Chi cuadrado, dependen de cómo fueron obtenidos los datos. Es decir, dependen de la intención del estudio lo cual determina el muestreo.

Los datos categóricos pueden ser presentados en tablas de tamaño  $r \times c$ , siendo  $r$  el número de filas y  $c$  el número de columnas. En algunas aplicaciones, las  $r$  filas pueden hacerse corresponder a resultados posibles de una variable categórica y las  $c$  columnas, a diferentes poblaciones muestradas.

Por ejemplo, interesa comparar el grado de satisfacción clasificado en nulo, regular o bueno, que experimentaron los clientes atendidos mediante dos sistemas diferentes. Las **poblaciones** en este caso quedan determinadas por los tratamientos, por lo tanto podemos pensar como si tuviéramos dos muestras, una de cada población. Estos resultados podrían extenderse a más tratamientos; es decir, a un número mayor de poblaciones.

En otras ocasiones, las filas y las columnas corresponden a dos criterios diferentes para clasificar los sujetos observados a partir de una única población.

### 4.2.1 Contraste de independencia

Veremos que existen esencialmente dos métodos de muestreo que dan lugar a las frecuencias de una tabla de doble entrada o de contingencia de tamaño  $r \times k$ .

En un **test de independencia** el tipo de muestreo, llamado *cross-sectional* o *transversal*, proviene de seleccionar una muestra aleatoria de  $n$  sujetos de una población y luego determinar para cada sujeto el nivel de la característica  $A$  y el nivel de la característica  $B$ . Sólo se especifica *a priori* el tamaño total de la muestra  $n$ . Muchos de los estudios que se realizan en investigaciones económicas, médicas y sociales pertenecen a esta categoría.

Los siguientes son algunos ejemplos ilustrativos.

- \* En un estudio sobre la calidad de los cuidados médicos de los distintos servicios de un Centro de Salud brindados a sus pacientes, todas las nuevas admisiones realizadas son clasificadas

según el servicio al que fueron derivadas y el nivel de satisfacción manifestado por el paciente respecto a la atención recibida.

- ✿ En un estudio para analizar si el consumo de alcohol está asociado con la edad, se seleccionan individuos y se los clasifica según la edad en tres categorías (jóvenes, adultos y de tercera edad) y según su consumo de alcohol en otras tres categorías( nada, poco o mucho).
- ✿ En un estudio para analizar si el hábito de fumar depende del sexo de un individuo, se toma una muestra de tamaño  $n$  y se clasifica a los individuos según el sexo y si tienen el hábito de fumar o no.

**Ejemplo 4.1.** Un estudio realizado con 80 personas, se refiere a la relación entre la cantidad de horas de programas con escenas de violencia vistas durante una semana en la televisión y la edad del televidente categorizada como joven, adulto y mayor. Los resultados obtenidos se muestran en la Tabla 4.2.

Nivel de violencia	Edad (en años)			Totales
	Joven (16-34)	Adulto (35-54)	Mayor (55 o más)	
Poca	8	12	20	40
Mucha	18	15	7	40
Totales	26	27	27	80

Tabla 4.2: Nivel de violencia según la edad

Nos interesa saber si los datos indican que el consumo de violencia en programas de televisión está asociados o no con la edad del televidente, con un nivel de significación del 5%.

Para la muestra aleatoria considerada en la cual se consultó sobre la edad y el consumo de programas televisivos con contenido de violencia, se calculan las frecuencias relativas y se presentan en la Tabla 4.3.

Nivel de violencia	Edad (en años)			Totales
	Joven (16-34)	Adulto (35-54)	Mayor (55 o más)	
Poca	0.1000	0.1500	0.2500	0.5
Mucha	0.2250	0.1875	0.0875	0.5
Totales	0.3250	0.3375	0.3375	1

Tabla 4.3: Frecuencias relativas del nivel de violencia según la edad

Simbolizando con  $PV$  a poca violencia y con  $MV$  a mucha violencia, la estructura general de la Tabla 4.3 se presenta en la Tabla 4.4.

Nivel de violencia	Edad (en años)			Totales
	Joven (16-34)	Adulto (35-54)	Mayor (55 o más)	
Poca	$P(16 - 34 \cap PV)$	$P(35 - 54 \cap PV)$	$P(\geq 55 \cap PV)$	$P(PV)$
Mucha	$P(16 - 34 \cap MV)$	$P(35 - 54 \cap MV)$	$P(\geq 55 \cap MV)$	$P(MV)$
Totales	$P(16 - 34)$	$P(35 - 54)$	$P(\geq 55)$	1

Tabla 4.4: Formato teórico del nivel de violencia según la edad

Las probabilidades  $P(PV)$ ,  $P(MV)$ ,  $P(16-34)$ ,  $P(35-54)$  y  $P(\geq 55)$  se denominan **probabilidades marginales** y las que aparecen en las celdas interiores de la tabla se llaman **probabilidades conjuntas**.

Con el objeto de comprender la lógica del test, recordemos el concepto de independencia entre eventos **y la definición de probabilidad condicional**.

**Definición 4.2.**  $P(A/B)$ : probabilidad de que ocurra A sabiendo que ocurrió B, o bien probabilidad de A condicional a la ocurrencia de B.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{con} \quad P(B) > 0$$

**Definición 4.3.** Decimos que dos eventos  $A$  y  $B$ , asociados a un mismo experimento, son **independientes** cuando la ocurrencia de uno de ellos no afecta la probabilidad de ocurrencia del otro; es decir, si  $P(B) > 0$ ,

$$P(A/B) = P(A) \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

En particular, diremos entonces que en nuestra población, la edad y la cantidad de horas de programas con escenas de violencia consumida serán independientes si y sólo si cada probabilidad conjunta (encontradas en las casillas interiores de la tabla) es el producto de las correspondientes probabilidades marginales (totales de las filas y de las columnas de la tabla). Los cálculos de este ejemplo se presentan en la Tabla 4.5.

Resulta evidente que las probabilidades observadas en esta muestra no coinciden con las esperadas bajo el supuesto teórico de independencia. La pregunta que debemos hacernos ahora es

*¿El apartamiento entre ambas es significativo o puede ser debido a la variabilidad muestral?*



A partir de la Tabla 4.5, es posible evaluar la hipótesis nula que plantea que para la población de interés las dos variables son independientes, versus la hipótesis alternativa de que no lo son; es decir, que las variables están asociadas de alguna manera. Como en todo test, el problema será decidir si la evidencia muestral es suficiente para rechazar la hipótesis nula o no. Debemos analizar entonces cuán probable es obtener una tabla como la obtenida a partir de los datos muestrales, o más alejada aún cuando la muestra se tomó de una población en las que las variables son independientes.

Probabilidad conjunta	Resultado	Producto marginal	Resultado
$P(16 - 34 \cap PV)$	0.1000	$P(16 - 34) \cdot P(PV)$	$0.3250 \cdot 0.5 = 0.16250$
$P(35 - 54 \cap PV)$	0.1500	$P(35 - 54) \cdot P(PV)$	$0.3375 \cdot 0.5 = 0.16875$
$P(\geq 55 \cap PV)$	0.2500	$P(\geq 55) \cdot P(PV)$	$0.3375 \cdot 0.5 = 0.16875$
$P(16 - 34 \cap MV)$	0.2250	$P(16 - 34) \cdot P(MV)$	$0.3250 \cdot 0.5 = 0.16250$
$P(35 - 54 \cap MV)$	0.1875	$P(35 - 54) \cdot P(MV)$	$0.3375 \cdot 0.5 = 0.16875$
$P(\geq 55 \cap MV)$	0.0875	$P(\geq 55) \cdot P(MV)$	$0.3375 \cdot 0.5 = 0.16875$

Tabla 4.5: Cálculos para el análisis de independencia

#### 4.2.2 Test Chi cuadrado de independencia

En esta sección vamos a presentar un estadístico que cuantifica el apartamiento entre las frecuencias observadas y las frecuencias esperadas bajo la hipótesis nula que establece en este caso la independencia.

Utilizaremos la siguiente notación:

- \*  $e_{ij}$  denota la frecuencia esperada bajo  $H_0$  en la celda de la  $i$ -ésima fila y la  $j$ -ésima columna.
- \*  $o_{ij}$  indica la observación en la celda de la  $i$ -ésima fila y la  $j$ -ésima columna.

##### 4.2.2.1 Hipótesis de interés

El test se plantea de manera conceptual como

$$\begin{cases} H_0 : & \text{las variables son independientes} \\ & \text{versus} \\ H_1 : & \text{las variables no son independientes} \end{cases}$$

En forma simbólica el planteo es:

$$\begin{cases} H_0 : & P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j), \forall(i, j) / 1 \leq i \leq r, 1 \leq j \leq k \\ & \text{versus} \\ H_1 : & \exists(i, j) / P(X = x_i, Y = y_j) \neq P(X = x_i)P(Y = y_j) \end{cases}$$

**Ejemplo 4.4.** La Tabla 4.2 contiene las frecuencias observadas para el estudio de la cantidad de horas consumidas en programas televisivos con cierto grado de violencia de los  $n$  televidentes encuestados clasificados de acuerdo con su edad.

Interesa calcular ahora las frecuencias esperadas bajo el supuesto de independencia que establece  $H_0$ .

Si la hipótesis nula es verdadera, entonces la probabilidad de que un individuo sea clasificado en una celda es el producto de las correspondientes probabilidades marginales.

Sin embargo, las probabilidades marginales son desconocidas por lo cual se estiman con las proporciones marginales observadas.

Por ejemplo, en el caso de un encuestado joven que consume poca violencia se tiene que, bajo el supuesto de independencia, la probabilidad conjunta estimada es el producto de las correspondientes probabilidades marginales estimadas :

$$\widehat{p}_{11} = \widehat{P}(16 - 34 \cap PV) = \widehat{P}(16 - 34) \cdot \widehat{P}(PV) = \frac{26}{80} \cdot \frac{40}{80} = \frac{13}{80}$$

Con lo cual, el número esperado de individuos jóvenes que consume poca violencia resulta

$$\widehat{e}_{11} = n \cdot \widehat{P}(16 - 34 \cap PV) = 80 \cdot \frac{13}{80} = 13$$

Del manera análoga se calculan las demás frecuencias esperadas ( $e_{ij}$ ) que se exhiben en la Tabla 4.6 junto con las frecuencias observadas ( $o_{ij}$ ).

Nivel de violencia	Edad (en años)						Totales
	Joven	Adulto	Mayor	$\widehat{o}_{ij}$	$\widehat{e}_{ij}$	$\widehat{o}_{ij}$	
Poca	8	(13)	12	(13.5)	20	(13.5)	40
Mucha	18	(13)	15	(13.5)	7	(13.5)	40
<b>Totales</b>	<b>26</b>	<b>(26)</b>	<b>27</b>	<b>(27)</b>	<b>27</b>	<b>(27)</b>	<b>80</b>

Tabla 4.6: Frecuencias observadas y esperadas del nivel de violencia según la edad

Se puede observar que las probabilidades marginales de las frecuencias esperadas son idénticas a las probabilidades marginales de los datos originales, salvo error de redondeo.

### 4.2.3 Test Chi cuadrado de homogeneidad

En esta sección, nos interesa estudiar si una variable aleatoria  $X$  sigue la misma distribución en distintos subgrupos de una población de estudio. Estos subgrupos serán denominados en lo sucesivo como *subpoblaciones*. En líneas generales, disponemos de:

- \*  $r$  muestras de tamaño  $n_j$  de una misma variable aleatoria ( $X$ ) y queremos comprobar si son homogéneas; es decir, si la variable tiene la misma distribución en las  $r$  subpoblaciones de interés. Las frecuencias observadas se exhiben en la Tabla 4.7, donde  $o_{ij}$  denota la frecuencia absoluta observada en la categoría  $j$  de la variable en la muestra  $i$ -ésima.

- \* el recorrido de la variable aleatoria  $X$  es  $X_1, X_2, \dots, X_k$ , pudiendo ser el nivel de medición de  $X$  nominal u ordinal.

	$X_1$	$X_2$	$\cdots$	$X_j$	$\cdots$	$X_k$	Totales
<b>Muestra 1</b>	$o_{11}$	$o_{12}$	$\cdots$	$o_{1j}$	$\cdots$	$o_{1k}$	$n_{1.}$
<b>Muestra 2</b>	$o_{21}$	$o_{22}$	$\cdots$	$o_{2j}$	$\cdots$	$o_{2k}$	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
<b>Muestra <math>i</math></b>	$o_{i1}$	$o_{i2}$	$\cdots$	$o_{ij}$	$\cdots$	$o_{ik}$	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
<b>Muestra <math>r</math></b>	$o_{r1}$	$o_{r2}$	$\cdots$	$o_{rj}$	$\cdots$	$o_{rk}$	$n_{r.}$
<b>Totales</b>	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.k}$	$n_{..}$

Tabla 4.7: Frecuencias teóricas de homogeneidad

En este caso tenemos de  $r$  subpoblaciones y una única variable observada que tiene  $k$  categorías distintas. El total de observaciones de la  $i$ -ésima muestra, para  $1 \leq i \leq r$ , es

$$n_{i.} = \sum_{j=1}^k o_{ij}$$

El total de observaciones de la categoría  $j$ -ésima de la variable en todas las muestras es

$$n_{.j} = \sum_{i=1}^r o_{ij}$$

El total de observaciones de las  $r$  muestras es

$$n_{..} = \sum_{i=1}^r n_{i.} = \sum_{j=1}^k n_{.j}$$

Se desea verificar si la distribución de las distintas categorías de la variable  $X$  es homogénea en las subpoblaciones muestreadas. Para comprender la idea de ‘homogeneidad’, consideremos la variable dada por el color en las tres subpoblaciones de la Figura 4.3 haciendo las siguientes preguntas

¿La distribución de la variable dada por el color es homogénea en las tres poblaciones representadas?

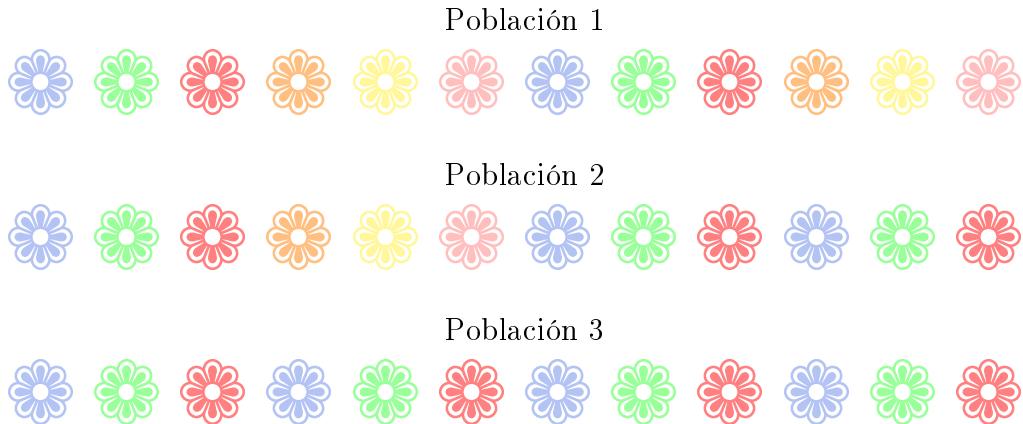


Figura 4.3: Poblaciones según variable de color

La respuesta a la pregunta planteada es sencilla y visual, dado que se trata de una variable simple y pocas subpoblaciones de tamaños reducidos. Sin embargo, este análisis para bases de datos grandes, no puede realizarse con un golpe de vista y es necesario cuantificar la situación.

Del mismo modo que hicimos en la prueba de independencia, debemos comparar las frecuencias observadas en cada una de las celdas con las frecuencias esperadas respectivamente, considerando ahora el supuesto de homogeneidad en la distribución de la variable de interés en las subpoblaciones. En este caso, las frecuencias observadas corresponden al número de individuos de la muestra  $i$  en la categoría  $X_j$ .

#### 4.2.3.1 Hipótesis de interés

Las hipótesis de interés pueden expresarse de la siguiente manera:

$$\begin{cases} H_0 : & P(X_j/m_i) = p_{j|i} = P(X_j) = p_j, \forall(i, j) / 1 \leq i \leq r, 1 \leq j \leq k \\ & \text{versus} \\ H_1 : & \exists(i, j) / P(X_j/m_i) = p_{j|i} \neq P(X_j) = p_j \end{cases}$$

siendo  $m_i$  la  $i$ -ésima muestra.

*¿Cuál es el valor esperado en cada casilla bajo la hipótesis de homogeneidad  $H_0$ ?*

Para responder a esta pregunta procedemos de la siguiente manera. Primero estimamos la probabilidad de la categoría  $j$  de la variable  $X$

$$\hat{p}_{.j} = P(X_j) = \frac{n_{.j}}{n_{..}}$$

Luego, estimamos la probabilidad condicional de la categoría  $j$  de  $X$  en la subpoblación  $i$

$$\hat{p}_{j|i} = \hat{P}(X_j/m_i) = \frac{n_{ij}}{n_i}$$

Como lo que esperamos bajo  $H_0$  es que  $p_{j|i} = p_j$  para todas las subpoblaciones teniendo en cuenta  $1 \leq i \leq r$ , podemos igualar sus estimaciones  $\hat{p}_{j|i} = \hat{p}_j$ . De esta igualdad se desprende que

$$\hat{e}_{ij} = \hat{p}_j n_i = \frac{n_j n_i}{n_{..}}$$

donde  $\hat{e}_{ij}$  es la frecuencia esperada bajo el supuesto de homogeneidad, que puede representarse como el producto entre el total de la  $i$ -ésima muestra y la probabilidad estimada de la categoría  $j$  en la población.

Es interesante observar que las frecuencias esperadas bajo independencia y bajo homogeneidad se calculan de la misma forma, sin embargo los tests son diferentes en cuanto a las hipótesis y al muestreo. Por ello las conclusiones deben redactarse en forma distinta en cada caso.

#### 4.2.4 Estadístico de contraste

Para las dos pruebas presentadas, podemos utilizar el siguiente estadístico de contraste que cuantifica la separación entre las frecuencias observadas y las esperadas cuando es cierta la hipótesis de nulidad:

$$\chi^2_{obs} = \sum_{i=1}^r \sum_{j=1}^k \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

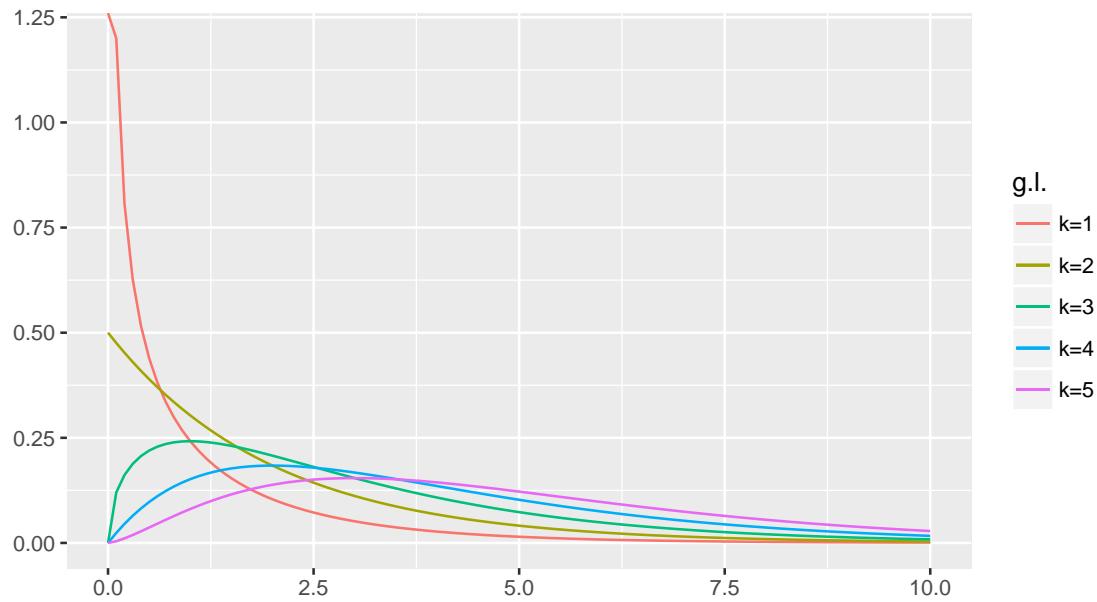
Este estadístico tiene distribución Chi cuadrado con  $(k - 1)(r - 1)$  grados de libertad lo cual se denota por  $\chi^2_{obs} \sim \chi^2_{(k-1)(r-1)}$ . Dicho de otra manera, los grados de libertad (g.l), del estadístico se obtienen multiplicando la cantidad de categorías de la primera variable menos uno por la cantidad de categorías de la segunda variable menos uno para el caso de independencia; y multiplicando la cantidad de categorías de la variable de estudio menos uno por la cantidad de subpoblaciones seleccionadas menos uno para el caso de homogeneidad.

**En ambos casos los grados de libertad son función de la cantidad de filas y columnas de la tabla.**

Esto está vinculado con la cantidad de datos que son necesarios para completar las tablas en cuestión, conocidos los totales de las filas y de las columnas.

#### 4.2.5 Región crítica

En ambos casos se trata de un test con región de rechazo unilateral a derecha; es decir, rechazamos  $H_0$  cuando los valores del estadístico son grandes y no se pueden atribuir al azar las diferencias entre los valores observados y los esperados.



**Figura 4.4:** Ejemplos de la distribución  $\chi^2$  según sus grados de libertad

La distribución Chi cuadrado es asimétrica por la derecha y su forma depende de los grados de libertad (ver Figura 4.4). Debido a ello, la región crítica variará en función de los grados de libertad de la variable y del nivel de significación establecido para el contraste .

**Ejemplo 4.5.** Interesa estudiar si cierta enfermedad ocurre con frecuencia similar o bien ocurre con frecuencia diferente en las poblaciones definidas por la adicción al tabaco. Para testear estas hipótesis se seleccionan dos muestras, una de 100 fumadores y otra de 50 no fumadores.

Destaquemos que en el test de homogeneidad los totales muestrales de cada una de las subpoblaciones se fijan a priori. En la Tabla 4.8 se muestra la cantidad de enfermos en cada una de las muestras.

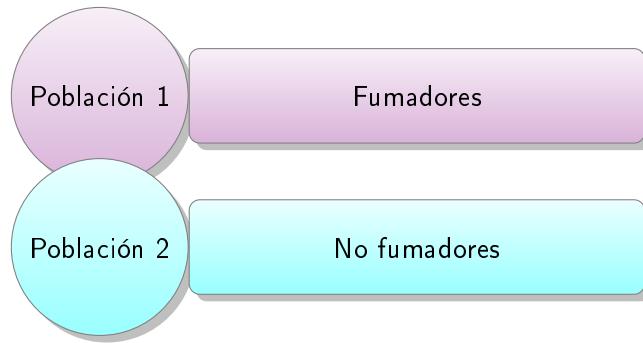
	No padece la enfermedad	Padece la enfermedad	Totales
Fumador	12	88	100
No fumador	25	25	50
Totales	37	113	150

**Tabla 4.8:** Datos enfermedad según tabaquismo

Se considera un nivel de significación del 5% para el ensayo. Realizamos un contraste de ho-

mogeneidad para responder a los interrogantes. Para comprender que se trata de una prueba de homogeneidad, debemos definir la variable de interés y las poblaciones en las cuales estamos comparando su distribución. La variable de interés es  $X$  que establece si un individuo padece la enfermedad. En este caso, la variable queda representada por dos niveles o categorías: ‘SI’ y ‘NO’.

Las poblaciones de estudio son las siguientes:



Debido a lo anterior, el estadístico de contraste tendrá  $(2 - 1) \cdot (2 - 1) = 1$  grado de libertad, y dado que el nivel de significación es del 5%, la región de rechazo unilateral a derecha queda definida por  $\{\chi^2_{obs} / \chi^2_{crit} > 3.841\}$  y está representada en la Figura 4.5.

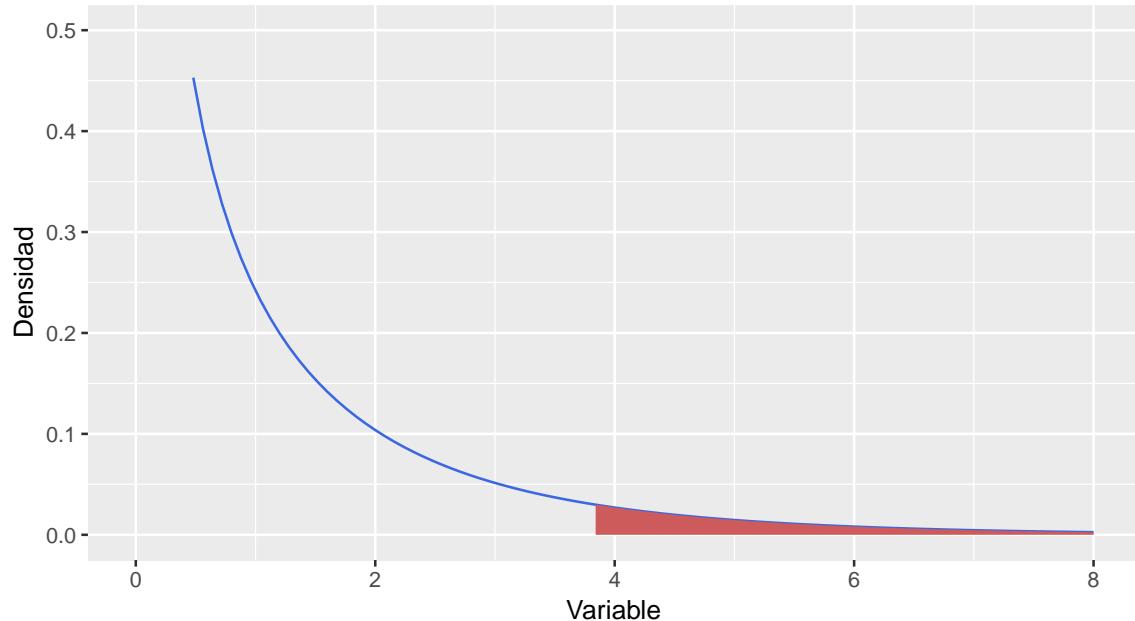


Figura 4.5: Distribución  $\chi^2$  y zona crítica

La hipótesis nula de este ensayo sostiene que las proporciones de enfermos en ambas poblaciones, ‘Fumadores’ y ‘No fumadores’, son iguales. Calculemos las frecuencias esperadas bajo la hipótesis

de homogeneidad a partir de las probabilidades estimadas para individuos enfermos ( $\hat{p}_e$ ) y para individuos sanos ( $\hat{p}_s$ ):

$$\hat{p}_e = \frac{113}{150} \quad \text{y} \quad \hat{p}_s = \frac{37}{150}$$

Denotando por  $n_f$  y  $n_{nf}$  a los totales de fumadores y no fumadores respectivamente, se tiene que:

$$\begin{aligned}\hat{e}_{11} &= \hat{p}_s \times n_f = \frac{37}{150} \cdot 100 = 24.67 \\ \hat{e}_{12} &= \hat{p}_e \times n_f = \frac{113}{150} \cdot 100 = 75.33 \\ \hat{e}_{21} &= \hat{p}_s \times n_{nf} = \frac{37}{150} \cdot 50 = 12.34 \\ \hat{e}_{22} &= \hat{p}_e \times n_{nf} = \frac{113}{150} \cdot 50 = 37.67\end{aligned}$$

En la Tabla 4.9 se agregaron las frecuencias esperadas.

		No padece la enfermedad	Padece la enfermedad			
		$o_{ij}$	$\hat{e}_{ij}$	$o_{ij}$	$\hat{e}_{ij}$	Totales
Fumador		12	24.67	88	75.33	100
No fumador		25	12.34	25	37.67	50
<b>Totales</b>		37	37	113	113	150

Tabla 4.9: Frecuencias observadas y esperadas de enfermedad según tabaquismo

En las casillas (1, 1) y (2, 1), se observan diferencias importantes entre los valores observados y la estimación de los valores esperados. Sin embargo, es importante notar que en todos los casos las pruebas de Chi cuadrado señalan que las distribuciones no son homogéneas pero no señalan a qué casillas o categorías se debe esta diferencia. Para poder determinar eso se deberán hacer otro tipo de pruebas como por ejemplo **diferencia de proporciones**.

El valor de estadístico de contraste en este caso resulta

$$\chi^2_{obs} = \frac{(12 - 24.67)^2}{24.67} + \frac{(25 - 12.34)^2}{12.34} + \frac{(88 - 75.33)^2}{75.33} + \frac{(25 - 37.67)^2}{37.67} = 25.88$$

Como el estadístico de contraste toma un valor muy superior al valor crítico establecido para este caso,  $\chi^2_{obs} = 25.88 >> 3.841 = \chi^2_{1,0.95}$ , la decisión es rechazar la hipótesis nula.

Es decir, existe evidencia en contra de la hipótesis de que la distribución de la variable que indica si un individuo padece la enfermedad, es similar en las dos poblaciones estudiadas.

El  $p$ -valor correspondiente a esta prueba es  $P(\chi_1^2 > 25.88) << 0.0001$ .

Recordemos una vez más que los resultados obtenidos no nos indican en qué radica la diferencia de las distribuciones o en qué son diferentes, solamente apoya la suposición de que no son iguales.

#### 4.2.6 Limitaciones

Recordemos que si una variable aleatoria tiene distribución Normal estándar, su cuadrado tiene distribución Chi cuadrado con 1 grado de libertad. Simbólicamente, si  $Z \sim N(0; 1)$  entonces  $U = Z^2 \sim \chi_1^2$

Además, la suma de dos variables aleatorias Chi cuadrado independientes, es una nueva variable aleatoria Chi cuadrado cuyos grados de libertad corresponden a la suma de los grados de libertad de los sumandos.

Simbólicamente, si  $U_1 \sim \chi_{\nu_1}^2$  y  $U_2 \sim \chi_{\nu_2}^2$  son independientes, entonces  $U = U_1 + U_2 \sim \chi_{\nu_1 + \nu_2}^2$ .

Basados en estos resultados y aplicando el Teorema del Límite Central se puede deducir la distribución del estadístico de contraste de la prueba de Pearson. Sin embargo, este resultado tiene validez asintótica por lo cual no es aplicable en todos los casos.

Para que sea válida la aplicación del test de Chi cuadrado, es necesario que todas las frecuencias esperadas resulten superiores a 1 y a lo sumo el 20% de las mismas inferiores a 5.

Cuando no puede aplicarse el test de Chi cuadrado, una alternativa disponible es el test exacto de Fisher [1].

*¿Qué similitudes y diferencias se pueden establecer entre los contrastes de homogeneidad e independencia?*

La respuesta a esta pregunta se muestra en la Tabla 4.10.

Prueba de independencia	Prueba de homogeniedad
Dos variables categóricas, nominales u ordinales	Una variable categórica, nominal u ordinal
Una sola población	Por lo menos dos subpoblaciones
$\hat{e}_{ij} = \frac{n_i \cdot n_j}{n_{..}}$	$\hat{e}_{ij} = \frac{n_i \cdot n_j}{n_{..}}$
$\chi_{obs}^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi_{(r-1)(k-1)}^2$	$\chi_{obs}^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi_{(r-1)(k-1)}^2$
Región de rechazo unilateral a derecha	Región de rechazo unilateral a derecha
Rechaza grandes diferencias entre frecuencias observadas y esperadas	Rechaza grandes diferencias entre frecuencias observadas y esperadas

Tabla 4.10: Similitudes y diferencias entre ambas pruebas

#### 4.2.7 Aplicación en R

En el Código 4.1 se muestra cómo aplicar el test Chi cuadrado al Ejemplo 4.5.

```
M=as.table(rbind(c(12,88), c(25,25)))
# Guarda los datos
dimnames(M)=list(Fumador=c('SI','NO'), Enfermedad=c("Padece","No_Padece"))
# Establece las poblaciones (filas) y las categorías (columnas) de estudio

Xsq=chisq.test(M) # Realiza el test Chi cuadrado\\
Xsq\\
Xsq$expected # Calcula las frecuencias esperadas
```

Código 4.1: Ejemplo de test de homogeneidad

En el Código 4.2 se realizan los cálculos del Ejemplo 4.4.

```
D=as.table(rbind(c(8,12,20), c(18,15,7)))
# Guarda los datos
dimnames(D)=list(Violencia=c('Poca','Mucha'),
Grupo_etáreo=c('Joven','Adulto','Mayor'))
# Establece las categorías de estudio

Xsq=chisq.test(D) # Realiza el test Chi cuadrado
Xsq$expected # Calcula las frecuencias esperadas
```

Código 4.2: Ejemplo de test de independencia

El comando `chisq.test` del Código 4.2 arroja como resultado lo siguiente:

```
Pearson's Chi-squared test
data: D
X-squared = 10.439, df = 2, p-value = 0.005411
```

Los valores esperados se almacenan en `Xsq$expected` del Código 4.2 y de la Tabla 4.6 se puede ver que son mayores que 5, por lo tanto es válida la distribución asintótica del estadístico y concluimos que se rechaza la hipótesis de nulidad.

#### 4.2.8 Test Exacto de Fisher

El test exacto de Fisher permite analizar si dos variables dicotómicas están asociadas cuando la muestra a estudiar es demasiado pequeña y no se cumplen los supuestos establecidos para la validez de la aplicación del test  $\chi^2$ .

Estas condiciones exigen que los valores esperados de al menos el 80% de las celdas en una tabla de contingencia sean mayores de 5 y que ninguno sea menor que 1. Así, en una tabla  $2 \times 2$  será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas muestre frecuencias esperadas ligeramente por debajo de este valor.

En situaciones como esta, una forma de plantear los resultados es su disposición en una tabla de contingencia de dos vías. Si las dos variables que se están considerando son dicotómicas, nos encontraremos con el caso de una tabla  $2 \times 2$  como la que se muestra en la Tabla 4.11.

Tabla 4.11: Tabla de contingencia de  $2 \times 2$

		Característica A		
Característica B		Presente	Ausente	Total
Presente	a	b	$a + b$	
Ausente	c	d	$c + d$	
Total	$a + c$	$b + d$	n	

El test exacto de Fisher se basa en evaluar la probabilidad asociada a cada una de las tablas de  $2 \times 2$  que se pueden formar manteniendo los mismos totales de filas y columnas de la tabla observada.

Cada una de estas probabilidades se obtiene bajo la hipótesis nula de independencia de las dos variables que se están considerando.

La probabilidad exacta de observar un conjunto concreto de frecuencias a, b, c y d en una tabla  $2 \times 2$  cuando se asume independencia y los totales de filas y columnas se consideran fijos viene dada por la distribución hipergeométrica.

Esta fórmula se obtiene calculando todas las posibles formas en las que podemos disponer n sujetos en una tabla de  $2 \times 2$  de modo tal que los totales de filas y columnas sean siempre los mismos,  $(a + b)$ ,  $(c + d)$ ,  $(a + c)$  y  $(b + d)$ .

$$p = \frac{C_{a+b,a} C_{c+d,c}}{C_{n,a+c}} = \frac{(a+b)!(c+d)!(a+c)!(a+d)!}{n!a!b!c!d!}$$

La probabilidad anterior deberá calcularse para todas las tablas de contingencia que puedan formarse con los mismos totales marginales que la tabla observada.

Luego, estas probabilidades se utilizan para calcular el p-valor asociado al test exacto de Fisher. Este valor de p indicará la probabilidad de obtener una diferencia entre los grupos mayor o igual a la observada, bajo la hipótesis nula de independencia.

Si esta probabilidad es pequeña ( $p < 0.05$ ) se deberá rechazar la hipótesis de partida y deberemos asumir que las dos variables no son independientes, sino que presentan asociación. En caso contrario, se dirá que no existe evidencia estadística de asociación entre ambas variables.

Existen dos métodos para el cómputo del valor p asociado al test exacto de Fisher. En primer lugar, podremos calcularlo sumando las probabilidades de aquellas tablas con una probabilidad asociada menor o igual a la correspondiente a los datos observados.

La otra posibilidad consiste en sumar las probabilidades asociadas a resultados al menos tan favorables a la hipótesis alternativa como los datos reales. Este cálculo proporcionaría el valor de p correspondiente al test en el caso de un planteamiento unilateral. Duplicando este valor se obtendría el p-valor correspondiente a un test bilateral.

Para ilustrar la explicación anterior veamos un ejemplo.

**Ejemplo 4.6.** *Se desea investigar entre los pacientes internados en cierto servicio del hospital, si existe asociación entre la aparición de síntomas de depresión y el sexo del paciente. Se observa una muestra de 14 pacientes de este servicio y se los clasifica por sexo y por la aparición de síntomas de depresión. Las observaciones se encuentran en la Tabla 4.12.*

Tabla 4.12: Depresión por Sexo

		Depresión		
Sexo	Sí	No	Total	
Mujeres	1 (a)	4 (b)	5 (a+b)	
Hombres	7 (c)	2 (d)	9 (c+d)	
Total	8 (a+c)	6 (b+d)	14 (n)	

En esta tabla los valores observados son

- \*  $a = 1$
- \*  $b = 4$
- \*  $c = 7$
- \*  $d = 2$

Los totales marginales son

- \*  $a + b = 5$
- \*  $c + d = 9$
- \*  $a + c = 8$
- \*  $b + d = 6$

La frecuencia esperada en tres de las cuatro celdas es menor de 5, por lo que no resulta adecuado aplicar el test de  $\chi^2$  por lo tanto, es recomendable aplicar el test exacto de Fisher.

Si las variables sexo y depresión fuesen independientes, la probabilidad asociada a los datos que han sido observados se calcula de la siguiente manera:

$$p = \frac{(a+b)!(c+d)!(a+c)!(a+d)!}{n!a!b!c!d!} = \frac{5!9!8!6!}{14!1!4!7!2!} \approx 0.0599$$

En la Tabla 4.13 se muestran todas las posibles tablas que mantienen la frecuencias marginales de nuestro ejemplo.

**Tabla 4.13:** Combinaciones para Fisher

	Depresión				Depresión			
	Si		No		Si		No	
(i)	Mujeres	0	5	5	(iv)	Mujeres	3	2
	Hombres	8	1	9		Hombres	5	4
		8	6	14			8	6
								14
(ii)	Mujeres	1	4	5	(v)	Mujeres	4	1
	Hombres	7	2	9		Hombres	4	5
		8	6	14			8	6
								14
(iii)	Mujeres	2	3	5	(vi)	Mujeres	5	0
	Hombres	6	3	9		Hombres	3	6
		8	6	14			8	6
								14

En la Tabla 4.14 se presentan las probabilidades asociadas a cada una de las combinaciones presentadas en la Tabla 4.13.

Sumando las probabilidades menores a la de la tabla observada, calculamos el valor p:

$$p.valor = 0.003 + 0.0599 + 0.028 = 0.0909$$

Considerando un nivel de significación del 5% no tenemos evidencia para rechazar la hipótesis de nulidad que sostiene que la depresión no se asocia con el sexo.

Para realizar el test en R se utiliza el siguiente Código:

```
tablita=cbind(c(1,7),c(4,2)) fisher.test(tablita)
```

La salida correspondiente es:

Fisher's Exact Test for Count Data

Tabla 4.14: Probabilidades asociadas a las combinaciones

	a	b	c	d	p
(i)	0	5	8	1	0,0030
(ii)	1	4	7	2	0,0599
(iii)	2	3	6	3	0,2797
(iv)	3	2	5	4	0,4196
(v)	4	1	4	5	0,2098
(vi)	5	0	3	6	0,0280

```
data: tablita
p-value = 0.09091
alternative hypothesis: true odds ratio is not equal to 1
```

**Observación:** R informa la decisión en función de una medida de asociación conocida como odds ratio que es cociente de chances.

## 4.3 Ejercitación

### Ejercicio 1.

Seleccionar la alternativa correcta en cada uno de los siguientes casos.

1.  El nivel de significación de un test de hipótesis suele ser pequeño y es fijado por el investigador o un convenio generalmente aceptado.  
 El nivel de significación de un test de hipótesis da la probabilidad de declarar significativo el resultado de un test cuando éste es falso.  
 Al disminuir el nivel de significación de un test de hipótesis, aumenta la probabilidad del error de tipo II.  
 Todo lo anterior es cierto.  
 Todo lo anterior es falso.
2. Un estudio sobre la efectividad de un tipo de campaña llega a la conclusión de que éste es significativamente distinto del tradicional con  $p < 0.05$ . ¿Cuál es la interpretación correcta de este resultado?  
 Con toda seguridad, el nuevo estilo supera al tradicional  
 La probabilidad de éxito con la nueva campaña supera a la probabilidad del anterior en un 95%.  
 El nuevo estilo es un 95% mejor que el tradicional.  
 Si la campana no fuese efectiva, existe menos del 5% de probabilidad de observar muestras tan contrarias a dicha hipótesis como las obtenidas.  
 Ninguna de las anteriores es correcta.
3. En una prueba de hipótesis el  $p$ -valor es  
 un número pequeño.  
 fijado antes de realizar la prueba.  
 la probabilidad de rechazar la hipótesis nula.  
 la probabilidad del error al rechazar la hipótesis alternativa.  
 conocido al extraer la muestra y calcular el estadístico experimental.
4.  Una prueba de hipótesis se considera significativa si una muestra aleatoria es coherente con la hipótesis nula.  
 Una prueba de hipótesis se considera significativa si una muestra aleatoria no es coherente con la hipótesis nula.

- Una prueba de hipótesis se considera significativa si la hipótesis alternativa es más probable que la nula.
  - Todo lo anterior es cierto.
  - Son ciertas la segunda y la tercera opción.
5. Se realizó un estudio para comparar la duración de lámparas de bajo consumo utilizando dos métodos de fabricación diferentes y no se encontró diferencia estadísticamente significativa. ¿Cuál de las siguientes razones podrían ser causantes del resultado?
- Los métodos ofrecen tiempos de duración muy diferentes.
  - El nivel de significación es demasiado alto.
  - Las muestras son demasiado numerosas.
  - Las muestras son demasiado pequeñas.
  - Nada de lo anterior.
6. La afirmación es falsa.
- El nivel de significación es normalmente un valor pequeño.
  - La significación de una prueba es conocida después de analizar los datos.
  - El nivel de significación de una prueba debe ser fijado antes de seleccionar la muestra.
  - Una prueba puede resultar significativa antes de recoger los datos.
  - Una prueba se señala como significativa cuando se obtiene una muestra que discrepa mucho de la hipótesis nula.
7. El error de tipo I consiste en
- Rechazar  $H_0$  cuando es falsa.
  - Rechazar  $H_0$  cuando es cierta.
  - No rechazar  $H_0$  cuando es cierta.
  - No rechazar  $H_0$  cuando es falsa
  - La probabilidad de rechazar  $H_0$  cuando es falsa.

### Ejercicio 2.

A un grupo de 350 adultos, quienes participaron en una encuesta, se les preguntó si accedían o no a *Twitter*. Las respuestas clasificadas por sexo fueron las que se muestran en la Tabla 4.15.

1. Representar gráficamente esta información e interpretar el gráfico.

Sexo	Femenino	Masculino	Totales
Usa Twitter	14	25	39
No usa Twitter	159	152	311
Totales	173	177	350

Tabla 4.15: Ingreso a *Twitter* según sexo

2. Obtener los porcentajes por filas y comparar las diferentes zonas.
3. Obtener los porcentajes por uso y comparar las diferentes usos.
4. Calcular las frecuencias esperadas bajo independencia y compararlas con las observadas.
5. ¿Sugieren estos datos que existe diferencia de proporciones entre mujeres y hombres que acceden o no a *Twitter*? (Considerar  $\alpha = 0.05$ .)

### Ejercicio 3.

Se clasificó en forma cruzada una muestra de 250 técnicos en telecomunicaciones en base a su especialidad y a la zona de la comunidad en que estaban trabajando. Los resultados están tabulados en la Tabla 4.16.

Zona	A	B	C	D	E	Totales
Norte	20	18	12	17	67	134
Sur	6	22	15	13	56	112
Este	4	6	14	11	35	70
Oeste	10	19	23	40	92	184
Totales	40	65	64	81	250	500

Tabla 4.16: Especialidad según zona

1. ¿Puede considerarse adecuado un test de homogeneidad o de independencia? Fundamentar la respuesta considerando el tipo de muestreo realizado.
2. Establecer las hipótesis de interés, realizar el contraste y concluir considerando un nivel de significación del 1%.

### Ejercicio 4.

	Con angioma	Sin angioma
Embarazo normal	37	1334
Embarazo patológico	11	223

Tabla 4.17: Presencia de angioma según tipo de embarazo

Entre 1605 recién nacidos registrados en una maternidad, se han presentado 48 con un angioma cuya presencia, se sospecha puede estar relacionada con el carácter (normal o patológico) del embarazo de la madre. Los resultados se muestran en la Tabla 4.17.

Plantear y testear las hipótesis correspondientes considerando un nivel de significación del 5%.

# Capítulo 5

## Análisis de correspondencias

*I can prove anything by statistics  
except the truth.*

— George Canning

### 5.1 Introducción

El **análisis de correspondencias** (AC) es una técnica descriptiva o exploratoria, cuyo objetivo es resumir una gran cantidad de datos en un número reducido de dimensiones con la menor pérdida de información posible [8]. Este análisis es una técnica multivariante que permite representar conjuntamente las categorías de las filas y columnas de una tabla de contingencia.

Este análisis constituye el equivalente de análisis de componentes principales para variables cualitativas.

Cuando nos referimos a una **técnica descriptiva o exploratoria**, hacemos referencia a que no se requiere el cumplimiento de ningún supuesto para poder aplicarla.

Si bien el objetivo de esta técnica es similar al de otros métodos factoriales, como componentes principales, en el caso del análisis de correspondencias el método se aplica sobre variables categóricas u ordinales. Más específicamente, busca una representación en coordenadas de las filas y columnas de una tabla de contingencia, de modo tal que los patrones de asociación presentes en la tabla se reflejen en dichas coordenadas.

Una *tabla de contingencia* es un arreglo matricial de números **no negativos** donde en cada casilla se presenta la frecuencia absoluta observada para esa combinación de categorías de las variables.

Este método trabaja con las proporciones que se encuentran en cada combinación de categorías de la variable *X* y de la variable *Y*.

El AC resulta adecuado también para trabajar con tablas de proximidad o distancia entre elementos.

Si nos centramos en una tabla de contingencia de dos variables cualitativas, las categorías de una de las variable aparecen en filas mientras que las de la otra variable en columnas.

El AC consiste en resumir la información presente en las filas y columnas de manera que pueda proyectarse sobre un subespacio reducido y representarse simultáneamente los puntos fila y los puntos columna, pudiéndose obtener conclusiones acerca de las relaciones entre estas variables.

Con el AC se construye una gráfica, llamada *mapa perceptual*, que señala la interacción de dos variables categóricas a través de la relación entre las filas y las columnas. Se cuantifica además el grado de asociación presente en un conjunto de variables.

Por ejemplo, consideremos la variable cualitativa fila que representa la bebida cuyos diferentes niveles en un mercado dado son:

- ✿ gaseosa
- ✿ sidra
- ✿ champaña
- ✿ cerveza
- ✿ leche
- ✿ agua
- ✿ jugo

Mientras que la variable columna es la percepción del cliente respecto de las bebidas consideradas, clasificándose en:

- ✿ sabroso
- ✿ seco
- ✿ fuerte
- ✿ empalagoso
- ✿ dulce

El análisis de correspondencias estudia las frecuencias de la distribución conjunta de ambas variables y produce un gráfico con dos ejes en los cuales cada categoría de la variable ubicada en las filas y cada categoría de la variable ubicada en las columnas está representada por un punto. Este gráfico podría sugerir, por ejemplo, que siguiendo la dirección de uno de los ejes, a la izquierda se encuentran las categorías-fila dadas por suave, dulce, empalagoso; mientras que a la derecha podemos encontrar las de seco, amargo, fuerte.

Se vería también que las categorías-columna de gaseosa y jugo se hallan a la izquierda y la de champaña **y de cerveza** a la derecha.

De esta manera se podrán establecer relaciones entre las categorías de las variables de las filas y las columnas.

La extensión del análisis de correspondencias simples al caso de varias variables nominales, representadas en tablas de contingencia multidimensionales, se denomina **análisis de correspondencias múltiples**, y utiliza los mismos principios generales que la técnica antes descripta.

Algunos ejemplos de aplicación del análisis de correspondencias simple y múltiple son

- ✿ Estudios de preferencias o estilos de consumo, muy usuales en Investigación de Mercados.
- ✿ Estudios que buscan tipologías de individuos respecto a variables cualitativas, como pueden ser el comportamiento de especies en Biología, los patrones de enfermedades en Medicina, los perfiles psicológicos, entre otros.
- ✿ Estudios de posicionamiento de empresas a partir de las preferencias de consumidores.
- ✿ Estudios para elegir tratamientos efectivos para una misma patología pero con diferentes etiologías.

La información disponible se puede organizar en una tabla con una estructura como la que se muestra en la Tabla 5.1. Siendo  $f_{ij}$  la cantidad de observaciones que tuvieron nivel  $i$  en la variable  $X$  y nivel  $j$  en la variable  $Y$ .

	$Y_1$	$Y_2$	$\cdots$	$Y_j$	$\cdots$	$Y_k$
$X_1$	$f_{11}$	$f_{12}$	$\cdots$	$f_{1j}$	$\cdots$	$f_{1k}$
$X_2$	$f_{21}$	$f_{22}$	$\cdots$	$f_{2j}$	$\cdots$	$f_{2k}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X_i$	$f_{i1}$	$f_{i2}$	$\cdots$	$f_{ij}$	$\cdots$	$f_{ik}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X_r$	$f_{r1}$	$f_{r2}$	$\cdots$	$f_{rj}$	$\cdots$	$f_{rk}$

Tabla 5.1: Tabla de contingencia

En la Tabla 5.1 encontramos valores enteros en las casillas, estos valores a veces no **son muy informativos**.

Si calculamos para cada casilla la proporción de observaciones que representa esa frecuencia absoluta; es decir, para cada combinación de categorías de la variable fila ( $X$ ) y la variable columna ( $Y$ ), obtenemos la frecuencia con la que se presenta esa combinación en el grupo general.

En realidad, como se trata de una muestra lo que obtenemos es la **estimación** de la frecuencia de esa combinación de categorías ( $i, j$ ) en la población de estudio.

De este modo, podemos estimar las probabilidades de la Tabla 5.2 mediante la fórmula

$$\hat{\pi}_{ij} = \frac{f_{ij}}{N..}$$

siendo  $N..$  el total de observaciones.

	$Y_1$	$Y_2$	...	$Y_j$	...	$Y_k$
$X_1$	$\hat{\pi}_{11}$	$\hat{\pi}_{12}$	...	$\hat{\pi}_{1j}$	...	$\hat{\pi}_{1k}$
$X_2$	$\hat{\pi}_{21}$	$\hat{\pi}_{22}$	...	$\hat{\pi}_{2j}$	...	$\hat{\pi}_{2k}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X_i$	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	...	$\hat{\pi}_{ij}$	...	$\hat{\pi}_{ik}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X_r$	$\hat{\pi}_{r1}$	$\hat{\pi}_{r2}$	...	$\hat{\pi}_{rj}$	...	$\hat{\pi}_{rk}$

Tabla 5.2: Tabla de probabilidades estimadas

A continuación, presentamos un ejemplo para ilustrar las ideas expuestas recientemente.

**Ejemplo 5.1.** El objetivo de este estudio es analizar la asociación entre el nivel cultural, representado en columnas, y el nivel del trastorno de la atención en los niños, representado en filas. Para ello, se observaron 1600 niños pacientes de un centro asistencial focalizando en su nivel cultural y en su respuesta a un test de trastornos de la atención. Los resultados obtenidos se muestran en la Tabla 5.3.

Nos interesa estudiar si existe alguna vinculación entre ambas variables.



<https://flic.kr/p/6hyajx>

En las Tablas 5.4, 5.5 y 5.6 se pueden apreciar, respectivamente, la presencia de la distribución marginal de la atención, la distribución marginal de del nivel cultural y la distribución conjunta de ambas variables.

	A	B	C	D	E	F	Totales
Atento	64	57	57	72	36	21	307
Síntomas leves	94	94	105	141	97	51	582
Síntomas moderados	58	54	65	77	54	34	342
Disperso	46	40	60	94	78	51	369
Totales	262	245	287	384	265	157	1600

Tabla 5.3: Nivel cultural según atención

	Frecuencia relativa	Probabilidad estimada
Atento	307/1600	0.1919
Síntomas leves	582/1600	0.3627
Síntomas moderados	342/1600	0.2138
Disperso	369/1600	0.2307

Tabla 5.4: Distribución marginal del nivel de atención

	A	B	C	D	E	F
Frec. relativa	262/1600	245/1600	287/1600	384/1600	265/1600	157/1600
Prob. estimada	0.1638	0.1531	0.1794	0.2400	0.1656	0.0981

Tabla 5.5: Distribución marginal del nivel de cultura

	A	B	C	D	E	F	Totales
Atento	0.04	0.0356	0.0356	0.045	0.0225	0.0131	0.1919
Síntomas leves	0.0588	0.0588	0.0656	0.0881	0.0606	0.0318	0.3638
Síntomas moderados	0.0363	0.0338	0.0406	0.0481	0.0338	0.0213	0.2138
Disperso	0.0288	0.025	0.0375	0.0588	0.0488	0.0319	0.2306
Totales	0.1638	0.1531	0.1794	0.24	0.1656	0.0981	1

Tabla 5.6: Distribución conjunta de niveles culturales y de atención

Al igual que en el caso de componentes principales, si las variables observadas fueran independientes, no podríamos asociar una categoría a la otra. Entonces cabe cuestionarnos

*Recordemos el concepto de independencia y pensemos:*

*¿Cómo se podría visualizar este hecho en una tabla?*

Recordemos que se dice que dos eventos son independientes cuando la ocurrencia de uno de ellos no altera la probabilidad de ocurrencia del otro.

Simbólicamente esto significa que

$$P(A/B) = P(A) \quad \text{cuando} \quad P(B) > 0.$$

Y aún cuando  $P(B)=0$ , vale que

$$P(A \cap B) = P(A)P(B)$$

Es decir que cuando dos eventos son independientes, la probabilidad de ocurrencia conjunta es el producto de las probabilidades marginales.

Este concepto se extiende a las variables aleatorias y sus distribuciones, diciendo que dos variables, digamos  $X$  e  $Y$ , son *independientes* si se verifica que

$$P(X = i/Y = j) = P(X = i)$$

Dicho de otra manera, la proporción de  $X = i$  es la misma para cualquiera de los niveles de la variable  $Y$ . O bien, la proporción de  $Y = j$  es la misma para cualquiera de los niveles de la variable  $X$ . De esta definición y por lo antes demostrado para eventos, se deduce que si  $X$  e  $Y$  son variables independientes entonces

$$P(X = i, Y = j) = P(X = i)P(Y = j)$$

Es decir que bajo el supuesto de independencia, se esperaría que en la casilla  $ij$  (fila  $i$ , columna  $j$ ), la probabilidad conjunta resulte el producto de las probabilidades marginales. Veamos qué consecuencia tiene este hecho en las frecuencias esperadas de cada casilla de la tabla dadas por

$$\hat{e}_{ij} = N..P(X = i)P(Y = j)$$

De lo que se desprende que deberíamos encontrar la manera de estimar  $\pi_{i.} = P(X = i)$  y  $\pi_{.j} = P(Y = j)$ . Podemos considerar el cociente entre el total de observaciones registradas en la categoría  $i$  (respectivamente  $j$ ) de la variable  $X$  (respectivamente  $Y$ ) y el total de registros obteniendo

$$\widehat{\pi}_{i.} = \frac{N_{i.}}{N..} \quad \text{y} \quad \widehat{\pi}_{.j} = \frac{N_{.j}}{N..}$$

De lo cual se infiere que

$$\hat{e}_{ij} = N_{..} \hat{\pi}_{i..} \hat{\pi}_{..j}$$

o equivalentemente,

$$\hat{e}_{ij} = N_{..} \frac{N_i}{N_{..}} \frac{N_j}{N_{..}} = \frac{N_i N_j}{N_{..}}$$

Retomemos el Ejemplo 5.1, calculamos las frecuencias de las casillas esperadas bajo independencia (ver Tabla 5.7) y comparamos estas frecuencias con las observadas para tener idea de si las variables  $X$  e  $Y$  son independientes.

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>Totales</b>
<b>Atento</b>	50.27	47.01	55.07	73.68	50.85	30.12	307
<b>Síntomas leves</b>	95.30	89.12	104.40	139.68	96.39	57.11	582
<b>Síntomas moderados</b>	56.00	52.37	61.35	82.08	56.64	33.56	342
<b>Disperso</b>	60.42	56.50	66.19	88.56	61.12	36.21	369
<b>Totales</b>	262	245	287	384	265	157	1600

Tabla 5.7: Frecuencias esperadas bajo independencia

Se puede apreciar que en algunas casillas los valores esperados bajo independencia son muy diferentes a los observados, citamos por ejemplo:

- ✿ Atento y F
- ✿ Atento y A
- ✿ Disperso y A
- ✿ Disperso y F

Podemos resumir cada observación en un número; es decir, cuantificar la discrepancia entre los valores observados y los valores esperados bajo independencia. Como ya hemos visto y ampliaremos a continuación, una alternativa disponible para cuantificar el apartamiento de la independencia de nuestras observaciones es el estadístico Chi cuadrado de Pearson.

Sin embargo, podemos pensar esta idea desde la siguiente perspectiva. Si las dos variables fueran independientes, la distribución marginal de una de ellas se repetiría para cada nivel de la otra variable. Podemos entonces, a partir de la distribución marginal de los niveles culturales, que se muestra en la Tabla 5.5, construir el perfil medio cultural.

Investigaremos ahora si la distribución de esos niveles en la población general es similar a la distribución condicional de estos niveles en cada una de las categorías de la variable atención. En

P(A/Atento)	P(B/Atento)	P(C/Atento)	P(D/Atento)	P(E/Atento)	P(F/Atento)
64/307	57/307	57/307	72/307	36/307	21/307
0.2085	0.1856	0.1856	0.2345	0.1172	0.0684

Tabla 5.8: Probabilidades condicionales dado el nivel ‘Atento’

la Tabla 5.8 se muestran las probabilidades condicionales estimadas de los niveles de atención dado el nivel superior de atención.

Si las variables que definen el nivel cultural y de atención no tuvieran influencia una sobre la otra, las distribuciones del nivel cultural serían muy similares para las diferentes categorías de la variable atención; es decir, reproducirían todas las filas el perfil medio de atención.



De la simple observación de una categoría pueden surgir diferencias pero aún faltaría analizar si estas diferencias pueden considerarse estadísticamente significativas o no.

### 5.1.0.1 Perfiles medios

El concepto de **perfil** es fundamental para el análisis de correspondencias.

Al analizar una tabla de frecuencias, uno se puede fijar en las frecuencias relativas de las filas o en las frecuencias relativas de las columnas, que llamaremos *perfíles fila* y **perfíles columna**, respectivamente.

Estos perfíles son vectores que pueden representarse como puntos en un espacio de perfíles.

Estos vectores tienen características geométricas especiales debido a que la suma de sus elementos es igual a 1, lo que representa el 100%.

En la Figura 5.2 se aprecia el aspecto que tienen los perfíles para el Ejemplo 5.1.

El cruce entre las líneas que representan los diferentes perfíles indica que existe asociación entre las variables, o bien que la distribución del nivel cultural no es la misma en todos los niveles de atención.

En el gráfico de la Figura 5.2 se aprecia la presencia de interacción entre el nivel cultural y el resultado del test de atención.

Las frecuencias observadas siempre suelen ser diferentes a las esperadas. Sin embargo, desde un punto de vista estadístico, se desea saber si estas diferencias son lo suficientemente grandes como para contradecir la hipótesis de independencia o bien estas diferencias son producto del muestreo.

Dicho de otra manera el objetivo es estudiar qué tan probable resulta ser que las discrepancias entre frecuencias observadas y esperadas se deban sólo al azar.

Una forma para analizar esto es definir una medida de la discrepancia entre las frecuencias observadas y esperadas.

Una forma de cuantificar la magnitud de las diferencias entre lo observado y lo esperado es el

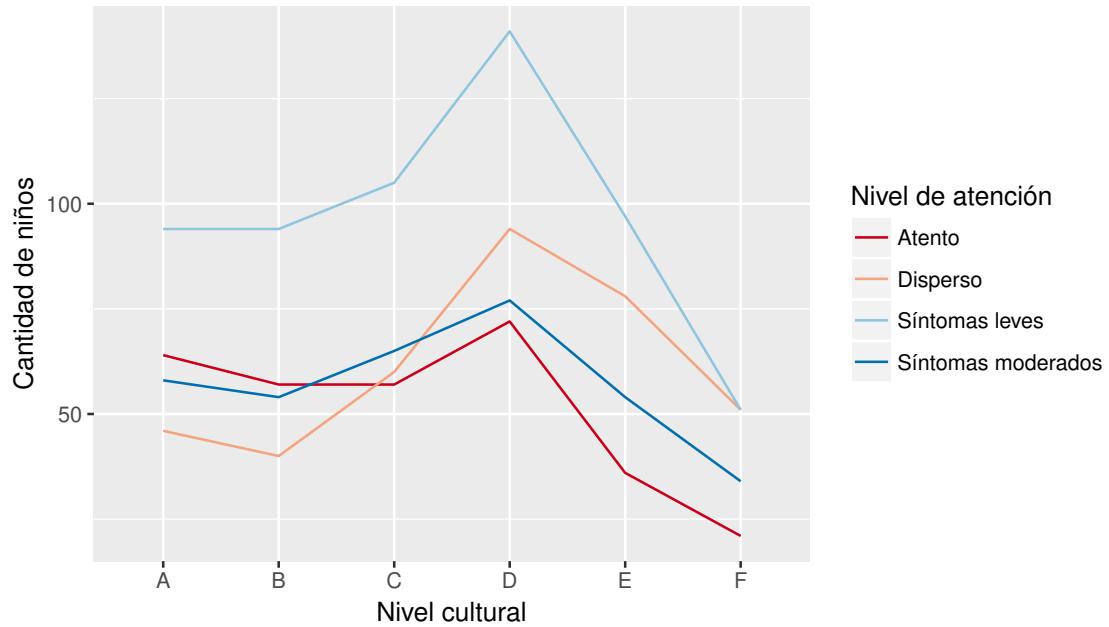


Figura 5.2: Perfiles de nivel cultural según atención

estadístico  $\chi^2$  de Pearson, definido como

$$\sum_{i=1}^r \sum_{j=1}^k \frac{(f_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

### 5.1.1 Inercia total

La *inercia total* de una tabla de contingencia se define como el cociente entre el estadístico Chi cuadrado y el tamaño muestral.

Una manera de construir una tabla de contingencia de  $r \times k$ , es definir  $r$  variables dicotómicas para las filas y  $k$  variables dicotómicas para las columnas. Luego, se disponen estas variables en matrices  $M_f$  con los datos de las filas y  $M_c$  con los datos de las columnas. Ambas matrices tienen tantas filas como cantidad total de observaciones.

La Tabla 5.9 correspondería a un modelo de tabla de contingencia para el Ejemplo 5.1 donde las matrices  $M_f$  y  $M_c$  están coloreadas en violeta y verde respectivamente.

En la misma se ve que el primer individuo es atento y tiene nivel cultural A, mientras que el segundo presenta síntomas leves y tiene nivel cultural F.

En nuestro caso se tiene lo siguiente:

- ✿ La matriz  $M_f$  (representada de la columna 2 a la 5) tiene dimensión  $n \times r = 1600 \times 4$  y describe las características de los individuos en función de su nivel de atención.

Individuo	Nivel de atención					Nivel cultural					
	Atento	Sínt. leves	Sínt. moderados	Disperso		A	B	C	D	E	F
1	1	0	0	0		1	0	0	0	0	0
2	0	1	0	0		0	0	0	0	0	1
⋮											
$k$	0	0	1	0		0	1	0	0	0	0
⋮											
$n$	0	0	0	1		0	0	1	0	0	0

Tabla 5.9: Representación de niveles como simulaciones (*dummies*)

- \* La matriz  $M_c$  (representada de la columna 6 a la 11) tiene dimensión  $n \times k = 1600 \times 6$  y describe el nivel cultural de los individuos.

Si realizamos el producto matricial

$$M_f^t M_c$$

obtendremos una matriz de dimensión  $r \times k$  que, en nuestro ejemplo particular, se corresponde con la tabla de contingencia de tamaño de  $4 \times 6$ . En esta matriz aparece en cada posición  $ij$  la frecuencia absoluta observada para cada combinación de niveles.

**Ejemplo 5.2.** Verifiquemos lo antes expuesto en un ejemplo sencillo a partir de los datos dados en la Tabla 5.10. Para este conjunto de observaciones  $M_f$  corresponde a la Característica  $A$  mientras que  $M_c$  corresponde a la Característica  $B$ .

El resultado del producto  $F = M_f^t M_c$  se exhibe en la Tabla 5.11.

La matriz  $F$  tiene las frecuencias absolutas y la matriz  $F_r$ , de la Tabla 5.12, las frecuencias relativas; es decir, el cociente entre las frecuencias absolutas observadas y el total de observaciones.

Observemos que  $F_r = \frac{1}{n}F$ .

La matriz  $F_r$  puede ser estudiada por filas o por columnas. De este modo, el análisis de  $F_r$  resulta análogo al de su traspuesta, dado que la elección de filas o columnas para cada una de las variables es arbitraria. ■

En lo que sigue vamos a estudiar cómo representar las filas, luego el razonamiento se sigue de manera similar para las columnas. Las  $r$  filas pueden pensarse como  $r$  puntos en el espacio  $\mathbb{R}^r$ . El propósito es representar estos  $r$  puntos en un espacio de dimensión menor de forma tal que nos permita apreciar sus distancias relativas. El objetivo es el mismo que en componentes principales, pero ahora se deben considerar las peculiaridades de los datos.

Individuo	Característica $A$		Característica $B$		
	$A_1$	$A_2$	$B_1$	$B_2$	$B_3$
1	1	0	1	0	0
2	1	0	1	0	0
3	1	0	0	1	0
4	1	0	0	1	0
5	1	0	0	0	1
6	1	0	0	0	1
7	0	1	1	0	0
8	0	1	0	1	0
9	0	1	0	1	0
10	0	1	0	1	0
11	0	1	0	0	1
12	1	0	1	0	0
13	1	0	1	0	0
14	1	0	0	1	0
15	0	1	0	1	0
16	0	1	0	1	0
17	0	1	0	0	1
18	0	1	0	0	1

Tabla 5.10: Representación de niveles para un ejemplo sencillo

	$B_1$	$B_2$	$B_3$	Totales
$A_1$	4	3	2	9
$A_2$	1	5	3	9
Totales	5	8	5	18

Tabla 5.11: Tabla de contingencia y matriz  $F$  para un ejemplo sencillo

	$B_1$	$B_2$	$B_3$	Totales
$A_1$	0.22	0.17	0.11	0.5
$A_2$	0.05	0.28	0.17	0.5
Totales	0.27	0.45	0.28	1

Tabla 5.12: Matriz  $F_r$  para un ejemplo sencillo

Estas peculiaridades provienen de que las frecuencias relativas de cada fila son distintas. Las filas tienen distintos pesos debido a que algunas tienen más datos que otras. Esto implica que la distancia euclídea, siendo una de las medidas más usadas, no sea una buena medida para cuantificar la proximidad entre las filas. Luego, es conveniente elegir otra forma de cuantificar esta distancia.

Observemos que la fila  $i$  tiene como frecuencia

$$f_{i \cdot} = \sum_{j=1}^r f_{ij}$$

Notando por  $\mathbb{1}_n$  al vector columna de que tiene un 1 en cada una de sus  $n$  componentes, se definen el *vector totales fila* y el *vector totales columna* respectivamente, de la siguiente manera:

$$f_T = F\mathbb{1}_r \quad \text{y} \quad c_T = \mathbb{1}_k^t F$$

donde  $F$  denota la matriz de frecuencias absolutas.

**Ejemplo 5.3.** Siguiendo los datos del Ejemplo 5.1, para calcular los vectores totales fila y columna, procedemos de la siguiente manera

$$f_T = F\mathbb{1}_3 = \begin{pmatrix} 4 & 3 & 2 \\ 1 & 5 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 9 \\ 9 \end{pmatrix}$$

$$c_T = \mathbb{1}_2^t F = (1 \ 1) \begin{pmatrix} 4 & 3 & 2 \\ 1 & 5 & 3 \end{pmatrix} = (5 \ 8 \ 5)$$

Definimos entonces las siguientes matrices diagonales:

$$D_f = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix} \quad \text{y} \quad D_c = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

Para efectuar estos cálculos en R nos referimos al Código 5.1 con datos extraídos de <https://goo.gl/KeE74T>.

```
library(readxl) # Permite leer archivos xlsx
m=read_excel("C:/.../ejemplosimple.xlsx")
# Importa la base con la cual se va a trabajar

M=as.matrix(m)
Mf=m[,2:3] # Guarda la matriz que caracteriza las filas
Mc=m[,4:6] # Guarda la matriz que caracteriza las columnas
F=t(Mf)%*%Mc # Arma la tabla de contingencia
totalf=F%*%rep(1,3) # Calcula el vector totales fila
```

```

totalc=rep(1,2)%*%F # Calcula el vector totales columna
n=sum(totalf) # Calcula el total de observaciones
Fr=F/n # Calcula las frecuencias relativas al total de observaciones
round(Fr,2) # Exhibe el resultado con 2 decimales

```

Código 5.1: Cálculos del ejemplo

Para dar a cada fila un peso proporcional a su frecuencia relativa, los componentes del vector  $f_T$  pueden considerarse como pesos. Podemos observar que en nuestro ejemplo, las filas tienen el mismo peso.

Con el fin de estudiar la distancia entre filas, llamaremos  $R$  a la matriz de frecuencias relativas condicionadas al total de la fila, que se obtiene calculando

$$R = D_f^{-1} F$$

donde  $D_f$  es una matriz diagonal de  $r \times r$  cuyos elementos diagonales son las componentes del vector  $f_T = f_{i\cdot}$ , las frecuencias relativas de los totales de las filas.

**Ejemplo 5.4.** Siguiendo con nuestro Ejemplo 5.1 y aplicando el Código 5.2 con datos extraídos de <https://goo.gl/KeE74T>, tenemos que

$$R = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}^{-1} \begin{pmatrix} 4 & 3 & 2 \\ 1 & 5 & 3 \end{pmatrix} = \begin{pmatrix} 1/9 & 0 \\ 0 & 1/9 \end{pmatrix} \begin{pmatrix} 4 & 3 & 2 \\ 1 & 5 & 3 \end{pmatrix} = \begin{pmatrix} 4/9 & 1/3 & 2/9 \\ 1/9 & 5/9 & 1/3 \end{pmatrix}$$

```

library(readxl) # Permite leer archivos xlsx

m=read_excel("C:/.../ejemplosimple.xlsx")
# Importa la base con la cual se va a trabajar

M=as.matrix(m)
Mf=m[,2:3] # Guarda la matriz que caracteriza las filas
F=t(Mf)%*%Mc # Arma la tabla de contingencia
totalf=F%*%rep(1,3) # Calcula el vector totales fila
Df=diag(as.vector(totalf))
# Arma la matriz diagonal con las frecuencias de las filas
R=solve(Df)%*%F # Calcula la matriz R
round(R,3) # Exhibe el resultado con 3 decimales

```

Código 5.2: Más cálculos del ejemplo

Observar que las filas de la matriz  $R$  suman 1. Es más, cada fila de esta matriz representa la distribución de la variable de las columnas condicionada a la presencia del atributo que representa

cada fila. Denotamos la fila  $i$ -ésima de la matriz  $R$  de frecuencias relativas condicionadas por filas como  $R_{i\cdot}$ , que puede ser considerada como un punto o como un vector del espacio  $\mathbb{R}^k$ .

Como la suma  $\sum_{j=1}^k R_{ij} = 1$ , todos los puntos se encuentran en un espacio de dimensión  $k - 1$ .

Nuestro objetivo es proyectar estos puntos sobre un espacio de dimensión menor de manera tal que las filas que tengan estructuras similares aparezcan próximas y las que tengan estructuras diferentes aparezcan alejadas.

Definimos la **distanzia Chi cuadrado** como

$$D^2(R_{a\cdot}, R_{b\cdot}) = \sum_{j=1}^r \frac{1}{f_{\cdot j}} \left( \frac{f_{aj}}{f_a} - \frac{f_{bj}}{f_b} \right)^2$$

siendo  $f_{aj}$  (resp.  $f_{bj}$ ) el elemento de la fila  $a$  (resp.  $b$ ) en la columna  $j$ ,  $f_a$  el total de la fila  $a$  y  $f_{\cdot j}$  el total de la columna  $j$ . Que puede expresarse matricialmente como

$$D^2(R_{a\cdot}, R_{b\cdot}) = (R_{a\cdot} - R_{b\cdot}) D_c^{-1} (R_{a\cdot} - R_{b\cdot})^t$$

donde  $D_c^{-1}$  es la inversa de la matriz diagonal cuyos elementos en la diagonal coinciden con los totales de las columnas.

**Ejemplo 5.5.** Estos cálculos para el Ejemplo 5.1 se realizan con las instrucciones del Código 5.3 con datos extraídos de <https://goo.gl/KeE74T>. A modo de ejemplo, calculamos la distancia entre las dos primeras filas

$$d(R_{1\cdot}, R_{2\cdot}) = d[(4, 3, 2), (1, 5, 3)] = \frac{1}{5} \left( \frac{4}{9} - \frac{1}{9} \right)^2 + \frac{1}{8} \left( \frac{1}{3} - \frac{5}{9} \right)^2 + \frac{1}{5} \left( \frac{2}{9} - \frac{1}{3} \right)^2 = 0.0308642$$

```
library(readxl) # Permite leer archivos xlsx

m=read_excel("C:/.../ejemplosimple.xlsx")
# Importa la base con la cual se va a trabajar

M<-as.matrix(m)
Mf=m[,2:3] # Guarda la matriz que caracteriza las filas
Mc=m[,4:6] # Guarda la matriz que caracteriza las columnas
F=t(Mf)%*%Mc # Arma la tabla de contingencia
totalf=F%*%rep(1,3) # Calcula el vector totales fila
totalc=rep(1,2)%*%F # Calcula el vector totales columna
Df=diag(as.vector(totalf))
# Arma la matriz diagonal con las frecuencias de las filas
R=solve(Df)%*%F # Calcula la matriz R
Dc=diag(as.vector(totalc))
# Arma la matriz diagonal con las frecuencias de las filas
```

```

distchi12=(R[1,]-R[2,])%*%solve(Dc)%*%(R[1,]-R[2,])
# Calculamos la distancia chi cuadrado entre las filas 1 y 2

```

Código 5.3: Ejemplo de distancia chi cuadrado



La distancia chi cuadrado es equivalente a la distancia euclídea entre los vectores transformados de la siguiente forma

$$Y_i = D_c^{-1/2} R_i.$$

Es decir que podemos construir la matriz de datos transformados y calcular la distancia euclídea entre las filas de esta matriz, siendo

$$Y = RD_c^{-1/2} = D_f^{-1} FD_c^{-1/2}$$

de donde se deduce que

$$Y_{ij} = \frac{f_{ij}}{\sqrt{f_i f_j}}$$

Cabe observar que los elementos transformados ya no suman 1, ni por filas ni por columnas.

**Ejemplo 5.6.** En nuestro ejemplo, tenemos que

$$Y = RD_c^{-1/2} = \begin{pmatrix} 4/9 & 1/3 & 2/9 \\ 1/9 & 5/9 & 1/3 \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 5 \end{pmatrix}^{-1/2} = \begin{pmatrix} 0.1987 & 0.1178 & 0.0993 \\ 0.0496 & 0.1964 & 0.1490 \end{pmatrix}$$



Las componentes de la matriz  $Y$  se corresponden con las frecuencias relativas condicionadas a las filas estandarizadas por su variabilidad, que depende del total de la columna. La distancia euclídea entre las filas de la matriz  $Y$  coincide con la calculada en la definición de distancia chi cuadrado. Consideremos la matriz

$$Z = D_f^{-1/2} FD_c^{-1/2}$$

de donde

$$Z_{ij} = \frac{f_{ij}}{\sqrt{f_i f_j}}$$

Buscamos ahora una representación para las filas de la matriz  $Z$  mediante una proyección en un espacio de dimensión menor. Es decir, buscamos una dirección  $\vec{w}$  de norma unitaria tal que  $\vec{w}^t \vec{w} = 1$  y la proyección de  $Z$  de manera tal de maximizar la variabilidad. Dicho de otro modo, buscamos maximizar  $\vec{w}^t Z^t Z \vec{w}$ . Pero recordemos que este problema ya lo hemos resuelto en el análisis de componentes principales.

Tenemos que proyectar en la dirección de los autovectores de la matriz  $Z^tZ$ . Entonces las coordenadas de las filas de la representación vienen dadas por

$$C_f = YW_2 = D_f^{-1}FD_c^{-1/2}W_2$$

donde  $W_2 = (\vec{w}_1 \ \vec{w}_2)$  la matriz formada por los dos primeros autovectores de la matriz  $Z^tZ$ .

Resumiendo, este procedimiento puede esquematizarse en los siguientes tres pasos:

- ✿ Calculamos las frecuencias relativas condicionales, consideradas como puntos del espacio.
- ✿ Computamos la distancia Chi cuadrado entre estos puntos.
- ✿ Proyectamos los puntos en el espacio que maximiza la variabilidad de la proyección. Es decir, proyectamos en la dirección de los primeros dos autovectores de la matriz  $Z^tZ$ .

Análogamente, podemos aplicar a las columnas un análisis similar al de las filas, resultando la mejor representación de las columnas:

$$C_c = YU_2 = D_c^{-1}F^tD_f^{-1/2}U_2$$

donde  $U_2 = (\vec{u}_1 \ \vec{u}_2)$  la matriz formada por los dos primeros autovectores de la matriz  $ZZ^t$ .

**Ejemplo 5.7.** Retomando el Ejemplo 5.1, podemos representar los puntajes o *scores* en un gráfico denominado **biplot simétrico** como el de la Figura 5.7, para lo cual nos referimos al Código 5.4. R también nos permite visualizar diferentes gráficos, como por ejemplo las contribuciones a la inercia de las filas y de las columnas como en las Figuras 5.3 y 5.4. La representación en el *biplot* de las categorías de las filas y de las columnas se muestra en las Figuras 5.5 y 5.6.

```
library(ca)      # Paquete para análisis de correspondencias
library(FactoMineR) # Paquete con métodos de análisis exploratorio de datos
library(factoextra) # Paquete para análisis multivariado de datos
library(ggplot2) # Paquete para confeccionar dibujos

# Armamos la base de datos
atento=c(64,57,57,72,36,21)
leve=c(94,94,105,141,97,51)
moderado=c(58,54,65,77,54,34)
disperso=c(46,40,60,94,78,51)
base=rbind(atento, leve, moderado, disperso)
colnames(base)=c("A", "B", "C", "D", "E", "F")
rownames(base)=c("Atento", "Sínt. leves", "Sínt. moderados", "Disperso")

atencion.ac=CA(base, graph=FALSE) # Realiza el análisis de correspondencias
get_ca_row(atencion.ac) # Muestra lo que se guarda de las filas
get_ca_col(atencion.ac) # Muestra lo que se guarda de las columnas
```

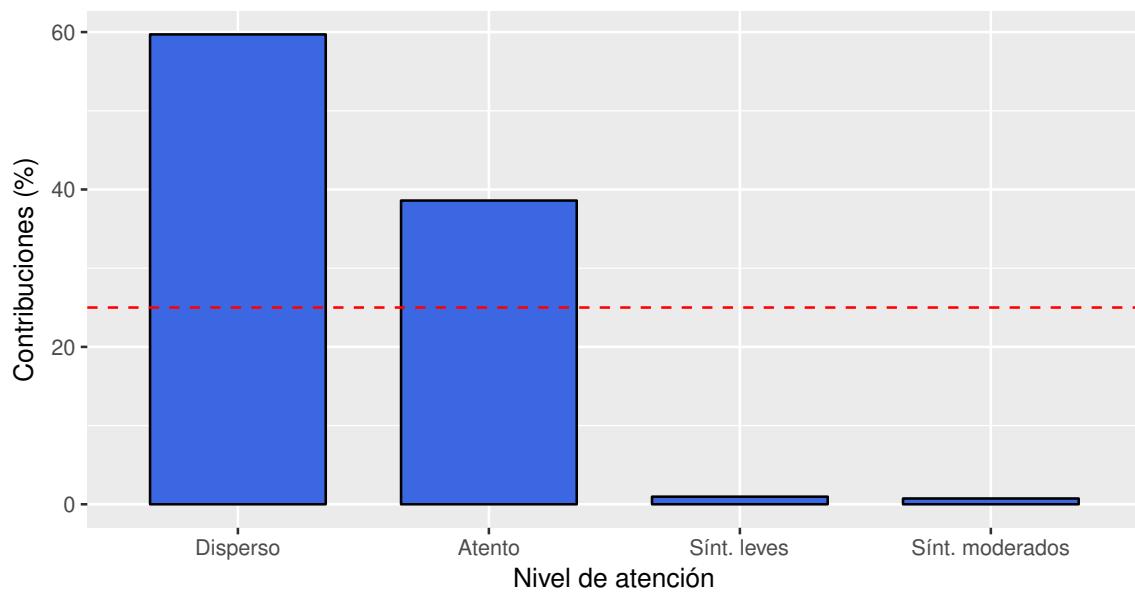


Figura 5.3: Contribución de filas a la dimensión 1

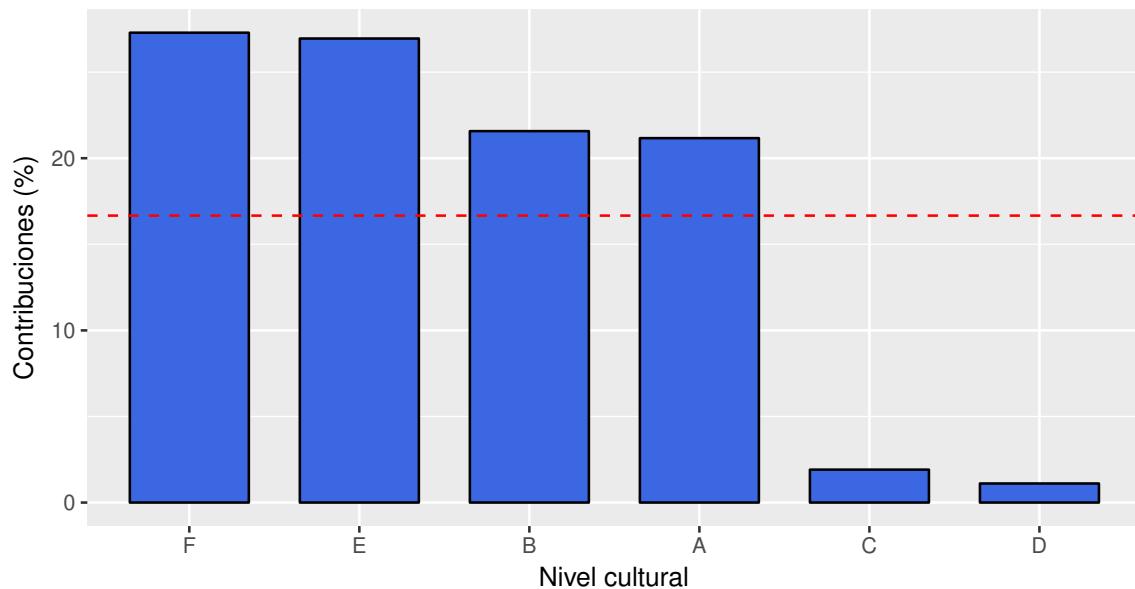


Figura 5.4: Contribución de columnas a la dimensión 1

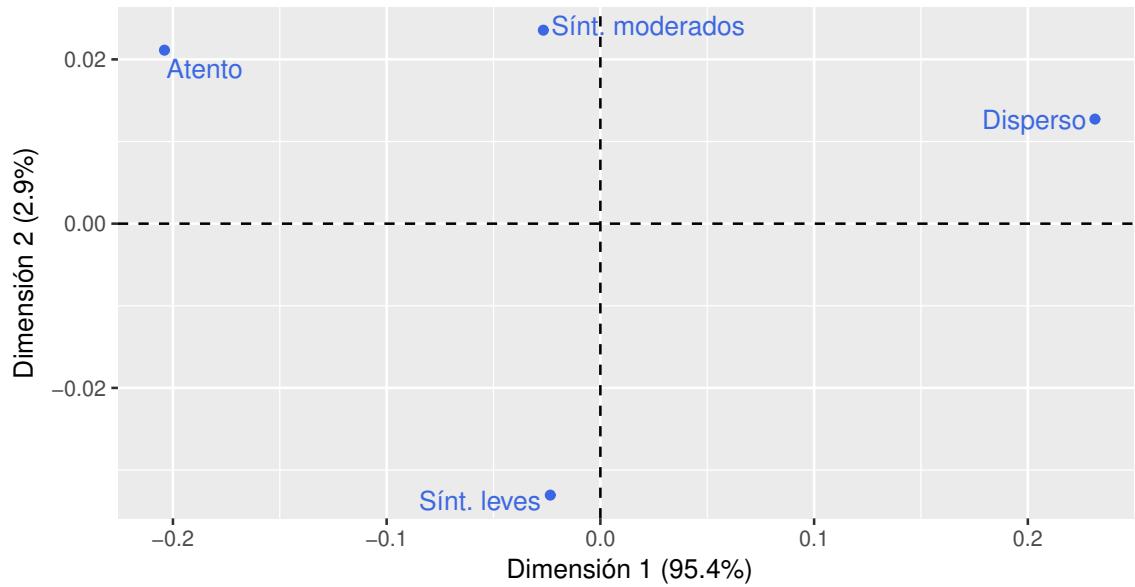


Figura 5.5: Puntos fila - AC

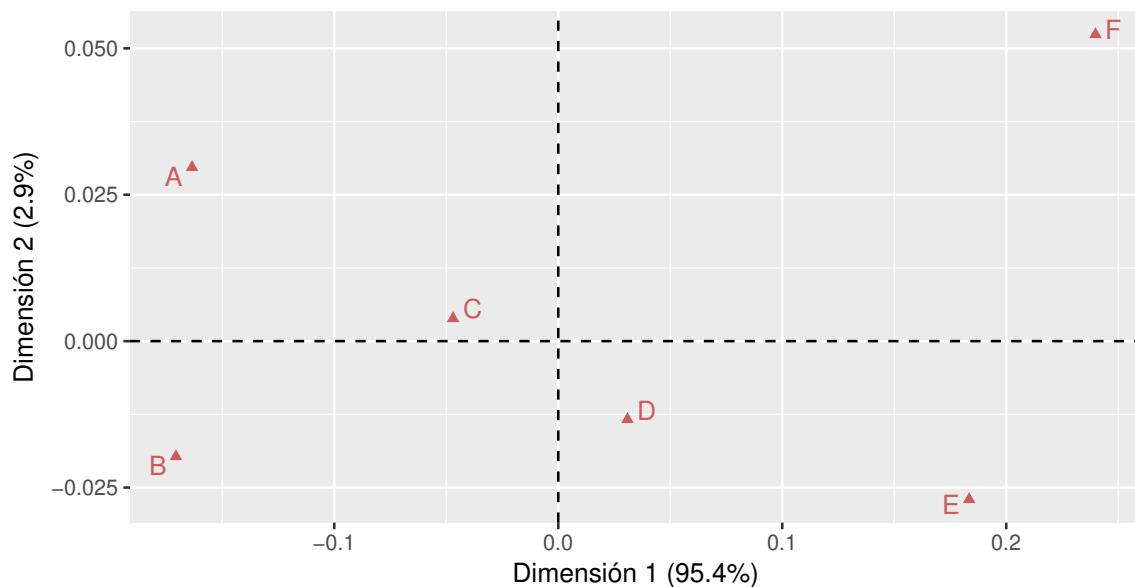


Figura 5.6: Puntos columna - AC

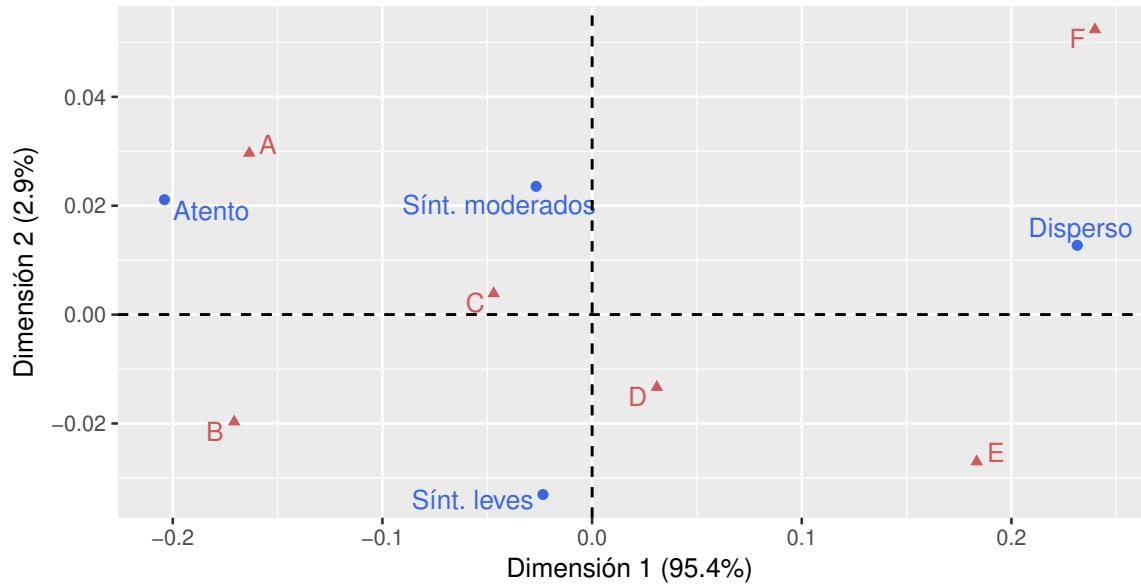


Figura 5.7: Biplot simétrico - AC

```

fviz_contrib(atencion.ac, choice="row", axes=1,
             fill="royalblue", color = "black") +
  theme_gray() +
  theme(axis.text.x = element_text(angle=0)) +
  xlab('Nivel_de_atención') +
  ylab('Contribuciones_(%)') +
  ggtitle('')
# Grafica las categorías de las filas

fviz_contrib(atencion.ac, choice="col", axes=1,
             fill="royalblue", color = "black") +
  theme_gray() +
  theme(axis.text.x = element_text(angle=0)) +
  xlab('Nivel_cultural') +
  ylab('Contribuciones_(%)') +
  ggtitle('')
# Grafica las categorías de las columnas

fviz_ca_row(atencion.ac, repel=TRUE, col.row="royalblue") +
  theme_gray() +
  xlab('Dimensión_1_(95.4%)') +
  ylab('Dimensión_2_(2.9%)') +
  ggtitle('')

```

```

# Grafica los puntos fila

fviz_ca_col(atencion.ac, repel=TRUE, col.col="indianred") +
  theme_gray() +
  xlab('Dimensión_1_(95.4%)') +
  ylab('Dimensión_2_(2.9%)') +
  ggtitle('')

# Grafica los puntos columna

fviz_ca_biplot(atencion.ac, repel=TRUE, col.row="royalblue",
  col.col="indianred") +
  theme_gray() +
  xlab('Dimensión_1_(95.4%)') +
  ylab('Dimensión_2_(2.9%)') +
  ggtitle('')

# Realiza el biplot simétrico

# Aplicamos ahora el paquete ca
atencion_ac=ca(base, graph = FALSE) # Realiza el análisis de correspondencias
summary(atencion_ac)
atencion_ac$rowcoord # Arroja las coordenadas del biplot de las filas
atencion_ac$colcoord # Arroja las coordenadas del biplot de las columnas

```

Código 5.4: Análisis de correspondencias para niveles cultural y de atención

### 5.1.1.1 Guía para la interpretación gráfica del *biplot* simétrico

Listamos a continuación algunas consideraciones a tener en cuenta.

- ✿ Las columnas (resp. filas) cercanas al origen reflejan categorías similares a la columna (resp. fila) promedio.
- ✿ Las columnas (resp. filas) cercanas entre sí reflejan categorías de similar perfil en términos de columnas (resp. filas).
- ✿ Las columnas (resp. filas) cercanas y lejanas al origen reflejan alta asociación positiva entre las categorías representadas.
- ✿ Los ejes representan ‘factores ocultos’ o ‘variables latentes’.
- ✿ En cada eje se indica el porcentaje de la inercia que logra representar el mismo.

Teniendo en cuenta esto, realizamos las siguientes observaciones para el Ejemplo 5.7:

- ✿ Lo más usual en cuanto al nivel de atención son los síntomas leves.
- ✿ El nivel cultural  $A$  es el más próximo a la categoría de ‘atento’.
- ✿ Las categorías de cultura  $C$  y  $D$  son las más similares al promedio en nivel cultural.
- ✿ Los síntomas leves en grado de atención se asocian con la categoría de cultura  $D$ .

El programa R guarda la siguiente información en el ‘objeto’:

- ✿ `nd` la dimensión de la solución
- ✿ `rownames` (resp. `colnames`) los nombres de las filas (resp. de las columnas)
- ✿ `rowinertia` (resp. `colinertia`) la cantidad de inercia de cada fila (resp. de cada columna)
- ✿ `rowdist` (resp. `coldist`) la distancia Chi cuadrado de las filas (resp. de las columnas) al centroide
- ✿ `rowcoord` (resp. `colcoord`) las coordenadas para representar las categorías de filas (resp. de columnas)

La salida correspondiente al análisis de correspondencias simples para el Ejemplo 5.1, aplicando el paquete `ac` (ver Código 5.4), se muestra en las Tablas 5.13, 5.14 y 5.15.

	1	2	3
Valor	0.020682	0.000639	0.000369
Porcentaje	95.35%	2.95%	1.70%

Tabla 5.13: Inercias principales (autovalores)

	Atento	Síntomas leves	Síntomas moderados	Disperso
Masa	0.191875	0.363750	0.213750	0.230625
Distancia Chi	0.206391	0.040579	0.047857	0.232132
Inercia	0.008173	0.000599	0.000490	0.012427
Dimensión 1	-1.418216	-0.162976	-0.185590	1.608987
Dimensión 2	-0.835542	1.307591	-0.931956	-0.503463

Tabla 5.14: Perfiles de las filas

A medida que las tablas de contingencia crecen en tamaño, debido al aumento de niveles considerados dentro de cada una de las variables, resulta difícil detectar la presencia de patrones desde

	A	B	C	D	E	F
Masa	0.163750	0.153125	0.179375	0.240000	0.165625	0.098125
Distancia Chi	0.166708	0.172160	0.054983	0.042676	0.185584	0.245543
Inercia	0.004551	0.004538	0.000542	0.000437	0.005704	0.005916
Dimensión 1	-1.136894	-1.186945	-0.326477	0.214821	1.275644	1.667705
Dimensión 2	-1.173725	0.778810	-0.153506	0.527975	1.068973	-2.071700

Tabla 5.15: Perfiles de las columnas

la mera observación de los perfiles o del apartamiento de los mismos respecto del perfil medio o de la distancia entre pares de ellos. Prácticamente, resultaría imposible resaltar las características esenciales de estos datos. Por tal motivo, deberíamos buscar una alternativa a los diagramas de dispersión, que ha sido el instrumento para la descripción de datos que hemos utilizado hasta ahora. Esa alternativa, precisamente, es la que nos ofrece el análisis de correspondencias.

Consideremos un nuevo ejemplo para clarificar el concepto de **perfil promedio**.

**Ejemplo 5.8.** En la Tabla 5.16 se muestran los datos que la Secretaría de Ciencia y Técnica de la Facultad de Ciencias Sociales registró de las universidades con las que se realizaron viajes de intercambio en el contexto del Proyecto Redes durante los últimos meses del año 2017. También consignó conjuntamente las actividades que se desarrollaron durante los días del intercambio, clasificadas en docencia, investigación y extensión.

Universidad	Docencia	Investigación	Extensión	Totales
UTN	18	3	33	54
UNC	3	9	33	45
UNL	12	75	0	87
UNS	6	6	60	72
<b>Totales</b>	<b>39</b>	<b>93</b>	<b>126</b>	<b>258</b>

Tabla 5.16: Registro viajes de intercambio

En el AC, como ya hemos destacado, el concepto de perfil se refiere al conjunto de frecuencias divididas por su total y el mismo resulta de fundamental importancia en la comprensión de esta técnica. Para obtener los valores de la Tabla 5.17 se divide cada fila de la Tabla 5.16 por su propio total y, para obtener el perfil medio de cada filas se divide a los totales de las columnas por el total general. Observar que, salvo diferencias por redondeo, la suma de cada una de las filas es 1.

Con el Código 5.5 se genera la Figura 5.8.

```
| library(ggplot2) # Paquete para confeccionar dibujos |
```

Universidad	Docencia	Investigación	Extensión	Totales
UTN	0.33	0.06	0.61	1
UNC	0.07	0.20	0.73	1
UNL	0.14	0.86	0.00	1
UNS	0.08	0.08	0.83	1

Tabla 5.17: Perfiles fila de los viajes de intercambio

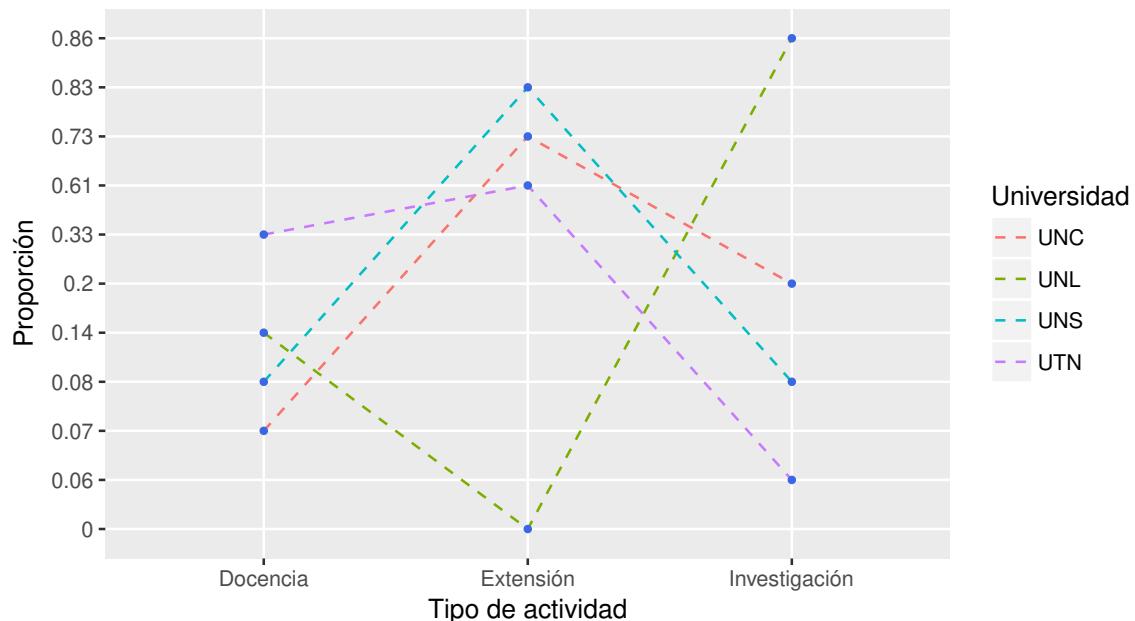


Figura 5.8: Perfiles fila de las actividades universitarias

```

# Cargamos los datos
Universidad=rep(c("UTN", "UNC", "UNL", "UNS"), 3)
actividad=c(rep("Docencia", 4), rep("Investigación", 4), rep("Extensión", 4))
prop=c(0.33, 0.07, 0.14, 0.08, 0.06, 0.2, 0.86, 0.08, 0.61, 0.73, 0, 0.83)
viajes=data.frame(cbind(Universidad, actividad, prop))

ggplot(data=viajes, aes(x=actividad, y=prop, group=Universidad, color=Universidad)) +
  geom_line(linetype="dashed")+
  geom_point(color="royalblue", size=1)+
  xlab("Tipo_de_actividad") +
  ylab("Proporción")

```

Código 5.5: Análisis de perfiles fila de las actividades universitarias

Analizando la Figura 5.8, se puede ver que tanto en el perfil de UTN como en el de UNC, existe una mayor concentración de frecuencia en actividades de extensión, si bien las otras dos frecuencias están invertidas. Mientras que los perfiles de UNS y de UNL concentran sus actividades en extensión e investigación respectivamente.

*¿Qué perfiles interesa comparar?*

Podemos estar interesados, por ejemplo, en comparar los perfiles de dos universidades, o bien en comparar el perfil de una universidad específica con el perfil medio dado por el perfil de la fila final de los totales de cada columna de la Tabla 5.16; es decir, el perfil de todas las actividades desarrolladas considerando a todas las universidades conjuntamente. Este cálculo resulta:

$$(\text{Docencia } \text{Investigación } \text{Extensión}) = \frac{1}{258} (39 \ 93 \ 126) = (0.1512 \ 0.3605 \ 0.4884)$$

Hasta el momento, nos hemos concentrado en observar los perfiles fila con el objetivo de comparar las modalidades de intercambio entre las diferentes universidades. Sin embargo, también podemos comparar los perfiles columna para ver de qué manera se distribuyen las modalidades de actividad en las distintas universidades. Esto se exhibe en la Tabla 5.18.

El perfil columna medio se obtiene realizando el cociente entre el total de las filas y el total general, obteniéndose por resultado

$$\begin{pmatrix} \text{UTN} \\ \text{UNC} \\ \text{UNL} \\ \text{UNS} \end{pmatrix} = \frac{1}{258} \begin{pmatrix} 54 \\ 45 \\ 87 \\ 72 \end{pmatrix} = \begin{pmatrix} 0.2093 \\ 0.1744 \\ 0.3372 \\ 0.2791 \end{pmatrix}$$

Podemos comparar los valores de los perfiles de los tipos de actividad con los valores del perfil columna medio, para ver si sus valores están por encima o por debajo de los del perfil medio. De

Universidad	Docencia	Investigación	Extensión
UTN	0.4615	0.0323	0.2619
UNC	0.0769	0.0968	0.2619
UNL	0.3077	0.8065	0.0000
UNS	0.1538	0.0645	0.4762
<b>Totales</b>	1	1	1

Tabla 5.18: Perfiles columna de los viajes de intercambio

esa manera surge con qué universidades el intercambio es diferente y de qué forma se manifiesta esa diferencia. Así, por ejemplo, el 46% de los intercambios de UTN son de docencia, y el 80% de los intercambios con la UNL son de investigación. Del mismo modo, la mayoría de los intercambios con UNS son de extensión. Veamos ahora lo siguiente:

- \* el promedio de UTN supera al promedio general de dedicación en docencia.
- \* los promedios de UNC y UNS superan al promedio general de dedicación media en extensión.
- \* el promedio de UNL supera al promedio general de dedicación en investigación.

Veamos si se aprecia alguna similitud en las caritas de Chernoff de la Figura 5.9.



### 5.1.1.2 Otra representación gráfica

En esta sección vamos a mostrar una manera completamente distinta de representación basándonos en el Ejemplo 5.8. Para representar los cuatro perfiles, ahora proponemos utilizar tres ejes, que corresponden a los tres tipos de actividad de intercambio, a modo de un diagrama de dispersión tridimensional como el de la Figura 5.10 generado por el Código 5.6 con datos extraídos de <https://goo.gl/NfwNwK>. En cada uno de estos tres ejes podríamos situar a uno de los tres elementos del perfil. Luego, podemos considerar estos tres elementos como las coordenadas de un único punto que represente todo el perfil, tomando las observaciones porcentuales por fila como las componentes de los puntos. Etiquetamos a estos tres ejes como docencia, investigación y extensión, calibrándolos de 0 a 1.

Este tipo de representación sólo será factible cuando se disponga de observaciones realizadas en tres categorías.

```
library(scatterplot3d) # Paquete para generar gráficos en 3D
library(readxl) # Permite leer archivos xlsx

universidades=read_excel("C:/.../universidades.xlsx")
```

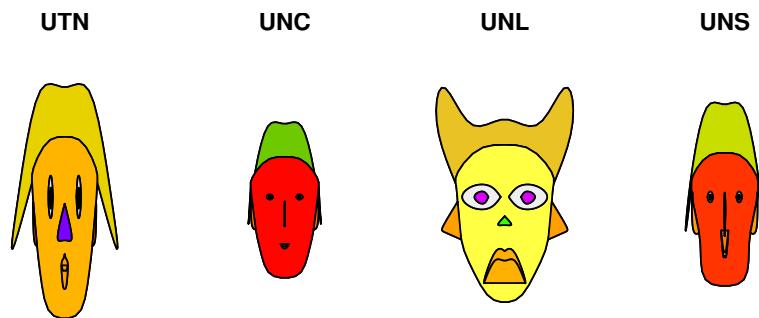


Figura 5.9: Caras de Chernoff para fila de las actividades universitarias

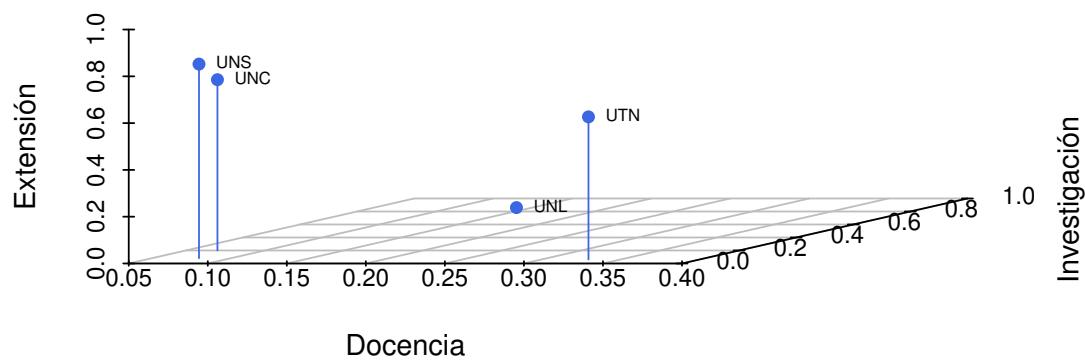


Figura 5.10: Representación en 3D de las actividades universitarias

```

# Importa la base con la cual se va a trabajar

with(universidades, {
s3d <- scatterplot3d(Docencia, Investigación, Extensión,
color="royalblue", pch=16, box=FALSE, angle=25,
type="h", xlab="Docencia", ylab="Investigación",
zlab="Extensión")
s3d.coords <- s3d$xyz.convert(Docencia, Investigación, Extensión)
text(s3d.coords$x, s3d.coords$y,
labels=universidades$Universidad,
cex=.6, pos=4)
})
# Realiza un diagrama de dispersión en 3D

with(universidades, {
s3d <- scatterplot3d(Docencia, Investigación, Extensión,
color="royalblue", pch=16, box=FALSE, angle=25)
s3d.coords <- s3d$xyz.convert(Docencia, Investigación, Extensión)
text(s3d.coords$x, s3d.coords$y,
labels=universidades$Universidad,
cex=.6, pos=4)
fit <- lm(Extensión ~ Docencia + Investigación)
s3d$plane3d(fit, col="indianred") # Agrega un plano
})

```

**Código 5.6:** Código para generar un diagrama de dispersión en 3D de las actividades universitarias

Los perfiles de las universidades, representados por las filas, se pueden llevar a dos dimensiones dado que, si bien tienen tres componentes, los mismos están restringidos debido a que la suma de las tres componentes para cualquiera de las  $n$  observaciones es igual a 1. Luego, en un espacio tridimensional el espacio ocupado es bidimensional, siendo el plano de ecuación  $x+y+z=1$ . Esto se puede ver en la Figura 5.11.

El comando `summary(univ.ac)` del Código 5.7 (con datos disponibles en <https://goo.gl/NfwNwK>) da como resultado los autovalores y autovectores considerados, el aporte a la inercia de las filas y de las columnas y el porcentaje de representación. Con el mismo código se generan los gráficos correspondientes al *biplot* simétrico (Figura 5.12), al aporte de las filas (Figura 5.13) y al aporte de las columnas (Figura 5.14).

```

library(readxl) # Permite leer archivos xlsx
library(FactoMineR) # Paquete con métodos de análisis exploratorio de datos
library(factoextra) # Paquete para análisis multivariado de datos
library(ggplot2) # Paquete para confeccionar dibujos

universidades=read_excel("C:/.../universidades.xlsx")
# Importa la base con la cual se va a trabajar

```

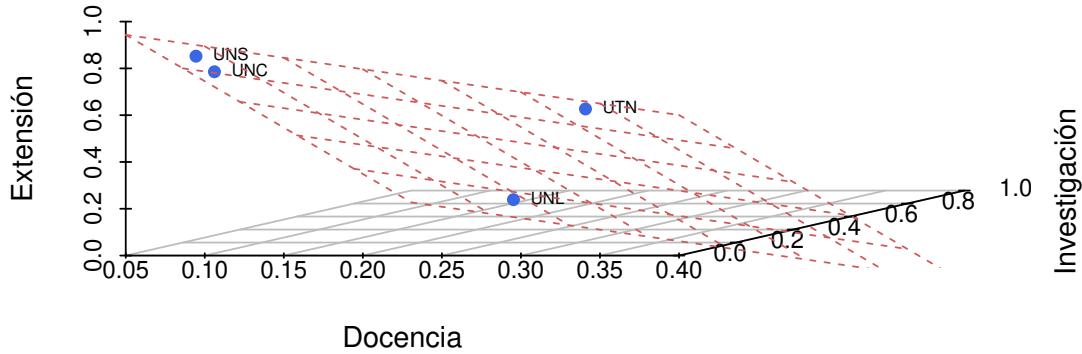


Figura 5.11: Plano de representación de las actividades universitarias

```
# Armamos la base de datos
base=as.matrix(universidades[1:4,2:4])
colnames(base)= c("Docencia", "Investigación", "Extensión")
row.names(base)= c("UTN", "UNC", "UNL", "UNS")

univ.ac=CA(base, graph = FALSE) # Realiza el análisis de correspondencias
summary(univ.ac) # Muestra el resultado del análisis de correspondencias

fviz_contrib(univ.ac, choice="row", axes=1,
fill="royalblue", color = "black") +
theme_gray() +
theme(axis.text.x = element_text(angle=0)) +
xlab('Universidad') +
ylab('Contribuciones (%)') +
ggtitle('')

# Grafica las categorías de las filas

fviz_contrib(univ.ac, choice="col", axes=1,
fill="royalblue", color = "black") +
theme_gray() +
theme(axis.text.x = element_text(angle=0)) +
xlab('Tipo de actividad') +
ylab('Contribuciones (%)') +
ggtitle('')
```

```
# Grafica las categorías de las columnas

fviz_ca_biplot(univ.ac, repel=TRUE, col.row="royalblue",
col.col="indianred") +
theme_gray() +
xlab('Dimensión_1_(86.7%)') +
ylab('Dimensión_2_(13.3%)') +
ggttitle('')

# Realiza el biplot simétrico
```

Código 5.7: Código para el análisis de correspondencias de las actividades universitarias

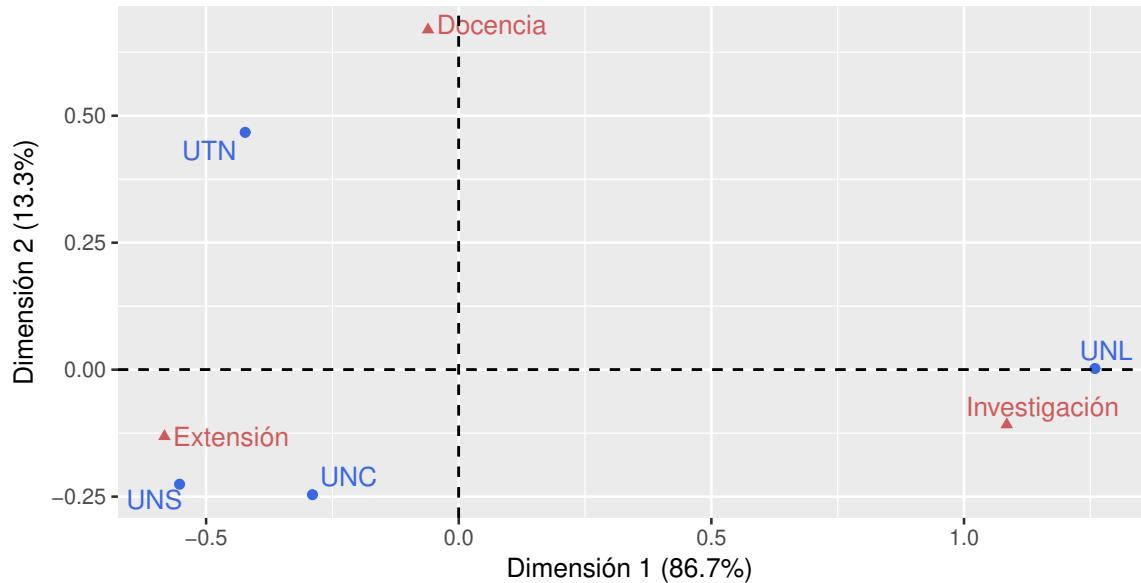


Figura 5.12: Biplot simétrico de las actividades universitarias

Presentamos a continuación un nuevo ejemplo.

**Ejemplo 5.9.** El objetivo de interés consiste en estudiar si, para la población de jóvenes estudiantes universitarios, existe una asociación entre la práctica de algún deporte y la ausencia de depresión. Para tal fin, se ha seleccionado una muestra aleatoria simple de 100 jóvenes universitarios. Sobre cada uno de estos jóvenes se observaron conjuntamente la presencia de depresión y la frecuencia con la que se realiza alguna práctica deportiva. Utilizando un nivel de significación del 5%, vamos a contrastar estas hipótesis.

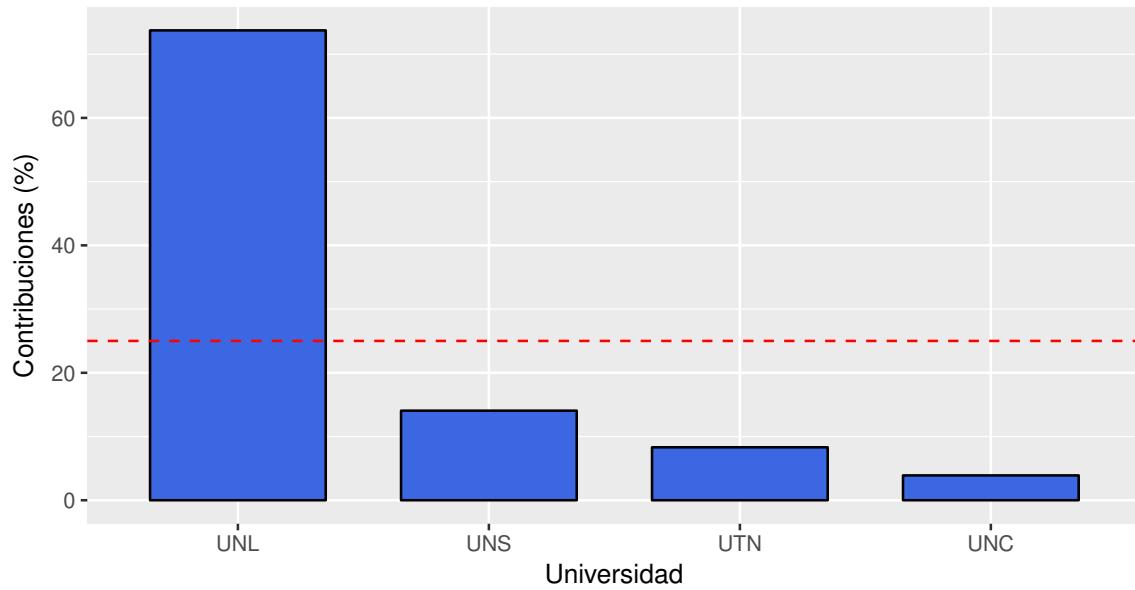


Figura 5.13: Contribución de las filas de las actividades universitarias

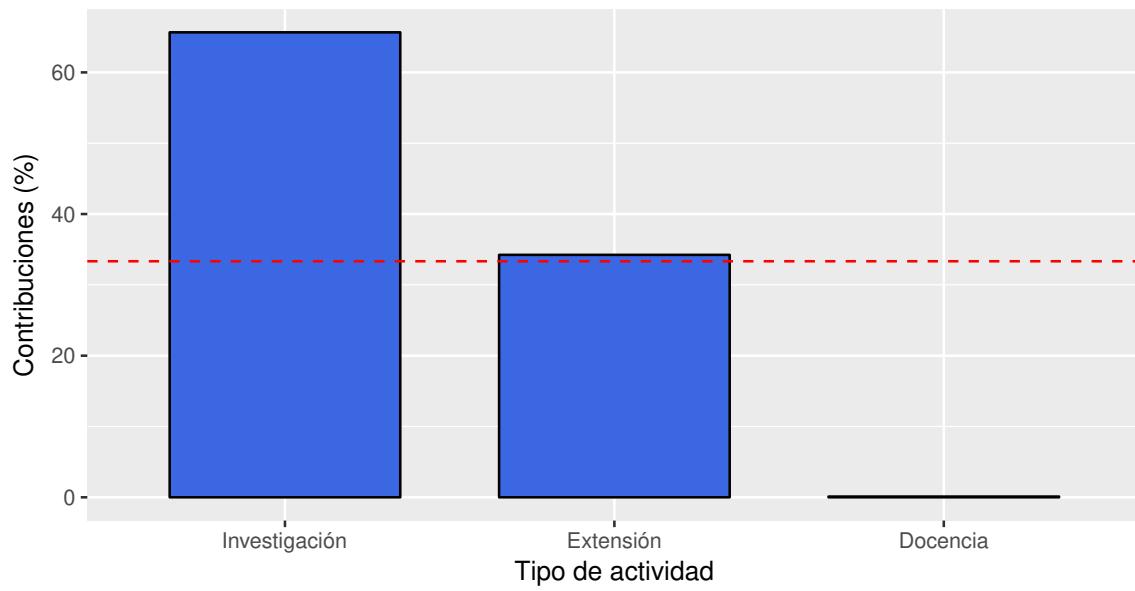


Figura 5.14: Contribución de las columnas de las actividades universitarias



<https://flic.kr/p/qGEafE>

Los datos obtenidos se presentan en la Tabla 5.19.

Frecuencia en la práctica deportiva	Ausencia de depresión	Presencia de depresión	Totales
Sin práctica	31	22	53
Hasta tres veces por semana	38	10	48
Más de tres veces por semana	40	6	46
<b>Totales</b>	<b>109</b>	<b>38</b>	<b>147</b>

**Tabla 5.19:** Datos para analizar ausencia de depresión según práctica deportiva

Recordemos que en el caso de la prueba de independencia tenemos una sola población y dos variables que se observan simultáneamente sobre cada individuo de la población. En nuestro caso, las variables a observar están dadas por

✿  $X$ : 'frecuencia en práctica deportiva'

✿  $Y$ : 'estado de depresión'

Entonces, tenemos que la variable  $X$  presenta tres niveles, equivale a la existencia de tres filas, y la variable  $Y$  presenta dos niveles, dando lugar a dos columnas.

Calculamos las frecuencias esperadas bajo independencia para este caso y anotamos los valores

esperados en la Tabla 5.20.

$$\begin{aligned}\hat{e}_{11} &= \frac{109 \times 53}{147} = 39.3 & \hat{e}_{12} &= \frac{38 \times 53}{147} = 13.7 \\ \hat{e}_{21} &= \frac{109 \times 48}{147} = 35.6 & \hat{e}_{22} &= \frac{38 \times 48}{147} = 12.4 \\ \hat{e}_{31} &= \frac{109 \times 46}{147} = 34.1 & \hat{e}_{32} &= \frac{38 \times 46}{147} = 11.9\end{aligned}$$

Frecuencia en la práctica deportiva	Ausencia de depresión	Presencia de depresión	Totales
Sin práctica	39.3	13.7	53
Hasta tres veces por semana	35.6	12.4	48
Más de tres veces por semana	34.1	11.9	46
<b>Totales</b>	<b>109</b>	<b>38</b>	<b>147</b>

**Tabla 5.20:** Frecuencias esperadas bajo independencia para ausencia de depresión según práctica deportiva

Se debe comparar con el percentil 95 de la distribución  $\chi^2$  con  $(3 - 1)(2 - 1) = 2$  grados de libertad que vale 5.99; es decir, se rechaza la hipótesis nula si el estadístico de contraste supera este valor. El estadístico de contraste es

$$\chi^2_{obs} = \frac{(31 - 39.3)^2}{39.3} + \frac{(22 - 13.7)^2}{13.7} + \frac{(38 - 35.6)^2}{35.6} + \frac{(10 - 12.4)^2}{12.4} + \frac{(40 - 34.1)^2}{34.1} + \frac{(6 - 11.9)^2}{11.9} = 11.34$$

Como la decisión a un nivel de significación del 5% es rechazar la hipótesis de nulidad que afirma la independencia entre las dos variables, se asume que existe relación entre la ausencia de depresión y los hábitos deportivos del individuo. La salida de R es:

```
Pearson's Chi-squared test
data: deporte
X-squared = 11.346, df = 2, p-value = 0.003437
```



Cuando la hipótesis nula se rechaza, debe suponerse que las variables  $X$  e  $Y$  son dependientes. Sin embargo, el test Chi cuadrado no señala en qué sentido están asociadas. Vale decir que no nos indica en qué nivel una de ellas se comporta muy distinto de lo esperado ni en qué sentido. Si deseamos indagar al respecto podríamos:

- \* Analizar los perfiles condicionales fila y columna.

- Estudiar los residuos del modelo para estudiar qué tipo de dependencia existe entre las variables.

Los residuos más utilizados son los llamados **residuos tipificados corregidos o ajustados** que vienen dados por la expresión

$$r_{ij} = \frac{o_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij} \left(1 - \frac{n_{i\cdot}}{n}\right) \left(1 - \frac{n_{\cdot j}}{n}\right)}}$$

donde  $n_{i\cdot}$  es la suma de la fila  $i$ -ésima,  $n_{\cdot j}$  es la suma de la columna  $j$ -ésima y  $n$  es el total de datos.

Estos residuos tomarán valores absolutos grandes cuando la correspondiente celda registre valores observados muy diferentes de los esperados.

*¿Cuándo debe considerarse que un residuo es alto?*

Dado que los residuos tienen distribución asintótica Normal estándar bajo la hipótesis nula, un valor absoluto del residuo superior a 2 nos indica que debemos prestar atención a dicha casilla.

Los residuos correspondientes al Ejemplo 5.9 se muestran en la Tabla 5.21.

Frecuencia en la práctica deportiva	Ausencia de depresión	Presencia de depresión
Sin práctica	-3.256443	3.256552
Hasta tres veces por semana	0.964038	-0.964466
Más de tres veces por semana	2.397395	-2.396194

Tabla 5.21: Residuos para analizar ausencia de depresión según práctica deportiva

Podemos observar que el más alto de los residuos corresponde al caso de un individuo que no practica deporte y que tiene depresión.

Bajo independencia esperaríamos que la cantidad de individuos que tiene depresión se presente en igual proporción en los distintos niveles de práctica deportiva. Sin embargo, se da en mayor proporción entre los que no lo practican.

La pregunta que podríamos hacernos es cómo cuantificar la relevancia de esta diferencia entre los valores observados y los esperados. Hasta ahora, habíamos calculado el estadístico Chi cuadrado de Pearson, pero este estadístico está afectado por la cantidad de datos. Además, como veremos en el siguiente apartado, se puede vincular esta medida con la distancia de los perfiles (fila o columna) a su respectivo perfil medio.

### 5.1.1.3 Estadístico de Pearson y la inercia

La suma de todas las distancias entre los perfiles fila y el perfil fila promedio, ponderadas por su importancia (cantidad de observaciones) se conoce como **inercia total** de la tabla de contingencia. La inercia total se calcula como

$$I_T = \sum_{i=1}^r f_i (R_i - r_m)^t D_c^{-1} (R_i - r_m) = \frac{\chi^2}{n}$$

donde  $r_m$  es el perfil medio de las filas; es decir, la estimación del perfil esperado por filas.

Se puede demostrar que la inercia total es la suma de los autovalores de la matriz  $Z^t Z$ , que coincide con la suma de los autovalores de  $Z Z^t$  (debido a que una es la traspuesta de la otra), por lo cual el análisis de las filas o de las columnas es simétrico y puede verse como una descomposición de los componentes del estadístico  $\chi^2$  en sus fuentes de variación.

La distancia Chi cuadrado tiene una propiedad importante que se conoce como el **principio de equivalencia distribucional**, que implica que si dos filas tienen la misma estructura y se agrupan en una nueva fila, las distancias entre las restantes filas permanecen invariantes. Por supuesto esta misma propiedad siguen valiendo para las columnas. Esta característica es importante pues asegura la invarianza del procedimiento por agregación de categorías irrelevantes.

Observemos, para el Ejemplo 5.9, que el estadístico de Pearson  $\chi^2_{obs}$  puede expresarse también de la siguiente manera

$$\left[ \frac{(31 - 39.3)^2}{39.3} + \frac{(22 - 13.7)^2}{13.7} \right] + \left[ \frac{(38 - 35.6)^2}{35.6} + \frac{(10 - 12.4)^2}{12.4} \right] + \left[ \frac{(40 - 34.1)^2}{34.1} + \frac{(6 - 11.9)^2}{11.9} \right]$$

Dividiendo numerador y denominador por el cuadrado del total de la fila, obtenemos

$$\left[ \frac{\left( \frac{31}{53} - \frac{39.3}{53} \right)^2}{\frac{39.3}{53^2}} + \frac{\left( \frac{22}{53} - \frac{13.7}{53} \right)^2}{\frac{13.7}{53^2}} \right] + \left[ \frac{\left( \frac{38}{48} - \frac{35.6}{48} \right)^2}{\frac{35.6}{48^2}} + \frac{\left( \frac{10}{48} - \frac{12.4}{48} \right)^2}{\frac{12.4}{48^2}} \right] + \left[ \frac{\left( \frac{40}{46} - \frac{34.1}{46} \right)^2}{\frac{34.1}{46^2}} + \frac{\left( \frac{6}{46} - \frac{11.9}{46} \right)^2}{\frac{11.9}{46^2}} \right]$$

Y extrayendo el total de la fila como factor común, concluimos

$$53 \left[ \frac{\left( \frac{31}{53} - \frac{39.3}{53} \right)^2}{\frac{39.3}{53}} + \frac{\left( \frac{22}{53} - \frac{13.7}{53} \right)^2}{\frac{13.7}{53}} \right] + 48 \left[ \frac{\left( \frac{38}{48} - \frac{35.6}{48} \right)^2}{\frac{35.6}{48}} + \frac{\left( \frac{10}{48} - \frac{12.4}{48} \right)^2}{\frac{12.4}{48}} \right] + 46 \left[ \frac{\left( \frac{40}{46} - \frac{34.1}{46} \right)^2}{\frac{34.1}{46}} + \frac{\left( \frac{6}{46} - \frac{11.9}{46} \right)^2}{\frac{11.9}{46}} \right]$$

De esta expresión es fácil ver que  $\chi^2_{obs}$  es igual a la suma de

$$\text{total de la fila} \times \frac{(\text{perfil fila observado de la casilla} - \text{perfil fila esperado de la casilla})^2}{\text{perfil fila esperado de la casilla}}$$

Esta forma comienza a parecerse a una distancia entre perfiles esperados y observados ponderados por el peso de la cantidad de observaciones de la fila.

Dividiendo ambos miembros por el total de observaciones tenemos que

$$\frac{\chi^2_{obs}}{147} = \frac{53}{147} \left[ \frac{\left( \frac{31}{53} - \frac{39.3}{53} \right)^2}{\frac{39.3}{53}} + \frac{\left( \frac{22}{53} - \frac{13.7}{53} \right)^2}{\frac{13.7}{53}} \right] + \frac{48}{147} \left[ \frac{\left( \frac{38}{48} - \frac{35.6}{48} \right)^2}{\frac{35.6}{48}} + \frac{\left( \frac{10}{48} - \frac{12.4}{48} \right)^2}{\frac{12.4}{48}} \right] + \frac{46}{147} \left[ \frac{\left( \frac{40}{46} - \frac{34.1}{46} \right)^2}{\frac{34.1}{46}} + \frac{\left( \frac{6}{46} - \frac{11.9}{46} \right)^2}{\frac{11.9}{46}} \right]$$

Hemos dicho que el cociente entre el estadístico Chi cuadrado de Pearson y el total de observaciones se denomina **inercia**. Entonces, con el cálculo anterior, hemos expresado a la inercia como la suma de las distancias entre los perfiles observados y el perfil esperado ponderadas por los perfiles esperados y la masa de las filas (frecuencia relativa del total de las filas). La inercia es entonces una medida de la variabilidad total de la tabla, independientemente de su tamaño.

En Física, la inercia se define como la suma de los cuadrados de las distancias al centro de gravedad, para nosotros el ‘centro de gravedad’ es el perfil medio. Es una medida similar a la variabilidad total de las componentes principales y mide el grado total de dependencia existente entre las variables  $X$  e  $Y$ .

Los estadísticos han denominado a este valor de distintas maneras, una de ellas es **coeficiente medio cuadrático de contingencia**. Su raíz cuadrada se denominado **coeficiente  $\phi$** , por lo cual se puede decir que la inercia es igual a  $\phi^2$ .

#### 5.1.1.4 Interpretación geométrica de la inercia

Como ya hemos visto, la inercia mide la magnitud de la distancia entre los perfiles fila (resp. columna) y el perfil fila (resp. columna) medio. Es decir, mide la distancia entre los perfiles fila (resp. columna) observados y los perfiles fila (resp. columna) esperados bajo la hipótesis de independencia.

Cuando esta distancia es grande, significa que existe asociación entre alguno de los perfiles fila y alguno de los perfiles columna, lo que deriva en que podremos descubrir algunas relaciones presentes en la distribución conjunta de las dos variables de interés.

En la Figura 5.16 se grafican los perfiles fila de un conjunto de datos y la inercia de su respectiva tabla de contingencia.

Podemos observar la siguiente relación de orden entre las inercias

$$\text{Inercia 1} < \text{Inercia 2} < \text{Inercia 3} < \text{Inercia 4}$$

*¿Qué sucede con las posiciones relativas de los perfiles fila a medida que la inercia aumenta?*

*¿Cómo se vincula esto con la hipótesis de independencia?*

El análisis de correspondencias propone la construcción de un sistema de coordenadas (habitualmente bidimensional) asociado a las filas y a las columnas de una tabla de contingencia, que

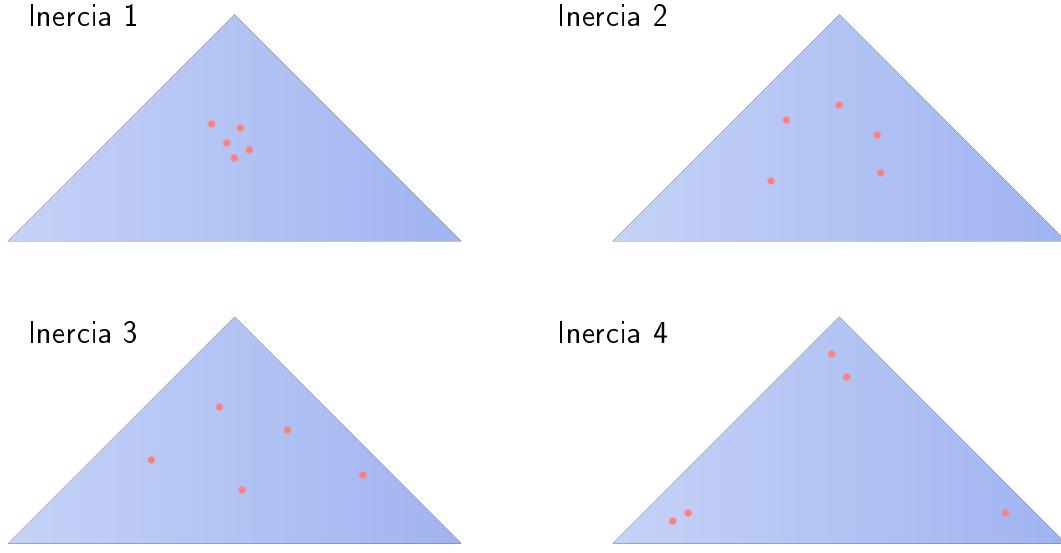


Figura 5.16: Ejemplos de inercias

refleje las relaciones existentes entre dichas filas y columnas. En dicha representación, juegan un papel importante las llamadas *distancias*  $\chi^2$  entre perfiles, que son las que el análisis de correspondencias intenta reproducir en su representación gráfica. Dichas distancias son distancias pitagóricas ponderadas entre perfiles que vienen dadas por las siguientes fórmulas:

- \*  $d_{ij} = \sum_{h=1}^r \frac{1}{n_{\cdot h}} \left( \frac{n_{ih}}{n_{\cdot i}} - \frac{n_{jh}}{n_{\cdot j}} \right)^2$  para los perfiles fila
- \*  $d_{ij} = \sum_{h=1}^k \frac{1}{n_{h \cdot}} \left( \frac{n_{hi}}{n_{\cdot i}} - \frac{n_{hj}}{n_{\cdot j}} \right)^2$  para los perfiles columna

donde usamos en forma indistinta, según el caso,  $n_{ij}$  para indicar  $O_{ij}$  o  $f_{ij}$ . En las Tablas 5.22 y 5.23 se calculan estas distancias para el Ejemplo 5.9.

Para la Tabla 5.22, calculemos a modo de ejemplo las distancias  $\chi^2$  entre los siguientes perfiles fila:

$$d_{12} = \frac{1}{109}(0.5849 - 0.7917)^2 + \frac{1}{38}(0.4151 - 0.2083)^2 = 0.001517$$

$$d_{13} = \frac{1}{109}(0.5849 - 0.8696)^2 + \frac{1}{38}(0.4151 - 0.1304)^2 = 0.002876$$

Para la Tabla 5.23, ejegimos calcular a modo de ejemplo la distancia  $\chi^2$  entre los siguientes

Frecuencia en la práctica deportiva	Ausencia de depresión	Presencia de depresión
Sin práctica	0.5849	0.4151
Hasta tres veces por semana	0.7917	0.2083
Más de tres veces por semana	0.8696	0.1304

Tabla 5.22: Distancias entre perfiles fila

Frecuencia en la práctica deportiva	Ausencia de depresión	Presencia de depresión
Sin práctica	0.2844	0.5789
Hasta tres veces por semana	0.3486	0.2632
Más de tres veces por semana	0.3670	0.1579

Tabla 5.23: Distancias entre perfiles columna

perfiles columna:

$$d_{12} = \frac{1}{53}(0.2844 - 0.5789)^2 + \frac{1}{48}(0.3486 - 0.2632)^2 + \frac{1}{46}(0.3670 - 0.1579)^2 = 0.002739$$

A partir de estas distancias, ¿cuáles serían las filas más similares? ¿La de no práctica deportiva con la de una frecuencia de menos de 3 veces por semana? O, ¿la de no práctica deportiva con la de una frecuencia de más de tres veces por semana?

Por ende las distancias  $\chi^2$  son invariantes a variaciones en la codificación de las categorías con comportamiento similar en cuanto a sus perfiles condicionales.

### 5.1.2 Análisis de correspondencias múltiples

Dado el nivel de complejidad del problema, el análisis de correspondencias puede subdividirse en dos categorías:

- ✿ Cuando se trata de tablas de contingencia de dos variables, estamos frente a un análisis de correspondencias simples (ACS), sin importar la cantidad de niveles que tengan estas dos variables.
- ✿ Cuando el número de variables registradas es superior a dos, diremos que el análisis es un análisis de correspondencias múltiples (ACM).

La idea en el análisis de correspondencias múltiples es la misma que en el análisis de correspondencias simples que hemos estado tratando en las secciones previas. El objetivo es reducir la

dimensión del problema y lograr una representación que, perdiendo la menor cantidad de información posible, represente nuestra tabla de contingencias.

**Ejemplo 5.10.** Consideremos un conjunto de 12 personas en las que hemos observado cuatro variables. En la Tabla 5.24 hemos registrado para cada una de estas 12 observaciones, sus categorías en las cuatro variables observadas.



<https://flic.kr/p/p1xq92>

Observación	Género	Edad	Estado civil	Color de cabello
1	M	joven	soltero	castaño
2	M	adulto	soltero	rojizo
3	F	mayor	casado	rubio
4	M	adulto	soltero	negro
5	F	mayor	casado	negro
6	F	mayor	soltero	castaño
7	M	joven	casado	rojizo
8	M	adulto	casado	rubio
9	M	mayor	soltero	castaño
10	F	joven	casado	negro
11	F	adulto	soltero	castaño
12	M	joven	casado	rubio

Tabla 5.24: Características observadas

Otra forma de presentar estos datos podría ser la que se muestra en la Tabla 5.25. Este tipo de matriz se denomina **matriz disyuntiva** y la vamos a designar con  $G$ .

La matriz traspuesta, será  $G^t$  y se muestra en la Tabla 5.26.

Obs.	Género		Edad			Estado civil		Color de Cabello			
	M	F	Jvn.	Adt.	Myr.	Slt.	Csd.	Rbo.	Cst.	Rjz.	Ngr.
1	1	0	1	0	0	1	0	0	1	0	0
2	1	0	0	1	0	1	0	0	0	1	0
3	0	1	0	0	1	0	1	1	0	0	0
4	1	0	0	1	0	1	0	0	0	0	1
5	0	1	0	0	1	0	1	0	0	0	1
6	0	1	0	0	1	1	0	0	1	0	0
7	1	0	1	0	0	0	1	0	0	1	0
8	1	0	0	1	0	0	1	1	0	0	0
9	1	0	0	0	1	1	0	0	1	0	0
10	0	1	1	0	0	0	1	0	0	0	1
11	0	1	0	1	0	1	0	0	1	0	0
12	1	0	1	0	0	0	1	1	0	0	0

Tabla 5.25: Matriz disyuntiva para las características observadas

1	1	0	1	0	0	1	1	1	0	0	1
0	0	1	0	1	1	0	0	0	1	1	0
1	0	0	0	0	0	1	0	0	1	0	1
0	1	0	1	0	0	0	1	0	0	1	0
0	0	1	0	1	1	0	0	1	0	0	0
1	1	0	1	0	1	0	0	1	0	1	0
0	0	1	0	1	0	1	1	0	1	0	1
0	0	1	0	0	0	0	1	0	0	0	1
1	0	0	0	0	1	0	0	1	0	1	0
0	1	0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	0	0	0	0	1	0	0

Tabla 5.26: Matriz  $G^t$

### 5.1.2.1 Matriz de Burt

Listaremos una serie de preguntas que intentaremos responder observando la matriz 5.26.

- ✿ ¿Cuántas variables participan de este análisis?
- ✿ ¿Qué aspecto tienen las matrices señaladas en los diferentes colores?
- ✿ ¿Cuáles son los elementos diagonales?
- ✿ ¿Cuánto valen los elementos no diagonales?

En el Ejemplo 5.10, las variables consideradas son las cuatro siguientes:

- ✿ **Género** con dos niveles: Masculino y Femenino
- ✿ **Edad** con tres niveles: Joven, Adulto y Mayor
- ✿ **Estado Civil** con dos niveles: Soltero y Casado
- ✿ **Color de cabello** con cuatro niveles: Rubio, Castaño, Rojizo y Negro

Si denotamos a la **matriz de Burt** con  $B$  y con  $B_{ij}$  al elemento de esta matriz correspondiente a la fila  $i$  y a la columna  $j$ , podemos hacer las siguientes observaciones.

- ✿ En la intersección de la  $j$ -ésima línea y de la  $j$ -ésima columna, dada por el valor  $B_{jj}$ , se encuentra el número de individuos que presentaron la  $j$ -ésima modalidad de una característica de ese bloque. En la Tabla 5.27,  $B_{33}$  indica la cantidad de jóvenes de la muestra que es 4.
- ✿ En la intersección de la  $i$ -ésima fila y de la  $j$ -ésima columna, dada por el valor  $B_{ij}$ , se encuentra un 0 si se refieren a la misma modalidad pero con distinto nivel,  $i \neq j$ . Es decir,  $B_{12} = B_{21} = 0$  en la Tabla 5.27, ya que no puede ser al mismo tiempo varón y mujer.
- ✿ En la intersección de la  $i$ -ésima fila con la  $j$ -ésima columna, el valor  $B_{ij}$  indica la cantidad de individuos que presentaron simultáneamente la  $i$ -ésima modalidad de una característica y la  $j$ -ésima modalidad de otra característica observada. Por ejemplo, en la Tabla 5.27  $B_{25} = B_{52} = 3$  significa que se observaron 3 mujeres mayores.
- ✿ La matriz  $B$  resulta del producto entre la matriz disyuntiva completa y su traspuesta. Este hecho implica que la matriz de Burt es una matriz simétrica. Además, es una matriz definida positiva como la matriz de varianzas y covarianzas.
- ✿ La matriz  $B$  puede no ser de rango completo.

¿Por qué puede suceder que  $B$  no sea de rango completo? ¿Qué impacto tiene en tal caso sobre el análisis de correspondencias?

	M	F	Jvn.	Adt.	Myr.	Slt.	Csd.	Rbo.	Cst.	Rjz.	Ngr.
M	7	0	3	3	1	4	3	2	2	2	1
F	0	5	1	1	3	2	3	1	2	0	2
Jvn.	3	1	4	0	0	1	3	1	1	1	1
Adt.	3	1	0	4	0	3	1	1	1	1	1
Myr.	1	3	0	0	4	2	2	1	2	0	1
Slt.	4	2	1	3	2	6	0	0	4	1	1
Csd.	3	3	3	1	2	0	6	3	0	1	2
Rbo.	2	1	1	1	1	0	3	3	0	0	0
Cst.	2	2	1	1	2	4	0	0	4	0	0
Rjz.	2	0	1	1	0	1	1	0	0	2	0
Ngr.	1	2	1	1	1	1	2	0	0	0	3

Tabla 5.27: Matriz de Burt para el Ejemplo 5.10

Las Figuras 5.18, 5.19, 5.20, 5.21 y 5.22 son distintas representaciones gráficas para las cuatro variables de interés. Todas fueron generadas por el Código 5.8 y con datos extraídos de <https://goo.gl/8pTyTW>.

```
library(readxl) # Permite leer archivos xlsx
library(FactoMineR) # Paquete con métodos de análisis exploratorio de datos
library(factoextra) # Paquete para análisis multivariado de datos

personas=read_excel("C:/.../personas.xlsx")
# Importa la base con la cual se va a trabajar

base=data.frame(personas)
personas.acm=MCA(base[2:5], quali.sup=1, graph=F)
# Realiza el análisis de correspondencias múltiple

# las variables deben ser introducidas como factores

fviz_contrib(personas.acm, choice="var", axes=1,
fill="royalblue", color = "black") +
theme_gray() +
xlab('') +
ylab('Contribuciones (%)') +
ggtitle('')
# Grafica las contribuciones de las variables

fviz_contrib(personas.acm, choice="ind", axes=1, top=5,
fill="royalblue", color = "black") +
theme_gray() +
xlab('')
```

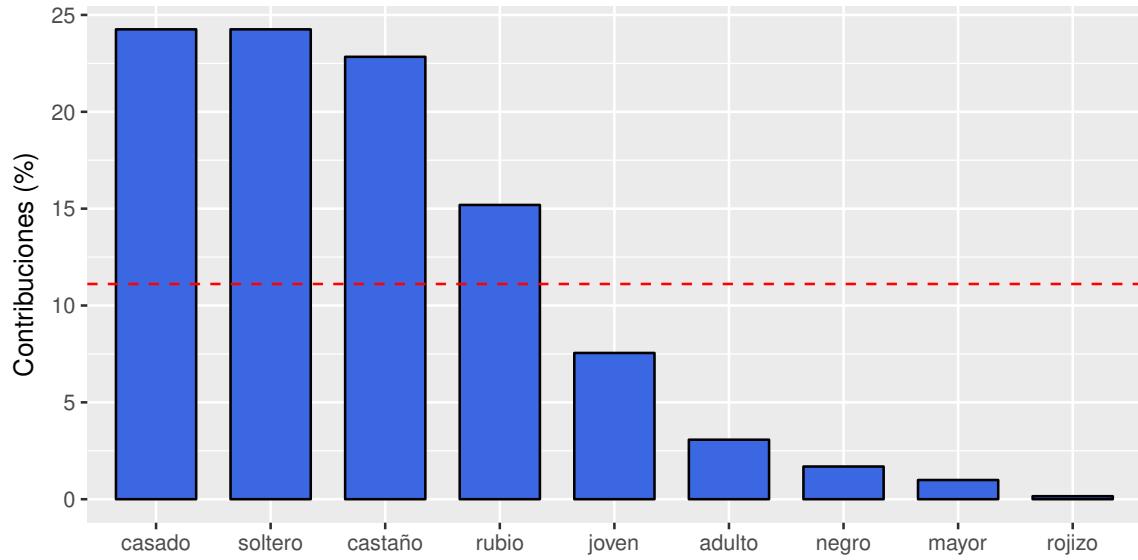


Figura 5.18: Contribución a la inercia de las variables para la dimensión 1

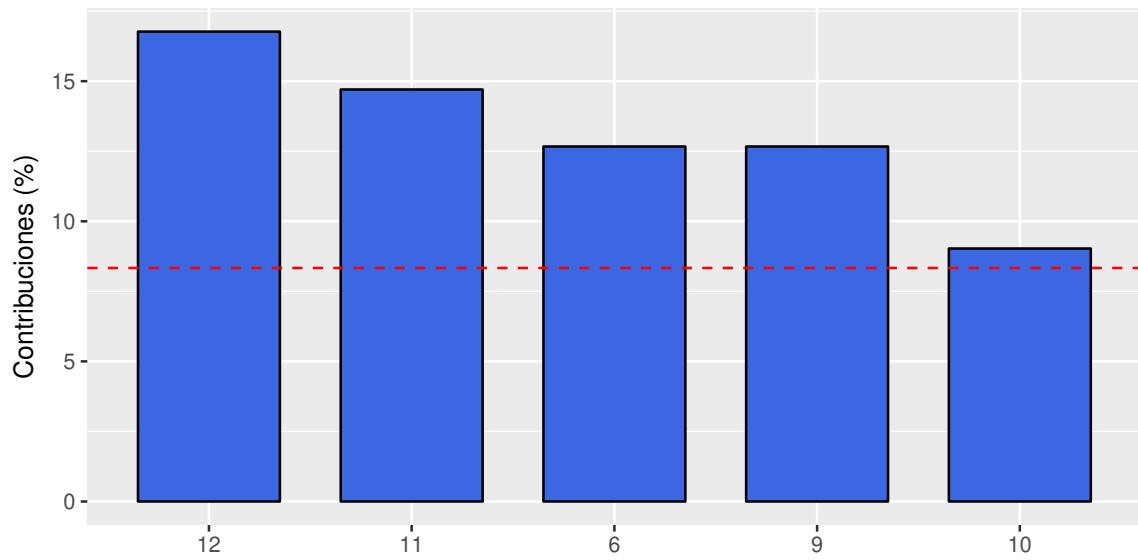


Figura 5.19: Contribución a la inercia de los individuos para la dimensión 1

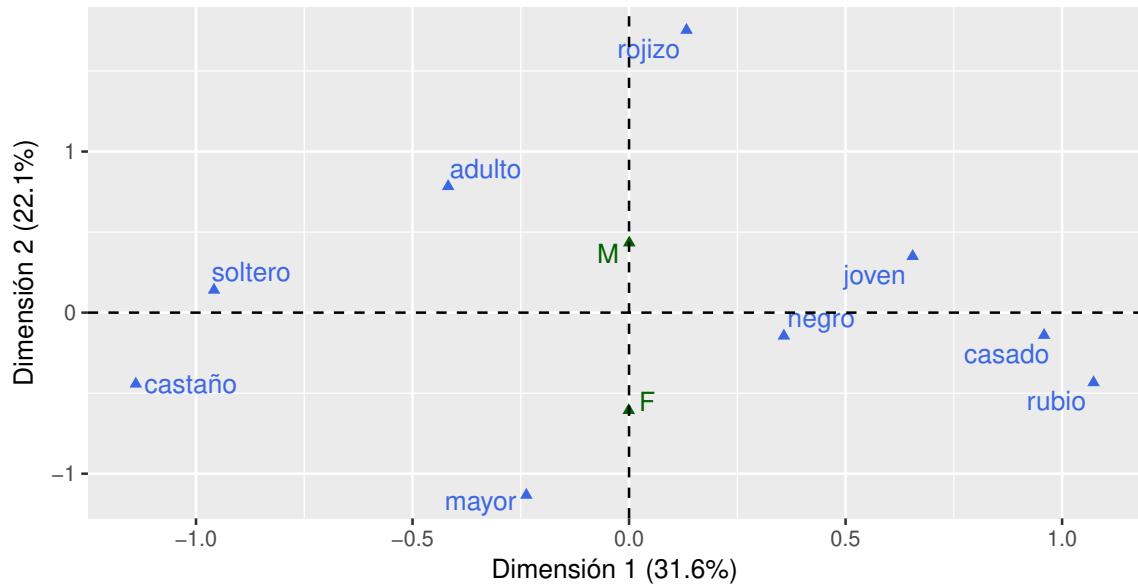


Figura 5.20: Categorías variables - ACM

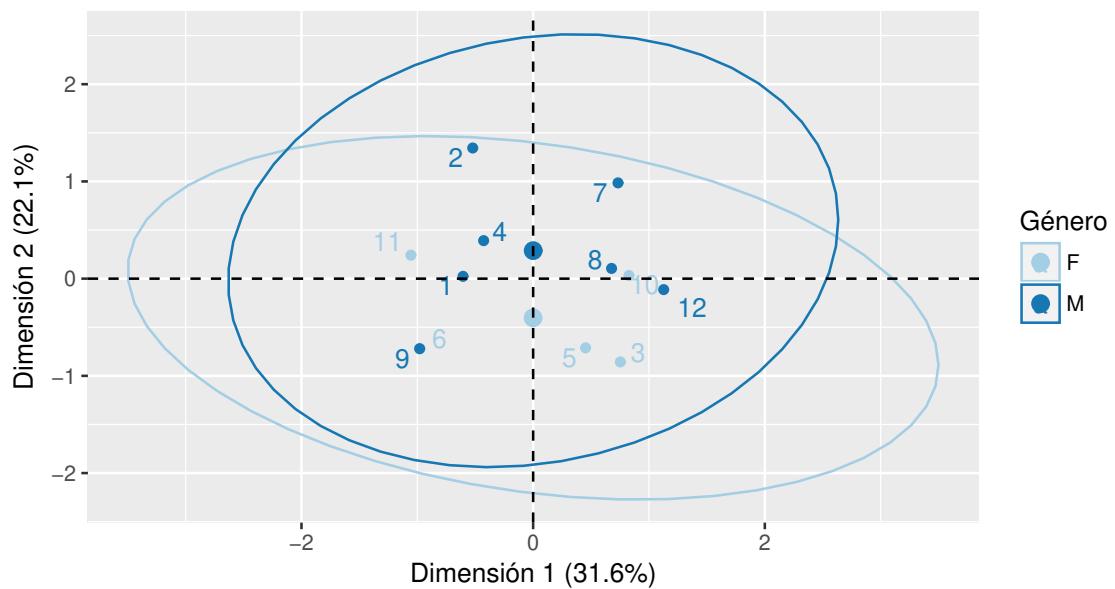


Figura 5.21: Individuos agrupados por género - ACM

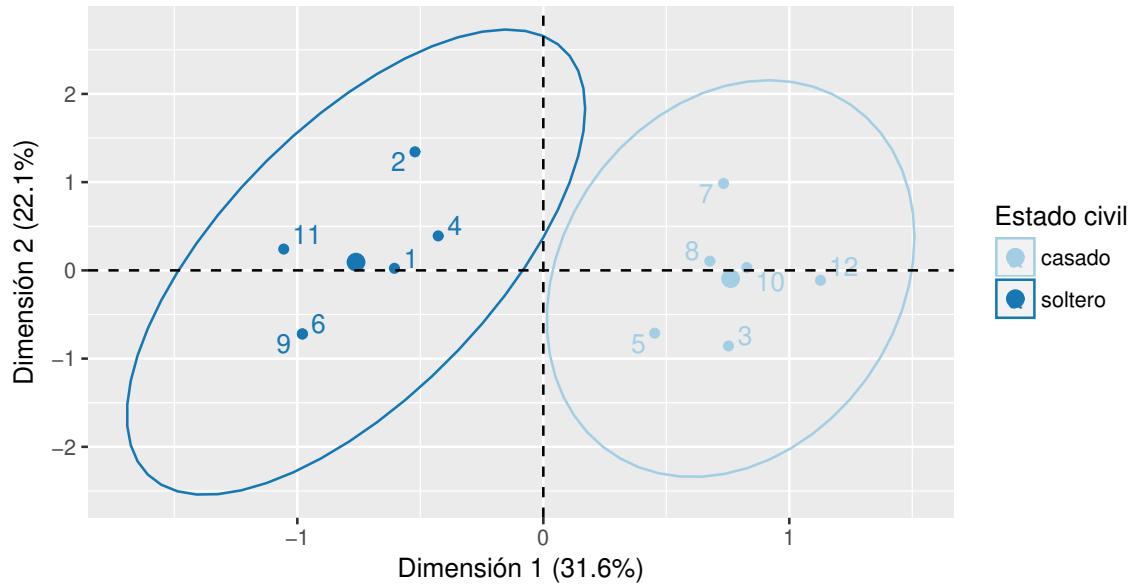


Figura 5.22: Individuos agrupados por estado civil - ACM

```

ylab('Contribuciones(%)') +
ggtitle('')
# Grafica las contribuciones de los individuos

fviz_mca_var(personas.acm, repel = TRUE, col.var="royalblue") +
theme_gray() +
xlab('Dimensión_1_(31.6%)') +
ylab('Dimensión_2_(22.1%)') +
ggtitle('')
# Realiza el biplot simétrico

fviz_mca_ind(personas.acm, habillage=factor(personas$Género),
addEllipses=TRUE, repel=TRUE, legend.title = "Género") +
theme_gray() +
xlab('Dimensión_1_(31.6%)') +
ylab('Dimensión_2_(22.1%)') +
ggtitle('') +
scale_color_brewer(palette="Paired")
# Realiza un agrupamiento por género

fviz_mca_ind(personas.acm, habillage=factor(personas$Estado),
addEllipses=TRUE, repel=TRUE, legend.title = "Estado_civil") +
theme_gray() +

```

```

xlab('Dimensión_1_(31.6%)') +
ylab('Dimensión_2_(22.1%)') +
ggtitle('') +
scale_color_brewer(palette="Paired")
# Realiza un agrupamiento por estado civil

```

**Código 5.8:** Código para el análisis de correspondencias múltiples de un grupo de personas

Citamos algunas claves para interpretar un *biplot* simétrico.

- ✿ Dos categorías de una variable se parecen si tienen casi las mismas frecuencias relativas en cada una de las modalidades de la otra.
- ✿ Dos modalidades de variables diferentes son cercanas si aparecen conjuntamente en los mismos individuos con mucha frecuencia.
- ✿ Dos modalidades de una misma variable son excluyentes por construcción. Si las mismas aparecen representadas de manera cercana, es porque presentan casi el mismo comportamiento respecto de las restantes variables.

Haciendo un análisis de la Figura 5.20, podemos sacar las siguientes conclusiones.

- ✿ Las modalidades ‘casado’ y ‘cabello rubio’ aparecen cercanas en la representación. Esto significa que para este conjunto de individuos, los casados tienen tendencia al cabello rubio. Si examinamos la base original veremos que se encuentran 4 casos de casados con cabello rubio.
- ✿ Conclusiones como la anterior son sencillas de examinar claramente en este caso en el que sólo se tiene una muestra de 12 individuos y unas pocas características observadas de cada uno de ellos. Sin embargo, en una gran tabla de datos esto puede hacernos encontrar patrones que a simple vista se nos escaparían casi con seguridad.
- ✿ Las modalidades para el color del cabello aparecen bien separadas en la representación.
- ✿ La modalidad ‘soltero’ aparece próxima a la modalidad ‘castaño’. Examinando la tabla original, encontramos 4 solteros y castaños.

Con el siguiente ejemplo revisaremos los conceptos del ACM.

**Ejemplo 5.11.** Una empresa desea analizar si los individuos que trabajan en ella, siguen algún patrón vinculado al género, sus ingresos y la antigüedad de los mismos en la empresa. Resulta de interés responder a preguntas como ¿podría asegurarse que la empresa paga la fidelidad?, ¿podría asegurarse que la empresa prefiere empleados de algún género en particular?, ¿podría asegurarse que los que eligen permanecer en la empresa son de un género determinado?



<https://flic.kr/p/dSG9KF>

Podríamos intentar responder a estas preguntas de manera sencilla realizando un test de independencia para cada par de variables involucradas en la pregunta. Sin embargo, con este procedimiento estaríamos cometiendo al menos dos errores importantes como pueden ser:

- ✿ realizar demasiadas pruebas con un sólo conjunto de datos y perder potencia.
- ✿ perder la riqueza del análisis conjunto de las tres variables con sus interacciones.

Debido a estas razones, la mejor de las opciones sería realizar un análisis de correspondencias múltiple.

En la Tabla 5.28 se encuentran las observaciones realizadas sobre 10 individuos de la empresa, en función del género, los años en la empresa y los ingresos mensuales.

Individuo	Género	Antigüedad	Ingresos
1	Mujer	5	Medio
2	Mujer	3	Alto
3	Hombre	4	Bajo
4	Mujer	1	Bajo
5	Mujer	2	Medio
6	Hombre	5	Alto
7	Mujer	2	Medio
8	Hombre	3	Bajo
9	Hombre	1	Alto
10	Mujer	4	Medio

Tabla 5.28: Situación de los empleados de una empresa

Individuo	Género		Antigüedad					Ingresos		
	Mujer	Hombre	1	2	3	4	5	Bajo	Medio	Alto
1	1	0	0	0	0	0	1	0	1	0
2	1	0	0	0	1	0	0	0	0	1
3	0	1	0	0	0	1	0	1	0	0
4	1	0	1	0	0	0	0	1	0	0
5	1	0	0	1	0	0	0	0	1	0
6	0	1	0	0	0	0	1	0	0	1
7	1	0	0	1	0	0	0	0	1	0
8	0	1	0	0	1	0	0	1	0	0
9	0	1	1	0	0	0	0	0	0	1
10	1	0	0	0	0	1	0	0	1	0

Tabla 5.29: Matriz disyuntiva para la situación de los empleados

A partir de la tabla original, se construye la tabla disyuntiva o matriz  $G$  que se presenta en la Tabla 5.29.



Es importante destacar lo siguiente.

- ✿ En la tabla disyuntiva completa  $G$ , si hay alguna variable continua debe transformarse en nominal, ordenándose en intervalos a los que se da un rango de valores. Esto es debido a que en el análisis de correspondencias se trabaja con variables categóricas.
- ✿ La tabla disyuntiva completa tiene tantas columnas como categorías y tantas filas como individuos de interés.
- ✿ Las frecuencias marginales de las filas son todas iguales al número de variables registradas sobre los individuos, y las frecuencias marginales de las columnas corresponden al número de sujetos que han elegido la modalidad  $j$  de la pregunta  $q$ . Para cada subtabla el número total de individuos es  $n$ .
- ✿ Relacionando cada variable con todas las demás la tabla disyuntiva se convierte a una tabla de Burt que contiene todas las tablas de contingencia simples entre las variables cruzadas dos a dos.
- ✿ A partir de la tabla disyuntiva completa se puede construir la tabla de contingencia de Burt  $B = G^t G$ , que es una tabla simétrica de orden  $p \times p$  donde  $p$  indica la suma de todos los niveles de las variables en cuestión.

- \*  $B$  es una yuxtaposición de tablas de contingencia y está formada por bloques. Cada bloque es una submatriz formada por tablas de contingencia correspondientes a las variables tomadas dos a dos, salvo los bloques que están en la diagonal que son las tablas de contingencia de cada variable consigo misma.

La tabla disyuntiva completa es equivalente a la tabla de Burt y ambas producen los mismos factores. Algunos programas de computación permiten ingresar la tabla disyuntiva y otros permiten ingresar la matriz de Burt. La matriz de Burt permite calcular las puntuaciones (distancias al centro de gravedad), las contribuciones absolutas de cada modalidad y variable a los ejes o factores obtenidos (contribución de cada modalidad o variable a la inercia de los nuevos ejes), las contribuciones relativas o correlaciones de cada modalidad con los nuevos ejes.

Como en la tabla de Burt las filas y las columnas representan las mismas modalidades, el estudio de ambas ofrece iguales resultados, por lo que sólo se representan los resultados obtenidos a partir de las filas.

### 5.1.2.2 Examen de los puntos

Es importante tener en cuenta los siguientes aspectos:

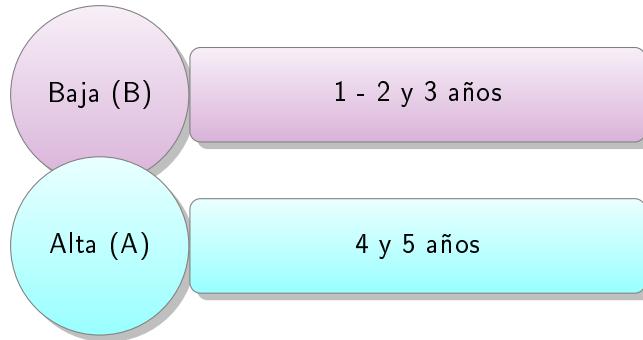
- \* Las distancias de las modalidades, mientras más alejadas se encuentren del origen, mejor representadas estarán. Cuanto más alejadas estén las modalidades entre sí en el gráfico, menor asociación existirá entre ellas; mientras que cuanto más cercanas se encuentren, más asociación existirá entre ellas.
- \* La contribución de los puntos a la inercia de cada dimensión, o contribución de cada una de las filas a la inercia o varianza explicada en cada uno de los ejes considerados.
- \* La contribución de las dimensiones a la inercia de cada punto, que se refiere a la correlación existente entre cada uno de los caracteres y los nuevos ejes.
- \* En el análisis de correspondencias múltiples, los valores propios generan una idea pesimista de la variabilidad explicada.

Por estos motivos, se recomienda medir la tasa de inercia realizando una modificación utilizando la corrección de Benzécri (1979). Se requiere de los siguientes pasos:

- \* Calcular  $b = 1/q$ , siendo  $q$  el número de variables.
- \* Seleccionar los valores propios  $VP$  iguales o superiores a  $b$ .
- \* Calcular los valores propios transformados  $VPT = (VP - b)^2$ .
- \* Calcular el porcentaje de varianza explicada  $VPE$  con los valores propios transformados, divididos por su suma.

Cada valor propio tiene una tasa de inercia sobre el total de varianza explicada por todos los valores propios transformados. Al calcular el porcentaje acumulado de varianza explicada, la parte de inercia debida a una modalidad de respuesta aumenta cuanto menor sea el número de personas de esta modalidad; es decir, cuanto menor sea su masa.

**Ejemplo 5.12.** Siguiendo el Ejemplo 5.11, definimos dos categorías para la variable dada por la antigüedad de cada empleado en la empresa:



Individuo	Género	Antigüedad	Ingresos	Categoría
1	Mujer		5	Medio
2	Mujer		3	Alto
3	Hombre		4	Bajo
4	Mujer		1	Bajo
5	Mujer		2	Medio
6	Hombre		5	Alto
7	Mujer		2	Medio
8	Hombre		3	Bajo
9	Hombre		1	Alto
10	Mujer		4	Medio

Tabla 5.30: Agregado de categoría para los empleados

El análisis de correpondencias múltiple se realiza mediante el Código 5.9 con datos extraídos de <https://goo.gl/Ca6NWu>, mediante el cual se generan las Figuras 5.24, 5.25, 5.26 y 5.27 y se obtienen las salidas dadas en la Tablas 5.31, 5.32, 5.33 y 5.34.

```

library(readxl) # Permite leer archivos xlsx
library(FactoMineR) # Paquete con métodos de análisis exploratorio de datos
library(factoextra) # Paquete para análisis multivariado de datos
library(ade4) # Paquete con herramientas para análisis multivariado de datos
library(anacor) # Paquete para análisis de correspondencias simple y canónico
  
```

```

empresa=read_excel("C:/.../empresa.xlsx")
# Importa la base con la cual se va a trabajar

Género=factor(empresa$Género)
Antigüedad=factor(empresa$Antigüedad)
Ingresos=factor(empresa$Ingresos)
Categoría=factor(empresa$Categoría)
base=data.frame(Género, Ingresos, Categoría)
# Armamos la base de datos con las variables como factores

empresa.acm=MCA(base, quali.sup=1, graph=F)
# Realiza el análisis de correspondencias múltiple

fviz_contrib(empresa.acm, choice="var", axes=1,
fill="royalblue", color = "black") +
theme_gray() +
xlab('') +
ylab('Contribuciones (%)') +
ggtitle('')
# Grafica las contribuciones de las variables

fviz_contrib(empresa.acm, choice="ind", axes=1, top=5,
fill="royalblue", color = "black") +
theme_gray() +
xlab('') +
ylab('Contribuciones (%)') +
ggtitle('')
# Grafica las contribuciones de los individuos

fviz_mca_var(empresa.acm, repel = TRUE, col.var="royalblue") +
theme_gray() +
xlab('Dimensión_1_(38.9%)') +
ylab('Dimensión_2_(33.3%)') +
ggtitle('')
# Realiza el biplot simétrico

fviz_mca_ind(empresa.acm, habillage=Género, addEllipses=TRUE,
repel=TRUE, legend.title = "Género") +
theme_gray() +
xlab('Dimensión_1_(38.9%)') +
ylab('Dimensión_2_(33.3%)') +
ggtitle('')
scale_color_brewer(palette="Paired")
# Realiza un agrupamiento por género

acm.disjonctif(base)
# Calcula la matriz disyuntiva

```

```

burtTable(base)
# Calcula la matriz de Burt

acm.empresa=dudi.acm(base, scannf = FALSE)
summary(acm.empresa)
# Calcula las inercias
round(acm.empresa$c1,3)
# Calcula las coordenadas para representar

```

Código 5.9: Código para el análisis de correspondencias múltiples para una empresa

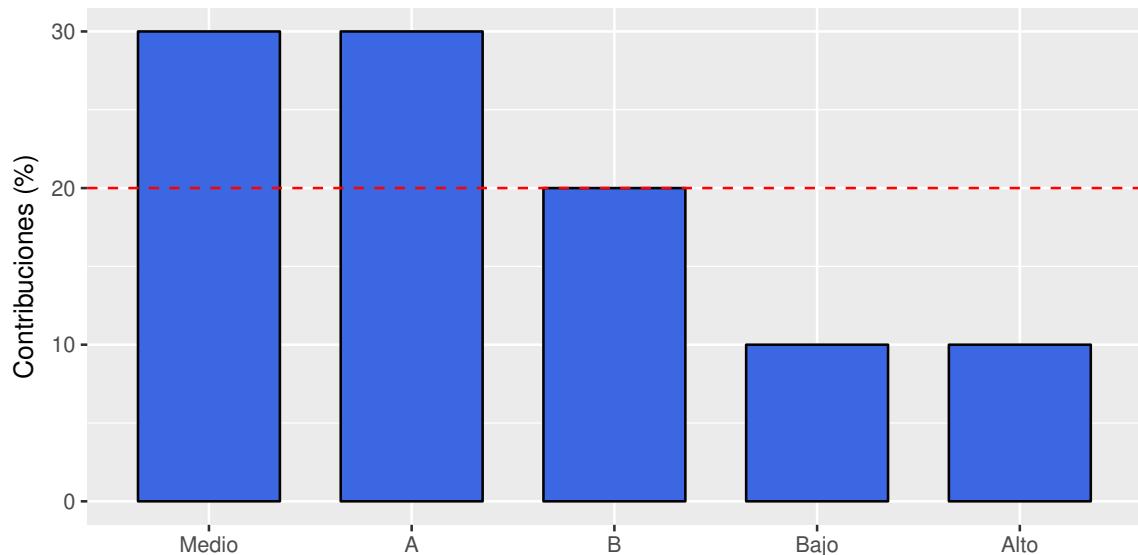


Figura 5.24: Contribución a la inercia de las variables



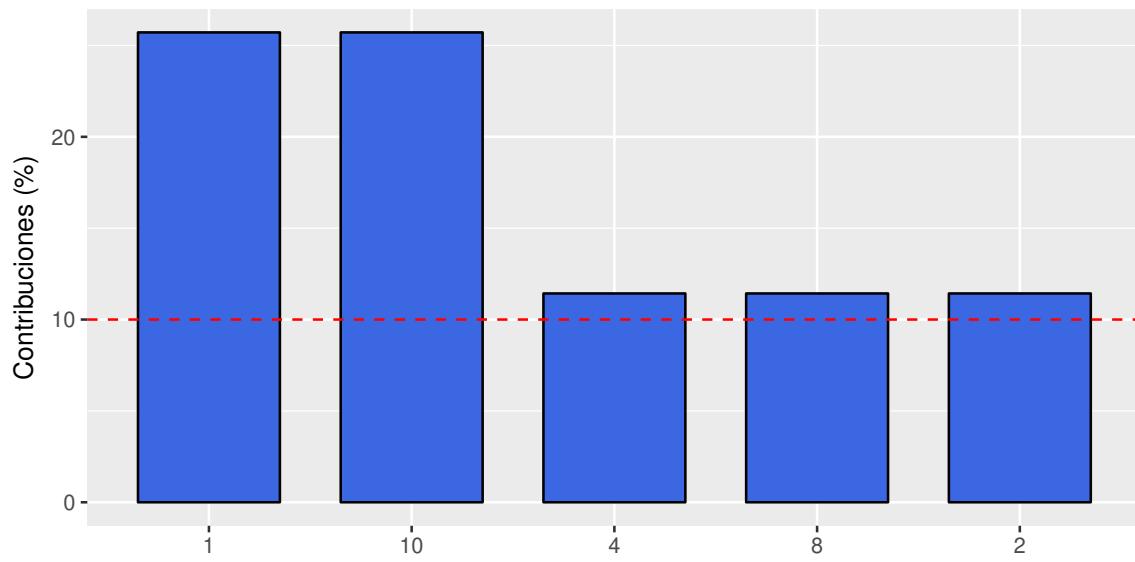


Figura 5.25: Contribución a la inercia de los individuos

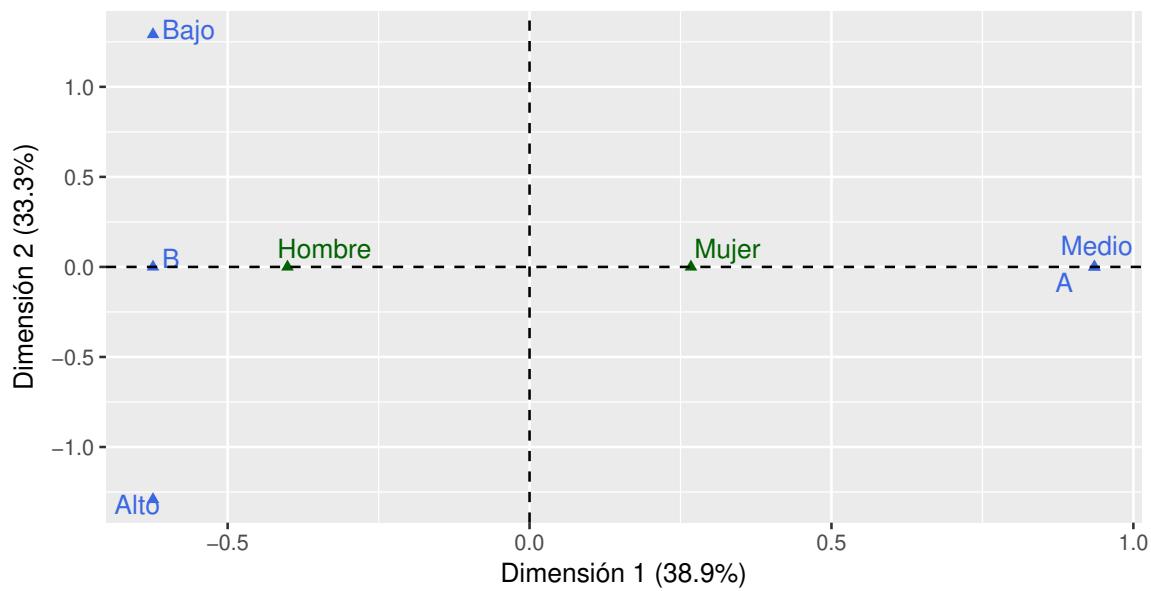


Figura 5.26: Biplot simétrico para la empresa

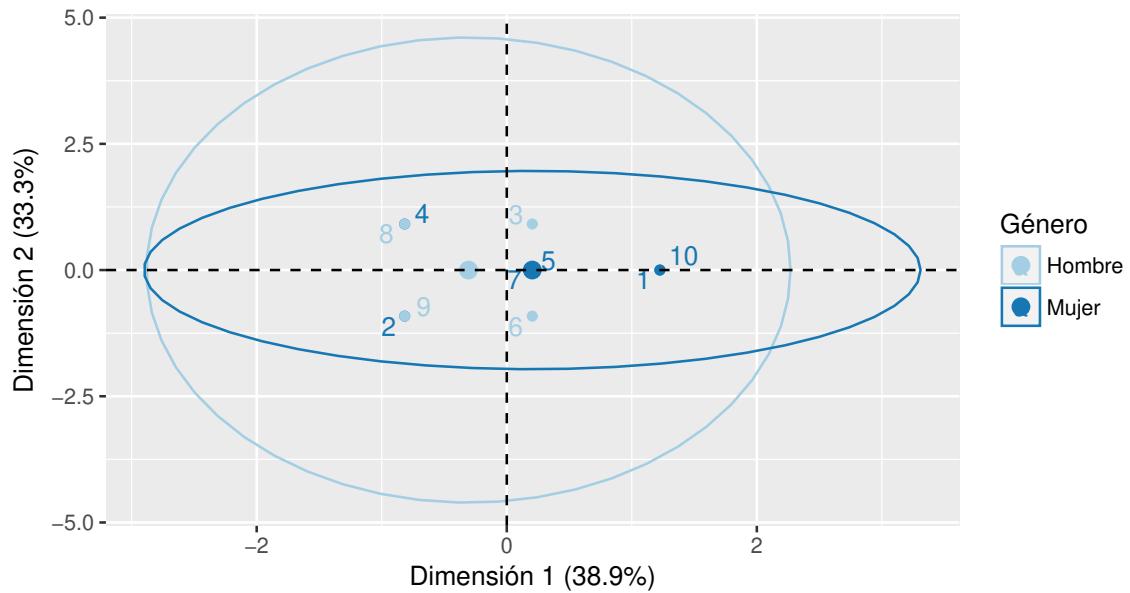


Figura 5.27: Empleados agrupados por género

Gén.Hombre	Gén.Mujer	Ing.Alto	Ing.Bajo	Ing.Medio	Cat.A	Cat.B
1	0	1	0	0	1	1
2	0	1	1	0	0	0
3	1	0	0	1	0	1
4	0	1	0	1	0	0
5	0	1	0	0	1	0
6	1	0	1	0	0	1
7	0	1	0	0	1	0
8	1	0	0	1	0	0
9	1	0	1	0	0	0
10	0	1	0	0	1	0

Tabla 5.31: Matriz disyuntiva para la empresa

	Gén.Hom.	Gén.Muj.	Ing.Alto	Ing.Bajo	Ing.Medio	Cat.A	Cat.B
Gén.Hom.	4	0	2	2	0	2	2
Gén.Muj.	0	6	1	1	4	2	4
Ing.Alto	2	1	3	0	0	1	2
Ing.Bajo	2	1	0	3	0	1	2
Ing.Medio	0	4	0	0	4	2	2
Cat.A	2	2	1	1	2	4	0
Cat.B	2	4	2	2	2	0	6

Tabla 5.32: Matriz de Burt para la empresa

	$A \times 1$	$A \times 2$	$A \times 3$	$A \times 4$
<b>Autovalores</b>	0.55556	0.35830	0.33333	0.08614
<b>Inercia proyectada (%)</b>	41.67	26.87	25.00	6.46
<b>Inercia proyectada acumulada (%)</b>	41.67	68.54	93.54	100.00
<b>Inercia acumulada</b>	0.55556	0.91386	1.24719	1.33333

Tabla 5.33: Inercias para la empresa

	CS1	CS2
<b>Género.Hombre</b>	1.5	-0.454
<b>Género.Mujer</b>	-1.0	0.303
<b>Ingresos.Alto</b>	1.0	0.303
<b>Ingresos.Bajo</b>	1.0	0.303
<b>Ingresos.Medio</b>	-1.5	-0.454
<b>Categoría.A</b>	0.0	-2.022
<b>Categoría.B</b>	0.0	1.348

Tabla 5.34: Coordenadas de representación para la empresa

## 5.2 Ejercitación

### Ejercicio 1.

Se ha realizado una encuesta entre el personal de una empresa al cual se le preguntó el cargo que desempeña y la cantidad de cigarrillos diarios que fuma. La frecuencia de fumador fue categorizada de la siguiente manera: No fuma - Fuma poco – Fuma moderadamente - Fuma mucho. En la Tabla 5.35 se resumen las respuestas obtenidas.

Puesto	Categoría de fumador				Totales
	No fuma	Poco	Moderado	Mucho	
Gerente Senior	4	2	3	2	11
Gerente Junior	4	3	7	4	18
Empleado Senior	25	10	12	4	51
Empleado Junior	18	24	33	13	88
Secretaria	10	6	7	2	25
<b>Totales</b>	<b>61</b>	<b>45</b>	<b>62</b>	<b>25</b>	<b>193</b>

Tabla 5.35: Hábito de fumar según puesto de trabajo

Estamos interesados en estudiar la relación, si existiera, entre las variables “puesto de trabajo” y “nivel de fumador” en el contexto de esta empresa.

1. Analizar si la distribución de la variable fumador es similar en todos los niveles de la variable puesto de desempeño, construyendo para eso las distribuciones condicionales de fumador por cada puesto de trabajo.
2. Realizar un análisis de correspondencias para estos datos. ¿Cuántos factores tiene sentido considerar?
3. Realizar los gráficos de perfiles que pueden considerarse adecuados.
4. Explicar la calidad de la representación y las relaciones entre las variables y los ejes (inerzia, calidad y cosenos).
5. Hacer una síntesis de las conclusiones obtenidas, inspeccionando relaciones entre perfiles fila, entre perfiles columna y asociaciones entre filas y columnas de manera adecuada.
6. ¿Cuál es la inercia total?

### Ejercicio 2.

En el archivo de datos disponible en <https://goo.gl/FeiXTg> (extraído de *Infostat*), se registran datos de 339 usuarios de auto. Las variables que se han preguntado son las siguientes:

**Origen:** del auto, que puede ser americano, japonés o europero

**Estado:** que puede ser soltero, soltero con hijos, casado o casado con hijos

**Casa:** que indica la relación con la casa en la que habita, siendo dueño o inquilino

**Tipo:** de auto que puede ser familiar, deportivo o para trabajo

**Sexo:** hombre o mujer

**Tamaño:** del auto distinguiendo entre chico, mediano y grande

**Ingreso:** familiar dividido en dos niveles

1. Elegir tres variables y construir la matriz disyuntiva y la matriz de Burt, explicando el significado de los valores diagonales y verificando las propiedades de la matriz.
2. Realizar un análisis de correspondencias múltiples con estas variables y explicar los resultados.

### **Ejercicio 3.**

La opinión que algunos ingleses tienen sobre algunos europeos, se encuentra disponible en <https://goo.gl/KDvuzn>. Se pide realizar un análisis de correspondencias para caracterizar a un grupo de europeos desde la mirada de los ingleses. Para ello, tener en cuenta las siguientes preguntas.

1. ¿Qué características son las más usuales?
2. ¿Qué características son las más raras?
3. En función de estos datos, ¿es justo decir que París es la ciudad del *glamour*?

### **Ejercicio 4.**

En el archivo disponible en <https://goo.gl/XcKCQN> se tabuló el consumo de distintos tipos de proteínas per cápita de los habitantes de distintos países de Europa. Conducir un análisis de correspondencias múltiples para describir el tipo de consumo de proteínas de los países vinculando este análisis con la posición geográfica de los mismos.

# Referencias

- [1] Alan Agresti et al., *A survey of exact inference for contingency tables*, Statistical science **7** (1992), no. 1, 131–153.
- [2] Kevin Ashton et al., *That ‘internet of things’ thing*, RFID journal **22** (2009), no. 7, 97–114.
- [3] Robert G Bland, Donald Goldfarb, and Michael J Todd, *The ellipsoid method: A survey*, Operations research **29** (1981), no. 6, 1039–1091.
- [4] Juan de Burgos Román, *Álgebra lineal y geometría cartesiana*, McGraw-Hill, 2006.
- [5] Herman Chernoff, *Chernoff faces*, International Encyclopedia of Statistical Science, Springer, 2011, pp. 243–244.
- [6] David L Donoho, Miriam Gasko, et al., *Breakdown properties of location estimates based on halfspace depth and projected outlyingness*, The Annals of Statistics **20** (1992), no. 4, 1803–1827.
- [7] John C Gower, *Some distance properties of latent root and vector methods used in multivariate analysis*, Biometrika **53** (1966), no. 3-4, 325–338.
- [8] Michael J Greenacre, *Correspondence analysis*, London: Academic Press, 1984.
- [9] Dominique Guinard, Vlad Trifa, Friedemann Mattern, and Erik Wilde, *From the internet of things to the web of things: Resource-oriented architecture and best practices*, Architecting the Internet of things, Springer, 2011, pp. 97–129.
- [10] Kenneth Hoffman, Ray Kunze, and Hugo E Finsterbusch, *Álgebra lineal*, Prentice-Hall Hispanoamericana, 1973.
- [11] Jan Holler, Vlasios Tsiatsis, Catherine Mulligan, Stamatis Karnouskos, Stefan Avesand, and David Boyle, *Internet of things*, Academic Press, 2014.
- [12] Harold Hotelling, *Analysis of a complex of statistical variables into principal components.*, Journal of educational psychology **24** (1933), no. 6, 417.

- [13] Prasanta Chandra Mahalanobis, *On the generalized distance in statistics*, ., National Institute of Science of India, 1936.
- [14] RARD Maronna, R Douglas Martin, and Victor Yohai, *Robust statistics*, vol. 1, John Wiley & Sons, Chichester. ISBN, 2006.
- [15] Mahdi H Miraz, Maaruf Ali, Peter S Excell, and Rich Picking, *A review on internet of things (iot), internet of everything (ioe) and internet of nano things (iont)*, 2015 Internet Technologies and Applications (ITA), IEEE, 2015, pp. 219–224.
- [16] Karl Pearson, *Principal components analysis*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **6** (1901), no. 2, 559.
- [17] Pedro R Peres-Neto, Donald A Jackson, and Keith M Somers, *How many principal components? stopping rules for determining the number of non-trivial axes revisited*, Computational Statistics & Data Analysis **49** (2005), no. 4, 974–997.
- [18] Robin L Plackett, *Karl pearson and the chi-squared test*, International Statistical Review/Revue Internationale de Statistique (1983), 59–72.
- [19] Peter J Rousseeuw and Katrien Van Driessen, *A fast algorithm for the minimum covariance determinant estimator*, Technometrics **41** (1999), no. 3, 212–223.
- [20] E Sathishkumar and K Thangavel, *A novel approach for outlier detection using rough entropy. department of computer science periyar university*, WSEAS TRANSACTIONS on COMPUTERS, E-ISSN (2015), 2224–2872.
- [21] Valentin Todorov, Peter Filzmoser, et al., *An object-oriented framework for robust multivariate analysis*, Citeseer, 2009.
- [22] Stefan Van Aelst and Peter Rousseeuw, *Minimum volume ellipsoid*, Wiley Interdisciplinary Reviews: Computational Statistics **1** (2009), no. 1, 71–82.
- [23] Stefan Van Aelst, Ellen Vandervieren, and Gert Willems, *A stahel-donoho estimator based on huberized outlyingness*, Computational Statistics & Data Analysis **56** (2012), no. 3, 531–542.