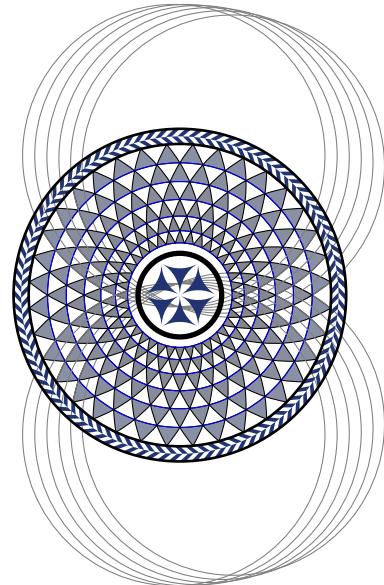


Análisis inteligente de datos con lenguaje R

Curso introductorio

✿ Débora Chan ✿ Cristina Badano ✿ Andrea Rey ✿



Índice de contenidos

1	Introducción a la minería de datos	1
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
2	Introducción al análisis de datos	9
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	16
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	24
2.2.5.3	Gráfico de bastones	27
2.2.5.4	Histograma y polígono de frecuencias	28
2.2.5.5	<i>Boxplot</i> o diagrama de caja	33
2.2.5.6	<i>Boxplots</i> comparativos	36
2.3	Información multivariada	37
2.3.1	Objetivos del análisis exploratorio	41
2.3.1.1	Tabla de clasificación cruzada	41
2.3.1.2	Gráfico de mosaicos	42
2.3.1.3	Diagrama de dispersión	43
2.3.1.4	Dispersograma	44

2.3.1.5	Gráfico de coordenadas paralelas	44
2.3.1.6	Gráfico de perfiles multivariados	46
2.3.1.7	Curvas de nivel	47
2.3.1.8	Gráficos de estrellas	49
2.3.1.9	Gráficos de caras de Chernoff	50
2.4	Medidas de posición y dispersión en datos multivariados	51
2.4.1	Propiedades del vector de medias	52
2.4.2	Propiedades de la matriz de varianzas y covarianzas	52
2.5	Transformación del conjunto de datos	53
2.5.1	Transformaciones por variables	53
2.5.1.1	Variables aleatorias estandarizadas	53
2.5.2	Transformaciones por individuo	53
2.6	Análisis multivariado	54
2.6.1	Covarianza y Correlación	56
2.7	Alternativas robustas para posición y escala	62
2.8	Ejercitación	66
3	Análisis de componentes principales	71
3.1	Nociones Previas	71
3.2	Transformaciones	75
3.2.1	Autovalores y Autovectores	80
3.2.1.1	Relación entre autovalores, traza y determinante	82
3.3	Motivación del problema de reducción de la dimensión	84
3.4	Análisis de componentes principales	91
3.4.1	Definición de las componentes	92
3.4.2	Variabilidad explicada por las componentes principales	93
3.4.3	Variabilidad de las componentes principales	95
3.4.4	Cantidad de componentes principales	96
3.4.4.1	Criterio 1: Porcentaje de variabilidad explicada	96
3.4.4.2	Criterio 2: Criterio de Kaiser	96
3.4.4.3	Criterio 3: Criterio del bastón roto	97
3.4.4.4	Criterio 4: Prueba de esfericidad	97
3.4.5	Estimación de las componentes principales	98
3.4.6	Escalas de medida	101
3.4.7	Cargas o <i>loadings</i>	101
3.4.8	Interpretación de las componentes principales	104
3.4.9	<i>Biplot</i>	106
3.5	Componentes principales robustas	119
3.6	Ejercitación	130

4 Contrastes de independencia y homogeneidad	135
4.1 Contraste de Hipótesis	135
4.1.1 Nivel de significación	140
4.1.2 Relaciones entre los errores de tipo I y II	141
4.1.3 Potencia de un contraste	141
4.1.4 Concepto de <i>p</i> -valor	142
4.2 Contrastes de homogeneidad e independencia	142
4.2.1 Contraste de independencia	143
4.2.2 Test Chi cuadrado de independencia	146
4.2.2.1 Hipótesis de interés	146
4.2.3 Test Chi cuadrado de homogeneidad	147
4.2.3.1 Hipótesis de interés	148
4.2.4 Estadístico de contraste	150
4.2.5 Región crítica	150
4.2.6 Limitaciones	154
4.2.7 Test exacto de Fisher	155
4.3 Ejercitación	161
5 Análisis de correspondencias	165
5.1 Perfiles medios	172
5.2 Inercia total	173
5.3 <i>Biplot</i> simétrico	180
5.3.1 Guía para la interpretación gráfica del <i>biplot</i> simétrico	184
5.3.2 Otra representación gráfica	189
5.4 Estadístico de Pearson e inercia	198
5.4.1 Interpretación geométrica de la inercia	199
5.5 Principio de equivalencia distribucional	202
5.6 Análisis de correspondencias múltiples	203
5.6.1 Matriz de Burt	206
5.6.2 Examen de los puntos	214
5.7 Ejercitación	221
6 Escalamiento multidimensional	223
6.1 Modelo general	223
6.2 Modelos particulares	225
6.2.1 Modelo de escalamiento métrico	225
6.3 Relación con otras técnicas	231

7 Comparación de medias en el caso univariado	233
7.1 Diferencia de medias de poblaciones normales para dos muestras independientes	233
7.1.1 Muestras normales independientes con varianzas conocidas	233
7.1.2 Muestras normales independientes con varianzas desconocidas	237
7.1.3 Muestras independientes de poblaciones cualesquiera	240
7.1.4 Muestras apareadas	242
7.2 Pruebas no paramétricas para dos muestras independientes	246
7.2.1 Test de Mann-Whitney-Wilcoxon	246
7.2.2 Test de la mediana	250
7.2.3 Tres o más grupos: análisis de la varianza de un factor	252
7.2.3.1 Descomposición de la suma de cuadrados totales	255
7.2.3.2 Estadístico del test	256
7.2.3.3 Diagnóstico del modelo	258
7.2.3.4 Test de Bartlett	258
7.2.3.5 Test de Levene	259
7.2.3.6 Tests de normalidad	259
7.2.3.7 Gráficos de cuantil-cuantil	260
7.2.3.8 Intervalo de confianza para la diferencia de dos medias	263
7.2.3.9 Test de Kruskal-Wallis no paramétrico para muestras independientes	270
7.3 Ejercitación	277
8 Comparación de medias en el caso multivariado	283
8.1 Distribución Normal univariada	283
8.2 Distribución Normal multivariada	284
8.2.1 Estimadores de máxima verosimilitud	287
8.2.2 Distribución de Wishart	288
8.2.3 Distribuciones muestrales de la media y la varianza	289
8.2.4 Distribución de Hotelling	290
8.2.5 Test del vector de medias para una población	291
8.2.6 Test para comparar medias de dos poblaciones	291
8.2.7 Análisis de perfiles	296
8.2.7.1 Caso de dos perfiles	297
9 Métodos de clasificación supervisada	303
9.1 Análisis discriminante	303
9.1.1 Reglas basadas en estimaciones de los parámetros	306
9.1.1.1 Estimación de las probabilidades de clasificación errónea	315
9.1.1.2 Casos de más de dos grupos	317
9.1.2 Validación de los supuestos del análisis discriminante	317
9.1.3 Interpretación de los coeficientes de la función discriminante	319

9.1.4	Costos de clasificación	320
9.2	Análisis discriminante cuadrático de Fisher	321
9.3	Alternativas robustas	332
9.4	Máquinas de soporte vectorial	334
9.4.1	Separabilidad lineal	335
9.4.1.1	Linealmente separable	335
9.4.1.2	No linealmente separable	337
9.4.2	<i>Kernels</i> y mapeo	338
9.5	Regresión logística	344
9.6	Ejercitación	354
10	Métodos de clasificación no supervisada	357
10.1	Distancias y medidas de proximidad	357
10.1.1	Medidas de distancia para vectores de observaciones continuas	358
10.1.2	Medidas de similaridad	362
10.2	Introducción al análisis de conglomerados o <i>clusters</i>	366
10.2.1	Análisis de <i>clusters</i> por individuos o variables	367
10.2.2	Métodos de agrupamiento	367
10.2.3	Algoritmos jerárquicos	369
10.2.4	Algoritmos para medir distancia entre <i>clusters</i>	380
10.2.4.1	Método de la media o <i>average linkage</i>	380
10.2.4.2	Método del vecino más próximo o <i>single linkage</i>	385
10.2.4.3	Método del vecino más lejano o <i>complete linkage</i>	388
10.2.4.4	Método del centroide o <i>unweighted centroid</i>	389
10.2.4.5	Método de Ward o varianza mínima	389
10.2.5	Métodos divisivos	390
10.2.6	Cantidad de <i>clusters</i>	391
10.2.7	Métodos de partición no jerárquicos	392
10.2.8	Otros métodos para elegir el número de <i>clusters</i>	401
10.3	Ejemplos	402
10.3.1	Ejemplo de agrupamiento jerárquico	402
10.3.2	Ejemplo de agrupamiento no jerárquico	405
10.3.3	Ejemplo de aplicación a <i>text mining</i>	407
10.3.4	Ejemplo de aplicación a imágenes	413
10.4	Ejercitación	421
A	Nociones elementales de Álgebra Lineal	425
B	Nociones de Estadística	427

A Nociones elementales de Álgebra Lineal	431
B Nociones de Estadística	433
Referencias	440

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 tortas anidadas	24
2.11	Diagrama de barras	25
2.12	Diagrama de barras superpuestas	26
2.13	Diagrama de barras adyacentes	27
2.14	Diagrama de bastones	28
2.15	Histograma	29
2.16	Polígono de frecuencias	30
2.18	Histogramas con distintos intervalos	31
2.19	Comparación de métodos para el cómputo de intervalos	32
2.20	Simetría en <i>boxplots</i>	35
2.22	<i>Boxplots</i> comparativos	37
2.24	Diagrama de mosaicos	42
2.25	Diagrama de dispersión para tres poblaciones	43
2.26	Dispersograma	45
2.27	Gráfico de coordenadas paralelas	46
2.28	Gráfico de perfiles	47
2.29	Gráfico de la distribución Normal Bivariada	48
2.30	Gráfico de las curvas de nivel de la distribución Normal Bivariada	49
2.31	Gráfico de estrellas	50
2.32	Gráfico de caras de Chernoff para galletitas saladas	51
2.34	Crontrol univariado	55
2.35	Control multivariado	57
2.36	Signo de la covarianza	59

2.37	Correlograma	61
2.38	Detección multivariada de <i>outliers</i>	64
3.1	Vectores en coordenadas	72
3.2	Dependencia lineal entre vectores	73
3.4	Bases para \mathbb{R}^2	76
3.5	Modelo de datos a proyectar	76
3.6	Simetría respecto del eje de abscisas	77
3.7	Rotación de ángulo π	78
3.8	Proyección ortogonal de un punto sobre el plano xy	79
3.9	Simetría respecto de la recta $y = x$	81
3.11	Dispersograma entre dos variables	86
3.12	Dispersograma 3D desde distintos puntos de vista	88
3.13	Dispersograma en 3D clasificado por grupos	89
3.14	Ejes principales	90
3.15	Direcciones principales en el espacio tridimensional	91
3.17	Gráfico de sedimentación	100
3.18	Cargas de la primera componente principal	102
3.19	Cargas de la segunda componente principal	103
3.20	<i>Biplot</i> para nadadores	106
3.21	Caras de Chernoff para nadadores	109
3.23	Gráfico de sedimentación para aspirantes	114
3.24	Cargas para los aspirantes	115
3.25	<i>Biplots</i> para los aspirantes	116
3.26	Comparación de <i>boxplots</i> para nadadores	120
3.27	Diagramas de dispersión para nadadores	121
3.28	Cargas ACP(clásico) para nadadores con datos agregados	122
3.29	Ánálisis clásico de <i>screeplot</i> para los nadadores con los datos agregados	123
3.30	<i>biplot</i> clásico para nadadores con datos agregados	123
3.31	<i>Screeplot</i> para MCD de nadadores con datos agregados	128
3.32	<i>Biplot</i> para MCD de nadadores con datos agregados	128
4.1	Ejemplo de regiones en un contraste bilateral	139
4.2	Representación de los errores de un test	141
4.4	Poblaciones según variable de color	149
4.5	Gráficos de la distribución χ^2 por grados de libertad	151
4.7	Distribución χ^2 y zona crítica	152
5.2	Perfiles de nivel cultural según atención	173
5.3	Contribución de filas a la dimensión 1	181

5.4	Contribución de columnas a la dimensión 1	181
5.5	Puntos fila - AC	182
5.6	Puntos columna - AC	182
5.7	<i>Biplot</i> simétrico - AC	183
5.8	Perfiles fila de las actividades universitarias	188
5.9	Caras de Chernoff universidades	190
5.10	Representación en 3D de las actividades universitarias	191
5.11	Plano de representación de las actividades universitarias	192
5.12	<i>Biplot</i> simétrico (actividades universitarias)	194
5.13	Contribución de las filas (actividades universitarias)	194
5.14	Contribución de columnas (actividades universitarias)	195
5.16	Ejemplos de inercias	200
5.18	Contribución de variables a la inercia (dimensión 1)	208
5.19	Contribución de individuos a la inercia (dimensión 1)	208
5.20	Categorías variables - ACM	209
5.21	Individuos agrupados por género - ACM	209
5.22	Individuos agrupados por estado civil - ACM	210
5.24	Contribución a la inercia de las variables	218
5.25	Contribución a la inercia de los individuos	218
5.26	<i>Biplot</i> simétrico para la empresa	219
5.27	Empleados agrupados por género	219
6.2	MDS aplicado a ciudades argentinas	231
7.2	Zonas de aceptación y de rechazo para esta prueba	235
7.4	Zonas de aceptación y de rechazo para esta prueba	239
7.8	<i>Boxplot</i> para las distintas variedades de oliva	250
7.10	<i>Boxplot</i> para vitamina B en distintas marcas de té	254
7.11	Zonas de rechazo para la prueba F	257
7.12	<i>QQ-plot</i> para vitamina B en distintas marcas de té	262
7.14	<i>Boxplot</i> comparativo para las distintas dietas	266
7.15	Salida del test de Box & Cox para el colesterol en conejos	268
7.17	Distribución del rendimiento por grupo	273
8.1	Distribución normal con varianzas distintas	284
8.2	Distribución normal con medias distintas	285
8.3	Ejemplo de distribución Normal bivariada	286
8.5	Diagrama de dispersión para las avispas	294
8.6	Ejemplo de comparación de perfiles	297
8.7	Perfiles según la especie	299

9.1	Perfiles según la especie con nuevas observaciones	307
9.2	Línea discriminante entre las especies de avispas	308
9.3	Zona problema en la clasificación	310
9.4	Diagrama de partición de las especies de las avispas	315
9.6	Análisis univariado por estado de billete	323
9.7	<i>QQ-plots</i> de las distintas medidas de billete	324
9.8	Correlogamas de los billetes según su estado	325
9.9	Partición por clases de billete	326
9.10	Partición por clases de billete (continuación)	327
9.11	Margen entre conjuntos linealmente separables	336
9.12	Ejemplo de envolventes conexas por clase	337
9.13	Ejemplo de conjunto no linealmente separable	338
9.14	Efecto de mapeos	339
9.15	Representación gráfica de los datos simulados	342
9.16	Representación gráfica de la clasificación por SVM	343
9.17	Curva logística	346
9.19	PSA por ruptura de la cápsula	349
9.20	Probabilidad de rotura capsular en función de PSA	350
9.21	Gleason por ruptura de la cápsula	351
10.1	Interpretación geométrica de distancias L_p	359
10.2	Ejemplo de la distancia <i>city blocks</i>	359
10.3	Número de características	363
10.4	Representación de las observaciones originales	372
10.5	Representación del primer paso de clusterización	374
10.6	Representación del segundo paso de clusterización	375
10.7	Representación del tercer paso de clusterización	376
10.8	Representación del cuarto paso de clusterización	377
10.9	Representación del quinto paso de clusterización	378
10.10	Dendograma	379
10.11	Dendograma con elección de cantidad de <i>clusters</i>	380
10.12	Dendograma con el método <i>average linkage</i>	384
10.13	Dendrograma de Ward para los datos del Ejemplo 10.5	390
10.14	¿Cuántos <i>clusters</i> se pueden ver?	392
10.15	Puntos originales	395
10.16	Representación de las observaciones originales tridimensionales	398
10.18	Agrupamiento de los países en el campeonato	406
10.19	Grupos de países con el algoritmo k- <i>means</i>	408
10.20	Grupos de países con el algoritmo k- <i>means</i> etiquetados	408
10.21	Otra manera de visualizar los grupos de países según k- <i>means</i>	409

10.22Relación entre goles a favor y en contra según agrupamiento	409
10.23Relación entre goles a favor y en contra según agrupamiento con etiquetas	410
10.24Palabras más frecuentes en poemas de Neruda	413
10.25Nube de palabras más frecuentes en poemas de Neruda	414
10.26Refinamiento de nube de palabras más frecuentes en poemas de Neruda	414
10.27Nube de palabras con frecuencia mayor a 4 en poemas de Neruda	415
10.28Otra forma para nube de palabras más frecuentes en poemas de Neruda	415
10.30Reconstrucción de la imagen	418

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	38
2.8	Base de datos para las galletitas	39
2.9	Cantidad de parámetros en función de las variables	40
2.10	Consideraciones para la compra de un auto	41
2.11	Distancias entre <i>outliers</i>	65
2.12	Distancias de Mahalanobis	65
2.13	Datos candidatas a recepcionistas	66
3.1	Tiempos por tramos en competencia de natación	74
3.2	Tiempos por tramos en competencia de natación ampliada	75
3.3	Análisis sobre riesgo cardíaco	85
3.4	Variabilidad de las componentes principales usando las variables estandarizadas	99
3.5	Variabilidad de las componentes principales usando las variables originales	99
3.6	Cargas para los nadadores	102
3.7	Datos de los nadadores estandarizados por columna	105
3.8	Puntajes (<i>scores</i>) de los nadadores	105
3.9	Estadística descriptiva univariada para los nadadores	106
3.10	Autovalores y autovectores	112
3.11	Esfericidad de Bartlett	112
3.12	Criterios de evaluación	113
3.13	Variabilidad explicada	114
3.14	Cargas de los datos de los aspirantes	115

3.15	Nuevos nadadores	120
3.16	PCA clásico con nuevos datos	121
3.17	Análisis de componentes principales usando MCD	127
4.1	Errores en un test	140
4.2	Nivel de violencia según la edad	144
4.3	Frecuencias relativas del nivel de violencia según la edad	144
4.4	Formato teórico del nivel de violencia según la edad	144
4.5	Cálculos para el análisis de independencia	145
4.6	Frecuencias observadas y esperadas (violencia por edad)	147
4.7	Frecuencias teóricas de homogeneidad	148
4.8	Datos enfermedad según tabaquismo	151
4.9	Frecuencias observadas y esperadas	153
4.10	Similitudes y diferencias entre ambas pruebas	154
4.11	Ejemplo de tabla de contingencia de 2×2	156
4.12	Depresión según sexo	157
4.13	Combinaciones para Fisher	158
4.14	Probabilidades asociadas a la Tabla 4.13	159
4.15	Ingreso a <i>Twitter</i> según sexo	163
4.16	Especialidad según zona	163
4.17	Presencia de angioma según tipo de embarazo	164
5.1	Tabla de contingencia	167
5.2	Tabla de probabilidades estimadas	167
5.3	Nivel cultural según atención	168
5.4	Distribución marginal del nivel de atención	169
5.5	Distribución marginal del nivel de cultura	169
5.6	Distribución conjunta de niveles cultural y de atención	169
5.7	Frecuencias esperadas bajo el supuesto de independencia	171
5.8	Probabilidades condicionales dado el nivel ‘Atento’	172
5.9	Representación de niveles como simulaciones (<i>dummies</i>)	174
5.10	Representación de niveles para un ejemplo sencillo	175
5.11	Tabla de contingencia y matriz F para un ejemplo sencillo	175
5.12	Matriz F_r para un ejemplo sencillo	175
5.13	Inercias principales (autovalores)	186
5.14	Perfiles de las filas	186
5.15	Perfiles de las columnas	186
5.16	Registro viajes de intercambio	187
5.17	Perfiles fila de los viajes de intercambio	187
5.18	Perfiles columna de los viajes de intercambio	189

5.19	Distribución de depresión según práctica deportiva	196
5.20	Frecuencias esperadas bajo independencia	196
5.21	Residuos correspondientes a depresión vs práctica deportiva	198
5.22	Inercia creciente - Asociación creciente	201
5.23	Perfiles fila	201
5.24	Perfiles columna	202
5.25	Primer ejemplo de equivalencia distribucional	202
5.26	Segundo ejemplo de equivalencia distribucional	203
5.27	Características observadas	204
5.28	Matriz disyuntiva para las características observadas	205
5.29	Matriz G^t	205
5.30	Matriz de Burt para el Ejemplo 5.13	207
5.31	Situación de los empleados de una empresa	213
5.32	Matriz disyuntiva para la situación de los empleados	213
5.33	Agregado de categoría para los empleados	216
5.34	Matriz disyuntiva para la empresa	217
5.35	Matriz de Burt para la empresa	220
5.36	Inercias para la empresa	220
5.37	Coordenadas de representación para la empresa	220
5.38	Hábito de fumar según puesto de trabajo	221
6.1	Ciudades argentinas	229
7.1	Observaciones del experimento del pH	234
7.2	Observaciones del experimento de las habichuelas	237
7.3	Distribución de Frecuencias para actividades físicas	241
7.4	Datos apareados para medición de plaquetas con ADP	243
7.5	Diferencias de datos apareados	245
7.6	Aceite por variedad	248
7.7	Tabla para el test de la mediana	251
7.8	Vitamina B en el té	253
7.9	Salida de ANOVA presencia de vitamina B en el té	257
7.10	Transformaciones de potencia	262
7.11	Comparaciones múltiples	264
7.12	Colesterol en conejos	265
7.13	Resúmenes para datos del colesterol en conejos	266
7.14	Calificaciones según los grupos	272
7.15	Prueba de normalidad de los datos	272
7.16	Datos de los puntajes ordenados y rankeados	274
7.17	Tiempos de reparación según grupos de técnicos	277

7.18	Cantidad de sodio por marca de cerveza	278
7.19	Eficiencia de conversión según suplemento	279
7.20	Comparación nuevo analgésico	279
7.21	Efectos gástricos colaterales	280
7.22	Tiempo de cocción por grupo	281
7.23	Porcentajes de alcohol según analista	281
8.1	Datos sobre las avispas	294
9.1	Salida medias	309
9.2	Clasificación discriminante lineal de las avispas	312
9.3	Matriz de confusión para las avispas	315
9.4	Matriz de confusión ingenua para los billetes	328
9.5	Matriz de confusión con una muestra de entrenamiento	328
9.6	Matriz de confusión para la alternativa robusta	333
9.7	Tabla de confusión para el modelo por SVM	342
9.8	Comparación entre probabilidades y <i>odds</i>	347
9.9	Tabla de confusión para el modelo logístico	350
10.1	Distintas medidas L_p	358
10.2	Tiempos para los nadadores	360
10.3	Distintas medidas con peso	361
10.4	Distintas medidas de similaridad	362
10.5	Entradas de vectores binarios	363
10.6	Coeficientes de similaridad	364
10.7	Registros para los pacientes	365
10.8	Entradas binarias para los pacientes	365
10.9	Métodos aglomerativos versus divisivos	369
10.10	Algoritmos jerárquicos	371
10.11	Observaciones originales	372
10.12	Distancias euclídeas: primer paso	373
10.13	<i>Clusters</i> luego del primer paso	374
10.14	Distancias euclídeas: segundo paso	374
10.15	<i>Clusters</i> luego del segundo paso	375
10.16	Distancias euclídeas: tercer paso	375
10.17	<i>Clusters</i> luego del tercer paso	376
10.18	Distancias euclídeas: cuarto paso	377
10.19	<i>Clusters</i> luego del cuarto paso	377
10.20	Distancias euclídeas: quinto paso	378
10.21	<i>Clusters</i> luego del quinto paso	379

10.22Promedio distancias: primer paso	381
10.23Distancias con <i>average linkage</i> : primer paso	381
10.24Promedio distancias: segundo paso	382
10.25Distancias con <i>average linkage</i> : segundo paso	382
10.26Promedio distancias: tercer paso	382
10.27Distancias con <i>average linkage</i> : tercer paso	382
10.28Promedio distancias: cuarto paso	383
10.29Distancias con <i>average linkage</i> : cuarto paso	383
10.30Promedio distancias: quinto paso	383
10.31Distancias con <i>average linkage</i> : quinto paso	384
10.32Mínimo de distancias: primer paso	385
10.33Distancias con <i>single linkage</i> : primer paso	386
10.34Mínimo de distancias: segundo paso	386
10.35Distancias con <i>single linkage</i> : segundo paso	386
10.36Mínimo de distancias: tercer paso	386
10.37Distancias con <i>single linkage</i> : tercer paso	387
10.38Mínimo de distancias: cuarto paso	387
10.39Distancias con <i>single linkage</i> : cuarto paso	387
10.40Mínimo de distancias: quinto paso	388
10.41Distancias con <i>single linkage</i> : quinto paso	388
10.42Posibles particiones en dos <i>subclusters</i>	391
10.43Ventajas y desventajas del método <i>k-means</i>	393
10.44Datos para el algoritmo <i>k-means</i>	395
10.45Distancia a los centroides en el primer paso	395
10.46Distancia a los centroides en el segundo paso	396
10.47Datos para aplicar el método de <i>k-means</i>	397
10.48Clasificación con $k = 2$ en el primer paso	397
10.49Clasificación con $k = 2$ en el segundo paso	398
10.50Clasificación con $k = 3$ en el primer paso	400
10.51Clasificación con $k = 3$ en el segundo paso	400
10.52Suma de cuadrados	417
10.53Estudiantes del Psicología	423

Capítulo 1

Introducción a la minería de datos

En algún lugar, algo increíble está esperando por ser descubierto.

— 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

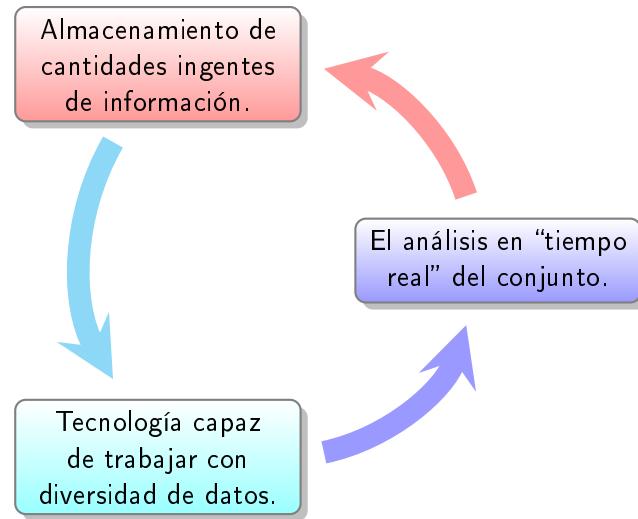
El *data mining* forma parte de un proceso conocido como “descubrimiento de conocimiento a partir de los datos” (*KDD: Knowledge Discovery in Databases*).

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 [23].

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 [5] 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 abarcaba 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 [21].

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 [34]. 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.

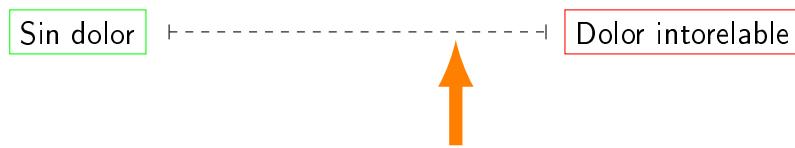


Figura 2.1: Escala visual

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 $f_1 + f_2 + \cdots + f_m = n$.

La frecuencia relativa se calcula dividiendo la frecuencia absoluta por la cantidad total de observaciones f_i/n y la frecuencia porcentual f_r 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 $Mo = 7$.

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 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 $Mo = 6$.



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 α -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}_α . 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,

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



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.

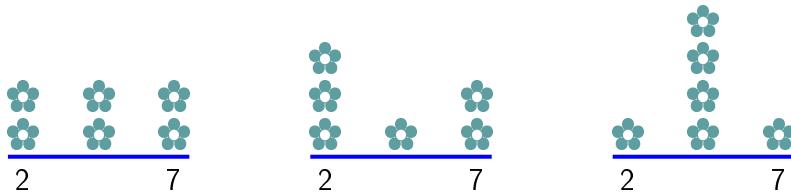


Figura 2.3: Variabilidad y rango

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 más 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.

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 $1.5, 1.5, 3.5, 4.5, 6.5, 20.5$.

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 [33].

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

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

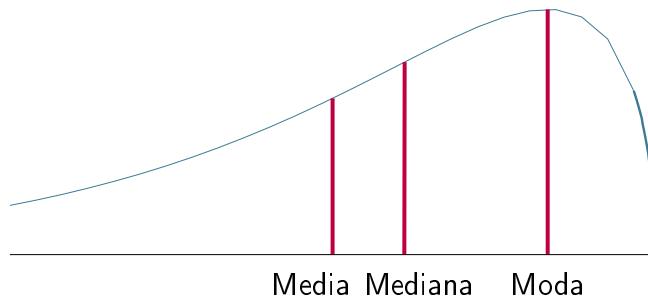


Figura 2.4: Asimetría negativa o a izquierda

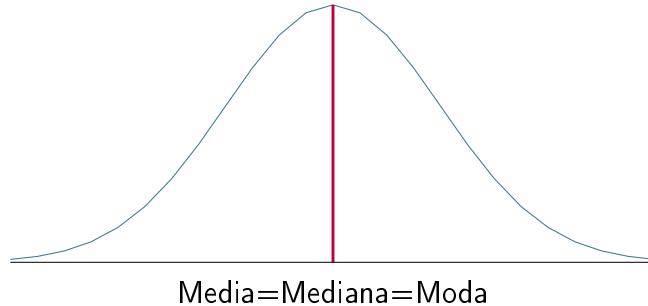


Figura 2.5: Simetría

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).

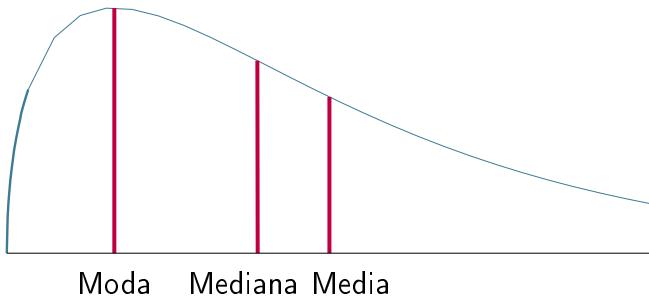


Figura 2.6: Asimetría positiva o a derecha

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.

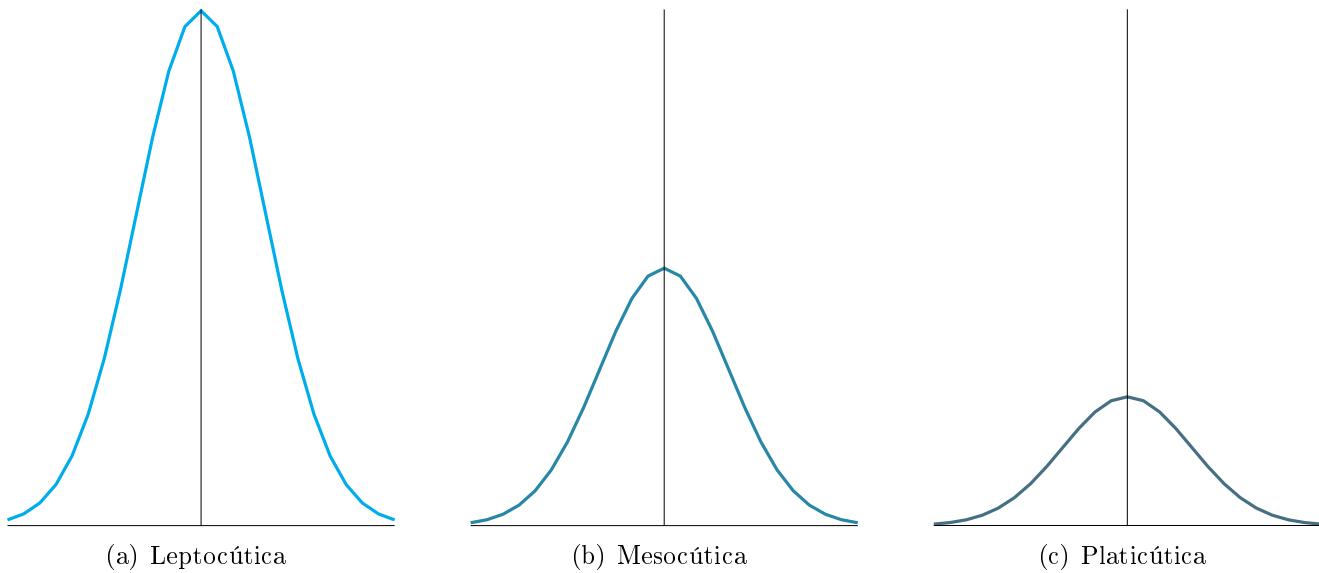
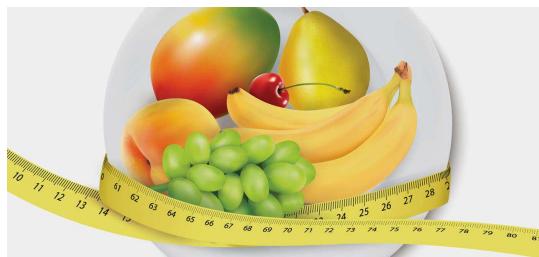


Figura 2.7: Distintos tipos de curtosis

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, mientras que el Código 2.2 produce un diagrama de tortas anidadadas como se muestra en la Figura 2.10. Los datos para ambas figuras son extraídos de <https://goo.gl/Dpnx9Z>.

```
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

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

pie3D(freq.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

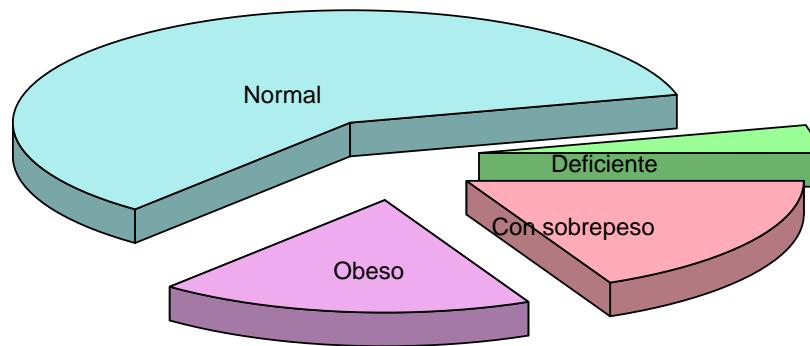


Figura 2.9: Diagrama circular con etiquetas

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

```

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

CatPeso <- IMCinfantil %>%
pull(CatPeso) %>%
plyr::mapvalues(c("D", "N", "OB", "SO"),
c("Deficiente", "Normal", "Obeso", "Sobrepeso"))
# Cambia el nombre a campos categóricos de la variable CatPeso
SEXO <- IMCinfantil %>%
pull(SEXO) %>%
plyr::mapvalues(c("M", "F"), c("Masc", "Fem"))
# Cambia el nombre a campos categóricos de la variable SEXO

IMCinfantil$CatPeso=CatPeso
IMCinfantil$SEXO=SEXO

interior <- IMCinfantil %>% group_by_(.dots=c("CatPeso")) %>%
tally() %>%
mutate(porcent_abs=round(n/sum(n)*100, 2))
# Produce la tabla para la torta interior

exterior <- IMCinfantil %>% group_by_(.dots=c("CatPeso", "SEXO")) %>%
tally() %>%
mutate(porcent_rel=round(n/sum(n)*100, 2))%>%
ungroup() %>%
mutate(porcent_abs=round(n/sum(n)*100, 2))
# Produce la tabla para la torta exterior

porcent_abs_ext=exterior$porcent_abs
tabla=table(exterior$CatPeso)[order(unique(exterior$CatPeso))]

colores=c("palegreen4", "paleturquoise4", "palevioletred4", "salmon3")
col_int=rep_len(colores, length(int_data$CatPeso))
col_ext=laply(Map(rep, colores[seq_along(tabla)]), tabla),
function(porcent_abs_ext) {
al <- head(seq(0, 1, length.out = length(porcent_abs_ext)+2L)[-1L], -1L)
Vectorize(adjustcolor)(porcent_abs_ext, alpha.f = al)}
) # Establece los colores

plot.new() # Borra gráficos anteriores

torta_ext=floating.pie(0.5, 0.5, exterior$porcent_abs, radius=0.25,
border="gray45", col=unlist(col_ext))
torta_int=floating.pie(0.5, 0.5, interior$porcent_abs, radius=0.2,
border="white", col=col_int)
# Produce los diagramas de tortas

```

```

pie.labels(x=0.5, y=0.5, torta_ext, paste0(exterior$SEXO, "\n",
exterior$porcent_rel, "% - ", exterior$n, " ind."),
minangle=0.2, radius=0.27, cex=0.6, font=1)
pie.labels(x=0.5, y=0.5, torta_int, paste0(interior$CatPeso, "\n",
interior$porcent_abs, "% - ", interior$n, " ind."),
minangle=0.2, radius=0.09, cex=0.6, font=1)
# Etiqueta las regiones

```

Código 2.2: Generación de un diagrama de tortas anidadas

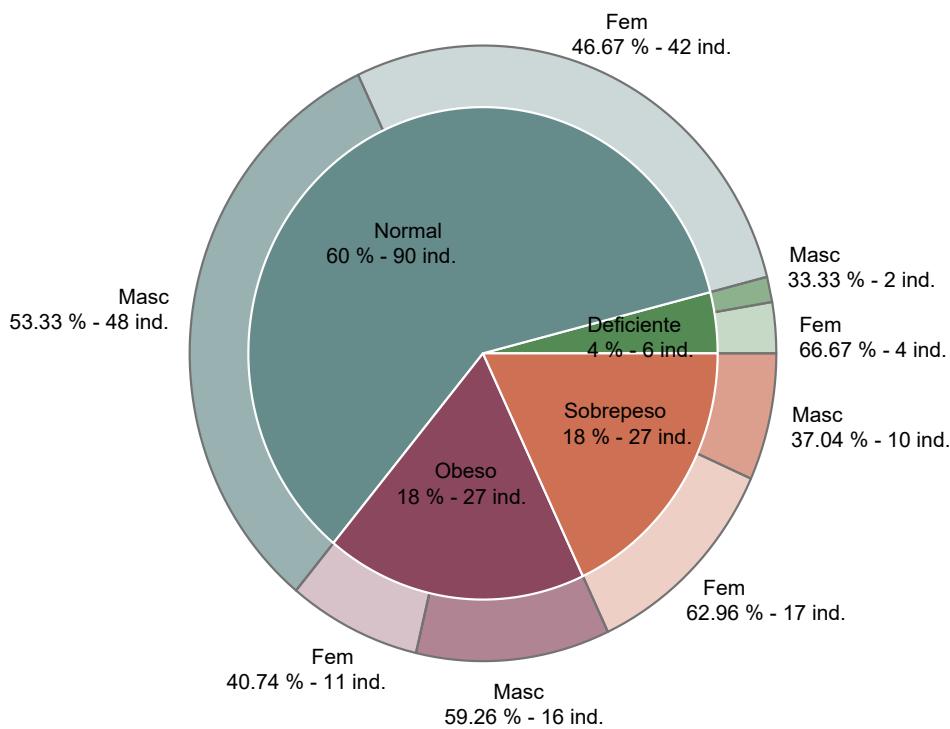


Figura 2.10: Diagrama de tortas anidadas

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 2.11 producida por el Código 2.3. Los datos son extraídos de <https://goo.gl/Dpnx9Z>.

```

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 2.3: Generación de un diagrama de barras

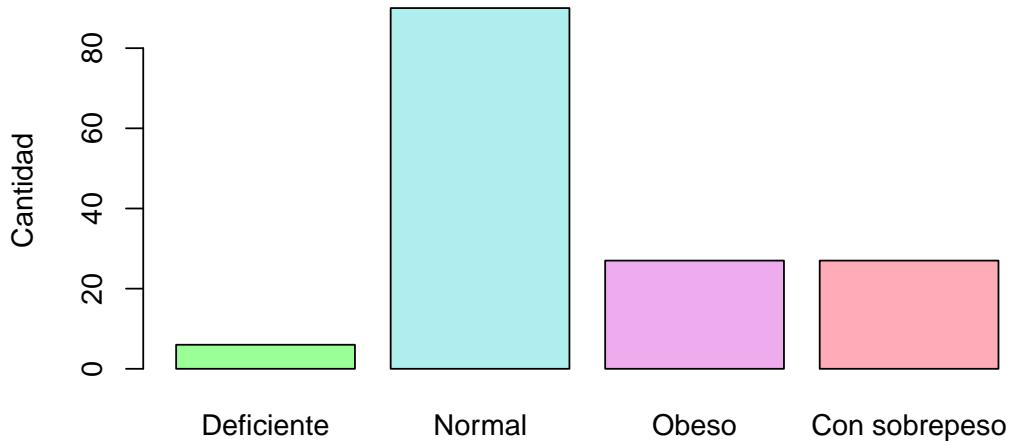


Figura 2.11: Diagrama de barras

Barras superpuestas

Este tipo de gráfico es útil cuando queremos apreciar la distribución en dos subconjuntos de individuos. A modo de ejemplo, la Figura 2.12 producida por el Código 2.4. Los datos son extraídos de <https://goo.gl/Dpxx9Z>.

```

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 2.4: Generación de un diagrama de barras superpuestas

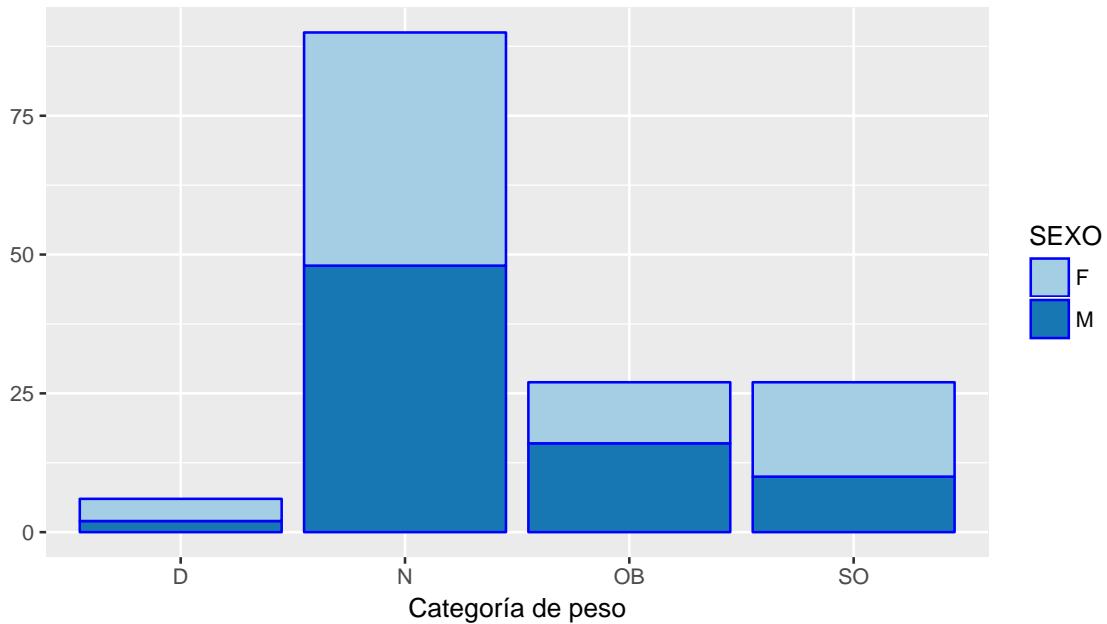


Figura 2.12: Diagrama de barras superpuestas

Barras adyacentes

En este tipo de esquemas, las barras pueden estar en posición vertical u horizontal. En la Figura 2.13, generada por el Código 2.5, 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 2.5: Generación de un diagrama de barras adyacentes

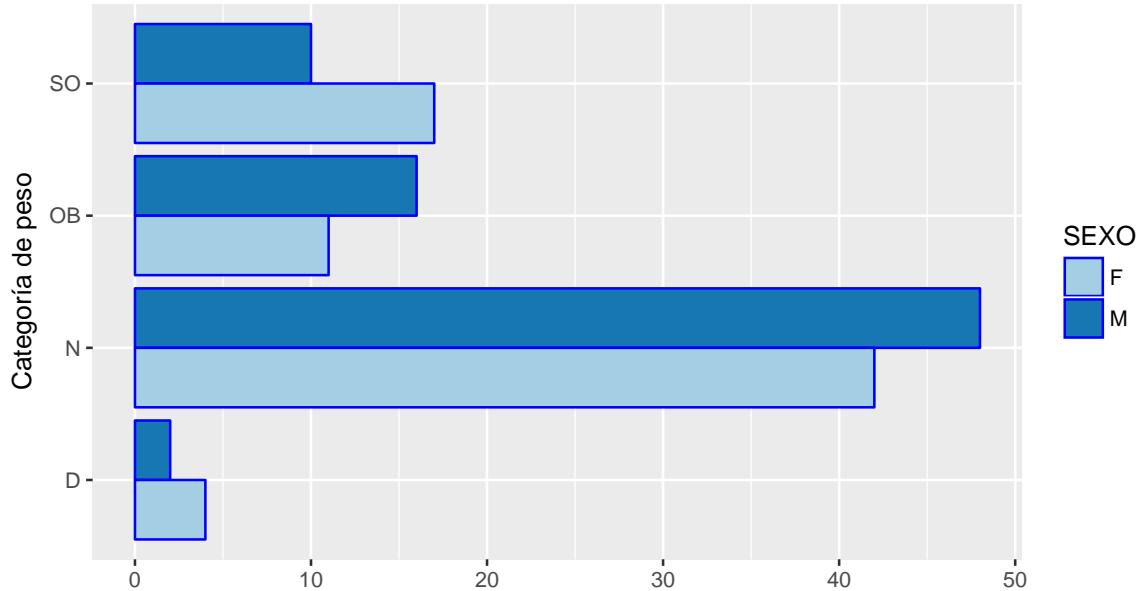


Figura 2.13: Diagrama de barras adyacentes

2.2.5.3 Gráfico de bastones

Es adecuado para representar la distribución de frecuencias de una variable discreta. Mostramos como el Código 2.6 genera la Figura 2.14.

```

Modelo=2010:2016 # Ingresa datos
Ventas=c(2,3,7,4,9,0,5) # Ingresa 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 2.6: Generación de un diagrama de bastones

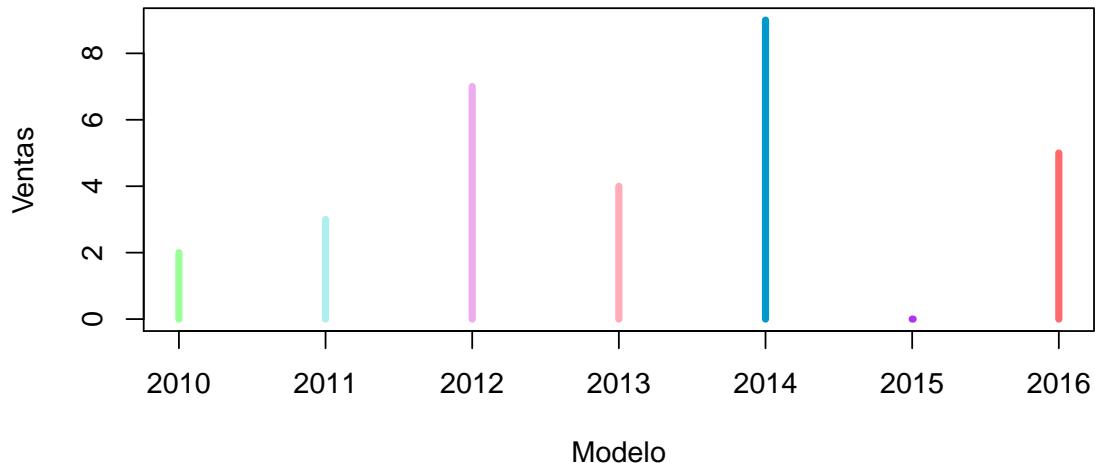


Figura 2.14: Diagrama de bastones

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 2.15 y 2.16 se obtienen mediante el Código 2.7. Los datos son extraídos de <https://goo.gl/Dpnx9Z>.

```
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 2.7: Generación de un histograma y su polígono de frecuencias

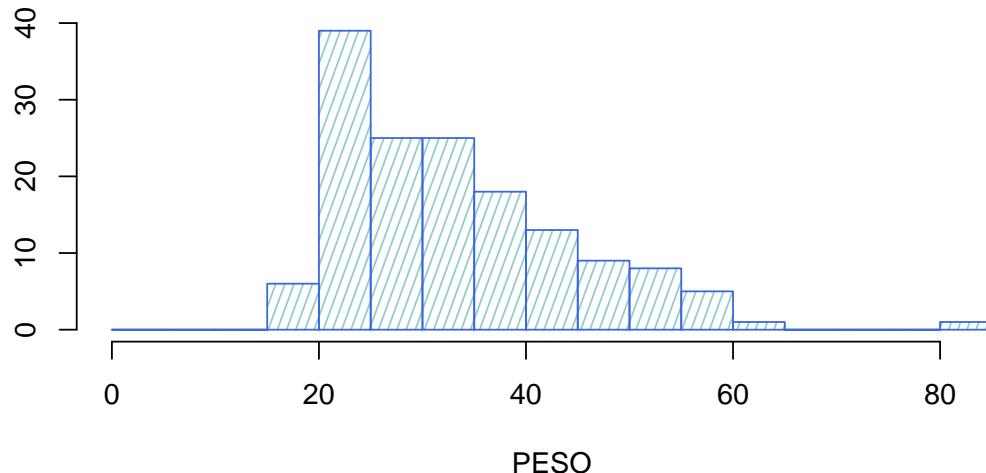


Figura 2.15: Histograma

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

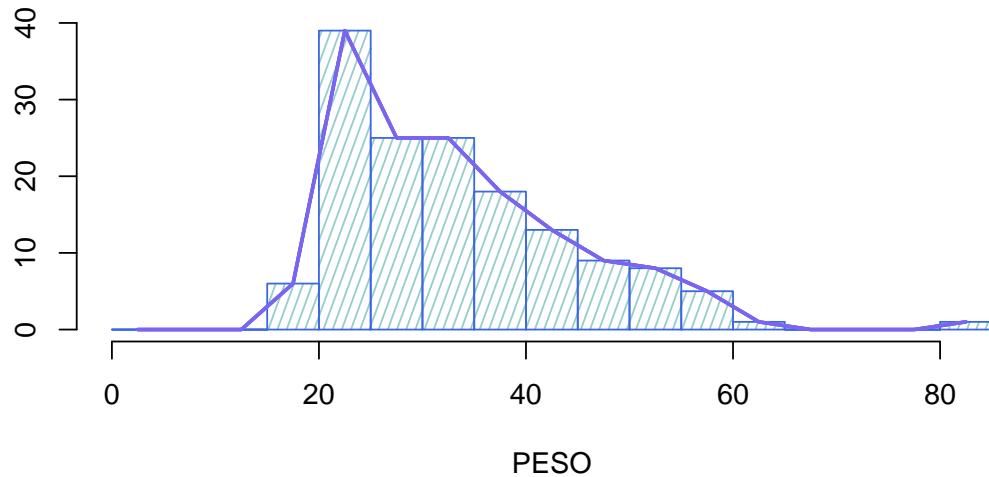


Figura 2.16: Polígono de frecuencias



<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 2.18 (ver Código 2.8) 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 2.8: Generación de un histogramas variando la cantidad de clases

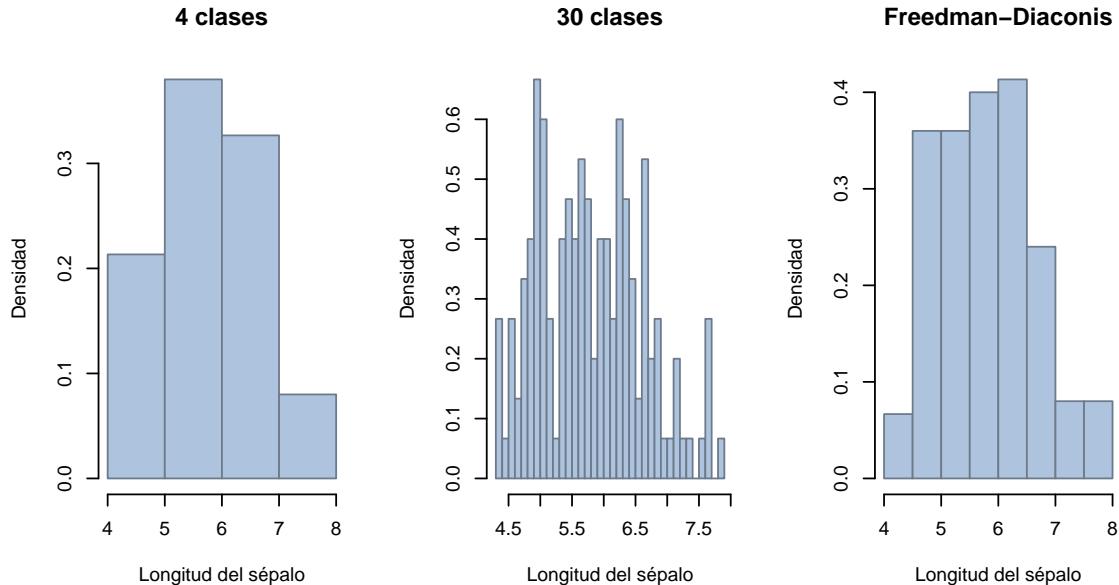


Figura 2.18: Histogramas con distintos intervalos

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 2.19 se muestra la comparación entre las tres opciones.

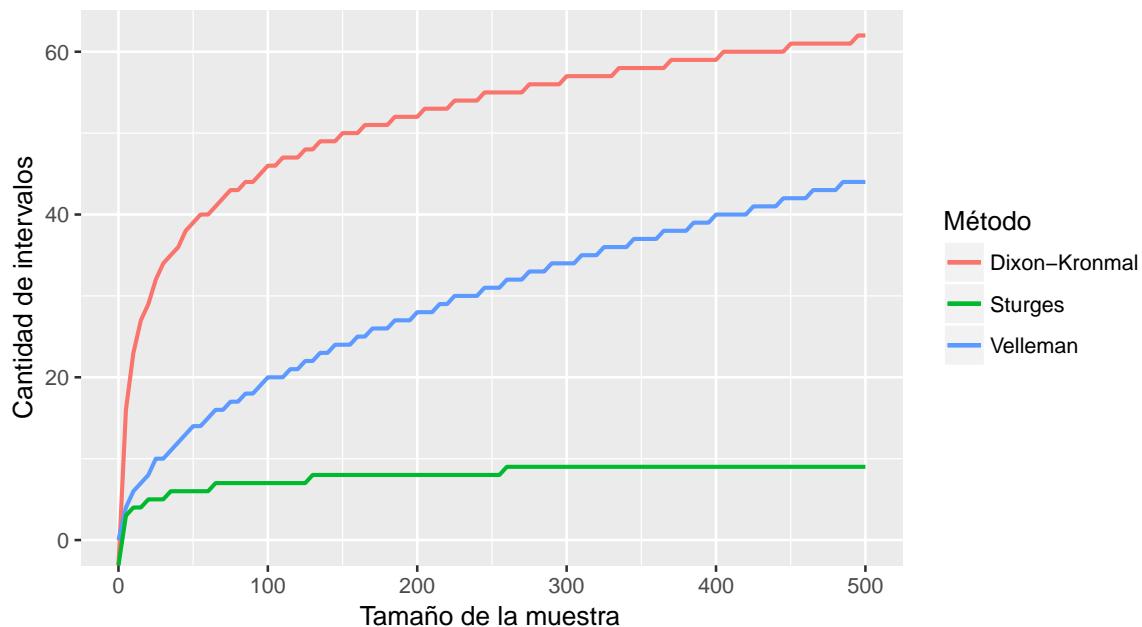


Figura 2.19: Comparación de métodos para el cómputo de intervalos

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 2.20 podemos apreciar el aspecto del *boxplot* para distribuciones simétricas y asimétricas.

Observaciones

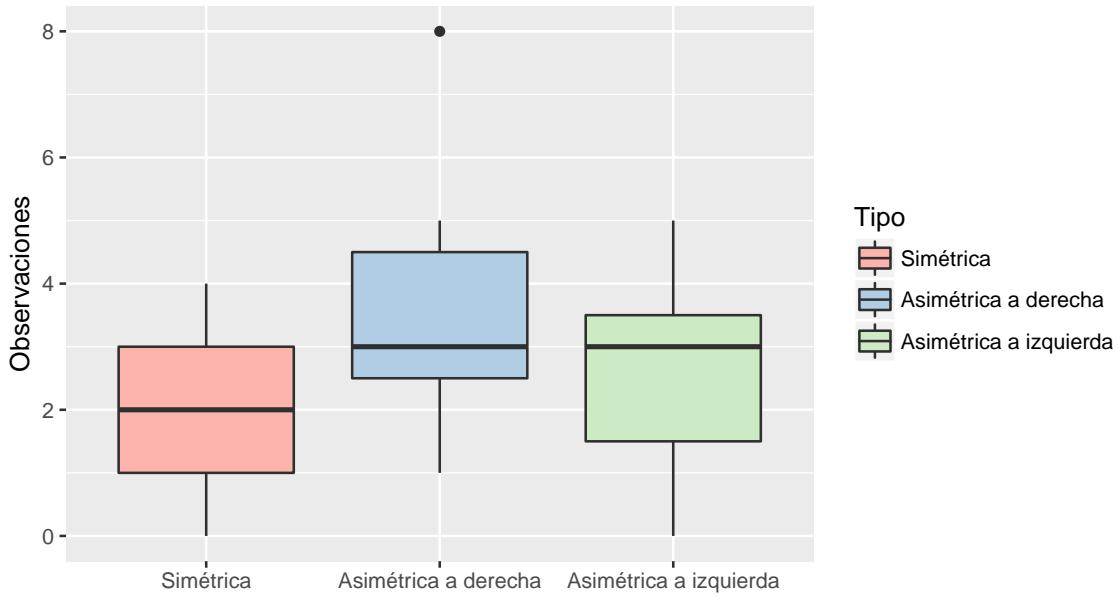


Figura 2.20: Simetría en *boxplots*

- ✿ Si la distribución es simétrica, vemos que la Mediana está ubicada en el centro de la caja y que los bigotes tienen longitudes similares.
- ✿ 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 2.22 generado por el Código 2.9 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 2.9: Generación de un *boxplot* comparativo

Observaciones

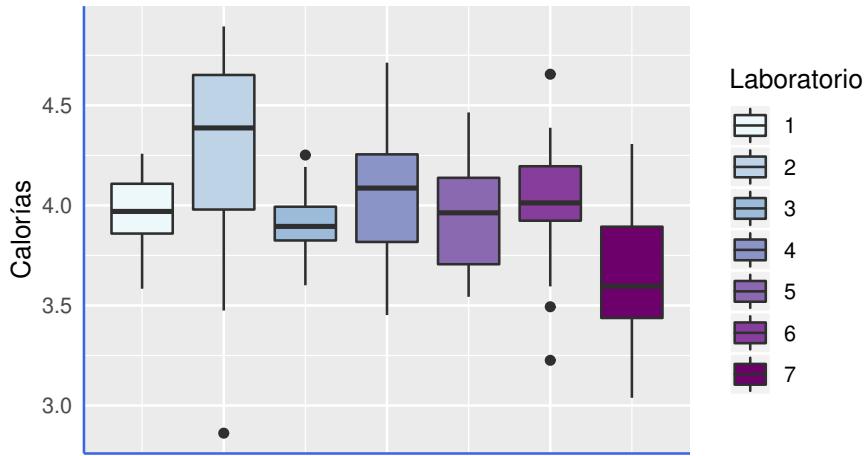


Figura 2.22: *Boxplots* comparativos

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

	Variable ₁	...	Variable _j	...	Variable _p
Individuo ₁	$X_{1,1}$...	$X_{1,j}$...	$X_{1,p}$
⋮	⋮		⋮		⋮
Individuo _i	$X_{i,1}$...	$X_{i,j}$...	$X_{i,p}$
⋮	⋮		⋮		⋮
Individuo _n	$X_{n,1}$...	$X_{n,j}$...	$X_{n,p}$

Tabla 2.7: Modelo de base de datos

- ✿ 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.
- ✿ 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

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

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 2.24 de mosaicos, producida mediante el Código 2.10, 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 2.10: Generación de un diagrama de mosaicos

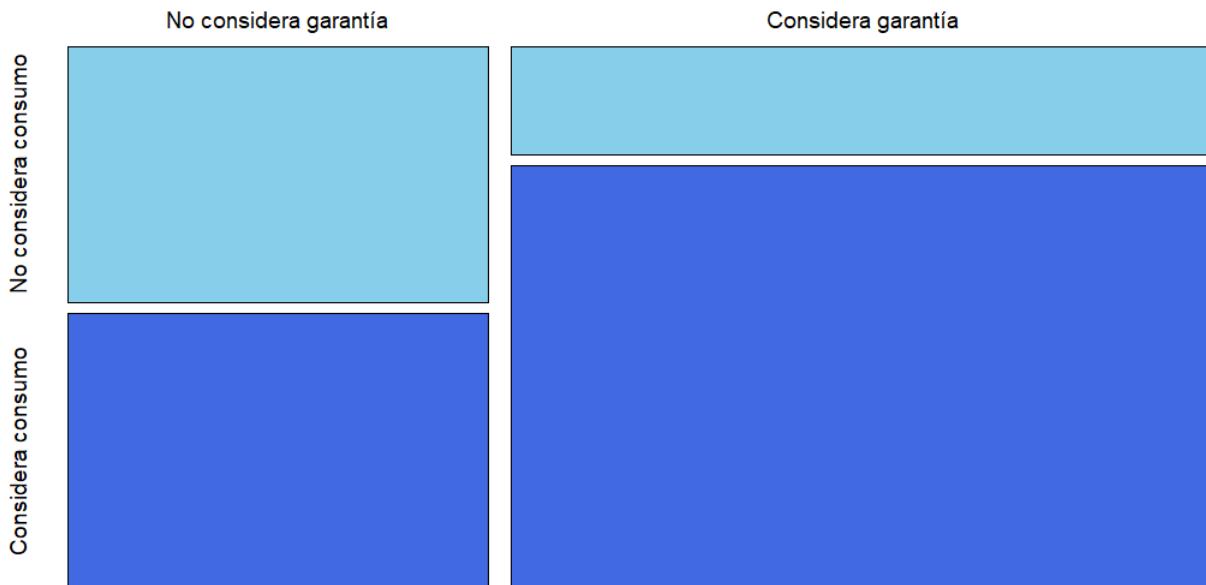


Figura 2.24: Diagrama de mosaicos

En la Figura 2.24 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 2.11 generamos el diagrama de dispersión de la Figura 2.25.

```
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 2.11: Generación de un diagrama de dispersión

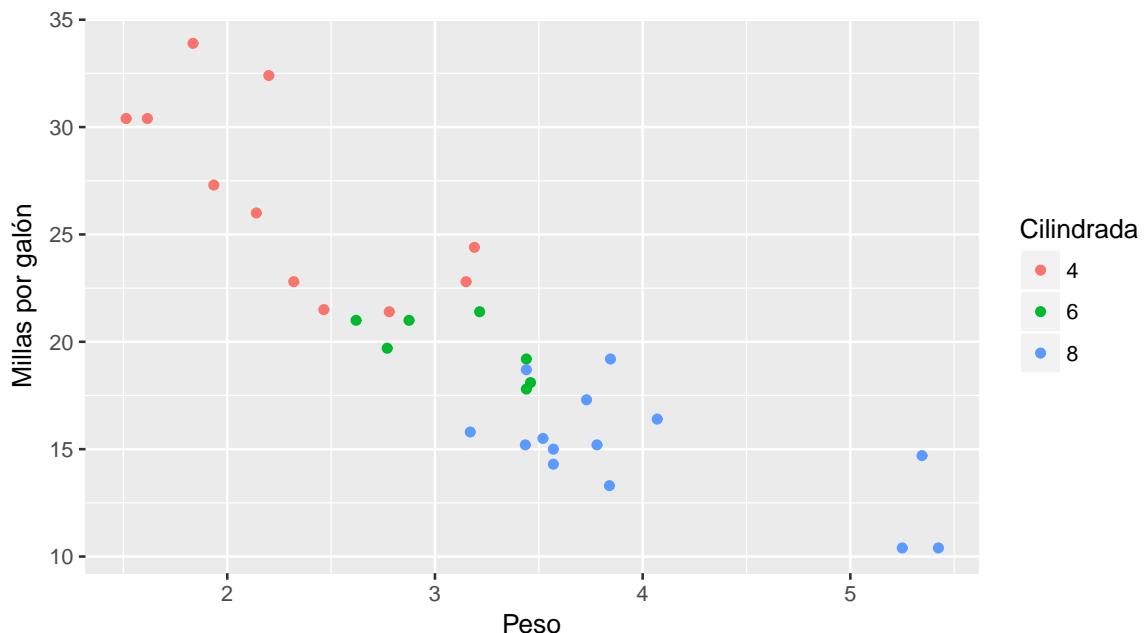


Figura 2.25: Diagrama de dispersión para tres poblaciones

En la Figura 2.25 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 2.12, generamos el dispersograma de la Figura 2.26.

```
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 2.12: Generación de un dispersograma

En el dispersograma de la Figura 2.26 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.

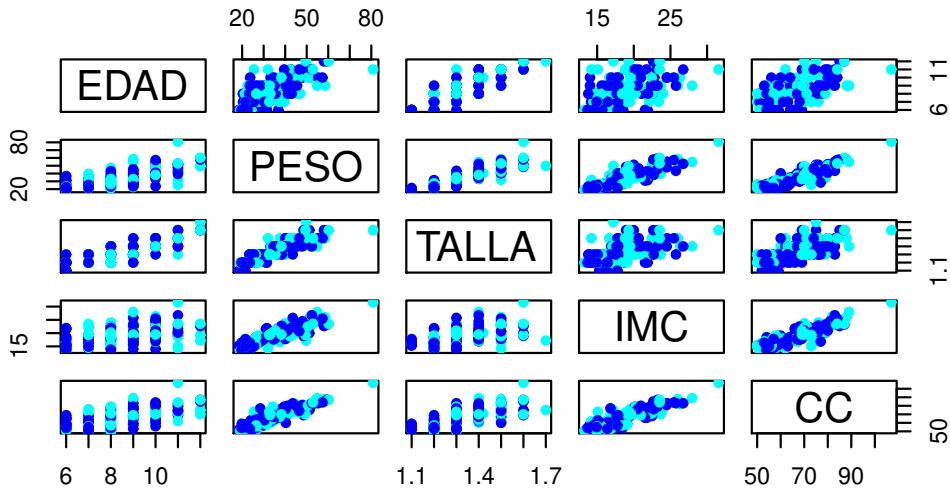


Figura 2.26: Dispersograma

- ✿ 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 2.27 mediante el Código 2.13.

```
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 2.13: Generación de un gráfico de coordenadas paralelas

En la Figura 2.27 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

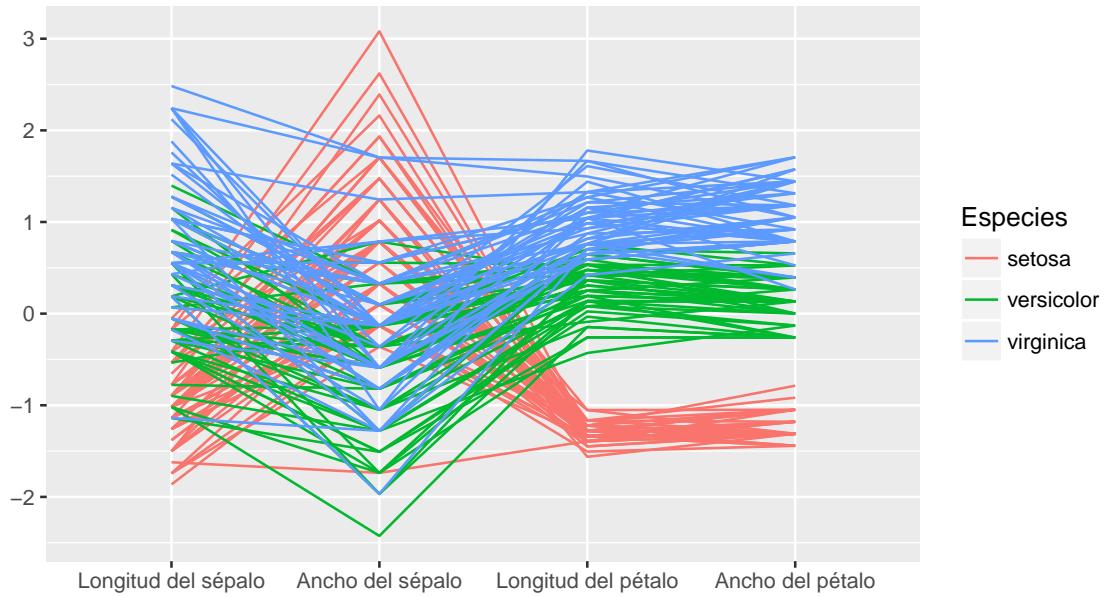


Figura 2.27: Gráfico de coordenadas paralelas

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 2.28 mediante el Código 2.14. 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 2.14: Generación de un gráfico comparativo de perfiles

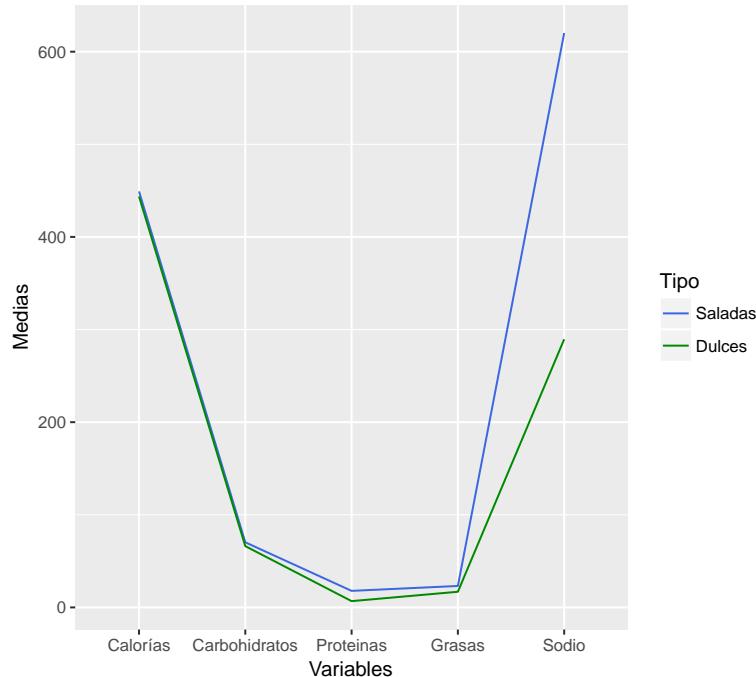


Figura 2.28: Gráfico de perfiles

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 2.29 y 2.30, ambas fueron generadas mediante el Código 2.15.

```
fun=function(x,y) exp(-x^2-y^2)
# Define la funcion 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 2.15: Generación de las curvas de nivel de la Normal Bivariada

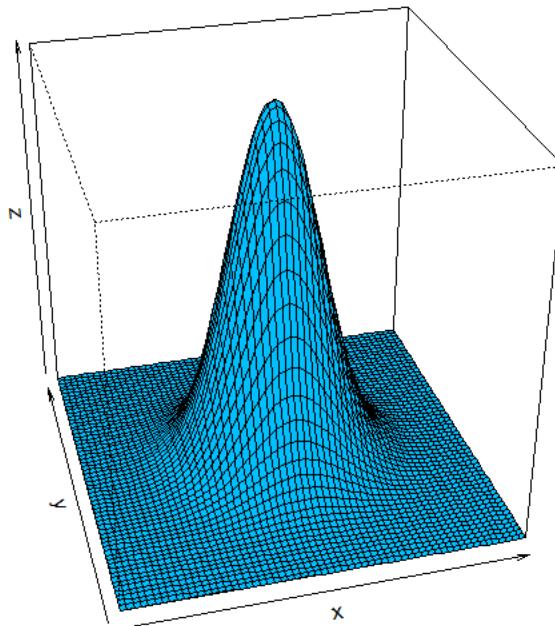


Figura 2.29: Gráfico de la distribución Normal Bivariada

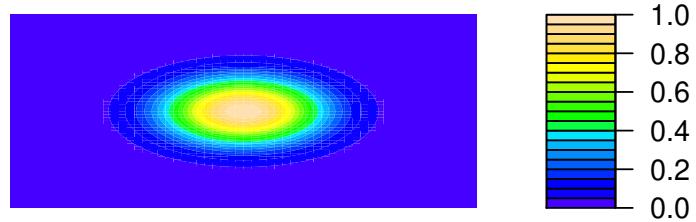


Figura 2.30: Gráfico de las curvas de nivel de la distribución Normal Bivariada

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 2.31, generada con el Código 2.16.

```
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 2.16: Generación de un gráfico de estrellas

En la Figura 2.31 se aprecia similitud en la estructura de los modelos *Mazda* y *Mazda Wag*, así como también son similares los modelos *Merc D* y *Merc*.

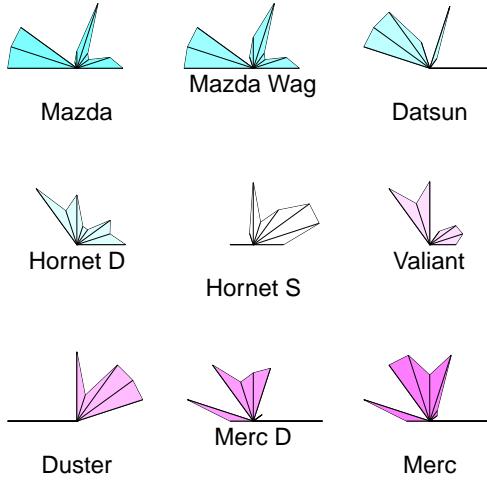


Figura 2.31: Gráfico de estrellas

2.3.1.9 Gráficos de caras de Chernoff

Las caritas de Chernoff [11] 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 2.32, generada con el Código 2.17 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 2.17: Generación de caras de Chernoff

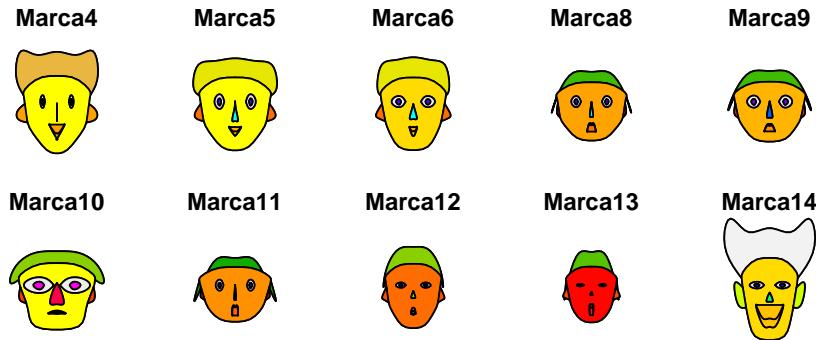


Figura 2.32: Gráfico de caras de Chernoff para galletitas saladas

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 2.18 generamos la Figura 2.34.



<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 2.18: Generación de un gráfico de control univariado

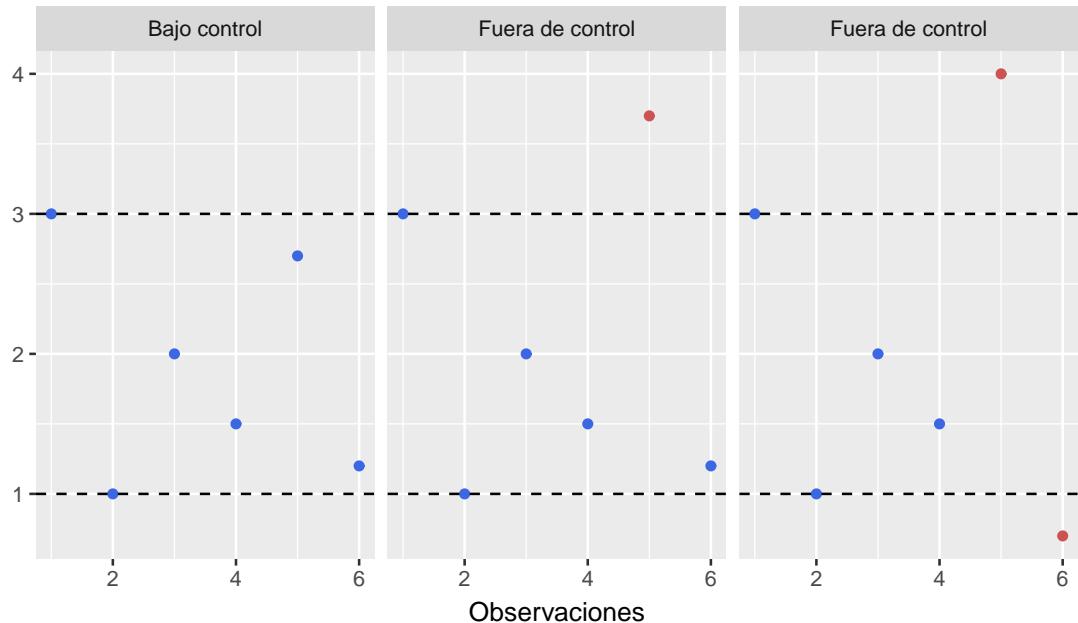


Figura 2.34: Control univariado

En la Figura 2.34 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 2.35 y Código 2.19), 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 2.19: Generación de un gráfico de control multivariado

Nos preguntamos ahora, ¿qué podemos observar en el dispersograma 2.26?

- ✿ 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

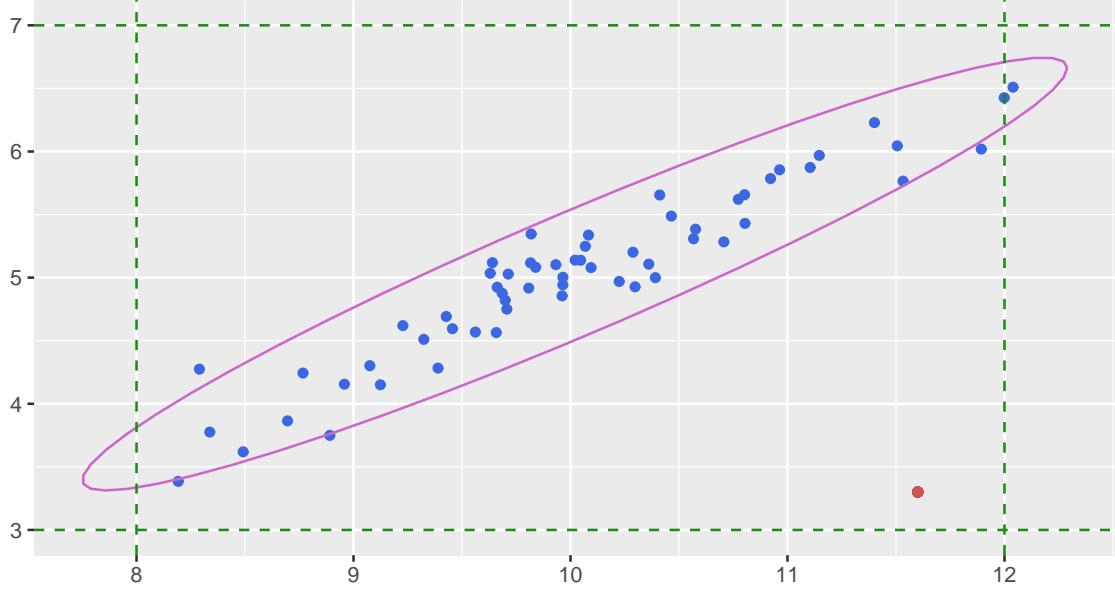


Figura 2.35: Control multivariado

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

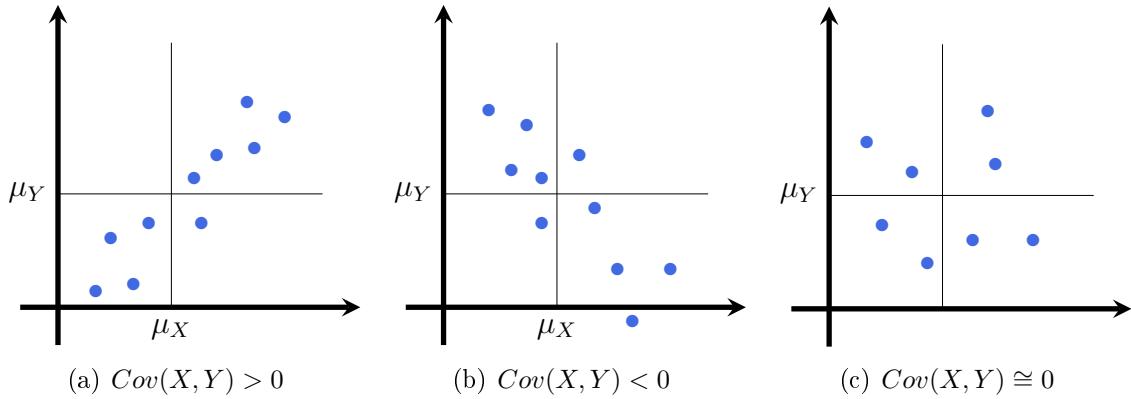


Figura 2.36: Signo de la covarianza

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 2.20, se genera la Figura 2.37.

- ✿ 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 2.20: Generación de un correlograma

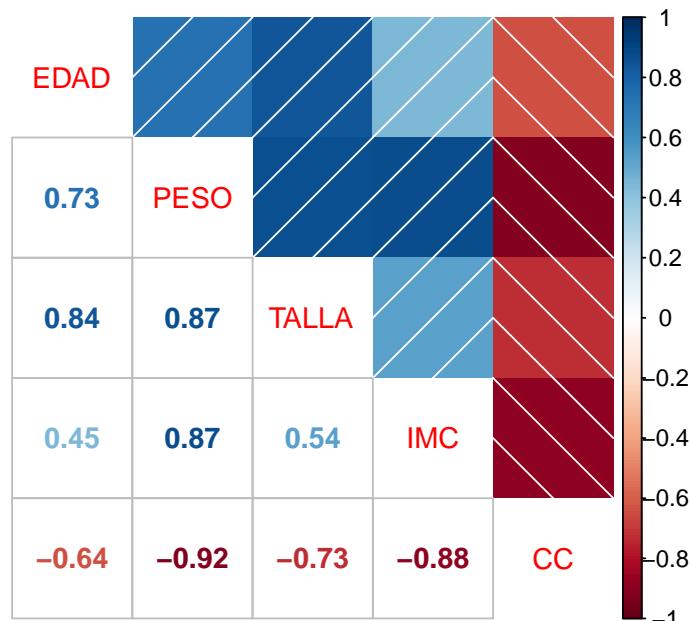


Figura 2.37: Correlograma

En la Figura 2.37 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 [32] 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 Σ 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 [42] 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 [46].

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 [7].

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 [41].

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 2.21 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 2.21: Cálculo en estadística robusta

La Figura 2.38 muestra la densidad estimada de las distancias de Mahalanobis de las observaciones realizadas.

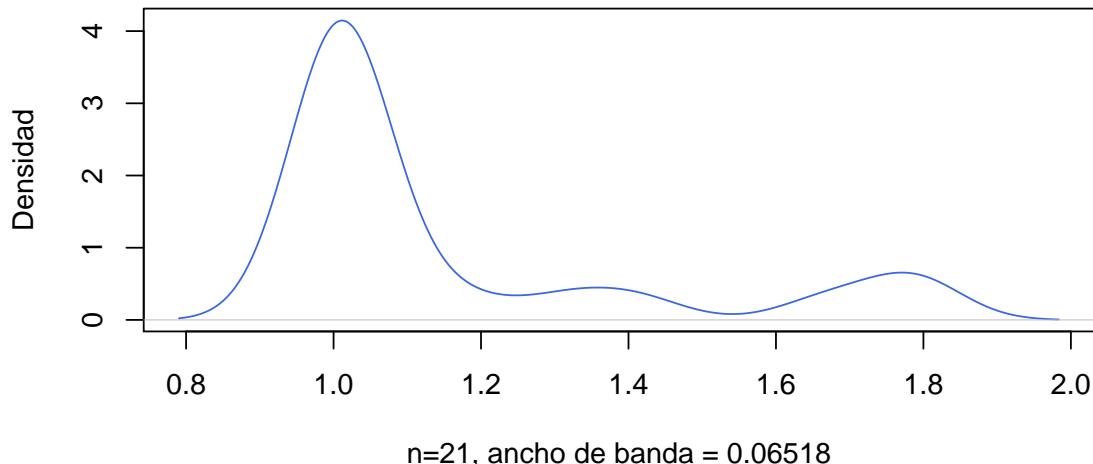


Figura 2.38: Detección multivariada de *outliers*

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	30.56	30.56	5.08
2	31.78	31.78	5.4
3	17.62	17.62	2.54
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	5.91	5.91	7.29
18	2.32	2.32	2.26
19	2.92	2.92	2.54
20	0.46	0.46	0.65
21	13.38	13.38	4.74

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

La vida es el arte de obtener conclusiones suficientes a partir de premisas insuficientes.

— Samuel Butler

3.1 Nociones Previas

En el Apéndice A 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 [22, 9]).

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 α y β 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 .

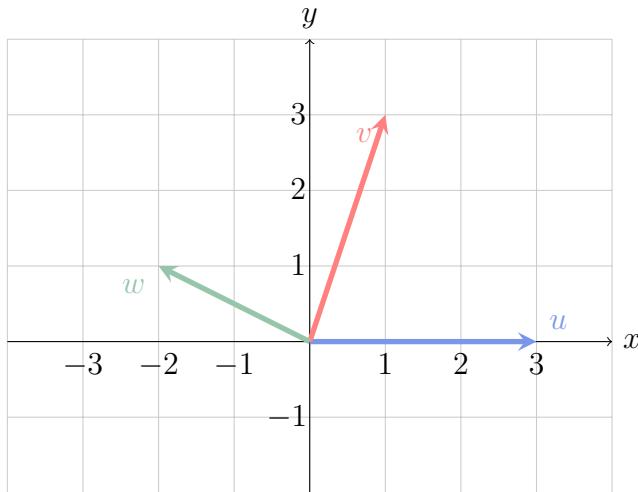


Figura 3.1: Vectores en coordenadas

- ✿ 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} .

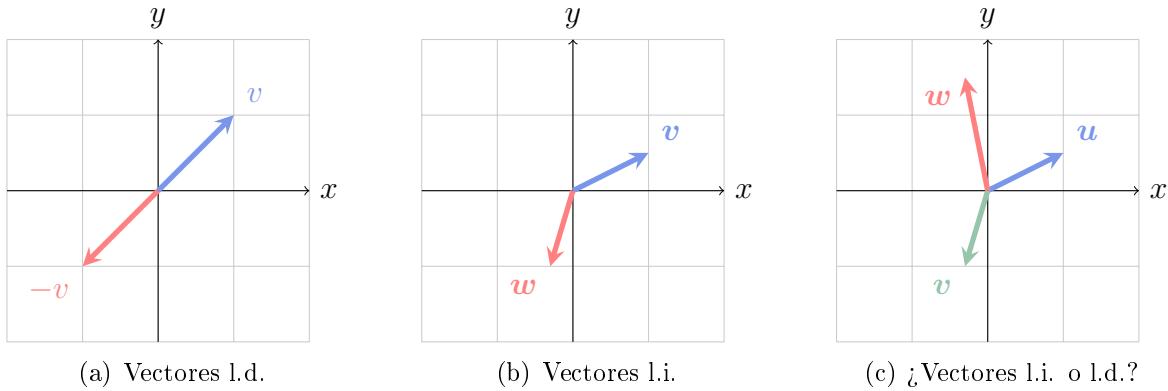


Figura 3.2: Dependencia lineal entre vectores

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

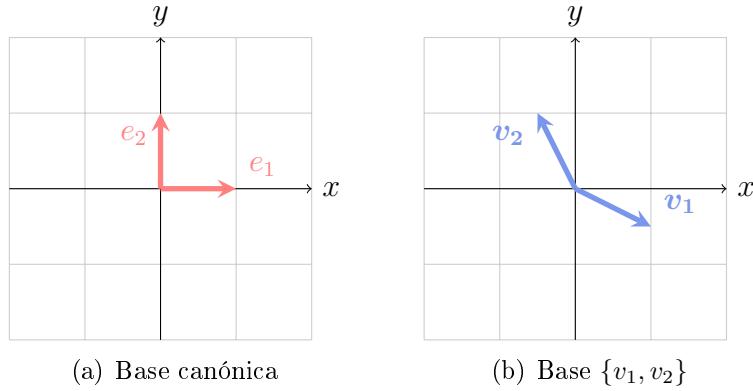


Figura 3.4: Bases para \mathbb{R}^2

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.

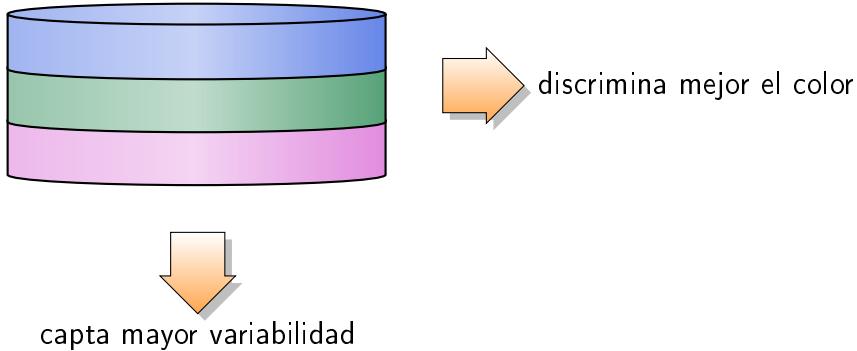


Figura 3.5: Modelo de datos a proyectar

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.

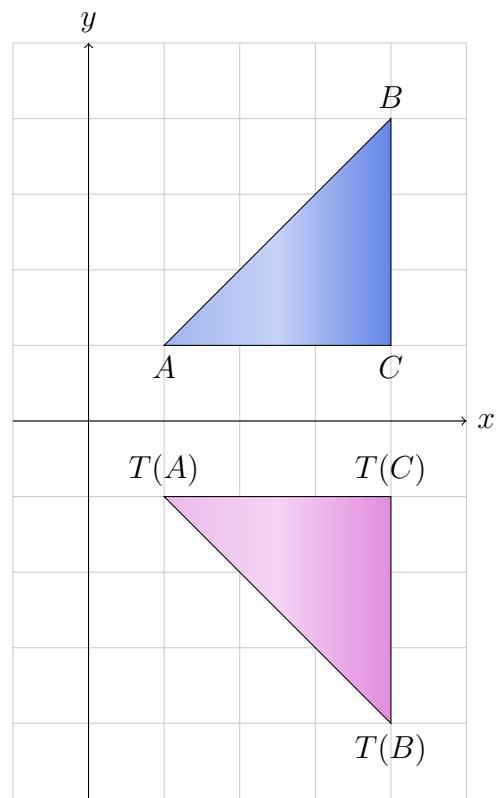


Figura 3.6: Simetría respecto del eje de abscisas

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.

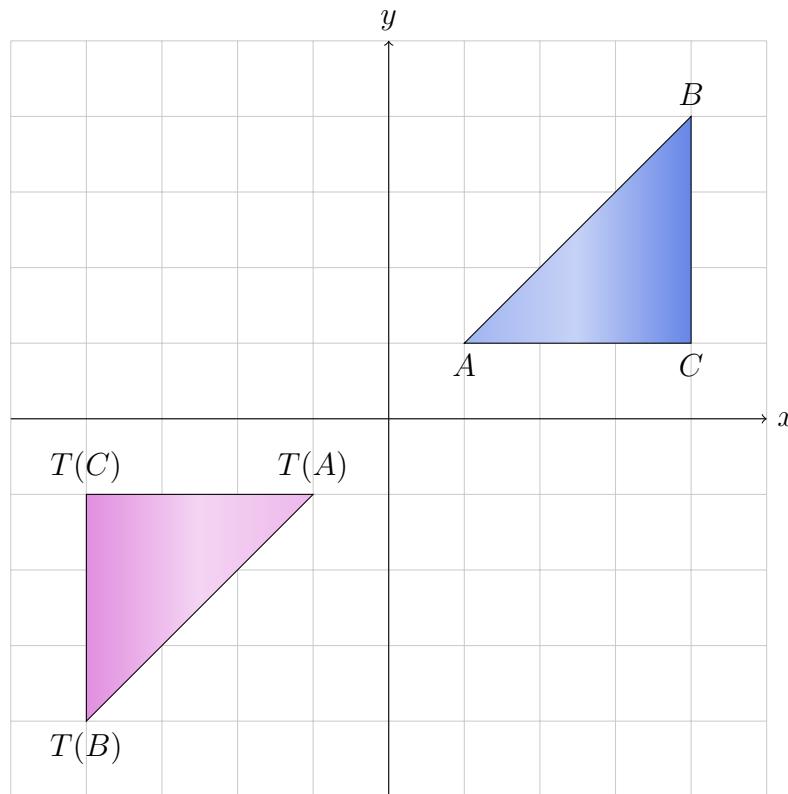


Figura 3.7: Rotación de ángulo π

En general, la matriz de rotación de ángulo θ 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

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

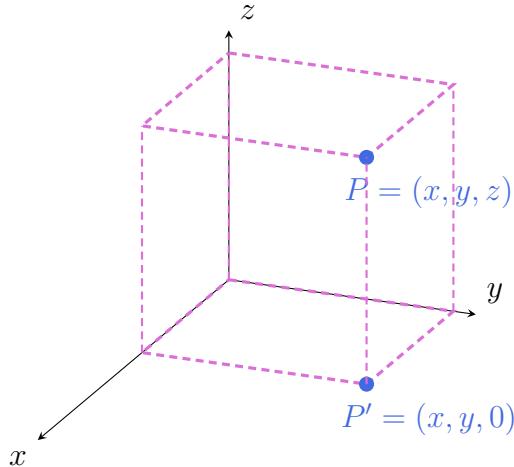


Figura 3.8: Proyección ortogonal de un punto sobre el plano xy

■

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 λ se denomina **autoespacio** asociado al autovalor λ . 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 λ 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 λ 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.

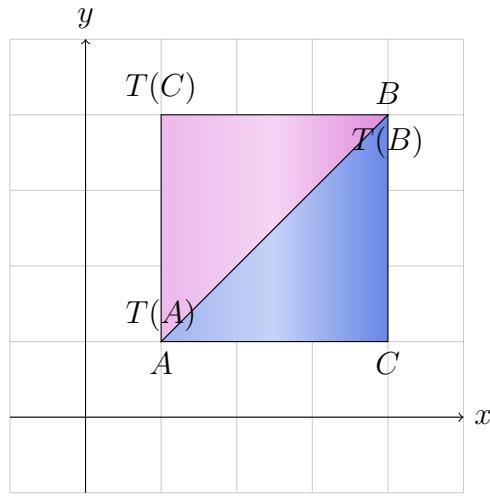


Figura 3.9: Simetría respecto de la recta $y = x$

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 2×2 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 λ_1 y λ_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

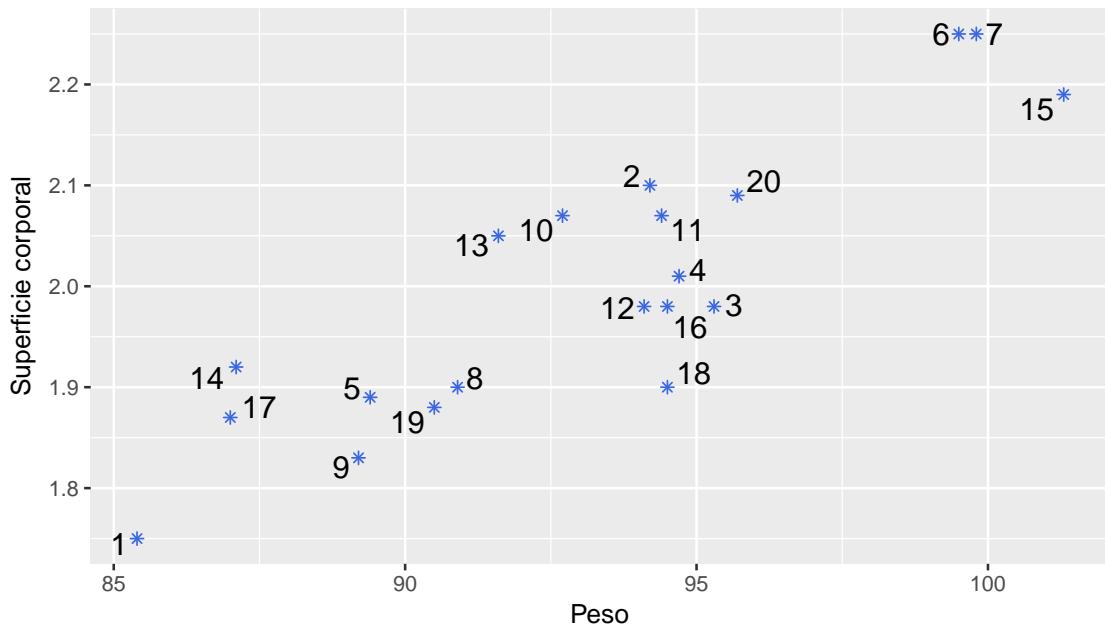


Figura 3.11: Dispersograma entre dos variables

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
```

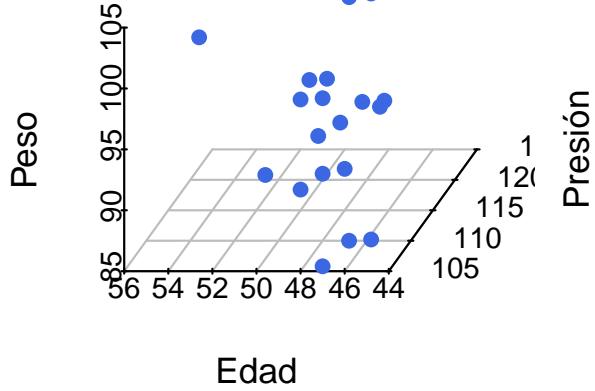
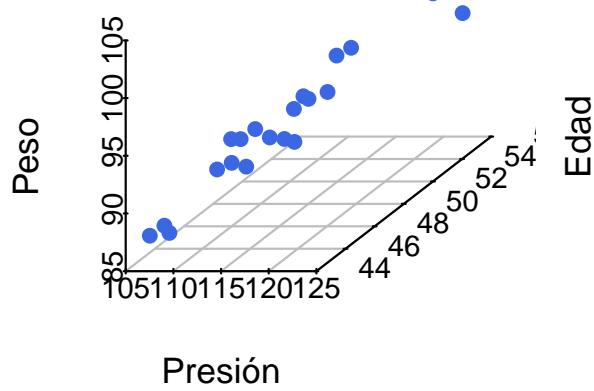


Figura 3.12: Dispersograma 3D desde distintos puntos de vista

```

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

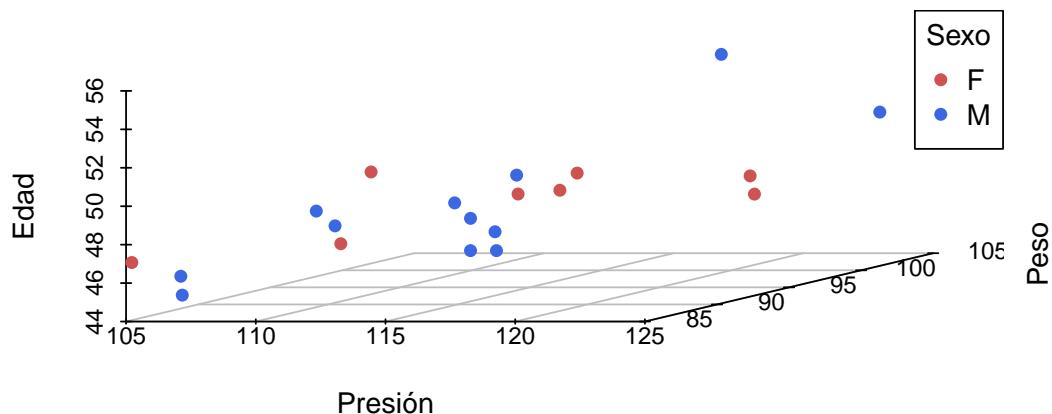


Figura 3.13: Dispersograma en 3D clasificado por grupos

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.

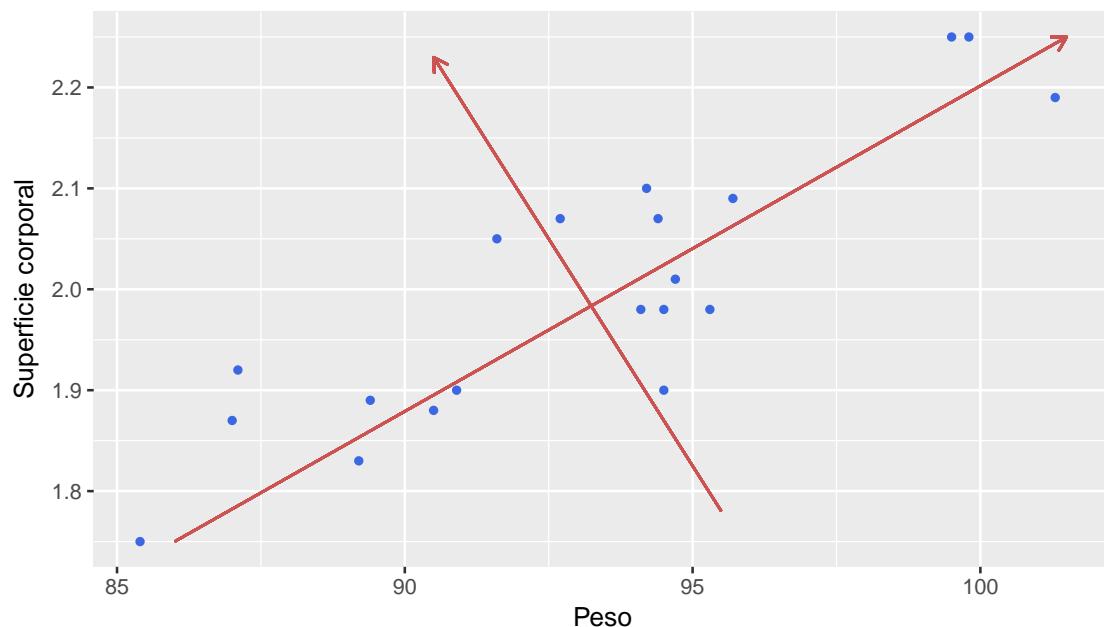


Figura 3.14: Ejes principales

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.

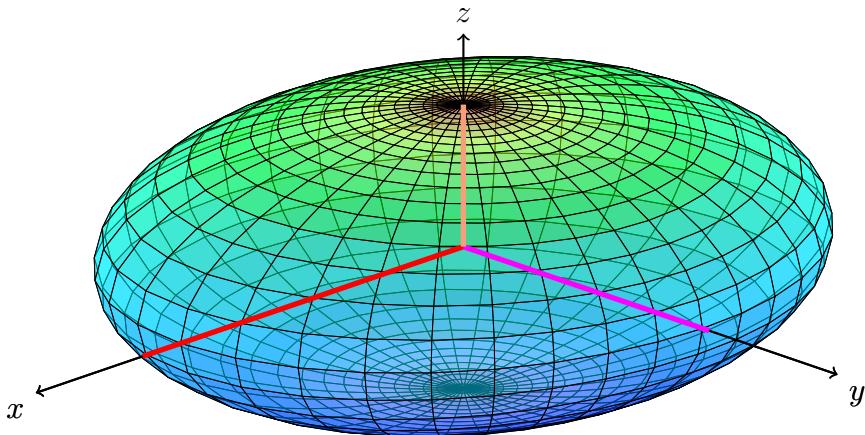


Figura 3.15: Direcciones principales en el espacio tridimensional

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** 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.

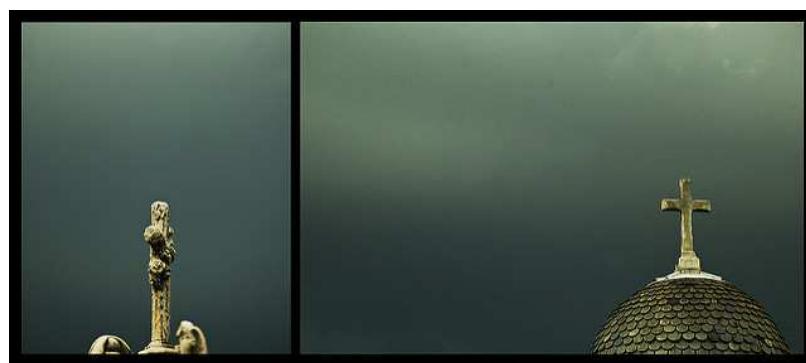
- ✿ sólo tendrá sentido su aplicación en el caso en que las variables originales estén fuertemente correlacionadas.
- ✿ dado que la variabilidad del conjunto está representada por la matriz de covarianzas y ésta se ve influenciada por las unidades de medición, es recomendable realizar el análisis de 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 [24]).
- ✿ 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 [37]).
- ✿ 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 [18]).



<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 Σ .

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 λ 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

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

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 λ que son solución de (3.3) son los autovalores de la matriz Σ y el vector a es el autovector asociado con el autovalor λ y de norma unitaria.

La matriz de covarianzas Σ es real, simétrica y de orden p , por lo cual tiene p autovalores reales. Además, Σ 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 λ_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

* $a_1 a_2^t = 0$,

* $a_2 a_2^t = 1$.

Repetiendo 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, λ_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 λ_i son los autovalores de la matriz Σ .

Mediante un razonamiento análogo a 3.4 se puede probar que $\lambda_i = Var(Y_i)$. Además, los autovalores λ_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 λ_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 [38]. 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

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

siendo

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

En general, no se trabaja con la matriz de covarianza de las variables Σ , 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 B).

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

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

y

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

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

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, χ^2 , y se aplica en forma secuencial. En el Apéndice B 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 Σ . Sin embargo, en general no se dispone de la matriz Σ 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 Σ 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

ggscreepplot(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
Desvío estándar	1.709	0.957	0.348	0.197
Proporción de variabilidad	0.731	0.229	0.030	0.009
Proporción acumulada	0.731	0.960	0.990	1.000

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

	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

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.
- ✿ los autovalores disminuyen considerablemente a partir de la tercer 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.

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.

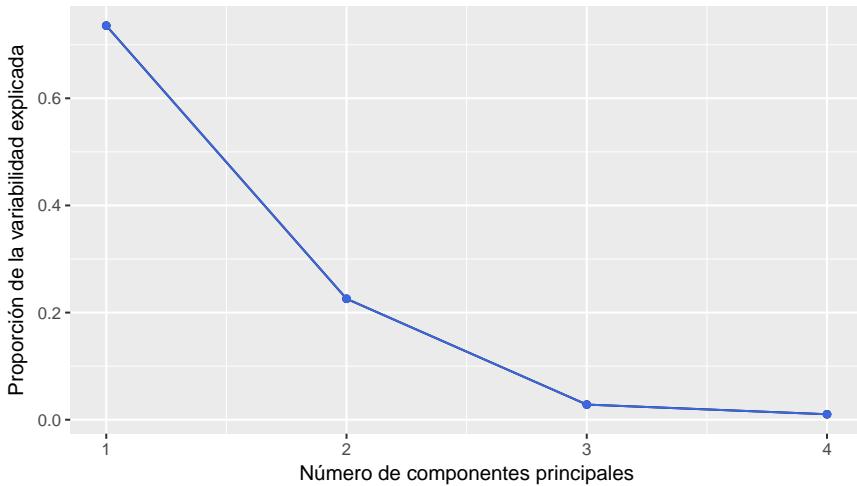


Figura 3.17: Gráfico de sedimentación

- ✿ **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>.

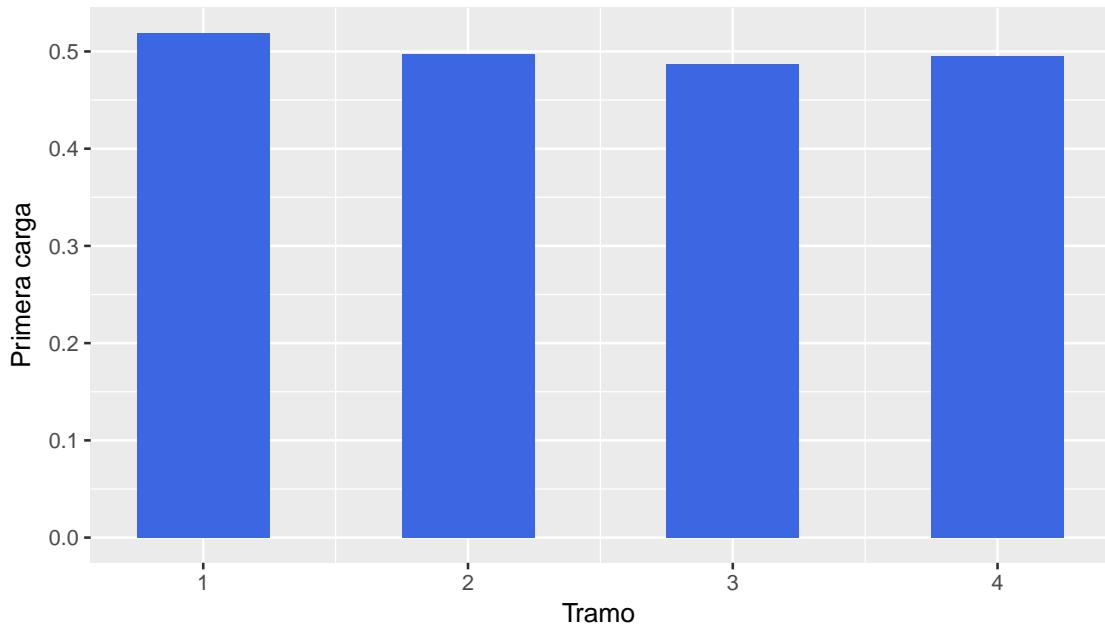


Figura 3.18: Cargas de la primera componente principal

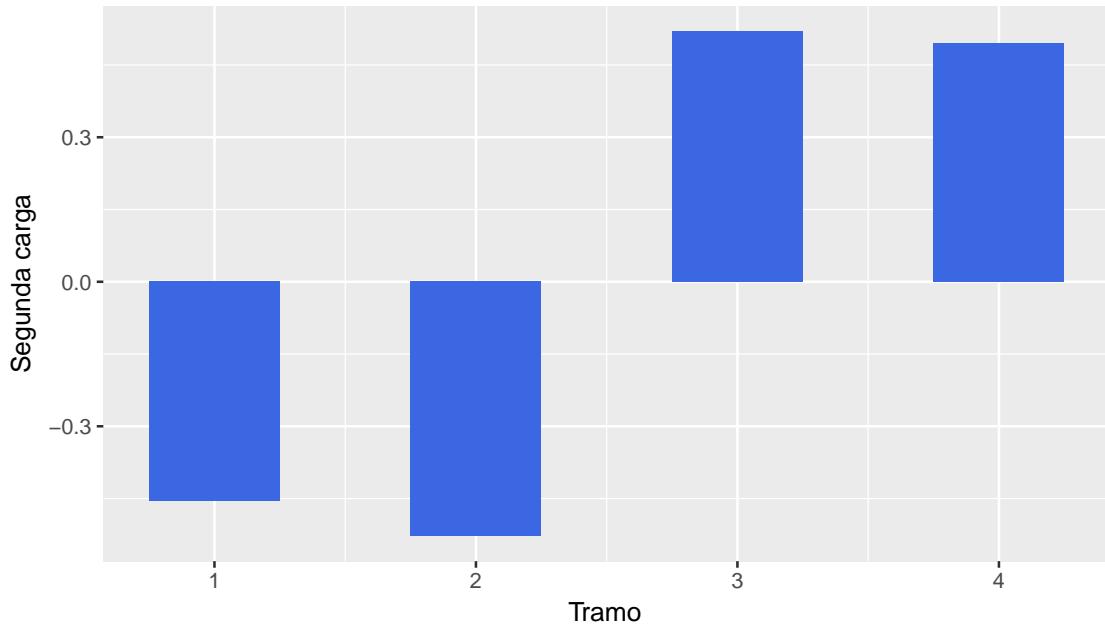


Figura 3.19: Cargas de la segunda componente principal

```

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])
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*, utilizando las expresiones de las componentes principales que se exhiben en la Tabla 3.8.

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.

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 gastó toda su energía en los dos primeros tramos y tarda mucho en los dos últimos porque está cansado, su segunda componente será baja. Esta componente podría denominarse ‘experiencia en carreras’.

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

PC1	PC2	PC3	PC4
-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

Tramo	Media	Desvío	n
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

3.4.9 Biplot

Para poder interpretar los resultados del análisis, se dibuja un gráfico en el que se representan simultáneamente las variables y los valores de cada individuo en pares de componentes principales entre las consideradas. 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>.

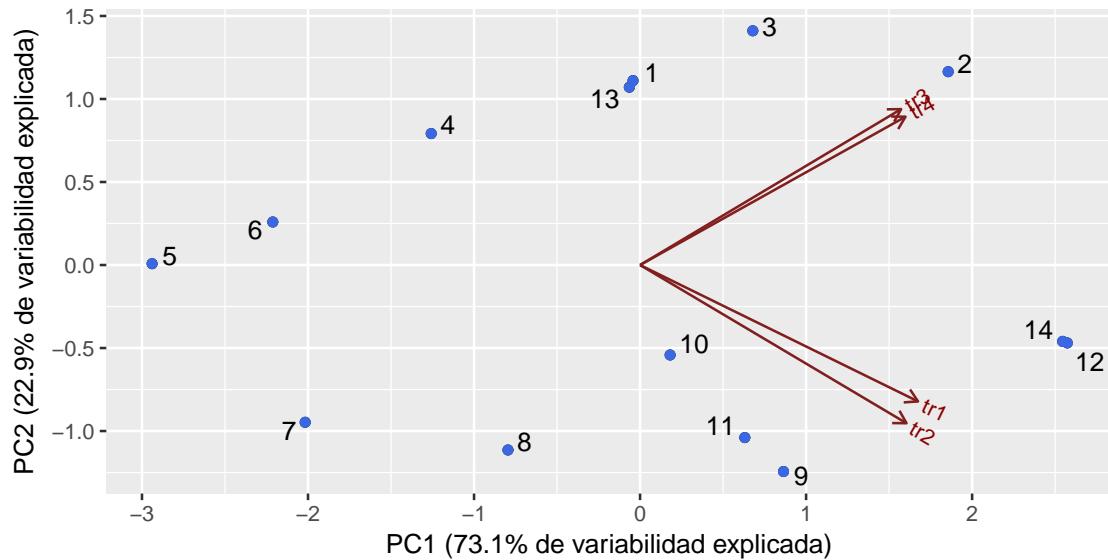


Figura 3.20: *Biplot* para nadadores

```
| 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") +
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

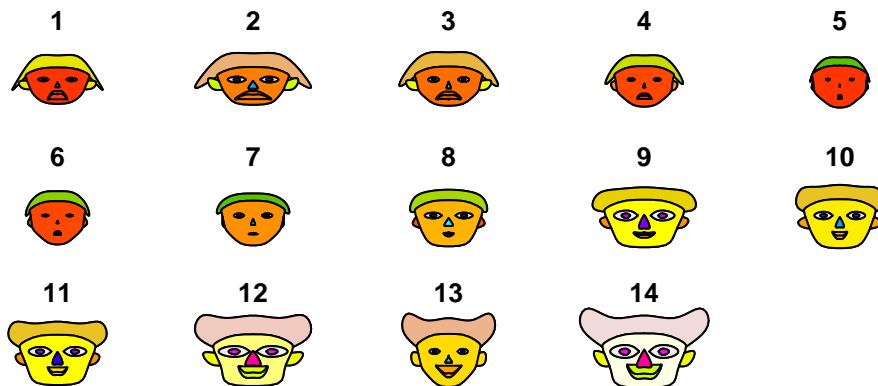


Figura 3.21: Caras de Chernoff para nadadores

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

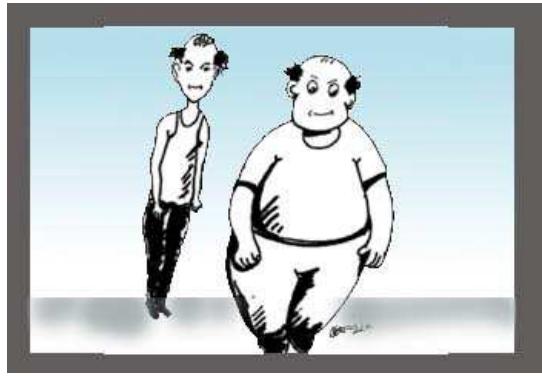
son más expertos. Además sólo tiene sentido realizar esta representación si se dispone de una base relativamente chica de datos.

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 λ_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 $0.7\bar{V} = 13.78$.

	X1	X2	X3	X4
	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
Autovalor	58.49	15.47	2.54	2.24
Porcentaje acumulado	74.27	93.92	97.15	100

Tabla 3.10: Autovalores y autovectores

m	χ^2	g.l.	p-valor
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>.

En la Tabla 3.13 se exhiben las proporciones de variabilidad explicadas por cada una de las componentes principales y la variabilidad explicada acumulada. Para la generación de estos valores ver el Código 3.10.

En la Figura 3.23 se presenta el gráfico de sedimentación o screeplot correspondiente a las componentes halladas generado dentro del Código 3.10.

En la Tabla 3.14 se encuentran los *loadings* o cargas de las primeras cuatro componentes principales calculados mediante el Código 3.10.

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

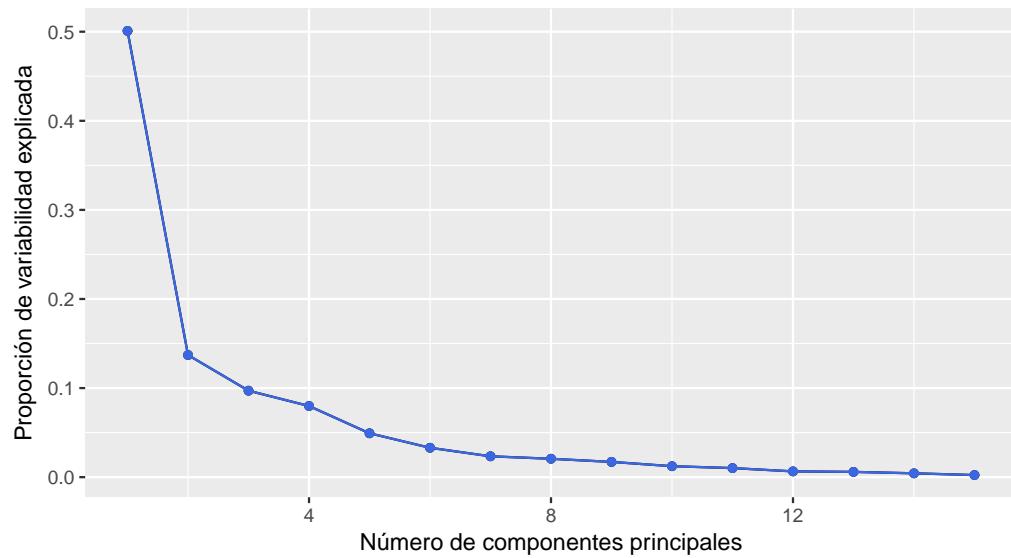


Figura 3.23: Gr\'afico de sedimentaci\'on para aspirantes

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

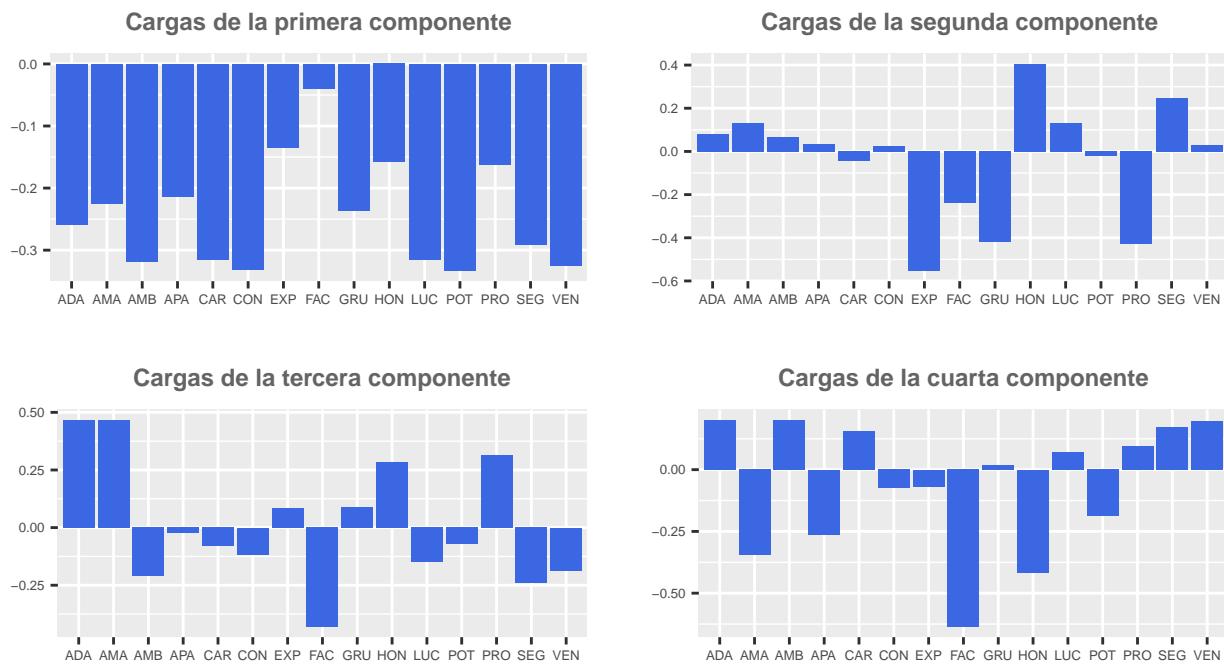


Figura 3.24: Cargas para los aspirantes

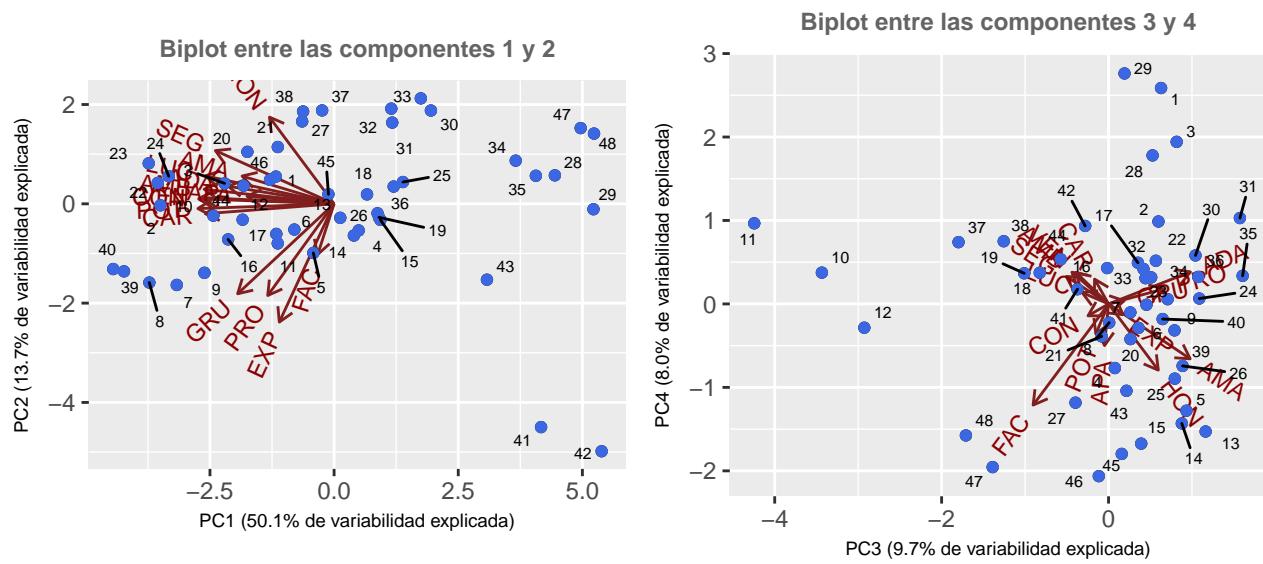


Figura 3.25: Biplots para los aspirantes

```

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 hacernos las siguientes preguntas:

- ✿ ¿Cuántas componentes principales sería conveniente considerar? ¿Qué criterio se utiliza para responder a esta pregunta?
- ✿ ¿Es pertinente la aplicación de esta técnica en este caso?
- ✿ ¿Las componentes principales son de tamaño o de forma?
- ✿ ¿Qué implica un valor alto en la primera componente? ¿Y en la segunda?
- ✿ ¿Qué nombres serían adecuados para las primeras dos componentes principales?



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) [44], otra es el estimador de Stahel-Donoho [14] y una tercera propuesta es el *Minimum volume ellipsoid* (MVE) [46].

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 [47].

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.

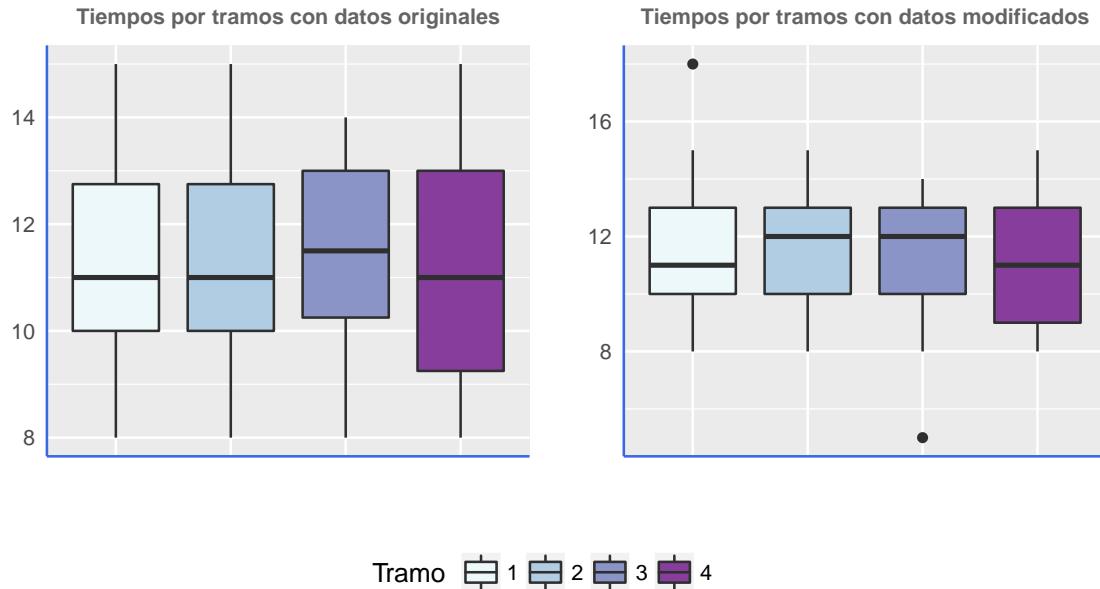


Figura 3.26: Comparación de boxplots para nadadores

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.

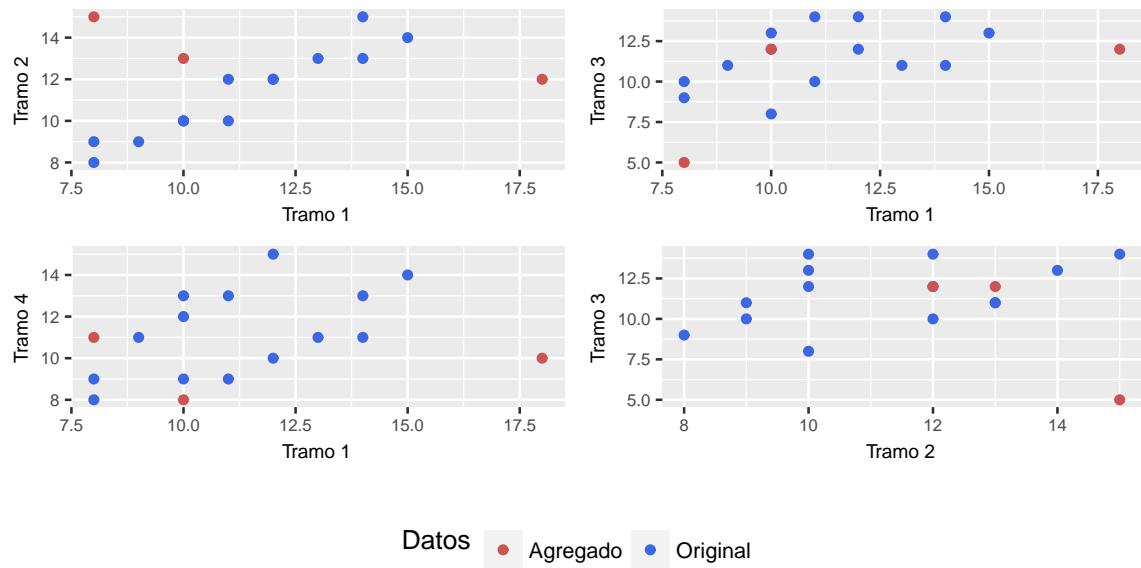


Figura 3.27: Diagramas de dispersión para nadadores

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.

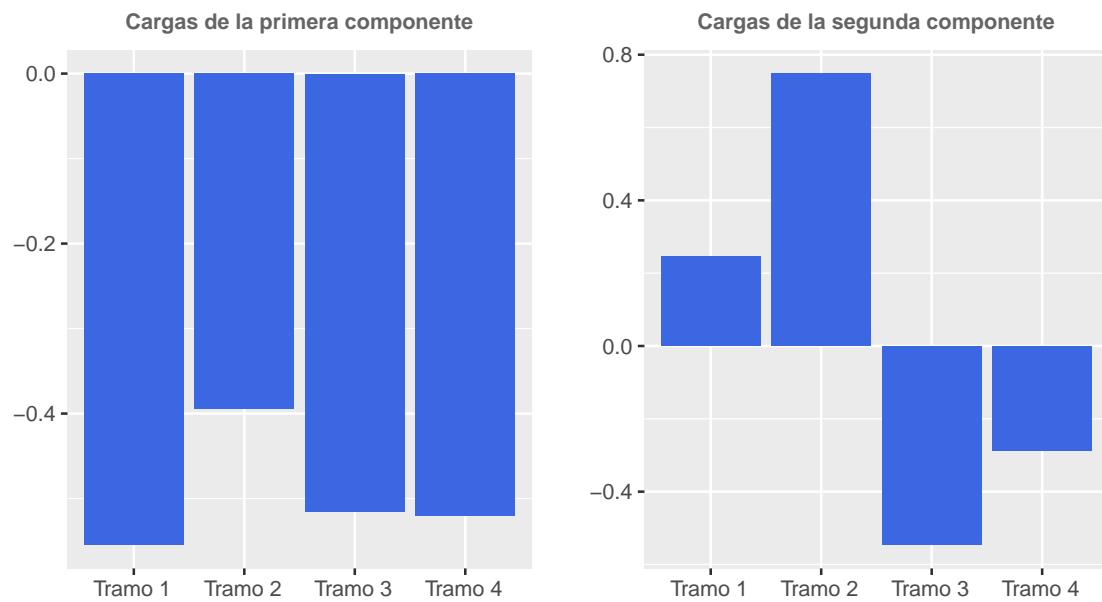


Figura 3.28: Cargas ACP(clásico) para nadadores con datos agregados

```
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
```

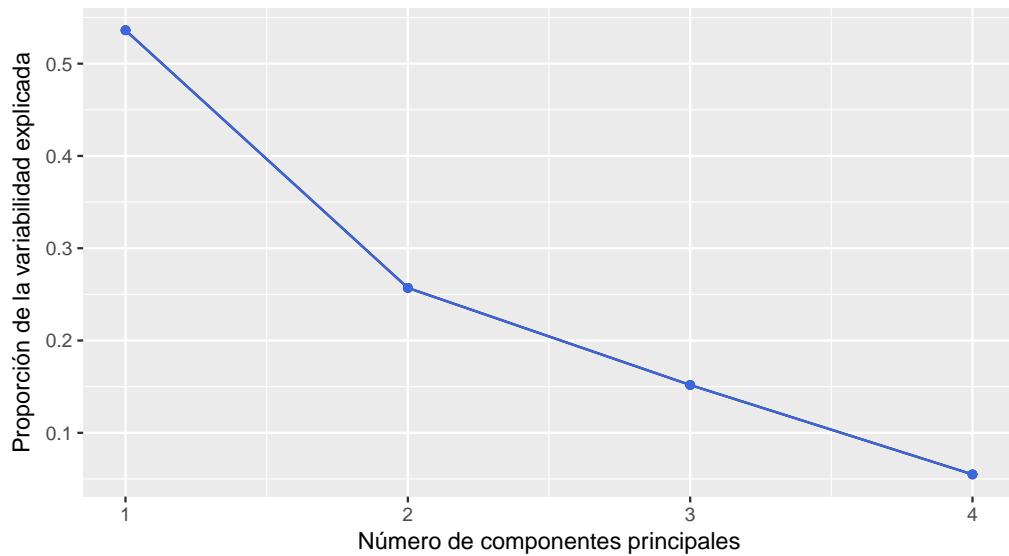


Figura 3.29: Análisis clásico de *screeplot* para los nadadores con los datos agregados

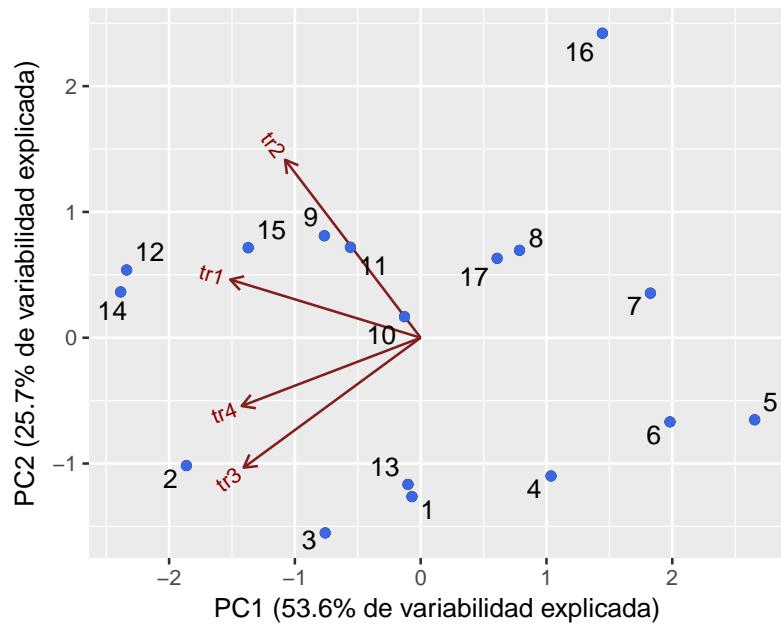


Figura 3.30: *biplot clásico* para nadadores con datos agregados

```

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_primer_a_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: ACP nadadores con datos agregados

A continuación veremos cómo actúan las diferentes alternativas robustas sobre 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
Desviación estandar	1.670	1.031	0.337	0.181
Proporción de variabilidad	0.698	0.266	0.028	0.008
Variabilidad acumulada	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
nad.cont=rbind(nad,c(15,18,12,12,10),c(16,8,15,5,11),c(17,10,13,12,8))

```

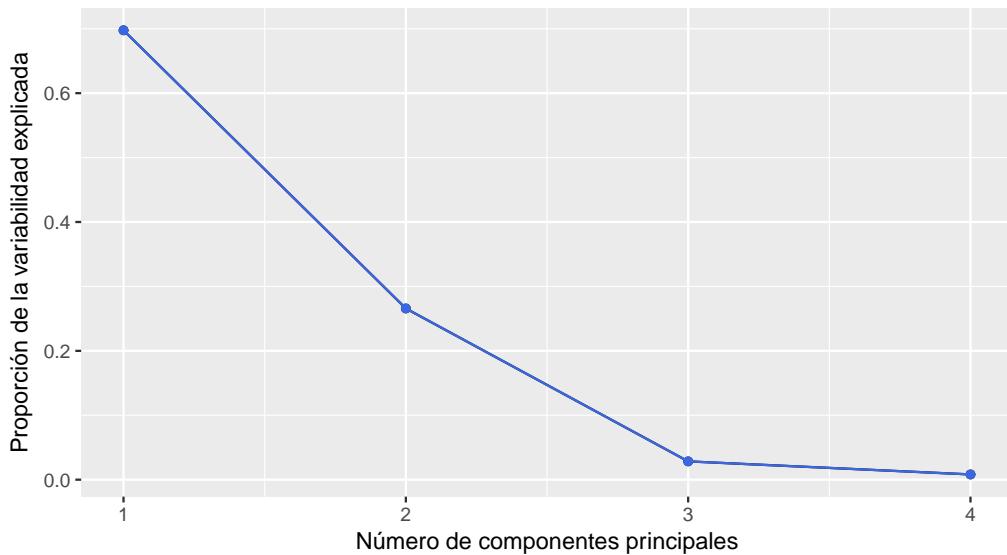


Figura 3.31: *Screeplot* para MCD de nadadores con datos agregados

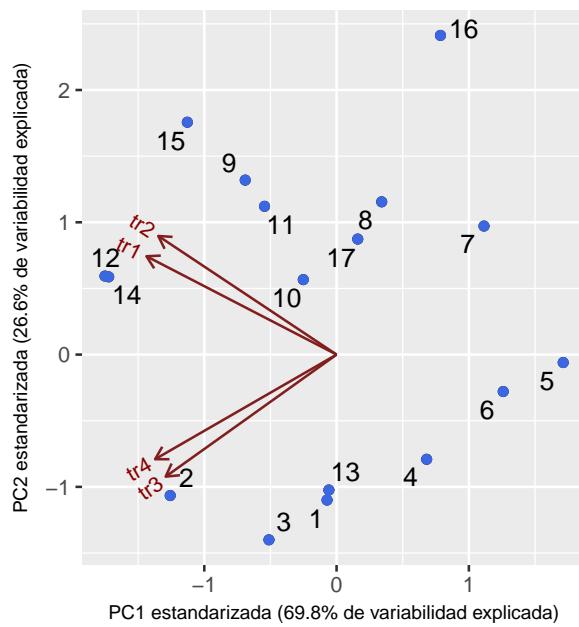


Figura 3.32: *Biplot* para MCD de nadadores con datos agregados

```

# 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:

X₁: puntuación media obtenida en las asignaturas de Econometría

X₂: puntuación media obtenida en las asignaturas de Derecho

X₃: 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 Σ .
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 son.
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**, aunque puede ser rechazada 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 del azar en la selección de la muestra.

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: μ que indica la media, y σ 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, para 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.

La probabilidad de la región crítica o de rechazo se denomina **nivel de significación** o **nivel de riesgo** y se representa con la letra α . De esta manera, la probabilidad 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 las zonas de rechazo o crítica y de no rechazo, se determina fijando el valor de α ; 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 α . 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.



Figura 4.1: Ejemplo de regiones en un contraste bilateral

Una vez planteada la 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 .
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 α de equivocarnos. Por el contrario, si no se rechaza, no significa que estamos 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. Podemos concluir que, si se mantiene o no se rechaza 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.

	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

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.

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, β , es un valor desconocido que depende de los siguientes factores:

- ✿ la hipótesis H_1 que se considere verdadera,
- ✿ el valor de α ,
- ✿ 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 β marcada con rojo en la Figura 4.2. Así, el valor de β depende del valor concreto que se haya supuesto en H_1 para el parámetro de interés. Es más, α y β se relacionan de forma inversa, siendo menor β cuanto mayor es α . 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 (μ_0 y μ_1 de la Figura 4.2).

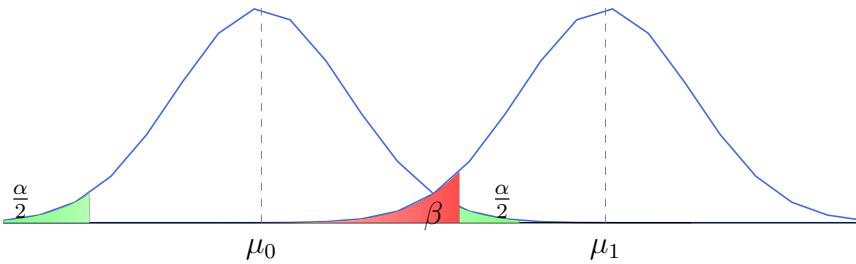


Figura 4.2: Representación de los errores de un test

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, α , se establece ‘a priori’ por lo tanto es único. Sin embargo, hay un valor diferente de β por cada valor del parámetro escogido en H_1 .

Como α 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 π , a la probabilidad de decidir por H_1 cuando ésta es cierta; es decir,

$$\pi = P(\text{Rechazar } H_0) = \begin{cases} 1 - \beta & \text{si } \mu \notin H_0 \\ \alpha & \text{si } \mu \in H_0 \end{cases}$$

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 (α), conduce a incrementar la probabilidad del error de tipo II (β).

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 α establecido, aquel que tenga mayor potencia; esto es, menor probabilidad de incurrir en el error de tipo II (β).

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 [39]. 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 muestreadas.

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 televíidente categorizada como joven, adulto y mayor. Los resultados obtenidos se muestran en la Tabla 4.2.



<https://flic.kr/p/s2iGbo>

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. Se denota por $P(A/B)$ a la probabilidad de que ocurra A sabiendo que ocurrió B , o bien, probabilidad de A condicionada 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. Se dice 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 $P(B) > 0$,

$$P(A/B) = P(A) \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

Ejemplo 4.4. En particular, diremos entonces que en la población del Ejemplo 4.1, 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.

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

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 las variables 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.

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 H_0 versus H_1 , siendo

$$\begin{cases} H_0 : \text{ las variables son independientes} \\ H_1 : \text{ las variables no son independientes} \end{cases}$$

En forma simbólica el planteo es:

$$\begin{cases} H_0 : P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j), \forall(i, j) / 1 \leq i \leq r, 1 \leq j \leq k \\ H_1 : \exists(i, j) / P(X = x_i, Y = y_j) \neq P(X = x_i)P(Y = y_j) \end{cases}$$

Ejemplo 4.5. Siguiendo los datos de la Tabla 4.2, que contiene las frecuencias observadas para el estudio de la cantidad de horas consumidas en programas televisivos con cierto grado de violencia de los $n = 80$ televidentes encuestados y clasificados de acuerdo con su edad.

Resulta ahora de interés, calcular 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:

$$\hat{p}_{11} = \hat{P}((16 - 34) \cap PV) = \hat{P}(16 - 34) \cdot \hat{P}(PV) = \frac{26}{80} \cdot \frac{40}{80} = \frac{13}{80}$$

Con lo cual, el número esperado de individuos jóvenes que consume poca violencia resulta

$$\hat{e}_{11} = n \cdot \hat{P}((16 - 34) \cap PV) = 80 \cdot \frac{13}{80} = 13$$

		Edad (en años)						
		Joven		Adulto		Mayor		Totales
Nivel de violencia		o_{ij}	\hat{e}_{ij}	o_{ij}	\hat{e}_{ij}	o_{ij}	\hat{e}_{ij}	
Poca		8	(13)	12	(13.5)	20	(13.5)	40
Mucha		18	(13)	15	(13.5)	7	(13.5)	40
Totales		26	(26)	27	(27)	27	(27)	80

Tabla 4.6: Frecuencias observadas y esperadas (violencia por edad)

De 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} .

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 poblaciones 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.

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

$$n_{i \cdot} = \sum_{j=1}^k o_{ij}$$

El total de observaciones de la categoría j -ésima de la variable en todas las muestras es

$$n_{\cdot j} = \sum_{i=1}^r o_{ij}$$

	X_1	X_2	...	X_j	...	X_k	Totales
Muestra 1	o_{11}	o_{12}	...	o_{1j}	...	o_{1k}	$n_{1..}$
Muestra 2	o_{21}	o_{22}	...	o_{2j}	...	o_{2k}	$n_{2..}$
⋮	⋮	⋮		⋮		⋮	⋮
Muestra i	o_{i1}	o_{i2}	...	o_{ij}	...	o_{ik}	$n_{i..}$
⋮	⋮	⋮		⋮		⋮	⋮
Muestra r	o_{r1}	o_{r2}	...	o_{rj}	...	o_{rk}	$n_{r..}$
Totales	$n_{.1}$	$n_{.2}$...	$n_{.j}$...	$n_{.k}$	$n_{...}$

Tabla 4.7: Frecuencias teóricas de homogeneidad

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 4.4 haciéndonos la siguiente pregunta

¿La distribución de la variable dada por el color es homogénea en las tres poblaciones representadas?

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:

$$\begin{cases} H_0 : P(X_j/m_i) = p_{j|i} = P(X_j) = p_j, \forall(i, j) / 1 \leq i \leq r, 1 \leq j \leq k \\ H_1 : \exists(i, j) / P(X_j/m_i) = p_{j|i} \neq P(X_j) = p_j \end{cases}$$

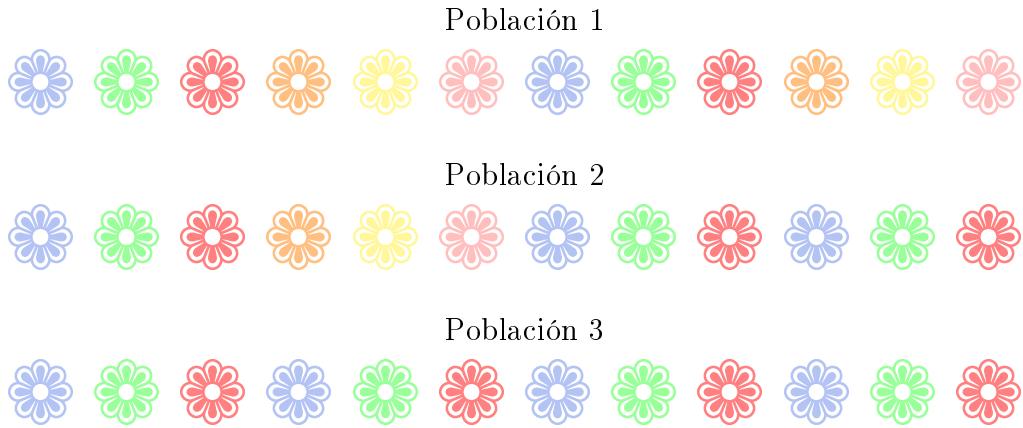


Figura 4.4: Poblaciones según variable de color

¿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. Es por ello que 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 poblaciones seleccionadas menos uno para el caso de homogeneidad.

En ambos casos los grados de libertad están en 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.

La distribución Chi cuadrado es asimétrica por la derecha y su forma depende de los grados de libertad (ver Figura 4.5). 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 4.6. 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.



<https://flic.kr/p/21MtHhT>

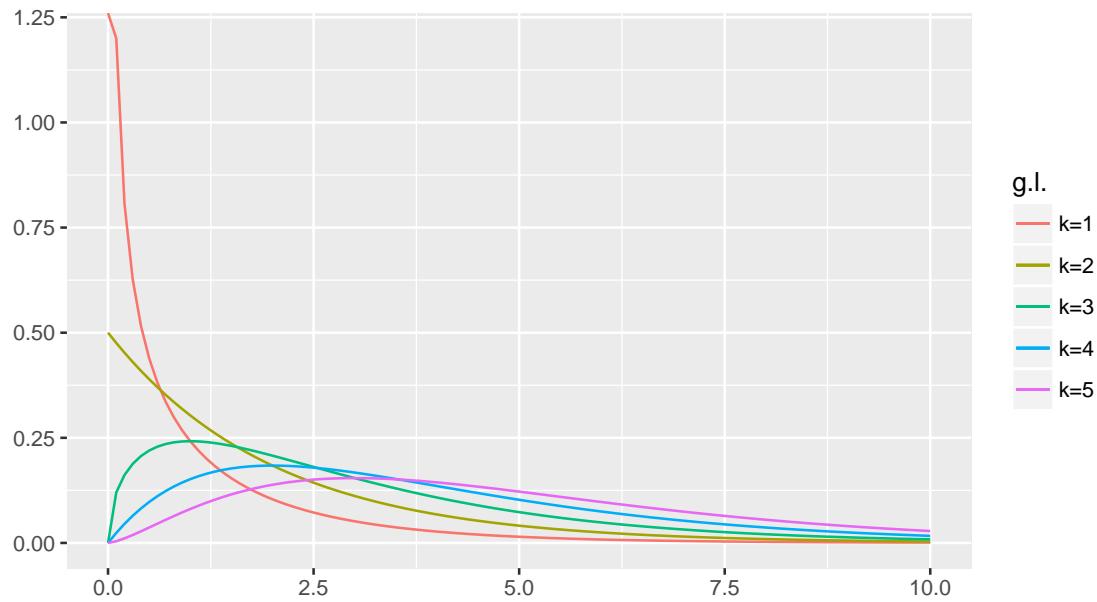


Figura 4.5: Gráficos de la distribución χ^2 por grados de libertad

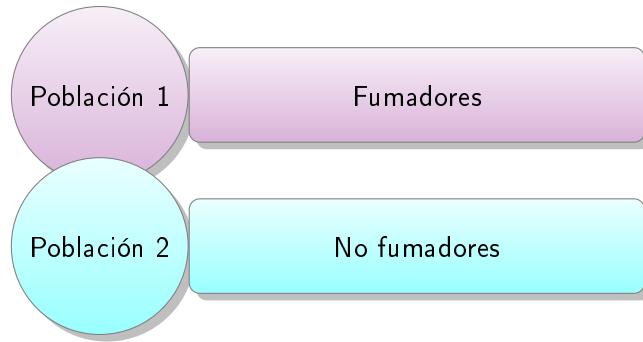
Destaquemos que en el test de homogeneidad los totales muestrales de cada una de las subpoblaciones se fijan *a priori*. En la Tabla 4.8 se muestra la cantidad de enfermos en cada una de las muestras.

	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 homogeneidad 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 4.7.

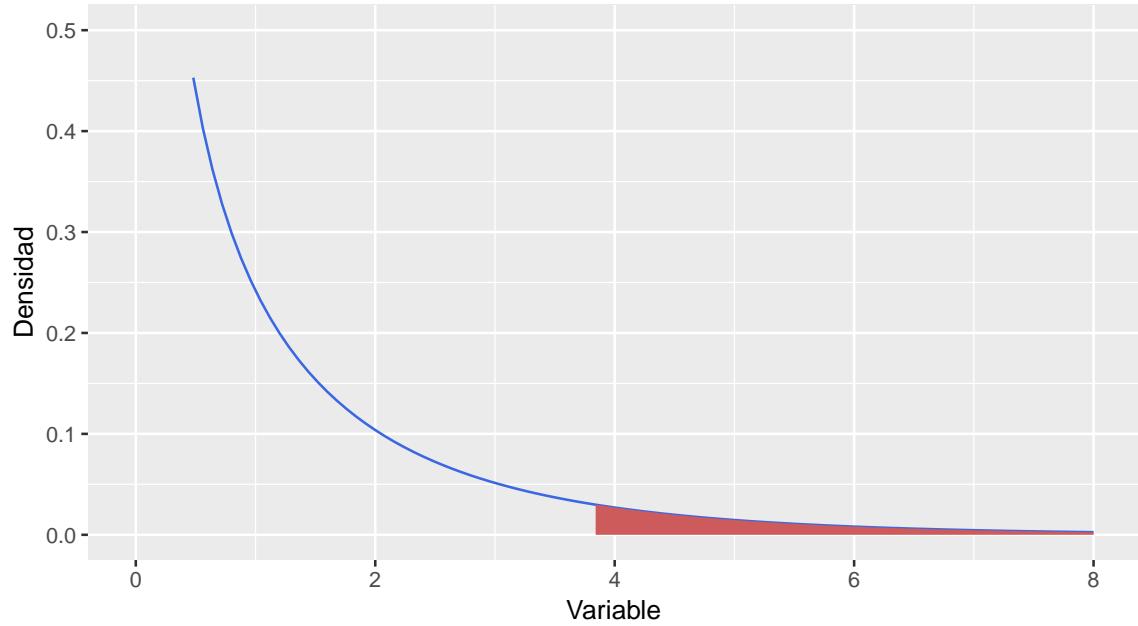


Figura 4.7: Distribución χ^2 y zona crítica

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
Totales	37	37	113	113	150

Tabla 4.9: Frecuencias observadas y esperadas

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 $P(\chi^2_1 > 25.88) << 0.0001$.

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 [2].

¿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

Ejemplo 4.7. En los Códigos 4.1 y 4.2 se muestra cómo aplicar el test Chi cuadrado a los Ejemplos 4.6 y 4.1, respectivamente.

```
mas.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$expected # Calcula las frecuencias esperadas

```

Código 4.1: Ejemplo de test de Chi cuadrado para Ejemplo 4.1

```

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 Chi cuadrado para Ejemplo 4.1

El comando `chisq.test` arroja como resultado lo siguiente:

```

Pearson's Chi-squared test with Yates' continuity correction
data: M
X-squared = 23.898, df = 1, p-value = 1.016e-06

Pearson's Chi-squared test
data: D
X-squared = 10.439, df = 2, p-value = 0.005411

```

Los valores esperados se calculan con el comando `Xsq$expected`, y se puede ver que, en ambos ejemplos, son mayores que 5 y por lo tanto es válida la distribución asintótica del estadístico, concluyendo que se rechaza la hipótesis de nulidad que sostiene la independencia entre las variables.



4.2.7 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 cuadrado. Estas condiciones exigen que los valores esperados de al menos el 80% de las celdas en una tabla de contingencia sean mayores que 5. Así, si consideramos una tabla de tamaño 2×2 , 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 ésta, una forma de plantear los resultados es su disposición en una tabla de contingencia de dos vías. Si las dos variables de consideración son dicotómicas, nos encontraremos con el caso de una tabla 2×2 como la que se muestra en la Tabla 4.11.

		Característica A		
		Presente	Ausente	Totales
Característica B	Presente	a	b	$a + b$
	Ausente	c	d	$c + d$
Totales		$a + c$	$b + d$	n

Tabla 4.11: Ejemplo de tabla de contingencia de 2×2

El test exacto de Fisher se basa en evaluar la probabilidad asociada a cada una de las tablas de tamaño 2×2 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 de 2×2 cuando se asume independencia y los totales de filas y columnas se consideran fijos, está dada por la distribución Hipergeométrica. Esto se obtiene calculando todas las posibles formas en las que podemos disponer n sujetos en una tabla de 2×2 de modo tal que los totales de filas y columnas sean siempre los mismos: $a + b$ y $c + d$ para las filas, $a + c$ y $b + d$ para las columnas. Simbólicamente,

$$p = \frac{C_{a+b,a} C_{c+d,c}}{C_{n,a+c}} = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}}$$

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 indica 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, se considera si $p < 0.05$, 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 p -valor asociado al test exacto de Fisher. En primer lugar, se puede 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 4.8. 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.



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

Se observa una muestra de 14 pacientes de este servicio y se los clasifica por sexo y según la aparición de síntomas de depresión. Las observaciones se encuentran en la Tabla 4.12.

		Síntomas de depresión		Totales
		Presente	Ausente	
Sexo	Mujer	1	4	5
	Hombre	7	2	9
Totales		8	6	14

Tabla 4.12: Depresión según sexo

Los valores observados son: $a = 1$, $b = 4$, $c = 7$ y $d = 2$. Mientras que los totales marginales son: $a + b = 5$, $c + d = 9$, $a + c = 8$ y $b + d = 6$.

En todo lo que sigue, nos referimos al Código 4.3.

La frecuencia esperada en tres de las cuatro celdas es menor que 5, por lo que no resulta adecuado aplicar el test de Chi cuadrado, por lo que 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:

$$p = \frac{C_{5,1}C_{9,7}}{C_{14,8}} = \frac{\binom{5}{1}\binom{9}{7}}{\binom{14}{8}} = \frac{5 \cdot 36}{3003} \approx 0.0599$$

En la Tabla 4.13 se muestran todas las posibles tablas que mantienen la frecuencias marginales de nuestro ejemplo, mientras que en la Tabla 4.14 se presentan las probabilidades asociadas a cada una de estas combinaciones.

Caso 1

Depresión			
	Sí	No	Totales
Mujer	0	5	5
Hombre	8	1	9
Totales	8	6	

Caso 2

Depresión			
	Sí	No	Totales
Mujer	5	0	5
Hombre	3	6	9
Totales	8	6	

Caso 3

Depresión			
	Sí	No	Totales
Mujer	1	4	5
Hombre	7	2	9
Totales	8	6	

Caso 4

Depresión			
	Sí	No	Totales
Mujer	4	1	5
Hombre	4	5	9
Totales	8	6	

Caso 5

Depresión			
	Sí	No	Totales
Mujer	2	3	5
Hombre	6	3	9
Totales	8	6	

Caso 6

Depresión			
	Sí	No	Totales
Mujer	3	2	5
Hombre	5	4	9
Totales	8	6	

Tabla 4.13: Combinaciones para Fisher

Sumando las probabilidades que no superan a la de la tabla observada 4.12, calculamos el *p*-valor

$$p\text{-valor} = 0.0030 + 0.0280 + 0.0599 = 0.0909$$

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 este test en R nos referimos al Código 4.3, cuya salida es:

Fisher's Exact Test for Count Data

Caso	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>p</i>
1	0	5	8	1	$\frac{\binom{5}{0} \binom{9}{8}}{\binom{14}{8}} \approx 0.0030$
2	5	0	3	6	$\frac{\binom{5}{5} \binom{9}{3}}{\binom{14}{8}} \approx 0.0280$
3	1	4	7	2	$\frac{\binom{5}{1} \binom{9}{7}}{\binom{14}{8}} \approx 0.0599$
4	4	1	4	5	$\frac{\binom{5}{4} \binom{9}{4}}{\binom{14}{8}} \approx 0.2098$
5	2	3	6	3	$\frac{\binom{5}{2} \binom{9}{6}}{\binom{14}{8}} \approx 0.2797$
6	3	2	5	4	$\frac{\binom{5}{3} \binom{9}{5}}{\binom{14}{8}} \approx 0.4196$

Tabla 4.14: Probabilidades asociadas a la Tabla 4.13

```

data: B

p-value = 0.09091

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.001283434 1.558054487

sample estimates:

odds ratio

0.09106548

```

```

B=as.table(rbind(c(1,4), c(7,2)))
# Guarda los datos
dimnames(B)=list(Sexo=c('Mujer','Hombre'), Síntomas=c('Presente','Ausente'))
# Establece las categorías de estudio

Xsq=chisq.test(B) # Realiza el test Chi cuadrado
Xsq$expected # Calcula las frecuencias esperadas

fisher.test(B) # Realiza el test de Fisher

```

Código 4.3: Ejemplo para el test de Fisher

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 lamparas 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

Nuestra generación ha entregado el alma a los contables y todas las pasiones que hoy nos conviven se derivan de las estadísticas: para saber si somos felices, ahora se hacen encuestas.

—Manuel Vincent

En este capítulo expondremos una técnica multivariante que permite representar conjuntamente las categorías de las filas y columnas de una tabla de contingencia.

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 [20]. Al referimos a una **técnica descriptiva o exploratoria**, estamos haciendo 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, se 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 estas nuevas coordenadas. Recordemos que 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 análisis de correspondencias simple centra su estudio en una tabla de contingencia de dos variables cualitativas, donde las categorías de una de las variables aparecen en las filas de la tabla, mientras que las de la otra variable en las columnas. El AC consiste entonces 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 ellos.

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:

✿ sabrosa ✿ seca ✿ fuerte ✿ empalagosa ✿ dulce ✿ amarga

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, empalagosa; mientras que a la derecha podemos encontrar las de seca, amarga, fuerte. Se vería también que las categorías-columna de gaseosa y jugo se hallan a la izquierda y las de champaña y de cerveza a la derecha. De esta manera se podrán establecer relaciones entre las categorías de las variables ubicadas en las filas y en las columnas.

La extensión del análisis de correspondencias simples al caso de varias variables categóricas, representadas en tablas de contingencia multidimensionales, se denomina **análisis de correspondencias múltiples**, y utiliza los mismos principios generales que la técnica antes descripta.

Dentro de los ejemplos de aplicación del análisis de correspondencias simple y múltiple podemos citar los siguientes

- ✿ 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.

	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

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 .

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 categoría de la variable fila X y la variable columna Y , obtenemos la frecuencia relativa 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 probabilidad 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

$$\hat{p}_{ij} = \frac{f_{ij}}{n}$$

siendo n el total de observaciones.

	Y_1	Y_2	\cdots	Y_j	\cdots	Y_k
X_1	\hat{p}_{11}	\hat{p}_{12}	\cdots	\hat{p}_{1j}	\cdots	\hat{p}_{1k}
X_2	\hat{p}_{21}	\hat{p}_{22}	\cdots	\hat{p}_{2j}	\cdots	\hat{p}_{2k}
\vdots	\vdots	\vdots		\vdots		\vdots
X_i	\hat{p}_{i1}	\hat{p}_{i2}	\cdots	\hat{p}_{ij}	\cdots	\hat{p}_{ik}
\vdots	\vdots	\vdots		\vdots		\vdots
X_r	\hat{p}_{r1}	\hat{p}_{r2}	\cdots	\hat{p}_{rj}	\cdots	\hat{p}_{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 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 el nivel sociocultural del paciente y sus resultados en el test de atención.



<https://flic.kr/p/6hyajx>

	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

En las Tablas 5.4, 5.5 y 5.6 se pueden apreciar, respectivamente, la distribución marginal de la atención, la distribución marginal del nivel de cultura y la distribución conjunta de ambas variables.

	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 cultural 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

¿Cómo se reflejaría la independencia de las variables en la representación de AC?

¿Cómo se podría observar este hecho en una tabla?

Para responder a estos planteos, recordemos primero el concepto de **probabilidad condicional**

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

siendo $P(B) > 0$. La probabilidad condicional de que ocurra A , sabiendo que ocurrió B , es el cociente entre la probabilidad conjunta de ambas y la probabilidad marginal de B . En términos coloquiales, 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

$$P(A/B) = P(A)$$

Lo cual se traduce en que $\frac{P(A \cap B)}{P(B)} = P(A)$, y por lo tanto

$$P(A \cap B) = P(A)P(B)$$

Es decir que cuando dos eventos son independientes, la probabilidad de ocurrencia conjunta es el producto de las probabilidades marginales de cada uno de ellos. La última expresión obtenida suele considerarse como la definición de independencia de dos eventos porque incluye el caso de que B sea imposible.

Este concepto puede extenderse para 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

$$\hat{e}_{ij} = n \times P(X = i) \times P(Y = j)$$

De lo que se desprende que deberíamos encontrar la manera de estimar $p_{i\cdot} = P(X = i)$ y $p_{\cdot j} = P(Y = j)$. 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

$$\hat{p}_{i\cdot} = \frac{f_{i\cdot}}{n} \quad \text{y} \quad \hat{p}_{\cdot j} = \frac{f_{\cdot j}}{n}$$

De lo cual se infiere que

$$\hat{e}_{ij} = n \times \hat{p}_{i\cdot} \times \hat{p}_{\cdot j}$$

o equivalentemente,

$$\hat{e}_{ij} = n \times \frac{f_{i\cdot}}{n} \times \frac{f_{\cdot j}}{n} = \frac{f_{i\cdot} f_{\cdot j}}{n}$$

Ejemplo 5.2. Calculemos para el Ejemplo 5.1, las frecuencias de las casillas esperadas bajo independencia (ver Tabla 5.7) y comparemos estas frecuencias con las observadas para ver si las variables X e Y podrían o no ser independientes.

	A	B	C	D	E	F	Totales
Atento	50.27	47.01	55.07	73.68	50.85	30.12	307
Síntomas leves	95.30	89.12	104.40	139.68	96.39	57.11	582
Síntomas moderados	56.00	52.37	61.35	82.08	56.64	33.56	342
Disperso	60.42	56.50	66.19	88.56	61.12	36.21	369
Totales	262	245	287	384	265	157	1600

Tabla 5.7: Frecuencias esperadas bajo el supuesto de independencia

Se puede apreciar que en algunas casillas los valores esperados bajo independencia son muy diferentes de los observado, citamos por ejemplo:

- ✿ Atento y F
- ✿ Atento y A
- ✿ Disperso y A
- ✿ Disperso y F

Para analizar la independencia vamos a 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 esta discrepancia 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 condicional de una de ellas se repetiría para cada nivel de la otra variable y coincidiría con el perfil medio, que es la distribución marginal. A partir de la distribución marginal de los niveles culturales, que se muestra en la Tabla 5.5, se puede construir el perfil medio cultural.

Investiguemos 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 la Tabla 5.8 se muestran las probabilidades condicionales estimadas de los niveles de atención dado el nivel superior de atención.

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’

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 del nivel de 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 Perfiles medios

El concepto de **perfil** dado por un conjunto de frecuencias relativas, es fundamental para el análisis de correspondencias. Estos conjuntos de frecuencias relativas, también llamados vectores, tienen características geométricas especiales debido a que la suma de sus elementos es igual a 1 (lo que representa el 100%).

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 pueden representarse como puntos en un espacio de perfíles. En la Figura 5.2 se muestra el aspecto tienen los perfíles para el Ejemplo 5.1.

El cruce entre las líneas que representan los diferentes perfíles indica que no existe independencia total entre las variables, o bien que la distribución del nivel cultural no es la misma en todos los

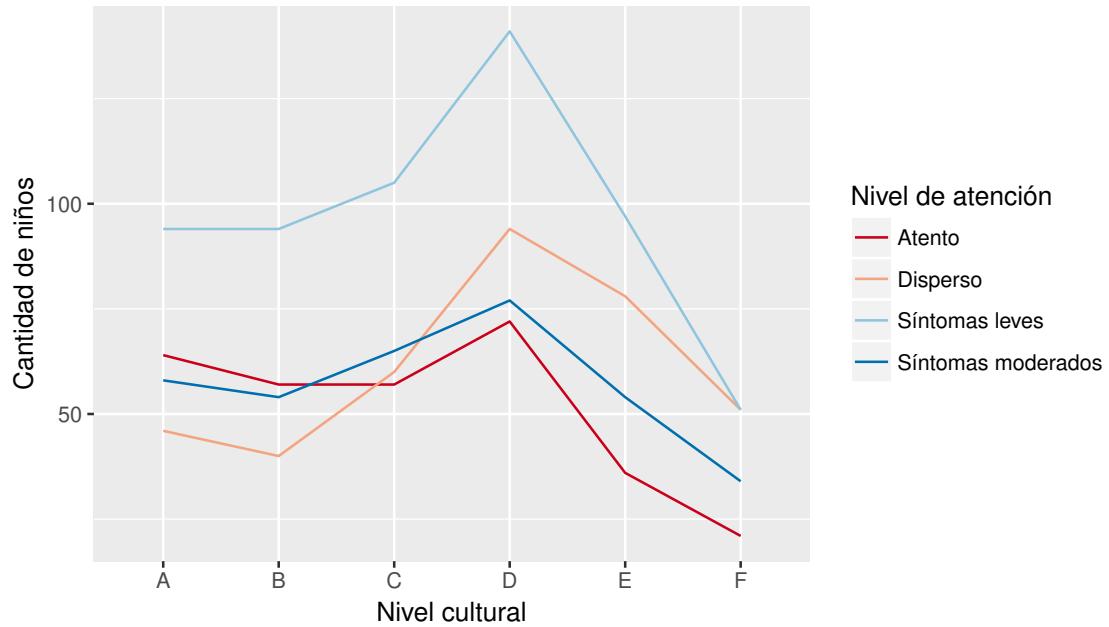


Figura 5.2: Perfiles de nivel cultural según atención

niveles de atención. Es decir, 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 las mismas resultan ser un 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. Para responder a esta pregunta, calcularemos una medida para 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 estadístico χ^2 de Pearson, definido como

$$\sum_{i=1}^r \sum_{j=1}^k \frac{(f_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

5.2 Inercia total

Vamos a definir un nuevo concepto, llamado **inercia total**, como el cociente entre el estadístico Chi cuadrado y el tamaño muestral.

Una manera de llegar a una tabla de contingencia de $r \times k$ columnas, es definir r variables binarias para las filas y k variables binarias 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.

Ejemplo 5.3. 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.

Individuo	Nivel de atención				Disperso	Nivel cultural					
	Atento	Sínt. leves	Sínt. moderados			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*)

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.
- ✿ 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 $F = M_f^t M_c$, obtendremos una matriz de dimensión $r \times k$ que, en el ejemplo anterior, se corresponde con la tabla de contingencia de tamaño de 4×6 . En esta matriz aparece en cada posición ij la frecuencia absoluta observada para cada combinación de características. Observar que si F_r es la matriz de frecuencias relativas; es decir, el cociente entre las frecuencias absolutas observadas y el total de observaciones, entonces $F_r = \frac{1}{n}F$.

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.

Ejemplo 5.4. 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

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

la Tabla 5.11. La matriz F tiene las frecuencias absolutas y la matriz F_r , de la Tabla 5.12, las frecuencias relativas.



En lo que sigue vamos a estudiar cómo representar las filas, luego el razonamiento se realiza 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.

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 relativa

$$f_{i\cdot} = \sum_{j=1}^r f_{ij}$$

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 5.5. 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 5.1, 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 $f_T = f_i$, las frecuencias relativas de los totales de las filas.

Ejemplo 5.6. 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
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
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 deben sumar 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

$$D^2(R_{a\cdot}, R_{b\cdot}) = \sum_{j=1}^r \frac{1}{f_{\cdot j}} \left(\frac{f_{aj}}{f_{a\cdot}} - \frac{f_{bj}}{f_{b\cdot}} \right)^2$$

siendo f_{aj} (resp. f_{bj}) el elemento de la fila a (resp. b) en la columna j , $f_{a\cdot}$ 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 5.7. Los cálculos anteriores en el caso del 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, 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 5.8. Para nuestro ejemplo 5.1, 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 $Z^t Z$. Entonces las coordenadas de las filas de la representación vienen dadas por

$$C_f = YW_2 = D_f^{-1} F D_c^{-1/2} W_2$$

donde $W_2 = (\vec{w}_1 \ \vec{w}_2)$ la matriz formada por los dos primeros autovectores de la matriz $Z^t Z$.

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 $Z^t Z$.

Análogamente, podemos aplicar a las columnas un análisis similar al de las filas, resultando la mejor representación de las columnas:

$$C_c = YU_2 = D_c^{-1} F^t D_f^{-1/2} U_2$$

donde $U_2 = (\vec{u}_1 \ \vec{u}_2)$ la matriz formada por los dos primeros autovectores de la matriz $Z Z^t$.

5.3 *Biplot* simétrico

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
```

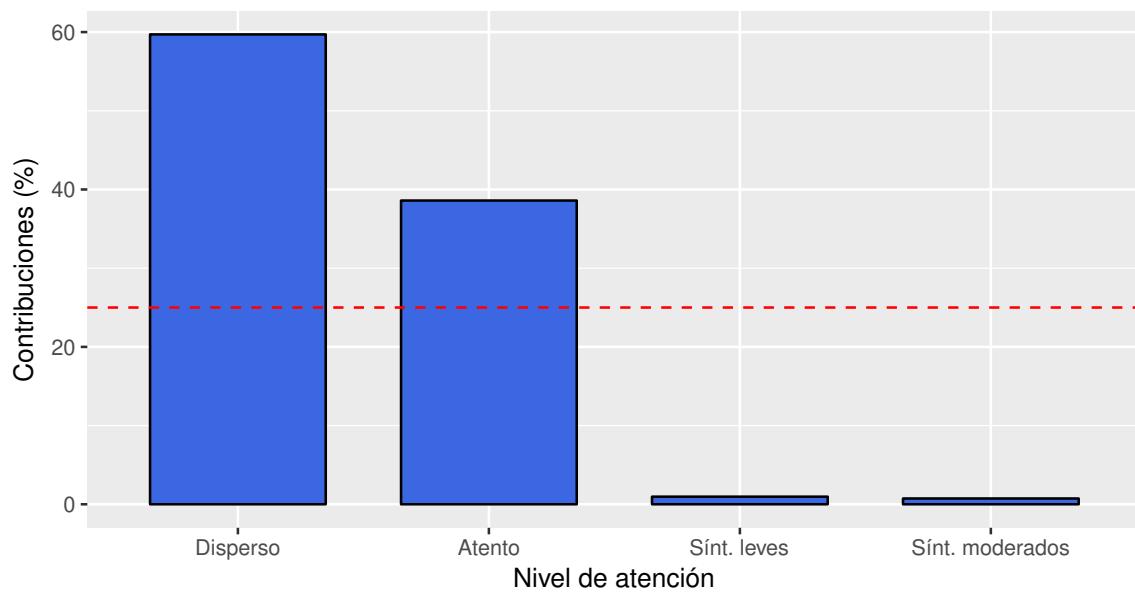


Figura 5.3: Contribución de filas a la dimensión 1

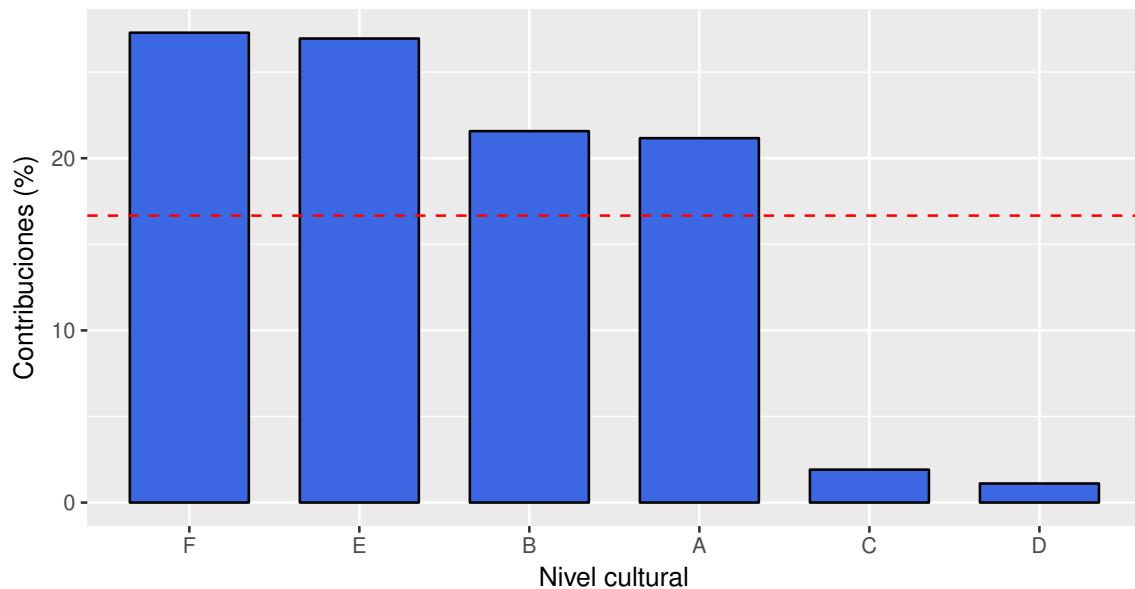


Figura 5.4: Contribución de columnas a la dimensión 1

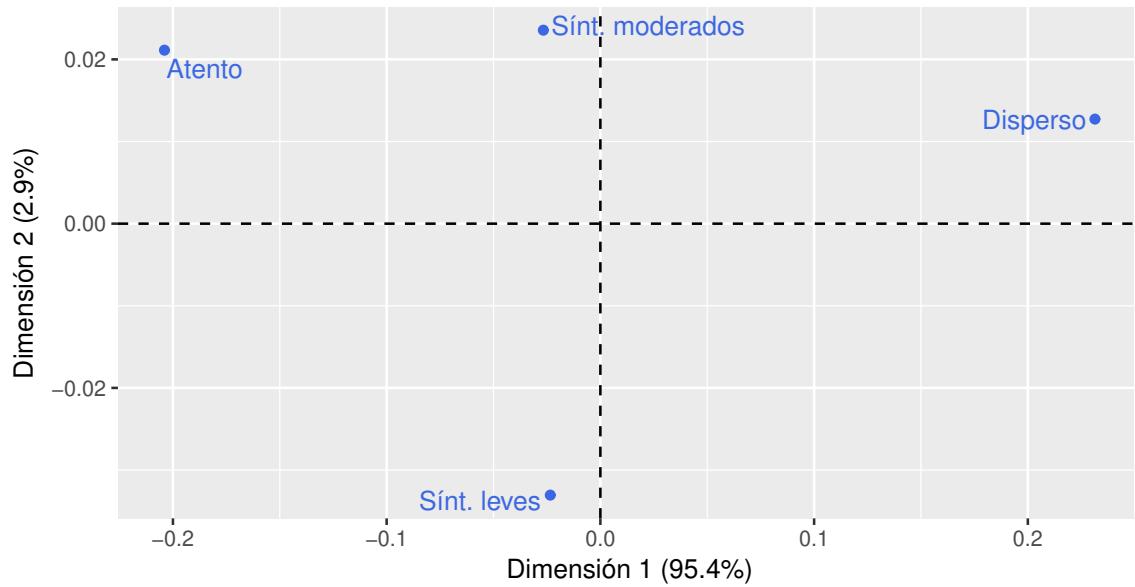


Figura 5.5: Puntos fila - AC

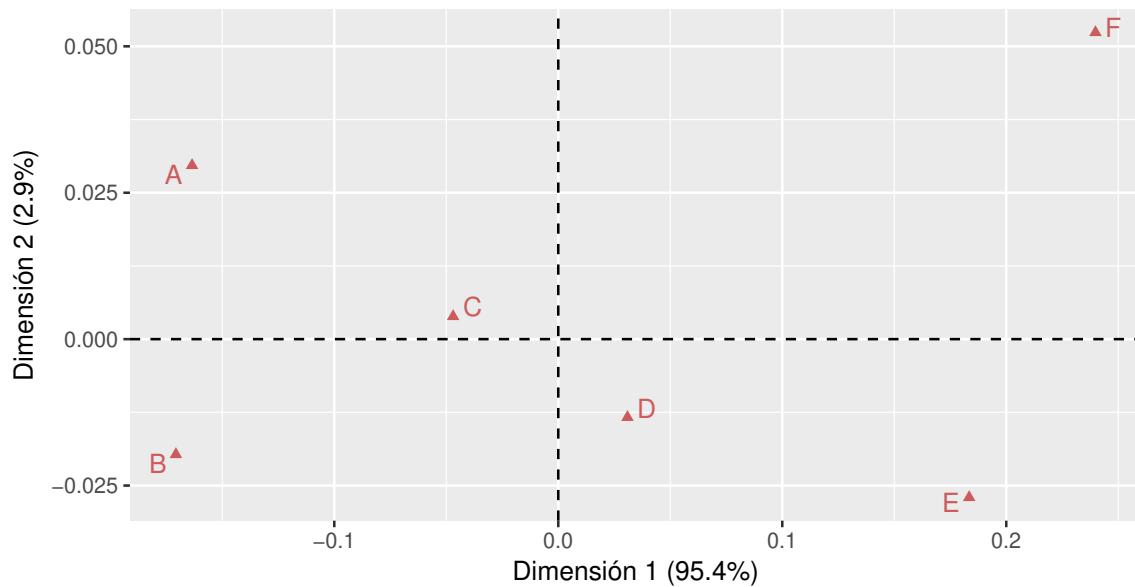


Figura 5.6: Puntos columna - AC

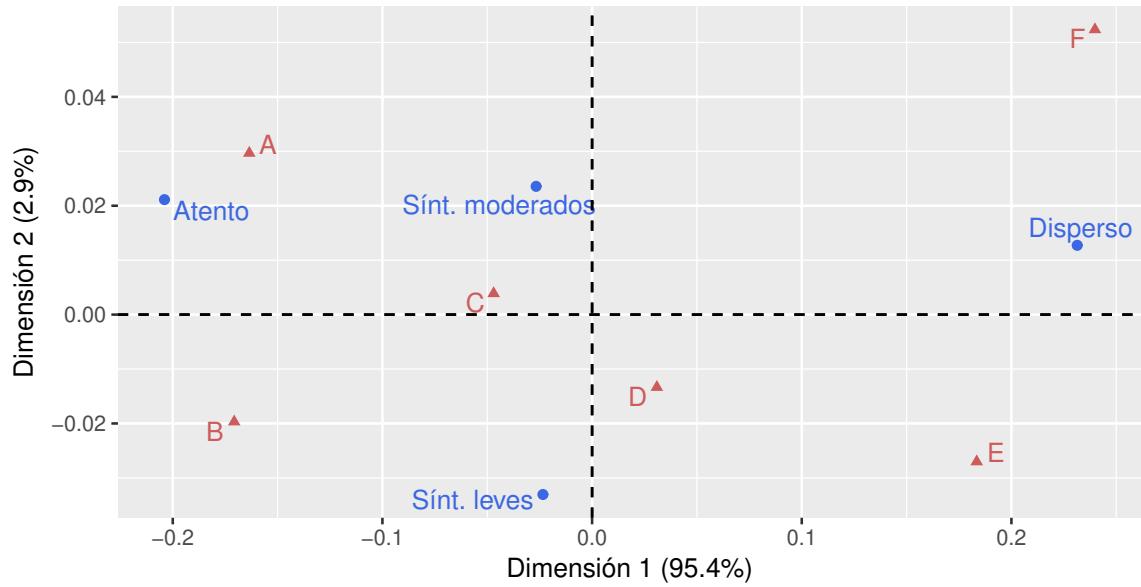


Figura 5.7: Biplot simétrico - AC

```
# 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

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: AC para niveles cultural y de atención

5.3.1 Guía para la interpretación gráfica del *biplot* simétrico

Listamos a continuación algunas consideraciones a tener en cuenta al momento de interpretar un *biplot* simétrico.

- ✿ Las columnas (respectivamente, filas) cercanas al origen reflejan categorías similares a la columna (respectivamente, fila) promedio.

- ✿ Las columnas (respectivamente, filas) cercanas entre sí reflejan categorías de similar perfil en términos de columnas (respectivamente, filas).
- ✿ Las columnas (respectivamente, 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 lo anterior, realizamos las siguientes observaciones para el *biplot* simétrico 5.7:

- ✿ Los niveles de atención más usuales son los que presentan síntomas leves y moderados.
- ✿ El nivel cultural *A* es el más próximo a la categoría de ‘atento’.
- ✿ Los niveles de cultura *C* y *D* son los más usuales.
- ✿ Los síntomas leves en grado de atención se asocian con la categoría de cultura *D*.

En general, el programa R guarda la siguiente información en el ‘objeto’:

- ✿ **nd**: dimensión de la solución
- ✿ **rownames** (respectivamente, **colnames**): nombres de las filas (respectivamente, de las columnas)
- ✿ **rowinertia** (respectivamente, **colinertia**): cantidad de inercia de cada fila (respectivamente, de cada columna)
- ✿ **rowdist** (respectivamente, **coldist**): distancia Chi cuadrado de las filas (respectivamente, de las columnas) al perfil medio de fila (respectivamente, de columna)
- ✿ **rowcoord** (respectivamente, **colcoord**): coordenadas para representar las categorías de filas (respectivamente, 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.

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 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 medio**.

	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	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

Ejemplo 5.9. En la Tabla 5.16 se muestran los datos que la Secretaría de Ciencia y Técnica de la Facultad de Ciencias Económicas registró de las universidades con las que se realizaron viajes de intercambio en el contexto del Proyecto Redes durante los últimos meses del año 2018. 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
Totales	39	93	126	258

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 las 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.

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

Con el Código 5.5 se genera la Figura 5.8.

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

# 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") +
```

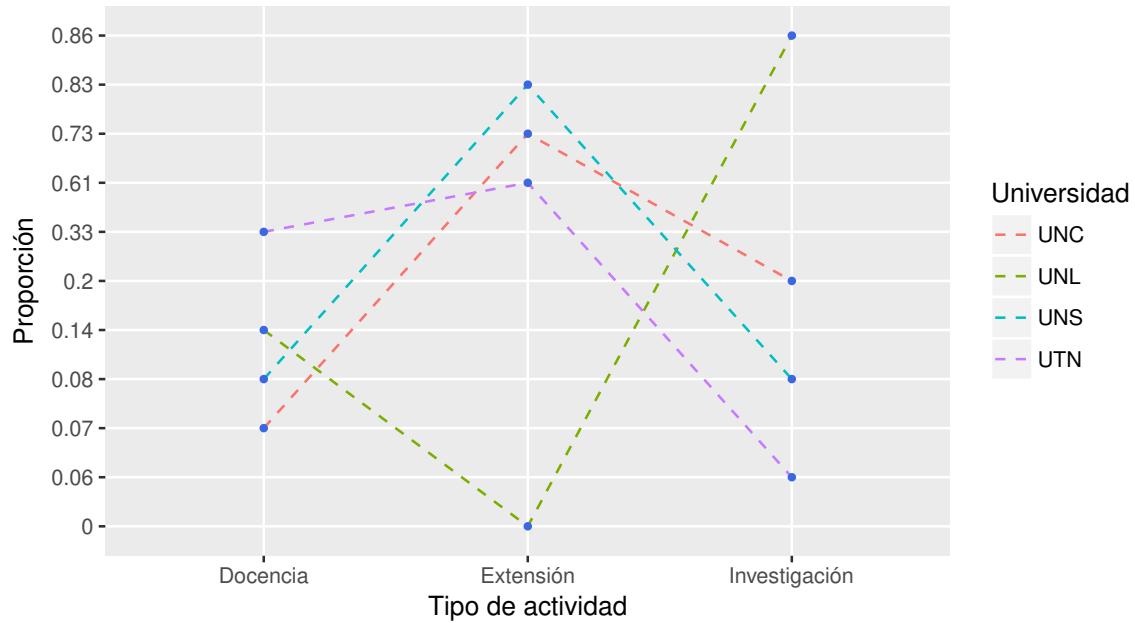


Figura 5.8: Perfiles fila de las actividades universitarias

```
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:

$$(\text{Docencia} \quad \text{Investigación} \quad \text{Extensión}) = \frac{1}{258} (39 \quad 93 \quad 126) = (0.1512 \quad 0.3605 \quad 0.4884)$$

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.

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
Totales	1	1	1

Tabla 5.18: Perfiles columna de los viajes de intercambio

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 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.



5.3.2 Otra representación gráfica

En esta sección vamos a mostrar una manera completamente distinta de representación basándonos en el Ejemplo 5.9. 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

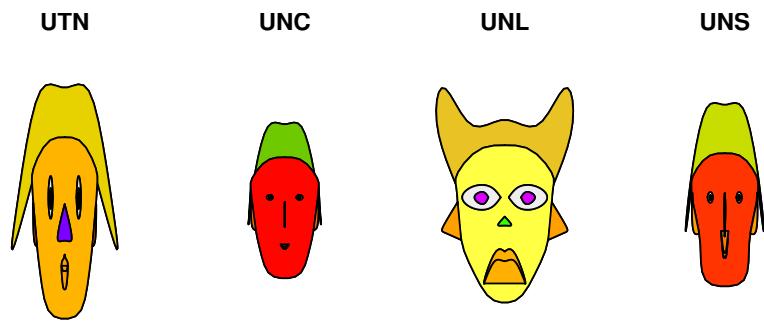


Figura 5.9: Caras de Chernoff universidades

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")
# 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,
```

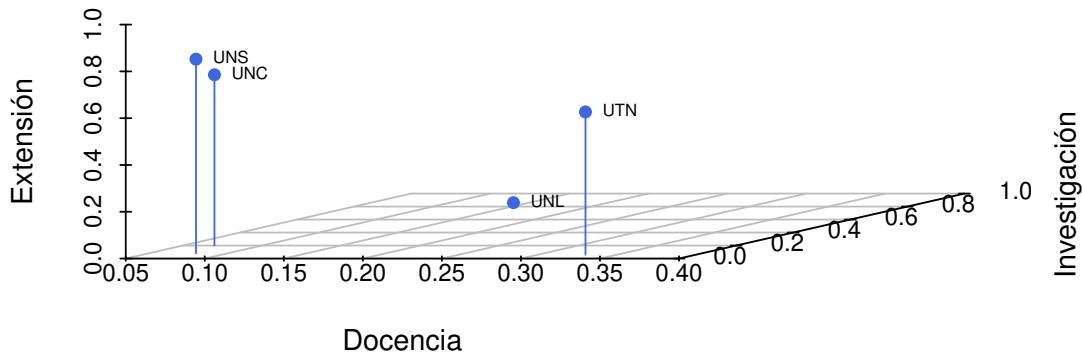


Figura 5.10: Representación en 3D de las actividades universitarias

```
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 diagrama de dispersión 3D (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 cuatro universidades 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.

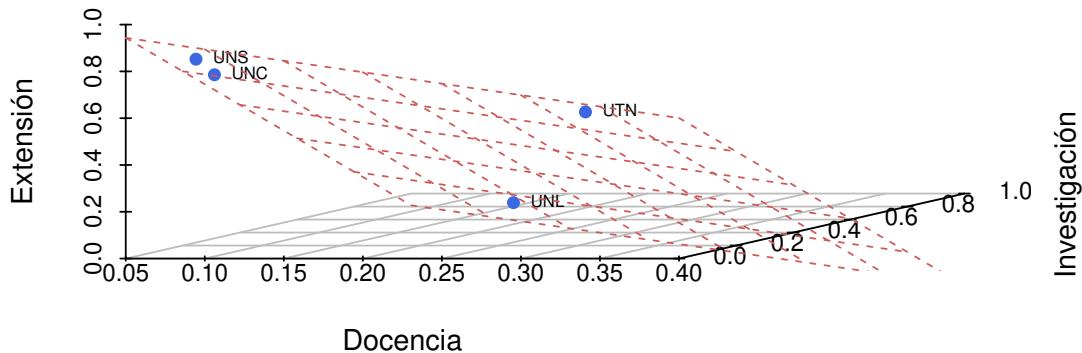


Figura 5.11: Plano de representación de las actividades universitarias

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

# 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%)') +
ggtitle('')
# Realiza el biplot simétrico

```

Código 5.7: Código para AC (actividades universitarias)

Presentamos a continuación un nuevo ejemplo.

Ejemplo 5.10. 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.

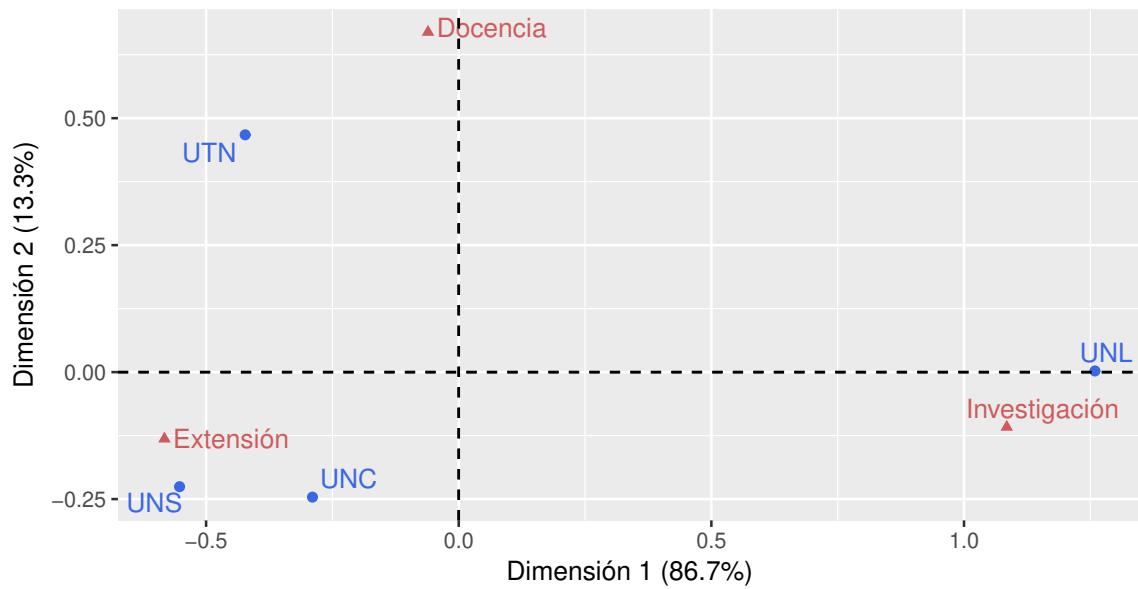


Figura 5.12: Biplot simétrico (actividades universitarias)

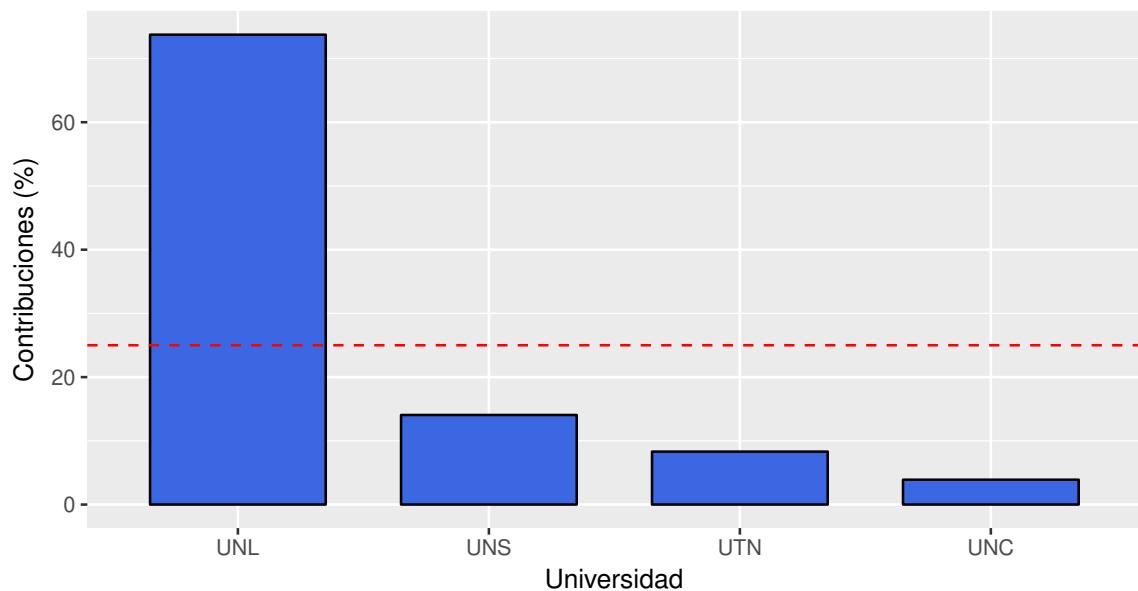


Figura 5.13: Contribución de las filas (actividades universitarias)

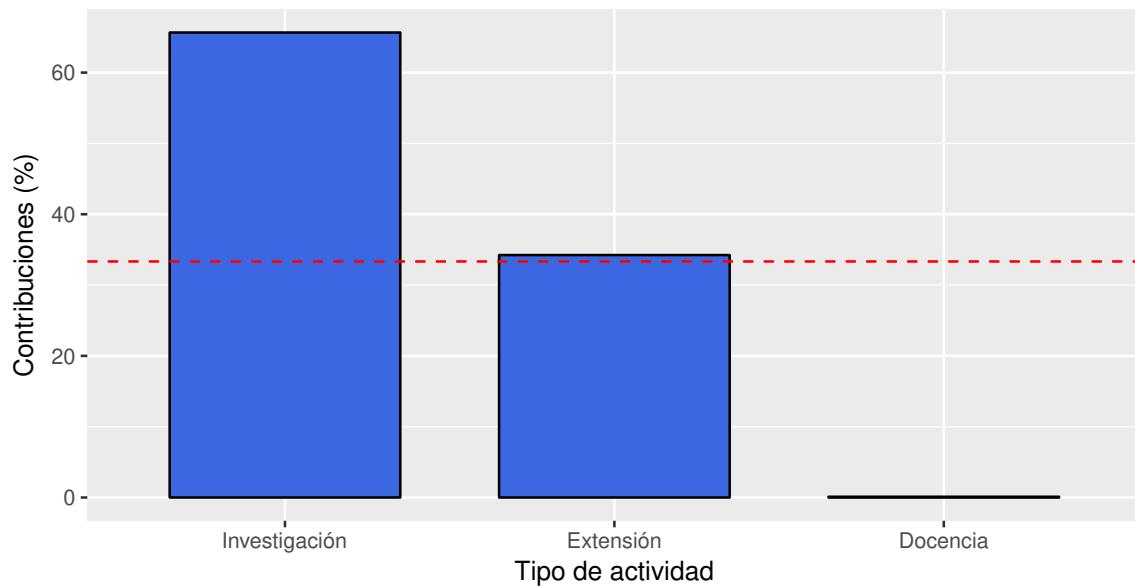


Figura 5.14: Contribución de columnas (actividades universitarias)



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

Los datos obtenidos se presentan en la Tabla 5.19.

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'

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
Totales	109	38	147

Tabla 5.19: Distribución de depresión según 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
Totales	109	38	147

Tabla 5.20: Frecuencias esperadas bajo independencia

Se debe comparar con el percentil 95 de la distribución χ^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

$$r_{ij} = \frac{f_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij} \left(1 - \frac{n_{i\cdot}}{n}\right) \left(1 - \frac{n_{\cdot j}}{n}\right)}}$$

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 5.10 se muestran en la Tabla 5.21.

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.

Frecuencia en la práctica deportiva	Ausencia de depresión	Presencia de depresión
Sin práctica	-3.256	3.256
Hasta tres veces por semana	0.967	-0.967
Más de tres veces por semana	2.393	-2.393

Tabla 5.21: Residuos correspondientes a depresión vs práctica deportiva

5.4 Estadístico de Pearson e 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 ZZ^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 χ^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 5.10, que el estadístico de Pearson χ^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 χ^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

$$\begin{aligned} \frac{\chi^2_{obs}}{147} &= \frac{53}{147} \left[\frac{\left(\frac{31}{53} - \frac{39.3}{53} \right)^2}{\frac{39.3}{53}} + \frac{\left(\frac{22}{53} - \frac{13.7}{53} \right)^2}{\frac{13.7}{53}} \right] + \frac{48}{147} \left[\frac{\left(\frac{38}{48} - \frac{35.6}{48} \right)^2}{\frac{35.6}{48}} + \frac{\left(\frac{10}{48} - \frac{12.4}{48} \right)^2}{\frac{12.4}{48}} \right] + \\ &\quad \frac{46}{147} \left[\frac{\left(\frac{40}{46} - \frac{34.1}{46} \right)^2}{\frac{34.1}{46}} + \frac{\left(\frac{6}{46} - \frac{11.9}{46} \right)^2}{\frac{11.9}{46}} \right] \end{aligned}$$

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 ϕ** , por lo cual se puede decir que la inercia es igual a ϕ^2 .

5.4.1 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

$$\text{Inercia } 1 < \text{Inercia } 2 < \text{Inercia } 3 < \text{Inercia } 4$$

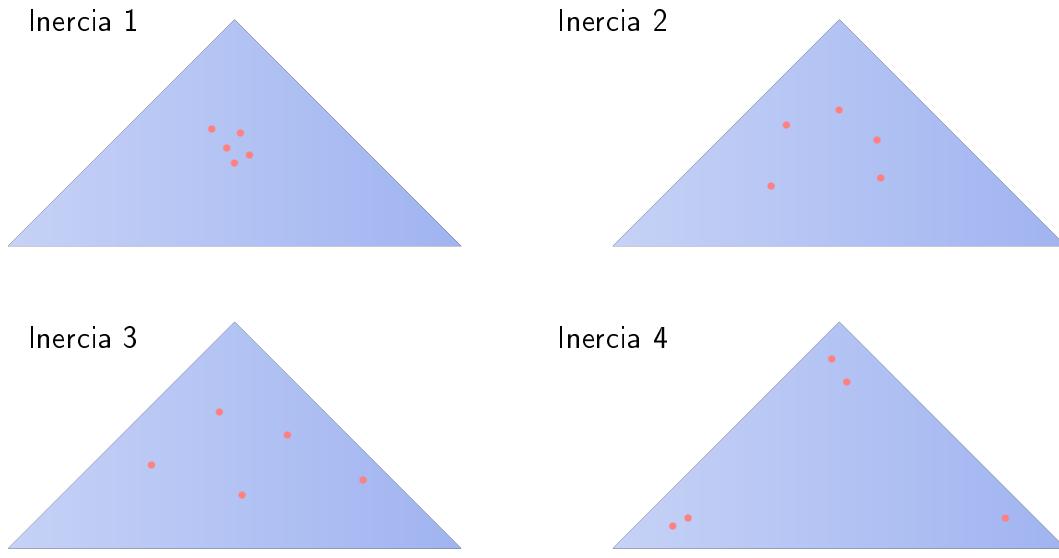


Figura 5.16: Ejemplos de inercias

¿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?

Ejemplo 5.11. En la Tabla 5.22 se puede visualizar lo que muestra la Figura 5.16, donde el total de observaciones se mantiene constante y se van polarizando las distribuciones. █

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 refleje las relaciones existentes entre dichas filas y columnas. En dicha representación, juegan un papel importante las llamadas *distancias* χ^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:

- * $d_{ij} = \sum_{h=1}^r \frac{1}{n_{.h}} \left(\frac{n_{ih}}{n_{i.}} - \frac{n_{jh}}{n_{j.}} \right)^2$ para los perfiles fila
- * $d_{ij} = \sum_{h=1}^k \frac{1}{n_{h.}} \left(\frac{n_{hi}}{n_{.i}} - \frac{n_{hj}}{n_{.j}} \right)^2$ para los perfiles columna

Tabla 1

	B1	B2	B3
A1	10	12	11
A2	12	10	13
A3	15	14	13

inercia1 = 0.0064

Tabla 2

	B1	B2	B3
A1	4	12	11
A2	12	10	13
A3	15	14	19

inercia2 = 0.0354

Tabla 3

	B1	B2	B3
A1	5	16	13
A2	12	5	13
A3	15	14	19

inercia3 = 0.0838

Tabla 4

	B1	B2	B3
A1	5	16	20
A2	2	5	19
A3	25	14	11

inercia4 = 0.2678

Tabla 5.22: Inercia creciente - Asociación creciente

donde usamos en forma indistinta, según el caso, n_{ij} para indicar O_{ij} o f_{ij} .

Ejemplo 5.12. En las Tablas 5.23 y 5.24 se calculan estas distancias para el Ejemplo 5.10.

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 5.23: Perfiles fila

Para la Tabla 5.23, calculemos a modo de ejemplo las distancias χ^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 5.24, elegimos calcular a modo de ejemplo la distancia χ^2 entre los siguientes perfiles columna:

$$d_{12} = \frac{1}{53}(0.2844 - 0.5789)^2 + \frac{1}{48}(0.3486 - 0.2632)^2 + \frac{1}{46}(0.3670 - 0.1579)^2 = 0.002739$$

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 5.24: Perfiles columna

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 χ^2 son invariantes a variaciones en la codificación de las categorías con comportamiento similar en cuanto a sus perfiles condicionales.



5.5 Principio de equivalencia distribucional

Veamos ahora que, efectivamente, la distancia Chi cuadrado satisface el principio de equivalencia distribucional. Para ello consideremos la tabla de contingencia 5.25 en la cual las dos primeras filas son proporcionales, por lo cual, los perfiles fila son iguales.

	B1	B2	B3	Totales
A1	10	12	20	42
A2	5	6	10	21
A3	8	4	5	17
Totales	23	22	35	80

Tabla 5.25: Primer ejemplo de equivalencia distribucional

Calculamos la distancia entre los perfiles fila aplicando el Código 5.8, y obtenemos que $d_{12} = 0$, $d_{13} = 2.13$ y $d_{23} = 2.13$.

Luego, colapsamos las dos primeras filas y tenemos una nueva tabla 5.26. Volvemos a calcular la distancia entre la fila colapsada y la tercera fila que en este caso pasa a ser la segunda fila (referimos nuevamente al Código 5.8).

Ahora sólo hay dos filas y la distancia entre ellas es: $D_{12} = 2.13$. Observar que la distancia entre la fila colapsada es igual a las distancias de las dos filas originales con la tercera fila de la tabla.

	B1	B2	B3	Totales
AI	15	18	30	63
AII	8	4	5	17
Totales	23	22	35	80

Tabla 5.26: Segundo ejemplo de equivalencia distribucional

```
# Guardamos los datos
A1=c(10,12,20)
A2=c(5,6,10)
A3=c(8,4,5)

# Calculamos los perfiles-fila
perf1=A1/sum(A1)
perf2=A2/sum(A2)
perf3=A3/sum(A3)

totcol=A1+A2+A3 # guarda los totales por columna

# Calculamos las distancias Chi cuadrado entre perfiles fila
d12=sum((perf1-perf2)*totcol)
d13=sum((perf1-perf3)*totcol)
d23=sum((perf2-perf3)*totcol)

# Colapsamos las dos primeras filas
AI=A1+A2
AII=A3
perfI=AI/sum(AI)
perfII=AII/sum(AII)
D12=sum((perfI-perfII)*totcol)
```

Código 5.8: Código para el análisis de equivalencia distribucional

5.6 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 5.13. Consideremos un conjunto de 12 personas en las que hemos observado cuatro variables. En la Tabla 5.27 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 5.27: Características observadas

Otra forma de presentar estos datos podría ser la que se muestra en la Tabla 5.28. Este tipo de matriz se denomina **matriz disyuntiva** y la vamos a designar con G .

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 5.28: 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 5.29: Matriz G^t

La matriz traspuesta, será G^t y se muestra en la Tabla 5.29.



5.6.1 Matriz de Burt

Listaremos una serie de preguntas que intentaremos responder observando la matriz 5.29.

- ✿ ¿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 5.13, 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 5.30, 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 5.30, 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 5.30 $B_{25} = B_{52} = 3$ 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 5.30: Matriz de Burt para el Ejemplo 5.13

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 5.9 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() +
```

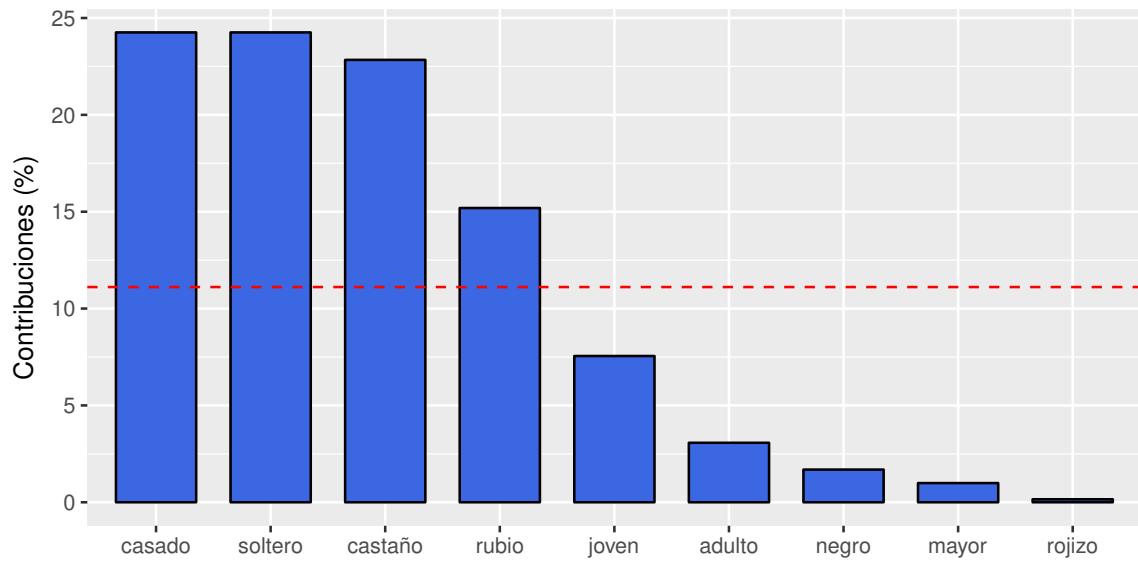


Figura 5.18: Contribución de variables a la inercia (dimensión 1)

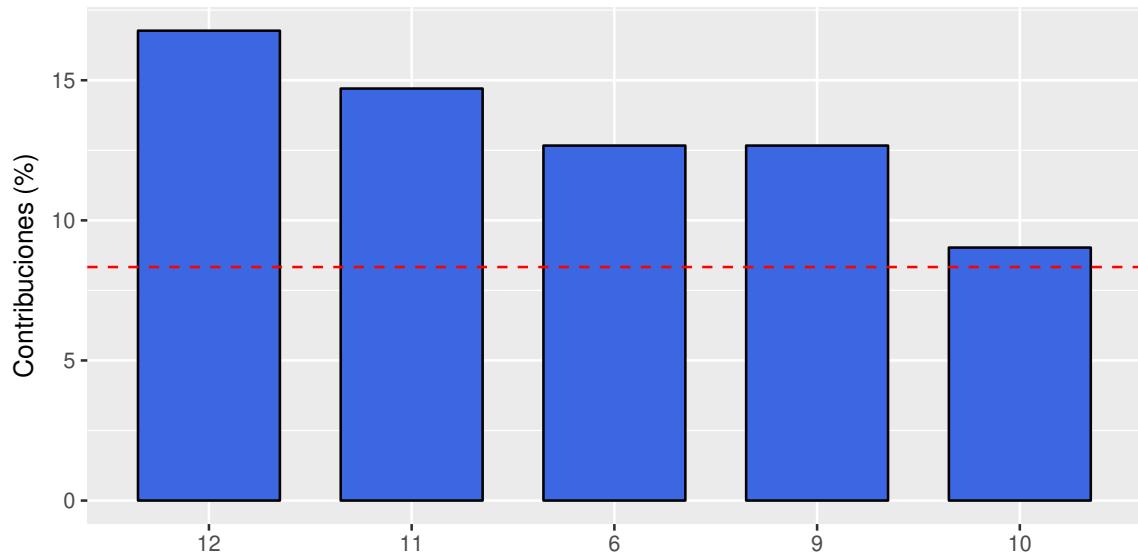


Figura 5.19: Contribución de individuos a la inercia (dimensión 1)

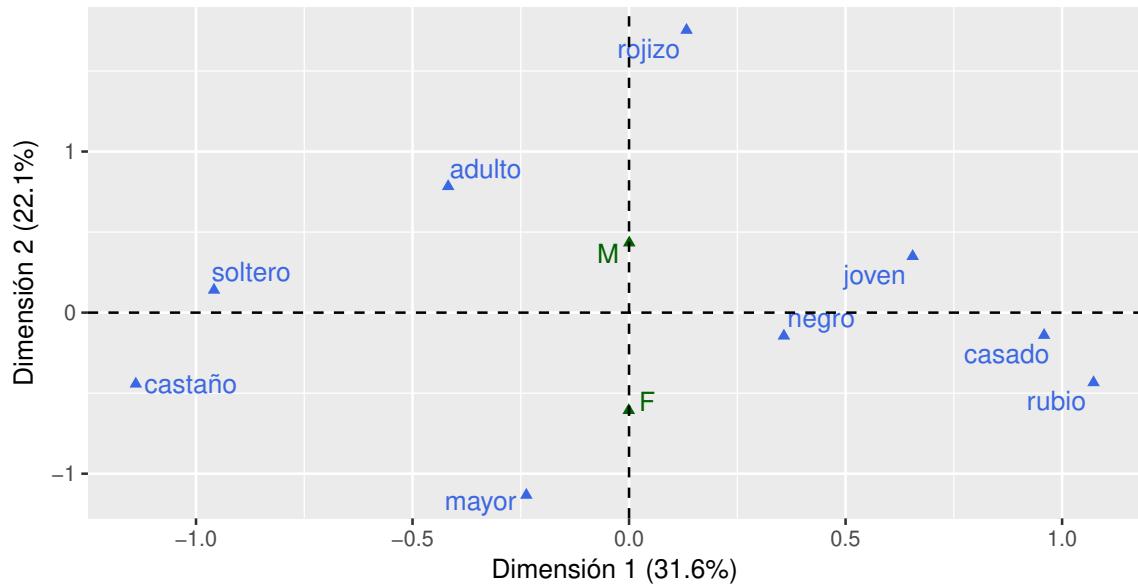


Figura 5.20: Categorías variables - ACM

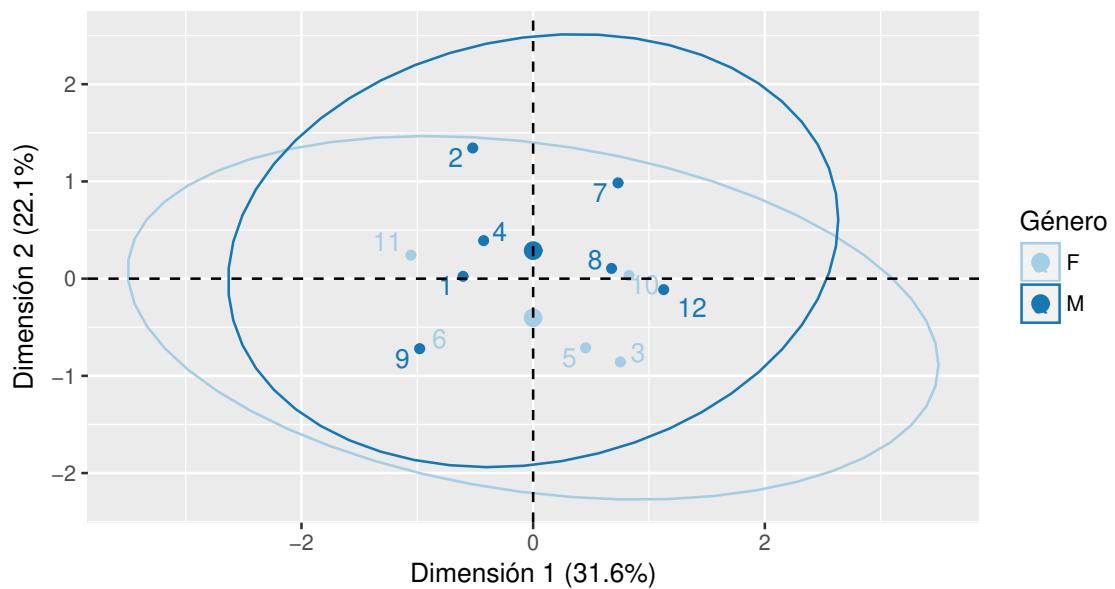


Figura 5.21: Individuos agrupados por género - ACM

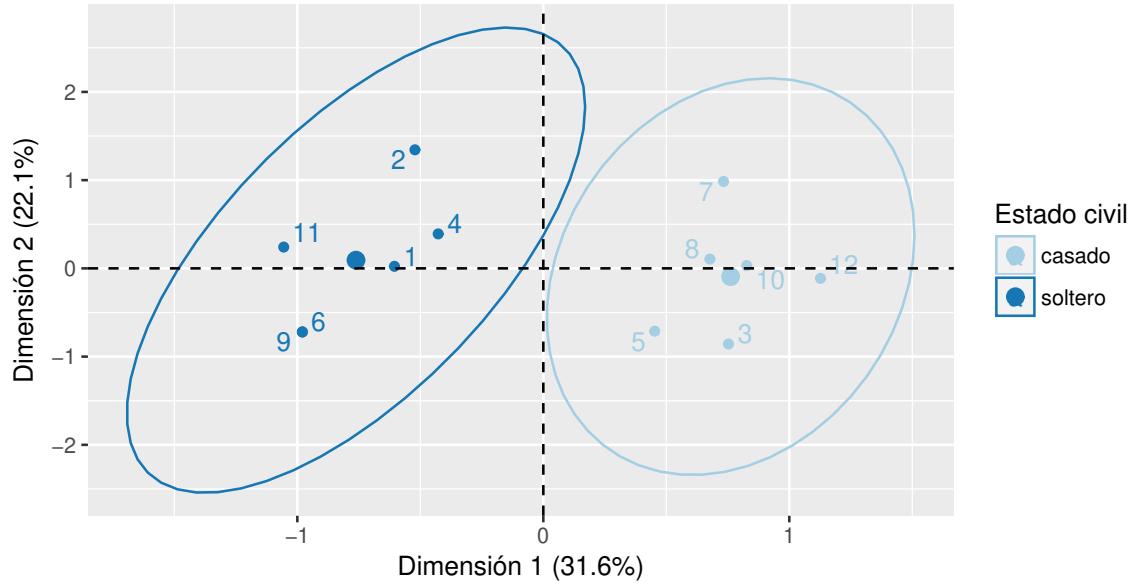


Figura 5.22: Individuos agrupados por estado civil - ACM

```

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('') +
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$Sexo),
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 5.9: Código para ACM del 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 5.14. 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 5.31 se encuentran las observaciones realizadas sobre 10 individuos de la empresa, en función del género, los años en la empresa y los ingresos mensuales.

A partir de la tabla original, se construye la tabla disyuntiva o matriz G que se presenta en la Tabla 5.32.



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.

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 5.31: 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 5.32: Matriz disyuntiva para la situación de los empleados

- ✿ 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.

5.6.2 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.

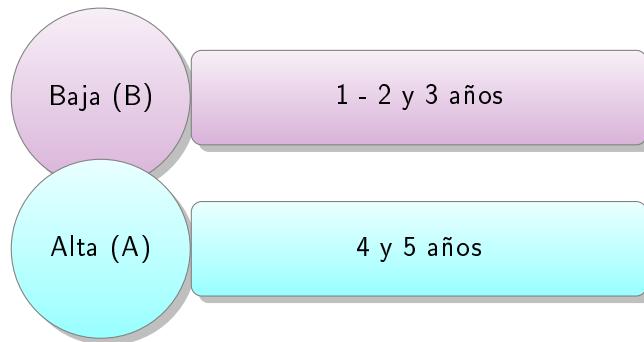
- ✿ En el análisis de correspondencias múltiples, los valores propios generan una idea pesimista de la variabilidad explicada.

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 5.15. Siguiendo el Ejemplo 5.14, definimos dos categorías para la variable dada por la antigüedad de cada empleado en la empresa:



El análisis de correspondencias múltiple se realiza mediante el Código 5.10 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 5.34, 5.35, 5.36 y 5.37.

```
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
```

Individuo	Género	Antigüedad	Ingresos	Categoría
1	Mujer	5	Medio	A
2	Mujer	3	Alto	B
3	Hombre	4	Bajo	A
4	Mujer	1	Bajo	B
5	Mujer	2	Medio	B
6	Hombre	5	Alto	A
7	Mujer	2	Medio	B
8	Hombre	3	Bajo	B
9	Hombre	1	Alto	B
10	Mujer	4	Medio	A

Tabla 5.33: Agregado de categoría para los empleados

```

Género=factor(empresas$Género)
Antigüedad=factor(empresas$Antigüedad)
Ingresos=factor(empresas$Ingresos)
Categoría=factor(empresas$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 5.10: Código para el análisis de correspondencias múltiples para una empresa

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	1
10	0	1	0	0	1	0

Tabla 5.34: Matriz disyuntiva para la empresa



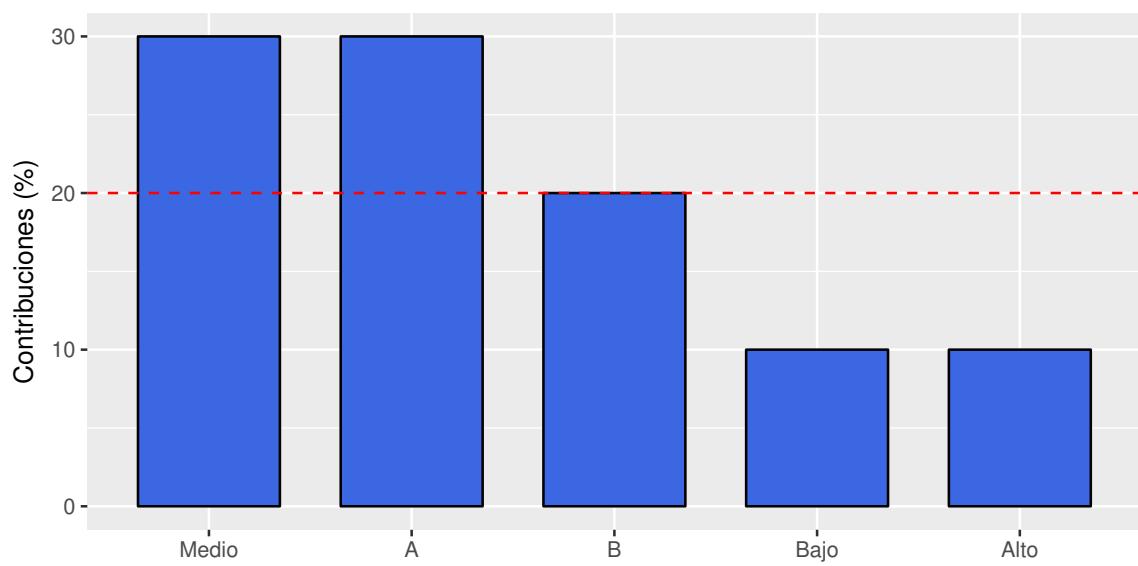


Figura 5.24: Contribución a la inercia de las variables

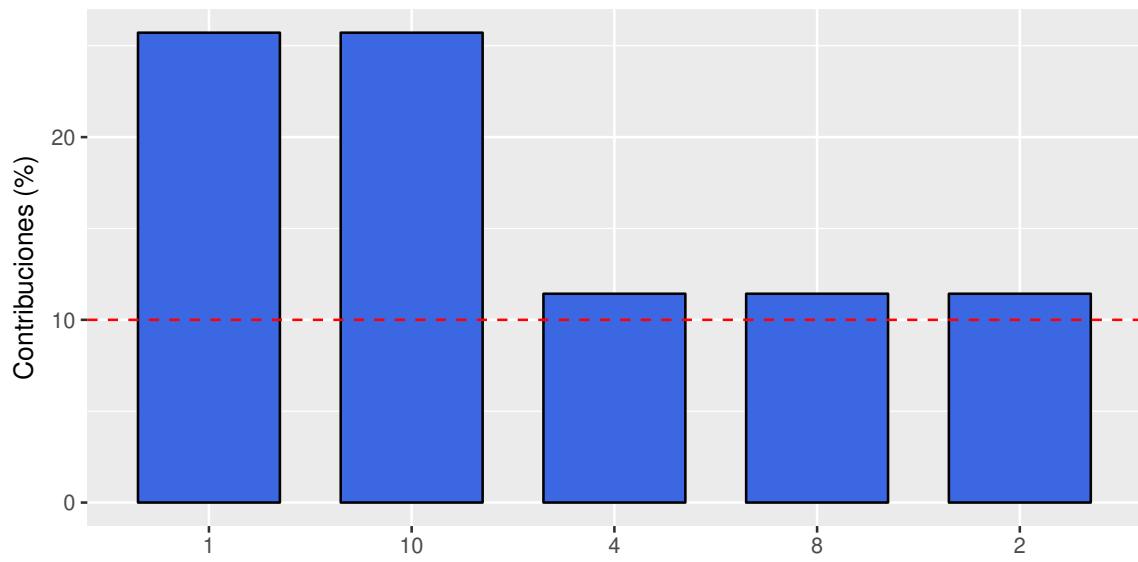


Figura 5.25: Contribución a la inercia de los individuos

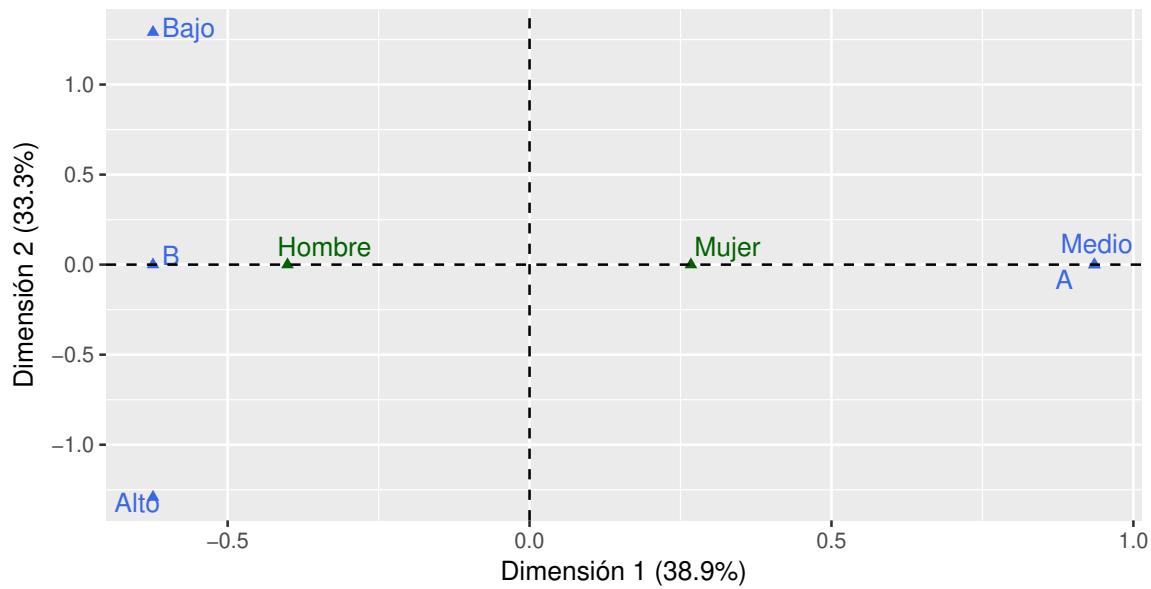


Figura 5.26: Biplot simétrico para la empresa

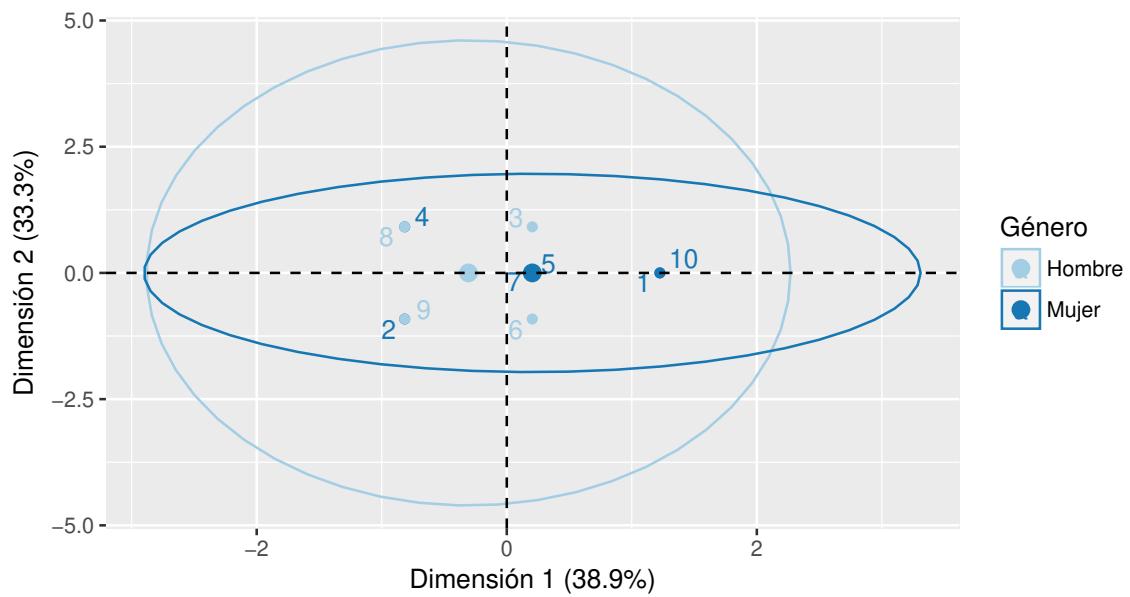


Figura 5.27: Empleados agrupados por género

	Gén.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 5.35: Matriz de Burt para la empresa

	$A \times 1$	$A \times 2$	$A \times 3$	$A \times 4$
Autovalores	0.556	0.358	0.333	0.086
Inercia proyectada (%)	41.67	26.87	25.00	6.46
Inercia proyectada acumulada (%)	41.67	68.54	93.54	100.00
Inercia acumulada	0.556	0.914	1.25	1.33

Tabla 5.36: Inercias para la empresa

	CS1	CS2
Género.Hombre	1.5	-0.454
Género.Mujer	-1.0	0.303
Ingresos.Alto	1.0	0.303
Ingresos.Bajo	1.0	0.303
Ingresos.Medio	-1.5	-0.454
Categoría.A	0.0	-2.022
Categoría.B	0.0	1.348

Tabla 5.37: Coordenadas de representación para la empresa

5.7 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 5.38 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
Totales	61	45	62	25	193

Tabla 5.38: 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 europeo

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.

Capítulo 6

Escalamiento multidimensional

Todas las verdades de las matemáticas están vinculadas entre sí.

— Adrien-Marie Legendre

El escalamiento multidimensional (MDS, del inglés *multidimensional scaling*) tiene sus orígenes a principios de siglo XX en el campo de la Psicología [19]. El mismo surge como respuesta al propósito de estudiar la relación que existía entre la intensidad física de ciertos estímulos y su intensidad subjetiva.

El MDS es una técnica multivariante de interdependencia, que trata de representar en un espacio geométrico de pocas dimensiones, las proximidades existentes entre un conjunto de objetos que pertenecen a un espacio de dimensión mayor [13]. Esta técnica de representación espacial, facilita la visualización de un conjunto de objetos en un mapa. Como ejemplos de estos objetos podemos considerar empresas, productos comerciales, candidatos políticos o ideas, cuya posición relativa interesa analizar.

Resulta conveniente relacionar esta técnica con otros modelos multivariados de modo que pueda servir como alternativa y/o complemento a los mismos para cierta investigación.

En la actualidad, el MDS es apto para diferentes tipos de datos de entrada, como por ejemplo tablas de contingencia, matrices de proximidad, datos de perfil o correlaciones.

En síntesis, es escalamiento multidimensional es una técnica multivariante que crea un gráfico a partir de un conjunto de similitudes o distancias definidas sobre un conjunto de objetos, tratando de preservar las distancias o similitudes originales en la nueva representación.

6.1 Modelo general

El MDS toma como insumo una **matriz de proximidades**, $\Delta \in \mathbb{R}^{n \times n}$, donde n es el número de objetos que se desea representar entre los cuales se ha definido una distancia o medida de

similaridad. Cada elemento δ_{ij} de Δ representa la proximidad entre los objetos i -ésimo y j -ésimo. Es decir: $\delta_{ij} = \text{dist}(I_i, I_j)$ siendo I_i el i -ésimo individuo e I_j el j -ésimo.

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1j} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2j} & \cdots & \delta_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ \delta_{i1} & \delta_{i2} & \cdots & \delta_{ij} & \cdots & \delta_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nj} & \cdots & \delta_{nn} \end{pmatrix}$$

Esta matriz tiene todos sus elementos mayores o iguales a 0, siendo su diagonal principal nula, debido a que la distancia de un objeto a sí mismo es 0.

$$\Delta = \begin{pmatrix} 0 & \delta_{12} & \cdots & \delta_{1j} & \cdots & \delta_{1n} \\ \delta_{21} & 0 & \cdots & \delta_{2j} & \cdots & \delta_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ \delta_{i1} & \delta_{i2} & \cdots & \delta_{ij} & \cdots & \delta_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nj} & \cdots & 0 \end{pmatrix}$$

Además, se trata de una matriz simétrica pues la distancia entre el objeto i y el objeto j , dada por δ_{ij} , es la misma que la distancia entre el objeto j y el objeto i , indicada por δ_{ji} .

A partir de esta matriz de proximidades, el MDS proporciona como salida una **matriz de coordenadas** $X \in \mathbb{R}^{n \times m}$,

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}$$

donde n es la cantidad de objetos, m es la dimensión de representación de los mismos y x_{ij} es la j -ésima coordenada del i -ésimo individuo. Usando ahora la matriz X , se puede calcular la distancia existente entre las representaciones de dos objetos i y j del conjunto considerado. Estas distancias

pueden presentarse en una **matriz de distancias** que denominamos $D \in \mathbb{R}^{n \times n}$,

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1j} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2j} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ d_{i1} & d_{i2} & \cdots & d_{ij} & \cdots & d_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nj} & \cdots & d_{nn} \end{pmatrix}$$

El MDS proporciona una solución que maximiza la correspondencia entre las distancias originales Δ y las representadas D .

Existen varias medidas que cuantifican esta correspondencia entre las distancias originales y las representadas, midiendo de alguna forma, la bondad de la representación.

6.2 Modelos particulares

Existen dos modelos básicos de escalamiento multidimensional que son el modelo de escalamiento métrico y el modelo de escalamiento no métrico. En el primero de ellos se considera que los datos están medidos en escala de razón o en escala de intervalo, mientras que en el segundo se supone que los datos están medidos en escala ordinal.

6.2.1 Modelo de escalamiento métrico

Todo modelo de escalamiento parte de la idea de que las distancias son una función de las proximidades; es decir, $d_{ij} = f(\delta_{ij})$. En el **modelo de escalamiento métrico** partimos del supuesto de que la relación entre las proximidades y las distancias es de tipo lineal:

$$d_{ij} = a + b \cdot \delta_{ij}$$

donde $a, b \in \mathbb{R}$.

El primer procedimiento de escalamiento métrico fue presentado por Torgerson [45], según el cual a partir de una matriz de distancias $D \in \mathbb{R}^{n \times n}$ se puede obtener una matriz $B \in \mathbb{R}^{n \times n}$ de productos escalares entre vectores.

El procedimiento en este caso consiste en transformar la matriz de proximidades $\Delta \in \mathbb{R}^{n \times n}$ en una matriz de distancias $D \in \mathbb{R}^{n \times n}$, de forma tal que verifique las tres condiciones que definen una distancia:

- **no negatividad** $d_{ij} \geq 0$
- **simetría** $d_{ij} = d_{ji}$

- **desigualdad triangular** $d_{ij} \leq d_{ik} + d_{kj}$

Los dos primeros axiomas se cumplen fácilmente, pero el tercer axioma, no siempre se cumple. Este problema se conoce con el nombre de **estimación de la constante aditiva**.

Togerson proporciona una solución para este problema estimando el valor mínimo de c que verifica la desigualdad triangular de la siguiente forma

$$c_{\min} = \max_{i,j,k} \{d_{ij} - d_{ik} - d_{kj}\}$$

De esta manera, las distancias se obtienen sumando la constante c a las proximidades.

Ejemplo 6.1. Sea la matriz de proximidades Δ para tres objetos A , B y C .

$$\Delta = \begin{pmatrix} 0 & 1 & 5 \\ 1 & 0 & 2 \\ 5 & 2 & 0 \end{pmatrix}$$

Se puede ver que no verifica la desigualdad triangular, puesto que $d(A,C) = 5$, mientras que $d(A,B) + d(B,C) = 1 + 2 = 3 < 5$.

En este caso el valor mínimo de la constante es $c = 2$. Sumando 2 a los elementos no diagonales de la matriz de proximidades Δ , obtenemos la matriz de distancias

$$D = \begin{pmatrix} 0 & 3 & 7 \\ 3 & 0 & 4 \\ 7 & 4 & 0 \end{pmatrix}$$



Una vez obtenida la matriz $D \in \mathbb{R}^{n \times n}$, se necesita transformarla en una matriz $B \in \mathbb{R}^{n \times n}$ de productos escalares entre vectores, para lo cual se aplica la siguiente transformación

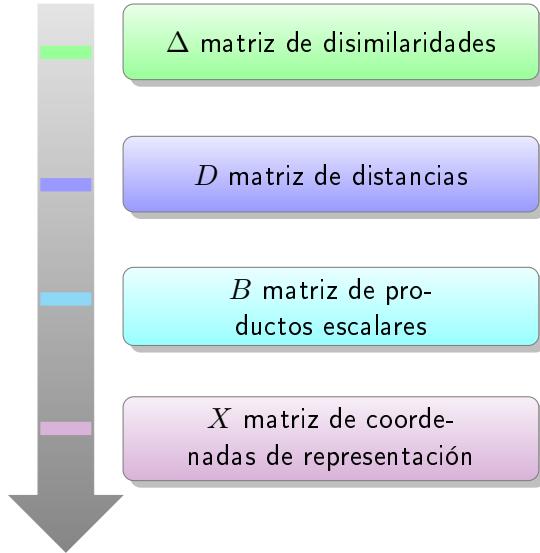
$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2)$$

donde

- ✿ $d_{i\cdot}^2 = \frac{1}{n} \sum_{j=1}^n d_{ij}^2$ es la distancia cuadrática media por fila,
- ✿ $d_{\cdot j}^2 = \frac{1}{n} \sum_{i=1}^n d_{ij}^2$ es la distancia cuadrática media por columna,
- ✿ $d_{\cdot\cdot}^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2$ es la distancia cuadrática media de la matriz.

Finalmente, se busca una matriz $X \in \mathbb{R}^{n \times m}$ tal que $B = XX^t$, siendo X la matriz que guarda las coordenadas de cada uno de los n objetos en cada una de las m dimensiones. Cualquier método de factorización permite transformar B en XX^t .

En resumen el método se describe a partir de los siguientes pasos:



De lo antes expuesto, tenemos que $d_{ij} = f(\delta_{ij})$. Si esto fuera exactamente así, no habría margen de error. Sin embargo, en las proximidades empíricas es difícil que valga la igualdad, con lo que generalmente ocurre que $d_{ij} \approx f(\delta_{ij})$.

A las transformaciones de las proximidades por f se las denomina **disparidades**. A partir de esto, se define el *error cuadrático* como

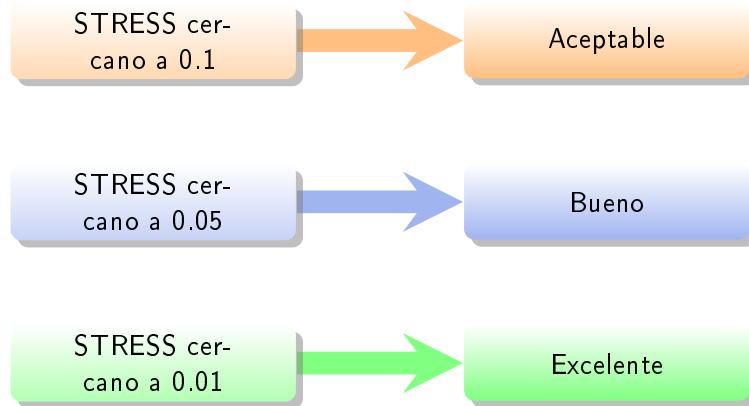
$$e_{ij}^2 = (d_{ij} - f(\delta_{ij}))^2$$

Mientras mayor sea la diferencia entre las disparidades y las distancias; es decir, las distancias entre $f(\delta_{ij})$ y d_{ij} , mayor será el *stress* y por tanto peor será el modelo. El ***stress*** es una medida de la falta de ajuste del modelo. El mismo toma como valor mínimo al 0, mientras que su límite superior es $1 - \frac{2}{n}$ donde n indica la cantidad de objetos. La manera de calcularlo es

$$STRESS = \sqrt{\frac{\sum_{i,j} (d_{ij} - f(\delta_{ij}))^2}{\sum_{i,j} d_{ij}^2}} = \sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}}$$

Kruskal [28] sugiere la siguiente clasificación del modelo según su *stress*





Una variante del *stress* es lo que se conoce como el *S-stress*, que resulta levemente distinto y la fórmula para calcularlo es

$$S\text{-STRESS} = \sqrt{\frac{\sum_{i,j} (d_{ij}^2 - f^2(\delta_{ij}))^2}{\sum_{i,j} (d_{ij}^2)^2}}$$

Otra medida que se suele utilizar es el **coeficiente de correlación al cuadrado** (RSQ), el cual nos informa sobre la proporción de variabilidad de los datos de partida que es explicada por el modelo. El mismo se define como

$$RSQ = \text{corr}^2(d_{ij}, f(\delta_{ij}))$$

El rango de variación de este coeficiente es $[0, 1]$ por ser un coeficiente de correlación al cuadrado. Valores cercanos a 1 indican que el modelo es bueno, mientras que valores cercanos a 0 indican que el modelo es malo.

El programa R tiene implementados tanto los algoritmos para obtener soluciones con MDS así como las medidas para determinar si el modelo es adecuado o no lo es. Los algoritmos implementados son reiterativos, de forma que se alcance la mejor solución posible.

Ejemplo 6.2. Consideremos un conjunto de ciudades de la República Argentina de las cuales tenemos información sobre su latitud y su longitud. Tengamos presente que cada una de ellas tiene también diferente altura respecto del nivel del mar. Esta información está consignada en la Tabla 6.1

En primera instancia calculamos la matriz de distancias euclídeas entre las posiciones, recordando que es simétrica. Esta distancia será medida en diferencias de latitud y longitud, pero puede no ser exactamente proporcional a la distancia en kilómetros entre las ciudades.

Número	Ciudad	Latitud	Longitud
1	Buenos Aires	-34.61	-58.38
2	Córdoba	-31.41	-64.18
3	Rosario	-32.95	-60.64
4	Mendoza	-32.89	-68.83
5	Tucumán	-26.82	-65.22
6	Salta	-24.79	-65.41
7	Santa Fe	-31.63	-60.70
8	San Juan	-31.54	-68.54
9	Resistencia	-27.46	-58.98
10	Santiago del Estero	-27.80	-64.26
10	Corrientes	-27.48	-58.83
12	Posadas	-27.37	-55.90
13	San Salvador de Jujuy	-24.19	-65.30
14	Bahía Blanca	-38.72	-62.27
15	Paraná	-31.73	-60.52
16	Neuquén	-38.95	-68.06

Tabla 6.1: Ciudades argentinas



<https://flic.kr/p/81zXt3>

La representación obtenida se muestra en la Figura 6.2 y es generada mediante el Código 6.1 con datos extraídos de <https://goo.gl/bby6JC>.

```

library(readxl) # Permite leer archivos xlsx
library(ggplot2) # Paquete para confeccionar dibujos
library(ggrepel) # Paquete que manipula etiquetas para gráficos
library(plotrix) # Paquete para gráficos requerido para la librería smacof
library(smacof) # Paquete para MDS basado en la minimización del stress

cities=read_excel("C:/.../ciudades.xlsx")
# Importa la base con la cual se va a trabajar
ciudades=data.frame(cities[,2:3])

D=dist(ciudades) # Calcula las distancias euclídeas entre las filas
MCD_D=cmdscale(D, eig=TRUE, k=2)
# Realiza el MCD de una matriz de datos con k dimensiones de representación
x=MCD_D$points[,1] # Guarda las abscisas de los puntos
y=MCD_D$points[,2] # Guarda las ordenadas de los puntos

# Preparamos base de datos para el gráfico
data=cbind(-x,-y)
datos=data.frame(data)
colnames(datos)=c("Latitud","Longitud")
rownames(datos)=c("Buenos_Aires","Córdoba","Rosario","Mendoza","Tucumán",
"Salta","Santa_Fe","San_Juan","Resistencia",
"Santiago_del_Estero","Corrientes","Posadas",
"San_Salvador_de_Jujuy","Bahía_Blanca","Paraná",
"Neuquén")

ggplot(datos, aes(x=Latitud, y=Longitud))+
geom_point(colour="royalblue") +
geom_text_repel(aes(label=rownames(datos))) +
theme_gray()
# Realiza un gráfico de puntos

MCD.D= smacofSym(D, ndim=2) # Realiza una escala multidimensional

```

```
MCD.D$stress # Calula el stress del ajuste
```

Código 6.1: Código para el análisis MDS de ciudades argentinas

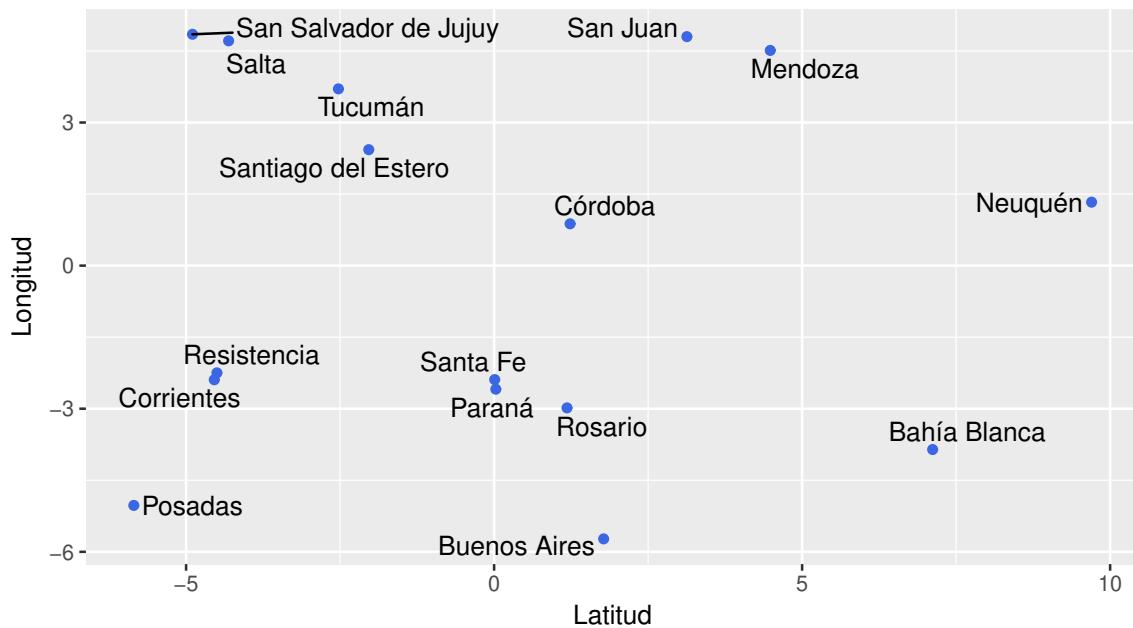


Figura 6.2: MDS aplicado a ciudades argentinas

6.3 Relación con otras técnicas

El MDS se utiliza en muchas investigaciones junto a otras técnicas multivariantes, bien como una alternativa a dichas técnicas o bien como un complemento de las mismas, dependiendo de los objetivos de la investigación. Aunque otras técnicas, como el análisis factorial, el análisis discriminante y el análisis conjunto, también sirven para reducir los datos a unos pocos factores o dimensiones, el MDS obtiene el grado de similaridad entre los datos, con el propósito de representar la información en un espacio de pocas dimensiones, preservando lo más posible las distancias originales.

Entre las ventajas del MDS podemos mencionar las siguientes:

- ✿ Los datos en MDS pueden estar medidos en cualquier escala, continua o discreta.
- ✿ El MDS proporciona soluciones para cada individuo.

- ✿ En el MDS el investigador no necesita especificar cuáles son las variables a emplear en la comparación de objetos.
- ✿ Las distancias en el MDS pueden ser interpretadas directamente entre todos los puntos, mientras que en el análisis de correspondencias solamente pueden ser interpretadas directamente las distancias entre las filas o bien entre las columnas.

Capítulo 7

Comparación de medias en el caso univariado

La estadística es el único tribunal de apelación para juzgar el nuevo conocimiento.

— Prasanta Chandra Mahalonibis

Es importante destacar que en este capítulo no trabajaremos con técnicas descriptivas o exploratorias, sino que emplearemos pruebas estadísticas tales como el contraste de hipótesis. Por esta razón, el cumplimiento de los supuestos de cada una de las pruebas es fundamental para la validez de las conclusiones.

7.1 Diferencia de medias de poblaciones normales para dos muestras independientes

7.1.1 Muestras normales independientes con varianzas conocidas

Ejemplo 7.1. Se sospecha que el pH (potencial de hidrógeno) de la superficie del suelo de dos regiones A y B es diferente. Un geólogo realiza mediciones y determina electromecánicamente el pH que se encuentra en la superficie del suelo en 20 puntos elegidos al azar para cada una de las dos regiones de interés.



<https://flic.kr/p/6W2onx>

Se supone que el pH en la superficie del suelo de ambas regiones, digamos A y B se distribuye normalmente, con varianzas 0.85 y 1.22 respectivamente. Los resultados muestrales se presentan en la Tabla 7.1.

	Región A	Región B
Media	6.58	5.74

Tabla 7.1: Observaciones del experimento del pH



Definimos el modelo a partir de dos muestras normales independientes:

- * X_1, X_2, \dots, X_{n_X} donde $X_i \sim N(\mu_X, \sigma_X^2)$ (región A)
- * Y_1, Y_2, \dots, Y_{n_Y} donde $Y_i \sim N(\mu_Y, \sigma_Y^2)$ (región B)

El interés radica en realizar inferencias acerca del parámetro diferencia de medias de las dos poblaciones dado por $\mu_X - \mu_Y$. Un **estimador puntual insesgado** para este parámetro es $\bar{X} - \bar{Y}$. Como ambas poblaciones son normales con varianzas conocidas, se sabe que

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_X}\right) \quad \text{y} \quad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_Y}\right)$$

y que estas variables son independientes. Luego,

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

Por lo tanto, estandarizando el estimador puntual, se tiene una distribución conocida y tabulada

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0, 1)$$

Las hipótesis para testear en nuestro ejemplo serán entonces:

$$\begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y \neq 0 \end{cases}$$

Como se trata de una hipótesis alternativa bilateral, valores muy grandes o muy pequeños del estadístico de contraste conducirán a rechazar la hipótesis de nulidad. Si se establece un nivel de significación $\alpha = 0.05$; es decir, una probabilidad máxima de 0.05 de rechazar H_0 siendo ésta cierta, la región de rechazo será:

$$RC = \{z_{obs} / z_{obs} \geq 1.96 \text{ o } z_{obs} \leq -1.96\}$$

Con lo cual la decisión será rechazar H_0 cuando z_{obs} resulte superior a 1.96 o inferior a -1.96. Esto queda representado en la Figura 7.2.

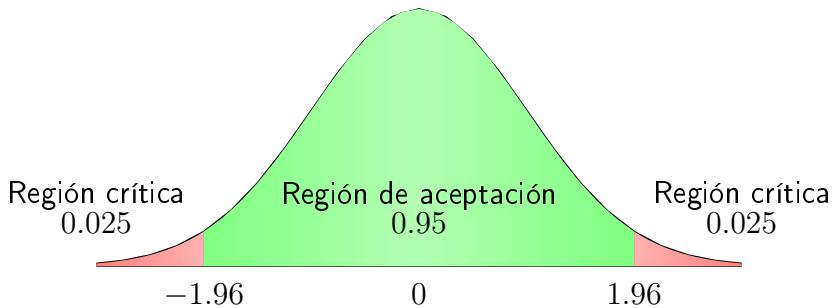


Figura 7.2: Zonas de aceptación y de rechazo para esta prueba

El valor de la variable pivotal o estadístico de contraste bajo la hipótesis nula para el Ejemplo 7.1 es

$$z_{obs} = \frac{6.58 - 5.74 - 0}{\sqrt{\frac{0.85}{20} + \frac{1.22}{20}}} = 2.61$$

Como $2.61 > 1.96$, $z_{obs} \in RC$ y entonces la decisión es rechazar H_0 . La conclusión es que existe evidencia empírica en contra de la hipótesis de que las medias poblacionales de los pH de los suelos de las dos regiones son iguales, con un nivel de significación del 5%.

Puede resultar de interés cuantificar la fuerza del rechazo de la hipótesis nula, o bien la probabilidad de encontrar un valor tan extremo o más que el hallado en esta muestra siendo cierta la hipótesis nula. A esta probabilidad se la denomina ***p*-valor** y en el Ejemplo 7.1 resulta ser

$$p\text{-valor} = P(|Z| > 2.61) = 2P(Z > 2.61) = 2 \cdot 0.0045 = 0.009$$

Cuando este valor es pequeño, indica que existe bastante seguridad en la decisión, ya que es muy poco probable que, siendo cierta H_0 , nos encontremos con una diferencia de medias como ésta.

Si se quiere construir un **intervalo de confianza** de nivel $1 - \alpha$ para la diferencia de las medias, se utiliza como variable pivotal

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0; 1)$$

Luego, podemos afirmar que:

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Realizando las operaciones algebraicas necesarias para despejar el parámetro diferencia de medias, tenemos la expresión del intervalo de confianza de nivel $1 - \alpha$ bajo el supuesto de poblaciones normales, independientes con varianzas conocidas. En forma general la expresión de este intervalo es:

$$\left[\bar{X} - \bar{Y} - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right]$$

Recordemos que un intervalo de confianza de nivel 95% para un parámetro se interpreta de la siguiente manera: de cada 100 intervalos construidos a partir de muestras de igual tamaño, alrededor de 95 contendrán el valor verdadero del parámetro.

Para el Ejemplo 7.1, el parámetro es la diferencia de medias poblacionales de pH correspondiente a dos regiones. Los límites del intervalo de confianza de nivel 95% para la diferencia de medias de pH resultan

$$\left[6.58 - 5.74 \mp 1.96 \sqrt{\frac{0.85}{20} + \frac{1.22}{20}} \right]$$

Con lo cual el intervalo de confianza es $[0.84 - 0.63, 0.84 + 0.63] = [0.21, 1.47]$. Se debe interpretar que con una confianza del 95%, que el intervalo $[0.21, 1.47]$ contiene al verdadero valor de la diferencia entre las medias de pH de las Regiones A y B. Se puede observar que como ambos extremos del intervalo son positivos, el cero no pertenece al intervalo. Entonces el test basado en el intervalo de confianza también rechaza la hipótesis de igualdad entre las medias.

7.1.2 Muestras normales independientes con varianzas desconocidas

Ejemplo 7.2. El tiempo que le toma a la habichuela en duplicar su peso es una medida de su calidad para enlatar.



<https://flic.kr/p/oALvQ6>

Un experimento con 15 repeticiones independientes de cada una de dos variedades produjo los resultados que se muestran en la Tabla 7.2.

Variedad A	Variedad B
$\bar{X} = 17.2$ horas	$\bar{Y} = 18.3$ horas
$s_X = 0.7$ horas	$s_Y = 0.8$ horas

Tabla 7.2: Observaciones del experimento de las habichuelas

Interesa decidir si la calidad de la variedad B (variable Y) es inferior a la calidad de la variedad A (variable X), utilizando para probar estas hipótesis un nivel de significación de 0.01.

Si se pudiera asegurar que ambas muestras provienen de distribuciones normales con la misma varianza, entonces el modelo podría definirse como dos muestras normales independientes con medias distintas y varianzas iguales pero desconocidas; es decir,

$$\textcircled{*} \quad X_1, X_2, \dots, X_{n_X} \text{ donde } X_i \sim N(\mu_X, \sigma^2)$$

$$\textcircled{*} \quad Y_1, Y_2, \dots, Y_{n_Y} \text{ donde } Y_i \sim N(\mu_Y, \sigma^2)$$

con σ^2 desconocida.

Las hipótesis de interés para este caso son

$$\begin{cases} H_0 : \mu_Y - \mu_X \geq 0 \\ H_1 : \mu_Y - \mu_X < 0 \end{cases}$$

o bien

$$\begin{cases} H_0 : \mu_Y - \mu_X = 0 \\ H_1 : \mu_Y - \mu_X < 0 \end{cases}$$

Se trata de una prueba unilateral a derecha, por lo cual se rechaza la hipótesis nula cuando el estadístico de contraste toma valores bajos. Estamos nuevamente interesados en realizar inferencias acerca del parámetro dado por la diferencia de medias poblacionales de las dos poblaciones, $\mu_Y - \mu_X$. Se sabe que

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n_X}\right) \quad \text{y} \quad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{n_Y}\right)$$

y que estas variables son independientes. Luego,

$$\bar{Y} - \bar{X} \sim N\left(\mu_Y - \mu_X, \frac{\sigma^2}{n_Y} + \frac{\sigma^2}{n_X}\right) = N\left(\mu_Y - \mu_X, \left(\frac{1}{n_Y} + \frac{1}{n_X}\right)\sigma^2\right)$$

Entonces, estandarizando el estimador puntual propuesto, se obtiene la expresión de la variable pivotal

$$Z = \frac{\bar{Y} - \bar{X} - (\mu_Y - \mu_X)}{\sigma \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}} \sim N(0, 1)$$

Debido a que la varianza es común a las dos poblaciones, tiene sentido construir un estimador insesgado de la varianza común, basado en ambas muestras. Este estimador usualmente se conoce como **varianza amalgamada** o *poolizada* y su fórmula es

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

Es sabido que las siguientes variables son independientes y tienen distribución Chi-cuadrado

$$\frac{(n_X - 1)S_X^2}{\sigma^2} \sim \chi_{n_X - 1}^2 \quad \text{y} \quad \frac{(n_Y - 1)S_Y^2}{\sigma^2} \sim \chi_{n_Y - 1}^2$$

Es un resultado conocido que la suma de variables aleatorias Chi-cuadrado independientes es otra variable aleatoria Chi-cuadrado cuyos grados de libertad son la suma de los grados de libertad de las variables sumadas. Entonces

$$U = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{\sigma^2} \sim \chi_{n_X + n_Y - 2}^2$$

Al ser Z y U independientes, se tiene que

$$\frac{Z}{\sqrt{\frac{U}{n_X + n_Y - 2}}} \sim t_{n_X + n_Y - 2}$$

donde la distribución es la t de Student con $n_X + n_Y - 2$ grados de libertad. Un simple cálculo algebraico aplicado a esta variable conduce a que

$$T = \frac{\bar{Y} - \bar{X} - (\mu_Y - \mu_X)}{S_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X+n_Y-2}$$

La región de rechazo del test de nivel 0.01 para el Ejemplo 7.2 es

$$RC = \{t_{obs}/t_{obs} > t_{28,0.99} = 2.467\}$$

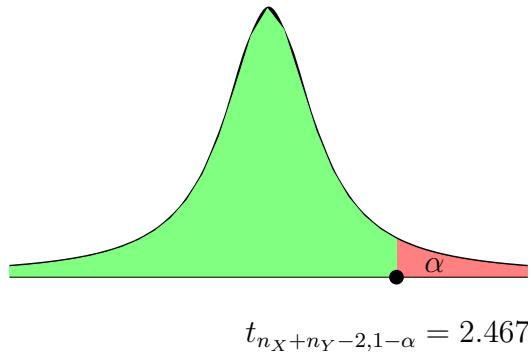


Figura 7.4: Zonas de aceptación y de rechazo para esta prueba

El valor observado del estadístico de contraste en el Ejemplo 7.2 es

$$t_{obs} = \frac{18.3 - 17.2}{\sqrt{\frac{14 \cdot 0.8^2 + 14 \cdot 0.7^2}{28} \left(\frac{1}{15} + \frac{1}{15} \right)}} = 4.008$$

Como $4.008 > 2.467$, rechazamos la hipótesis de nulidad con un nivel de significación del 1%, lo cual significa que hay evidencia en contra de la hipótesis nula que sostiene que el valor medio poblacional del tiempo que tarda la variedad A de habichuelas en duplicar su tamaño es igual o mayor que el tiempo medio poblacional que tarda la variedad B.

Si quisieramos construir un intervalo de confianza de nivel $1 - \alpha$ para la diferencia de medias de poblaciones normales provenientes de muestras independientes y con varianzas desconocidas, pero que pueden suponerse iguales, debemos utilizar la variable pivotal

$$\frac{\bar{Y} - \bar{X} - (\mu_Y - \mu_X)}{S_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X+n_Y-2}$$

y plantear

$$P\left(-t_{n_X+n_Y-2, \frac{\alpha}{2}} \leq \frac{\bar{Y} - \bar{X} - (\mu_Y - \mu_X)}{S_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \leq t_{n_X+n_Y-2, \frac{\alpha}{2}}\right) = 1 - \alpha$$

Realizando las operaciones algebraicas necesarias para despejar de esta expresión el parámetro de interés, obtenemos el intervalo de confianza de nivel $1 - \alpha$ para la diferencia de medias de poblaciones independientes con varianzas desconocidas pero iguales,

$$\left[\bar{Y} - \bar{X} - t_{n_X+n_Y-2, \frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, \bar{Y} - \bar{X} + t_{n_X+n_Y-2, \frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\right]$$

Aplicando esta fórmula para el Ejemplo 7.2, el intervalo de confianza para la diferencia de medias de nivel 99% resulta [0.342, 1.858].

Este intervalo tiene ambos extremos positivos lo que indica que la estimación de la diferencia de medias poblacionales es positiva. Este hecho implica que existe evidencia empírica en favor de que la media poblacional de la calidad de la variedad B de habichuelas es superior a la de la variedad A.

7.1.3 Muestras independientes de poblaciones cualesquiera

Si las muestras son suficientemente grandes, es posible aplicar la distribución Normal, basándonos en el **Teorema del Límite Central** que tiene un nivel aproximado o asintótico. Este es el caso más usual para *data mining* donde se dispone de mucha información y, en general, la información no satisface el supuesto de normalidad. Si se desea aplicar una prueba basada en el supuesto de normalidad y los datos disponibles no la satisfacen, una alternativa viable es aplicar **transformaciones de Box & Cox** [35] o **transformaciones de Jhonson** para normalizar los datos y que dichos test sean válidos [30].

Ejemplo 7.3. Los datos expuestos en la Tabla 7.3 corresponden a dos muestras aleatorias, una de varones y otra de mujeres, estudiantes universitarios cuyas edades oscilan entre los 20 y 30 años, que realizan algún tipo de actividad física y a los cuales se les preguntó sobre la cantidad promedio de horas semanales dedicadas a este tipo de actividades, que incluyen algún deporte, clases de gimnasia y caminata, entre otras.

	Varones (X)	Mujeres (Y)
Número de observaciones	124	110
Media muestral	6.6	5.4
Desvío estándar muestral	4.3	3.6

Tabla 7.3: Distribución de Frecuencias para actividades físicas



<https://flic.kr/p/Vi7F56>

Queremos contrastar las hipótesis de que la cantidad de horas semanales dedicadas a la actividad deportiva es la misma para ambos grupos definidos por la variable sexo, con un nivel de significación del 1%.



Para este tipo de experimento, las hipótesis de interés son:

$$\begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y \neq 0 \end{cases}$$

Para el Ejemplo 7.3, consideramos μ_X la media del tiempo semanal dedicado a la gimnasia por los hombres y μ_Y la media del tiempo semanal dedicado a la gimnasia por las mujeres.

En líneas generales, el modelo consiste de dos muestras aleatorias independientes de tamaños grandes (se considera grande cuando cada una de ellas es mayor a 30) cuya distribución no puede garantizarse que sea Normal. Es decir, se tiene las siguientes variables independientes:

- * X_1, X_2, \dots, X_{n_X} con $E(X_i) = \mu_X$ y $V(X_i) = \sigma_X^2$, para $i = 1, \dots, n_X$ (con $n_X > 30$)
- * Y_1, Y_2, \dots, Y_{n_Y} con $E(Y_i) = \mu_Y$ y $V(Y_i) = \sigma_Y^2$, para $i = 1, \dots, n_Y$ (con $n_Y > 30$)

Aplicando el Teorema Central del Límite, tenemos que la distribución aproximada o asintótica para los promedios muestrales es

$$\frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n_X}}} \approx N(0, 1) \quad \text{y} \quad \frac{\bar{Y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n_Y}}} \approx N(0, 1)$$

Al desconocer los valores que toman las varianzas poblacionales, podemos utilizar la propiedad que asegura que $\frac{\sigma_X}{S_X}$ y $\frac{\sigma_Y}{S_Y}$ convergen a uno conforme el tamaño muestral tiende a infinito. Con lo cual,

$$\frac{\bar{X} - \mu_X}{\frac{S_X}{\sqrt{n_X}}} \approx N(0, 1) \quad \text{y} \quad \frac{\bar{Y} - \mu_Y}{\frac{S_Y}{\sqrt{n_Y}}} \approx N(0, 1)$$

De donde se deduce

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \approx N(0, 1)$$

Al tratarse en este caso, de una hipótesis bilateral, se rechaza la hipótesis nula para valores muy altos o para valores muy bajos del estadístico de contraste. Si denotamos por z_α al percentil derecho α de la distribución Normal estándar; es decir, $P(Z > z_\alpha) = \alpha$, entonces la región de rechazo del test para un nivel de significación del 1% es

$$RC = \{z_{obs}/z_{obs} > z_\alpha \text{ o } z_{obs} < -z_\alpha\}$$

Refiriéndonos al Ejemplo 7.3, rechazaremos H_0 cuando los valores del estadístico de contraste resulten superiores a 2.58 o inferiores a -2.58. El valor del estadístico del test observado para estos datos es

$$z_{obs} = \frac{6.6 - 5.4 - 0}{\sqrt{\frac{4.3^2}{124} + \frac{3.6^2}{110}}} = 2.32$$

Por lo tanto, la decisión es no rechazar H_0 . Vale decir que no existe evidencia a favor de que los tiempos destinados semanalmente a la actividad deportiva sean diferentes de acuerdo al sexo. Concluimos que la media del tiempo semanal que dedican a actividades físicas los varones no es significativamente diferente de la media del tiempo que le dedican las mujeres.

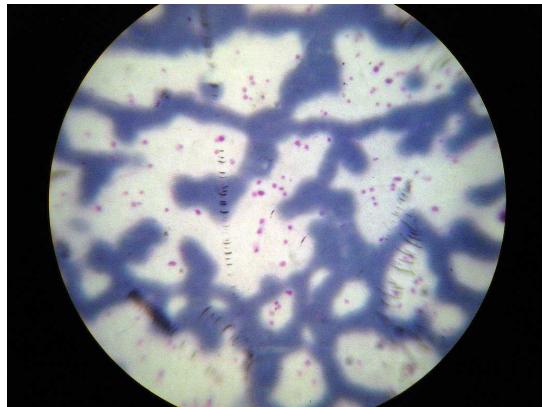
7.1.4 Muestras apareadas

Para analizar este caso, introducimos el siguiente ejemplo extraído de [40] (ver páginas 51-52).

Ejemplo 7.4. En un estudio para investigar si la acción de fumar afecta a la función plaquetaria, se reclutaron 12 sujetos a los cuales se les midió la agregación plaquetaria con adenosín difosfato (ADP) antes de fumar y después de fumar. Los resultados se muestran en la Tabla 7.4.

Sujeto	Posterior a fumar	Anterior a fumar
1	70	60
2	82	78
3	66	64
4	69	68
5	75	72
6	72	76
7	77	74
8	49	46
9	54	48
10	66	60
11	69	64
12	36	27

Tabla 7.4: Datos apareados para medición de plaquetas con ADP



<https://flic.kr/p/4CFf5u>

El objetivo es evaluar si el hecho de fumar aumenta la agregación plaquetaria.



La diferencia entre este caso y los anteriores radica en que las dos muestras de observaciones **no son independientes**.

El interés no se centra en evaluar si la media del primer conjunto de observaciones es mayor, menor o distinta de la media del segundo conjunto de observaciones, sino más bien interesa estudiar la media de las diferencias por individuo.

El **apareamiento** surge como una estrategia de investigación cuando las observaciones son realizadas en un mismo individuo en dos instantes de tiempo, mediando alguna intervención o bien

en individuos apareados con algún criterio en los que se aplican distintos tratamientos.

Por ejemplo, parejas de gemelos pueden ser asignadas al azar para que reciban dos tratamientos específicos, de tal manera que los miembros de una sola pareja reciban tratamientos distintos. Pueden asimismo ensayarse dos raciones distintas en dos lotes de terneros formando pares de raza de la misma edad, sexo, entre otras características, y ocurrir que al cabo de un tiempo exista diferencia significativa o no, entre los promedios de ganancia de peso de ambos lotes y de esta forma se logra eliminar la influencia de calidad inicial de los lotes considerados.

En el Ejemplo 7.4, se elimina la influencia de las características individuales de los sujetos dadas por la variación biológica.

¿Cómo se formaliza el modelo y las hipótesis?

El conjunto de observaciones ahora es de la forma $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ y se define la nueva variable **diferencia** como $D_i = X_i - Y_i$ para $i = 1, \dots, n$, donde $D_i \sim N(\mu_D, \sigma_D)$ son independientes para $1 \leq i \leq n$.

Consideramos el estadístico dado por

$$T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

siendo \bar{D} la media muestral de las diferencias y S_D su desvío estándar muestral.

Es importante observar que lo que se pide para aplicar este modelo es la normalidad en la distribución de las diferencias, esto implica que las observaciones X_i o Y_i podrían no ser normales. Además se trabaja con el supuesto de independencia entre los sujetos muestreados. Si las distribuciones de X e Y son normales, la distribución de las diferencias también lo es. Sin embargo, las condiciones requeridas son menos estrictas o más relajadas.

El estudio se realiza sobre esta nueva muestra formada por las diferencias entre los pares de observaciones, con un nivel de significación del 0.05.

Las hipótesis que se plantean son:

$$\begin{cases} H_0 : \mu_D \leq 0 \\ H_1 : \mu_D > 0 \end{cases}$$

Se trata en este caso de un test unilateral derecho y por lo tanto la región de rechazo está dada por

$$RC = \{t_{obs}/t_{obs} > t_{n-1, 1-\alpha}\}$$

En el Ejemplo 7.4, se rechaza H_0 si $t_{obs} > t_{11, 0.95} = 1.796$. En la Tabla 7.5 se calculan las diferencias de las observaciones.

Sujeto	Posterior a fumar	Anterior a fumar	Diferencia
1	70	60	10
2	82	78	4
3	66	64	2
4	69	68	1
5	75	72	3
6	72	76	-4
7	77	74	3
8	49	46	3
9	54	48	6
10	66	60	6
11	69	64	5
12	36	27	9

Tabla 7.5: Diferencias de datos apareados

Siguiendo con el Ejemplo 7.4, tenemos que $n = 12$, $\bar{D} = 4.17$ y $S_D = 3.97$. Entonces la variable pivotal toma el valor

$$t_{obs} = \frac{4.17 - 0}{\frac{3.97}{\sqrt{12}}} = 3.63$$

y como $3.63 > 1.7958$, se rechaza H_0 . Es decir: existe evidencia estadística con una significación del 5% de que fumar aumenta la agregación plaquetaria.

Si se quisiera estimar en cuánto aumenta la agregación plaquetaria debido al efecto de fumar durante este período, se podría estimar un intervalo de confianza para las diferencias.

Utilizando la información que portan las diferencias D_i y siguiendo el procedimiento habitual, se obtiene el siguiente intervalo de confianza de nivel $1 - \alpha$ para la media de las diferencias

$$\left[\bar{D} - t_{n-1,1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}, \bar{D} + t_{n-1,1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}} \right]$$

Este intervalo para el Ejemplo 7.4 está dado por $[2.11, 6.22]$. Es evidente observar que la estimación de esta diferencia es positiva debido a que ambos extremos del intervalo son positivos. Todo esto es válido cuando se satisface el supuesto de normalidad, pero cabe preguntarnos lo siguiente.

¿Qué sucede cuando no se puede asegurar el cumplimiento del supuesto de normalidad?

7.2 Pruebas no paramétricas para dos muestras independientes

Existen alternativas **no paramétricas**, o de **libre distribución** para los casos en que el supuesto de normalidad no se satisface. Estos tests se basan generalmente en rangos o *scores* [12].

Una aproximación intuitiva al problema de comparar dos muestras de observaciones, resulta de combinar ambas muestras en una única muestra ordenada y luego asignar a cada dato su rango correspondiente; es decir, su posición en el ordenamiento, sin tener en cuenta de cuál de las muestras proviene.

Si no existieran diferencias entre los valores medios de los dos grupos y ambos tuvieran la misma cantidad de observaciones, esperaríamos que la suma de rangos de la primera muestra resultara “similar” a la suma de rangos de la segunda muestra. Esto nos indicaría que los datos de las dos muestras aparecen alternadamente en la muestra que agrupa a las dos originales.

Si proponemos como estadístico del test la suma de rangos de una de las muestras, una suma demasiado grande o demasiado pequeña conducirá a rechazar la hipótesis nula de igualdad.

Si las formas de las distribuciones de ambas muestras son similares podemos pensar que el rechazo del test se origina en las diferencias entre las posiciones centrales de ambas poblaciones. Mientras que, si por el contrario, las distribuciones no fueran similares, dichas diferencias sólo indicarían que las distribuciones son distintas.

¿En qué casos se debe preferir un test de rangos?

- ✿ Si los datos no son numéricos y corresponden a categorías ordinales, los rangos contienen la misma información que los datos.
- ✿ Si la variable es numérica, su distribución no es Normal y la muestra es pequeña, no valen los *tests* que hemos presentado anteriormente.

La teoría de métodos basados en rangos es relativamente simple y no es necesario especificar una distribución para las variables.

7.2.1 Test de Mann-Whitney-Wilcoxon

Este test tiene dos modelos posibles. Cada uno de ellos, permite testear diferentes hipótesis respecto de las poblaciones de las cuales provienen los datos.

Modelo 1

Se considera lo siguiente:

- ✿ X_1, X_2, \dots, X_{n_X} observaciones independientes de una distribución F con mediana $\tilde{\mu}_X$

- * Y_1, Y_2, \dots, Y_{n_Y} observaciones independientes de una distribución F con mediana $\tilde{\mu}_Y$

Es importante destacar que ambas muestras provienen de poblaciones con la misma distribución F , pero no es necesario especificar cuál es esta distribución. Si hay una diferencia entre ellas se debe sólo a la posición central de la distribución.

Las hipótesis para este caso son

$$\begin{cases} H_0 : \tilde{\mu}_X - \tilde{\mu}_Y = 0 \\ H_1 : \tilde{\mu}_X - \tilde{\mu}_Y \neq 0 \end{cases}$$

Modelo 2

En el caso que no pueda asumirse que ambas variables tienen la misma distribución, que no es necesario especificar, se considera entonces

- * X_1, X_2, \dots, X_{n_X} observaciones independientes de una distribución F
- * Y_1, Y_2, \dots, Y_{n_Y} observaciones independientes de una distribución G

Y las hipótesis a contrastar son

$$\begin{cases} H_0 : \forall x : F(x) = G(x) \\ H_1 : \exists x : F(x) \neq G(x) \end{cases}$$

La hipótesis nula afirma que las dos distribuciones poblacionales son iguales, lo cual sería un caso equivalente a la H_0 del Modelo 1. Mientras que la hipótesis alternativa dice que las dos distribuciones difieren de algún modo, sin indicar de qué modo [48].

Estadístico de Contraste

En ambos modelos, se toma como estadístico de contraste a

$T =$ Suma de los rangos de la muestra con menor número de observaciones

Si estamos ante la presencia de muchos empates, no cambia el valor esperado de la suma de los rangos pero sí cambia su varianza.

La distribución de este estadístico para tamaños de muestra pequeños está tabulada. Cuando ambos tamaños muestrales son grandes (> 30) se usa la distribución asintótica del estadístico, que es la distribución Normal.

Ejemplo 7.5. En un estudio efectuado a fin de caracterizar la calidad y producción de aceite de oliva en la provincia de Catamarca de la República Argentina, se estudiaron dos de las variedades más conocidas. Para ello, se tomaron muestras de aceitunas de distintos ejemplares a una misma altura de copa de aproximadamente dos metros, y de todos los puntos cardinales de la misma, a efectos de evitar las variaciones debidas a la posición del fruto en la planta. Las aceitunas fueron

Arbequina	Carolea
34.5	16.4
20.1	14.8
21.8	17.8
18.2	12.3
19.5	11.9
20.2	15.5
22.5	13.4
23.9	16.0
22.1	15.8
24.2	16.2

Tabla 7.6: Aceite por variedad

secadas en estufa y se les determinó su contenido porcentual de aceite por extracción química, obteniéndose los resultados de la Tabla 7.6.



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

Aplicamos el Código 7.1 para analizar el test, siendo su salida

```
* Shapiro-Wilk normality test
data: Arbequina
W = 0.76828, p-value = 0.00596

* Shapiro-Wilk normality test
data: Carolea
W = 0.92481, p-value = 0.3989
```

```
* Wilcoxon rank sum test
data: Arbequina and Carolea
W = 100, p-value = 1.083e-05
alternative hypothesis: true location shift is not equal to 0
```

```
# Cargamos los datos
Arbequina=c(34.5,20.1,21.8,18.2,19.5,20.2,22.5,23.9,22.1,24.2)
Carolea=c(16.4,14.8,17.8,12.3,11.9,15.5,13.4,16,15.8,16.2)

shapiro.test(Arbequina) # Testea la normalidad de los datos
shapiro.test(Carolea) # Testea la normalidad de los datos

wilcox.test(Arbequina, Carolea, alternative="two.sided")
# Realiza el test de Mann-Whitney-Wilcoxon bilateral
```

Código 7.1: Código del Test de Mann-Whitney-Wilcoxon

Debido a que el p -valor de la variedad Arbequina es inferior a 0.05, no puede considerarse que la variable satisfaga el supuesto de normalidad distribucional, con lo cual no puede aplicarse un test t . Aplicamos entonces un test de Mann-Whitney-Wilcoxon bilateral, el cual rechaza la hipótesis de nulidad. Cabe ahora preguntarnos, ¿cuál es la hipótesis que podemos considerar en este caso?

En la Figura 7.8 se puede ver el *boxplot* paralelo de las dos distribuciones.

Debido a que las formas de los *boxplots* de las distribuciones son similares, se puede testear la igualdad de valores centrales o medianos. Para ello, denominamos

- * θ_1 a la mediana poblacional (posición central) del contenido de aceite de la variedad Arbequina

- * θ_2 a la mediana poblacional (posición central) del contenido de aceite de la variedad Carolea

Luego, las hipótesis a testear son

$$\begin{cases} H_0 : \theta_1 = \theta_2 \\ H_1 : \theta_1 \neq \theta_2 \end{cases}$$

Entonces la conclusión del test es que no se puede sostener la igualdad de valores medianos. ■

¿Es válido usar el test de Mann-Whitney-Wilcoxon si la distribución de las dos muestras es muy diferente?

Debemos tener en cuenta que en este caso, el test de Mann-Whitney-Wilcoxon no es un test para el parámetro de posición. Por lo tanto, si rechazamos la hipótesis nula, podemos concluir que las distribuciones difieren pero no sabemos de qué modo difieren.

En el caso en que no se puedan comparar los valores centrales mediante el test de Mann-Whitney-Wilcoxon por ser muy diferentes las distribuciones de ambas muestras, disponemos de un test que no tiene supuestos sobre las distribuciones de los grupos y que desarrollaremos en la siguiente sección.

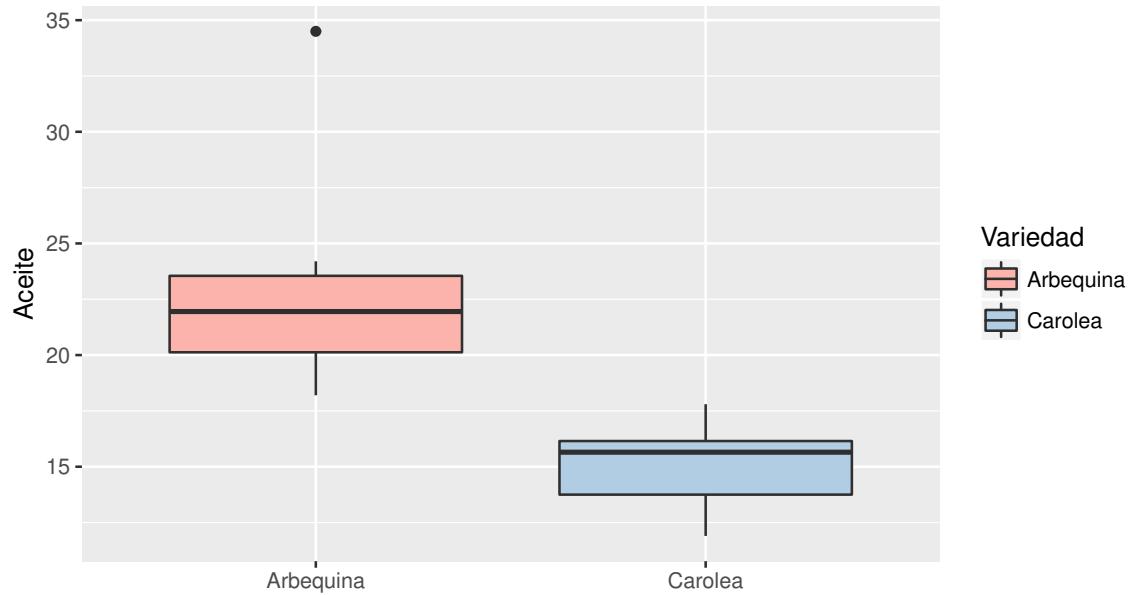


Figura 7.8: Boxplot para las distintas variedades de oliva

7.2.2 Test de la mediana

Este test posee las siguientes características:

- ✿ Puede generalizarse a más de dos grupos y resulta ser una alternativa al test de Mann-Whitney-Wilcoxon cuando interesa un test para el parámetro de posición.
- ✿ Puede aplicarse sin que se cumpla el supuesto de igualdad distribucional de las dos poblaciones.
- ✿ Puede ser usado con datos numéricos u ordinales.

El modelo consiste en:

- ✿ X_1, X_2, \dots, X_{n_X} observaciones independientes de una distribución F con mediana θ_X
- ✿ Y_1, Y_2, \dots, Y_{n_Y} observaciones independientes de una distribución G con mediana θ_Y

Las hipótesis a contrastar son

$$\begin{cases} H_0 : \theta_X = \theta_Y \\ H_1 : \theta_X \neq \theta_Y \end{cases}$$

Este test, tal como lo calculan la mayoría de los paquetes estadísticos, no acepta hipótesis alternativas unilaterales.

Para definir el estadístico, se ordenan los $n_X + n_Y$ datos y se calcula la mediana general de los datos agrupados de ambas muestras, digamos θ . Luego, se cuenta el número de observaciones menores o iguales que la mediana y el número de observaciones mayores que la mediana de cada una de las muestras. Estos datos se vuelcan a una tabla de doble entrada como la de la Tabla 7.7.

	Muestra X	Muestra Y
$\leq \theta$	m_X	m_Y
$> \theta$	M_X	M_Y
Totales	n_X	n_Y

Tabla 7.7: Tabla para el test de la mediana

Si H_0 fuera verdadera, las proporciones entre los datos menores que la mediana y los mayores que la mediana, deberían ser similares en las dos muestras; es decir, se esperaría que

$$\frac{m_X}{n_X} \cong \frac{m_Y}{n_Y} \cong \frac{M_X}{n_X} \cong \frac{M_Y}{n_Y}$$

El estadístico del test mide la distancia entre lo observado y lo esperado cuando H_0 es verdadera. Si las muestras son relativamente grandes, el estadístico tiene distribución aproximada Chi cuadrado con 1 grado de libertad, χ^2_1 [1].

Ejemplo 7.6. Aplicamos el test de la mediana para los datos del Ejemplo 7.5 usando el Código 7.2 con datos extraídos de <https://goo.gl/kkNzuB>, el cual arroja la siguiente salida

```
Mood's median test
data: Aceite by Variedad
p-value = 1.083e-05
```

```
library(RVAideMemoire)
# Paquete que contiene funciones misceláneas útiles en bioestadística
library(readxl) # Permite leer archivos xlsx

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

mood.medtest(Aceite~Variedad, data=aceite)
# Realiza el test de la mediana de Mood
```

Código 7.2: Código para aplicar el test de la mediana a los distintos tipos de oliva

Concluimos que rechazamos la hipótesis de la igualdad de las medianas de las dos variedades.

7.2.3 Tres o más grupos: análisis de la varianza de un factor

En esta sección nos concentraremos en responder la siguiente pregunta

¿Qué ocurre si se desea comparar las medias de varios grupos?

Si se realiza una comparación de a pares para k grupos, se deberían realizar $\frac{k(k-1)}{2}$ contrastes. Considerando un nivel de significación α , para cada uno de los contrastes, la probabilidad global de no cometer error de tipo I, que equivale a no cometer error de tipo I en ninguno de los contrastes, es: $1 - (1 - \alpha)^{0.5k(k-1)}$.

Por ejemplo, si el nivel de significación para cada comparación se establece en 0.05 y la cantidad de grupos es $k = 5$, el número de comparaciones es 10 y el nivel de significación global es de $1 - (1 - 0.05)^{10} = 0.4012$. Con lo cual se puede observar que la probabilidad de cometer algún error crece notablemente!

Una mejor respuesta para este problema, desarrollada por Fisher entre los años 1920 y 1930, es comparar las medias de tres o más poblaciones independientes con distribuciones normales de igual varianza. El análisis que desarrollaremos a continuación y se denomina Análisis de la varianza (ADEVA) o, en inglés, *analysis of variance* (ANOVA) [17].

Ejemplo 7.7. El té es la bebida más usual en el mundo entero después del agua, actualmente se ha difundido mucho el consumo del té verde, dado que se ha encontrado que contiene vitamina B. Recientes avances en métodos de ensayo han determinado de manera más precisa el contenido de esta vitamina.



<https://flic.kr/p/5Gq3Xm>

Consideremos los datos que se exhiben en la Tabla 7.8 acerca del contenido de vitamina B para especímenes recogidos al azar de las cuatro marcas de té verde más conocidas del mercado.

	Marca 1	Marca 2	Marca 3	Marca 4
	7.9	5.7	6.8	6.4
	6.2	7.5	7.8	7.1
	6.6	9.8	5.1	7.9
	8.6	6.1	7.4	4.5
	8.9	8.4	5.3	5.0
	10.1	7.2	6.1	4.0
	9.6			
Media	8.0500	7.4500	6.4167	5.8167
Desvío estándar	1.4680	1.5083	1.1053	1.551

Tabla 7.8: Vitamina B en el té

En primera instancia y a partir de la Figura 7.10, analicemos gráficamente si se observan diferencias importantes entre los contenidos medios de vitamina B de las distintas marcas.

Lo que deberemos investigar ahora es si estas diferencias que se aprecian visualmente en el *boxplot* comparativo son estadísticamente significativas o no.



Para generalizar lo antes expuesto, supongamos el siguiente modelo aplicado a la observación de k muestras normales independientes con varianzas iguales

$$\begin{aligned}
 \text{Muestra 1} \quad & X_{11}, X_{12}, \dots, X_{1n_1} \text{ v.a.i.i.d. } X_{1j} \sim N(\mu_1, \sigma^2) \\
 & \vdots \qquad \qquad \vdots \qquad \qquad \vdots \\
 \text{Muestra } i: \quad & X_{i1}, X_{i2}, \dots, X_{in_i} \text{ v.a.i.i.d. } X_{ij} \sim N(\mu_i, \sigma^2) \\
 & \vdots \qquad \qquad \vdots \qquad \qquad \vdots \\
 \text{Muestra } k: \quad & X_{k1}, X_{k2}, \dots, X_{kn_k} \text{ v.a.i.i.d. } X_{kj} \sim N(\mu_k, \sigma^2)
 \end{aligned}$$

donde v.a.i.i.d significa variables aleatorias independientes e idénticamente distribuidas o, equivalentemente, una muestra aleatoria.

El modelo considera

$$X_{ij} = \mu_i + \varepsilon_{ij} \text{ para } 1 \leq i \leq k, 1 \leq j \leq n_i$$

siendo $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes.

Las variables aleatorias observadas son normales, independientes entre sí dentro de las muestras y entre las muestras y homocedásticas, lo que significa que sus varianzas son iguales. Este es un

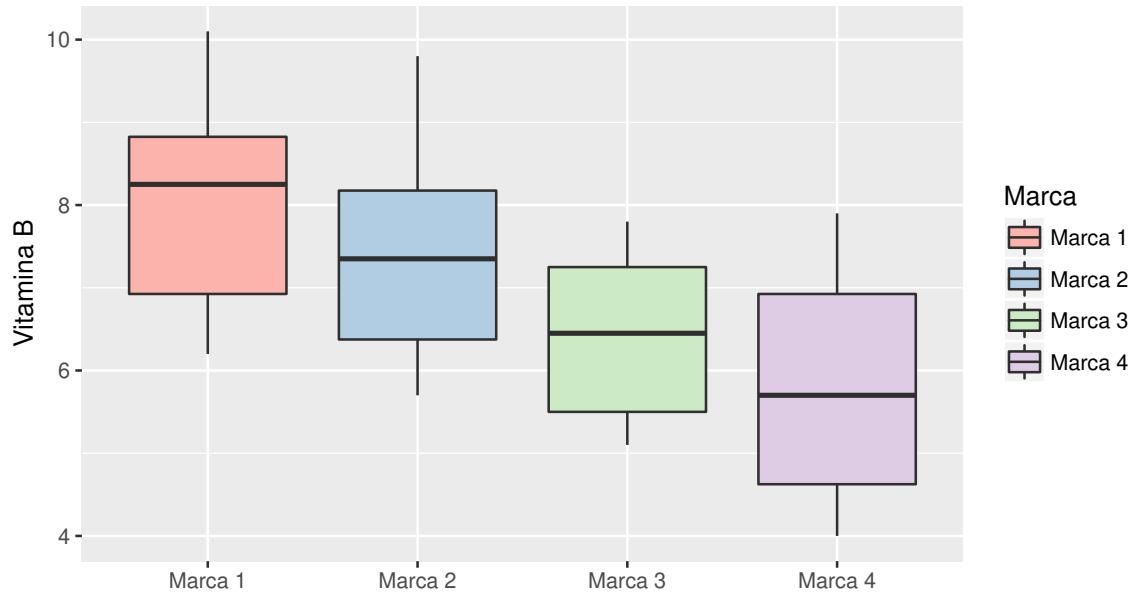


Figura 7.10: Boxplot para vitamina B en distintas marcas de té

supuesto bastante fuerte que, en caso de no satisfacerse, se deberá realizar una transformación de los datos o aplicar técnicas no paramétricas en las cuales no se supongan homocedasticidad ni normalidad.

Cuando las transformaciones disponibles no son efectivas para que los supuestos se satisfagan, veremos más adelante, una alternativa interesante conocida como el test de Kruskall-Wallis [29].

Introducimos la siguiente notación: \bar{X}_i y S_i^2 para indicar respectivamente las variables media y varianza de la i -ésima muestra, con $1 \leq i \leq k$.

Parece natural que el estimador de σ^2 se obtenga calculando un promedio ponderado de las varianzas de cada muestra s_i^2 , lo que es una generalización de la idea de la varianza amalgamada o *poolada*.

Se puede demostrar que el mejor estimador insesgado de σ^2 bajo este modelo es

$$S_P^2 = \frac{SSW}{n - k} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \dots + (n_k - 1)S_k^2}{n_1 + n_2 + \dots + n_k - k} = \frac{\sum_{i=1}^k (n_i - 1)S_i^2}{n - k}$$

donde SSW , del inglés *sum squares within*, indica la suma de cuadrados dentro de los grupos y $n = \sum_{i=1}^k n_i$.

La hipótesis a testear son

$$\begin{cases} H_0: \mu_1 = \mu_2 = \dots = \mu_k \\ H_1: \mu_i \neq \mu_j \text{ para algún par } (i, j) \end{cases}$$

La media general de todas las observaciones se calcula como

$$\bar{X}_{..} = \frac{1}{n} \sum_{i=1}^k n_i \bar{X}_i = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}$$

7.2.3.1 Descomposición de la suma de cuadrados totales

Se define la *suma de cuadrados totales* (SST del inglés *total sum of squares*) como la suma de los cuadrados de las diferencias a la media general de todas las observaciones,

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2$$

Esta suma se puede descomponer en la suma de cuadrados dentro de los grupos (SSW) y entre los grupos (SSB) como mostraremos a continuación

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} [(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})]^2$$

Desarrollando el cuadrado del binomio,

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..})$$

Puede probarse que el último sumando es nulo; utilizando

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..}) = \sum_{i=1}^k (\bar{X}_{i.} - \bar{X}_{..}) \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.}) \right]$$

Observar que la suma puede escribirse de esta forma ya que el segundo factor no depende del subíndice j , además la suma entre corchetes es la suma de los desvíos del i -ésimo grupo respecto de su media y por lo tanto, es nula.

Luego, podemos expresar a la suma de cuadrados totales como suma de cuadrados dentro y entre los grupos. Simbólicamente,

$$SST = SSW + SSB$$

7.2.3.2 Estadístico del test

El estadístico para este test es el cociente entre dos estimaciones de la varianza común de los grupos. La estimación del numerador considera la varianza entre grupos mientras que la del denominador considera la varianza dentro de los grupos. La distribución de este estadístico es F -Fisher Snedecor, por ser un cociente de variables aleatorias con distribución Chi cuadrado normalizadas por sus respectivos grados de libertad.

La suma de cuadrados entre los grupos, SSB del inglés *sum squares between* está dada por

$$SSB = \frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X}_{..})^2}{k - 1}$$

El estadístico del test se obtiene mediante el cociente $\frac{SSB}{SSW}$ y es

$$F = \frac{\frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{X}_{..})^2}{k - 1}}{S_P^2}$$

Para decidir si las medias son o no iguales en las distintas subpoblaciones, debemos aplicar un test F . Como las variables Chi cuadrado son positivas, la variable F asume solamente valores positivos también. Para tomar la decisión procedemos de la siguiente manera:

✿ **Primer paso:** se establecen la hipótesis de nulidad y la alternativa

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_k \\ H_1 : \exists(i, j) : \mu_i \neq \mu_j \end{cases}$$

✿ **Segundo paso:** se calcula el estadístico F el cual tiene distribución $F_{k-1, n-k}$ cuyos grados de libertad se corresponden con los grados de libertad de los estimadores de la varianza del numerador y del denominador.

✿ **Tercer paso:** se decide con la siguiente regla si $F_{obs} > F_{k-1, n-k, \alpha}$ entonces se rechaza H_0 con un nivel de significación α .

¿Por qué se rechaza para valores grandes del estadístico? O equivalentemente, ¿por qué se trata de una prueba unilateral derecha?

La respuesta se basa en que estamos comparando dos estimadores de la misma varianza, en el numerador utilizamos las diferencias entre las medias de los grupos y la media general, mientras que en el denominador amalgamamos las varianzas estimadas para cada subgroupo.

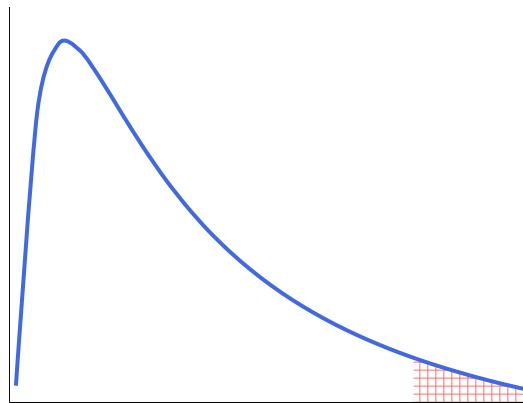


Figura 7.11: Zonas de rechazo para la prueba F

El hecho de que el numerador sea mucho mayor que el denominador indica que las medias son muy distintas entre sí.

En la Figura 7.11 se señala la región de rechazo de la prueba.

Suponiendo en primera instancia que se verifican los supuestos del modelo del análisis de la varianza para el Ejemplo 7.7, armamos la base de datos y aplicamos el test F para decidir si existen diferencias entre las medias del contenido de vitamina B en las distintas marcas de té a nivel 0.05. En el Código 7.3 con datos extraídos de <https://goo.gl/JCD1Uw>, se define el contraste y la salida del mismo se muestra en la Tabla 7.9.

```
library(readxl) # Permite leer archivos xlsx
te=read_excel("C:/.../te.xlsx")
# Importa la base con la cual se va a trabajar
te.anova=aov(VitaminaB~Marca, data=te) # Realiza el ANOVA
summary(te.anova) # Devuelve la síntesis de la prueba
```

Código 7.3: Código para ANOVA presencia de vitamina B en el té

	Grados de libertad	Suma de cuadrados	Media de cuadrados	F	$Pr(> F)$ (p -value)
Marca	3	22.93	7.645	3.791	0.0256*
Residuos	21	42.35	2.016		

Tabla 7.9: Salida de ANOVA presencia de vitamina B en el té

Donde los códigos para la significación son: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

De la Tabla 7.9, el test F rechaza la igualdad de medias a nivel 0.05. Ahora, antes de tomar una decisión, se debe estudiar si los supuestos del contraste se satisfacen con el objeto de ver si la conclusión es válida. Para ello se realiza el diagnóstico del modelo que será desarrollado en la próxima sección.

7.2.3.3 Diagnóstico del modelo

Para que el test F sea válido el modelo de k muestras normales independientes con varianzas iguales tiene que ser aproximadamente cierto. Al igual que con el test t , hay que observar los datos para detectar si existe alguna razón para pensar que este modelo es o no el adecuado.

Analicemos en primera instancia el supuesto de homocedasticidad que establece la igualdad de varianzas de los grupos. En el Ejemplo 7.7, los diagramas de caja de la Figura 7.10 aparecen a diferentes alturas pero el tamaño de las cajas se ve muy similar y tampoco se detecta la presencia de *outliers*. Por estas razones, no hay motivos para sospechar, a partir del gráfico, que no se cumple el supuesto de homocedasticidad.

Para realizar el análisis cuantitativo, existen diferentes pruebas alternativas para esta hipótesis.

7.2.3.4 Test de Bartlett

Las hipótesis a contrastar en el **test de Bartlett** son

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 = \dots = \sigma_k \\ H_1 : \exists(i, j) : \sigma_i \neq \sigma_j \end{cases}$$

En el Código 7.4 con datos disponibles en <https://goo.gl/JCD1Uw>, se aplica este test para el Ejemplo 7.7 y su salida es

```
Bartlett test of homogeneity of variances
data: te$Vitamina.B and Marca
Bartlett's K-squared = 0.6168, df = 3, p-value = 0.8926
```

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

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

te=data.frame(te)
Marca=factor(te$Marca) # Transforma en factor
bartlett.test(te$Vitamina.B, Marca) # Aplica el test de Bartlett
```

Código 7.4: Código del Test de Bartlett

El test de Bartlett no rechaza la hipótesis de nulidad; es decir, no hay evidencia estadística significativa de que la varianza de alguno de los subgrupos difiera de las otras.

El problema de este test es su sensibilidad a la falta de normalidad. Esto implica que puede ocurrir que el mismo rechace la hipótesis nula por no cumplirse el supuesto de normalidad en lugar de rechazarla por no cumplirse el supuesto de homocedasticidad.

Una alternativa más robusta, lo que significa que no es sensible a la falta de normalidad o a la presencia de algún valor atípico, la brinda el test de Levene.

7.2.3.5 Test de Levene

El **test de Levene** realiza un nuevo análisis de la varianza para los valores absolutos de los residuos de las observaciones respecto de la mediana, o la media, de su grupo.

En el Código 7.5 con datos disponibles en <https://goo.gl/JCD1Uw>, se aplica este test al Ejemplo 7.7, siendo su salida

```
Levene's Test for Homogeneity of Variance (center = median)
  Df   F value  Pr(>F)
group    3   0.2949  0.8286
          21
```

```
library(readxl) # Permite leer archivos xlsx
library(car) # Paquete con funciones que acompañan regresión aplicada

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

te=data.frame(te)
Marca=factor(te$Marca) # Transforma en factor
leveneTest(te$VitaminaB, Marca) # Aplica el test de Levene
```

Código 7.5: Código del Test de Levene

Como el p -valor de la prueba es 0.8286, no se rechaza la hipótesis de homocedasticidad. Esto significa que el test de Levene no rechaza la hipótesis nula de homocedasticidad, lo que brinda la misma conclusión que el test de Bartlett. Por lo tanto, podemos suponer que se cumple la hipótesis de homocedasticidad.

Faltaría analizar el cumplimiento del supuesto de normalidad de la distribución de los residuos, que es equivalente a analizar el supuesto de normalidad de la distribución de la variable original.

7.2.3.6 Tests de normalidad

Dentro de las herramientas conocidas, se dispone de distintos tests de normalidad así como de un gráfico que compara los cuantiles empíricos con los esperados, en el caso de que el supuesto se verifica. Este gráfico se denomina ***QQ-plot*** o **gráfico de cuantil-cuantil**.

El programa R tiene implementada una batería de tests de normalidad incluidos en la librería **nortest**. Dos de los más conocidos y potentes son el **test de Shapiro-Wilk** y el **test de Anderson-Darling**.

En el Código 7.6 con datos extraídos de <https://goo.gl/JCD1Uw>, se muestra como aplicarlos y las salidas correspondientes son:

```
* Shapiro-Wilk normality test  
data: residuals(te.anova)  
W = 0.95307, p-value = 0.2937  
  
* Anderson-Darling normality test  
data: residuals(te.anova)  
A = 0.36947, p-value = 0.3995  
  
* D'Agostino skewness test  
data: residuals(te.anova)  
skew = 0.065564, z = 0.160000, p-value = 0.8729  
alternative hypothesis: data have a skewness
```

```
library(readxl) # Permite leer archivos xlsx  
library(nortest)  
# Paquete con pruebas para probar la hipótesis compuesta de normalidad  
library(moments) # Paquete requerido para el test de D'agostino  
  
te=read_excel("C:/.../te.xlsx")  
# Importa la base con la cual se va a trabajar  
  
te.anova=aov(VitaminaB~Marca, data=te) # Realiza el ANOVA  
shapiro.test(residuals(te.anova)) # Aplica el test de Shapiro-Wilk  
ad.test(residuals(te.anova)) # Aplica el test de Anderson-Darlin  
agostino.test(residuals(te.anova)) # Aplica el test de D'Agostino
```

Código 7.6: Código para Tests de normalidad (té)

7.2.3.7 Gráficos de cuantil-cuantil

Con el Código 7.7 y datos disponibles en <https://goo.gl/JCD1Uw>, se genera el gráfico cuantil-cuantil de la Figura 7.12.

```

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

te=read_excel("C:/.../te.xlsx")
# Importa la base con la cual se va a trabajar
te=data.frame(te)
te.anova=aov(Vitamina.B~Marca, data=vitamina) # Realiza el ANOVA

y=quantile(te$Vitamina.B, c(0.25, 0.75), type=5)
# Encuentra los cuartiles 1 y 3 para la muestra
x<-qnorm(c(0.25, 0.75))
# Encuentra los cuartiles 1 y 3 para la distribución Normal
slope<-diff(y)/diff(x) # Calcula la pendiente de la recta de regresión
int<-y[1]-slope*x[1] # Calcula la constante de la recta de regresión

ggplot(te, aes(sample=residuals(te.anova))) +
stat_qq(alpha = 0.5, color="royalblue") +
xlab("Valores teóricos") +
ylab("Valores de la muestra") +
geom_abline(int=int, slope=slope, color="indianred")
# Realiza un qqplot

```

Código 7.7: Código para generar un QQ-plot de la presencia de vitamina B en el té

Si los datos fueran normales, los puntos que corresponden a las observaciones deberían posicionarse sobre la recta. Esto en la realidad no ocurrirá nunca, dado que se trata de una muestra aleatoria. Lo que debemos determinar es si el alejamiento observado de los puntos es significativo o no.

En base a las salidas de R ya estudiadas, no existe evidencia empírica en contra de la normalidad de la distribución de la variable o de los residuos. Luego, podemos dar por válido el rechazo de la hipótesis de nulidad del test F de análisis de la varianza. De esta manera, para el Ejemplo 7.7, concluimos que al menos una de las medias de los contenidos de vitamina B de las marcas de té es significativamente distinta de las demás.

¿Qué debe hacerse si los residuos no son normales o resultan heterocedásticos?

Una primera opción es transformar los datos para que se cumplan los dos supuestos. En este contexto disponemos de los lineamientos que se muestran en la Tabla 7.10 para elegir la transformación más adecuada. Analizando la relación entre la variabilidad y la media de los grupos y teniendo en cuenta la siguiente información se aplican transformaciones de Box-Cox.

En el programa R está implementado un conjunto de transformaciones de potencia denominadas de **Box y Cox**. Incluye además un gráfico que señala la potencia más adecuada.

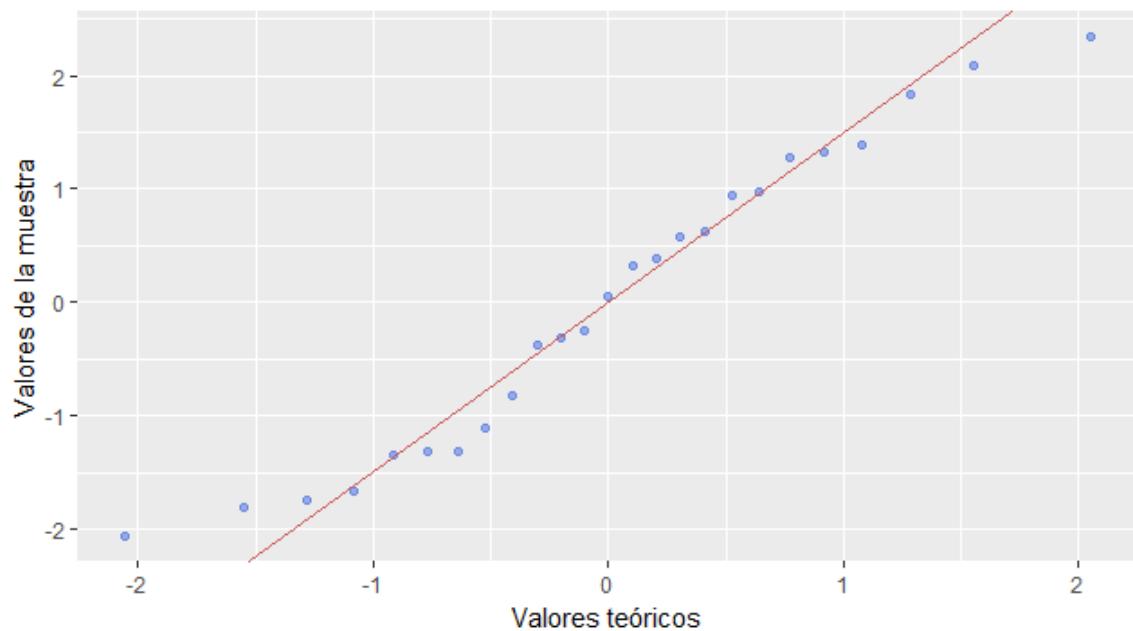


Figura 7.12: *QQ-plot* para vitamina B en distintas marcas de té

Relación	Transformación	y^p	p
σ no proporcional a μ	ninguna	y	1
σ proporcional a $\sqrt{\mu}$	raíz	\sqrt{p}	0.5
σ proporcional a μ	logaritmo	$\log(y)$	0

Tabla 7.10: Transformaciones de potencia

Cuando no se puede rechazar la hipótesis nula de ANOVA, generalmente el análisis finaliza en esa instancia. Sin embargo, cuando se rechaza, resulta lógico que el experimentador no se conforme con esta respuesta, sino que desee comparar las medias de a pares en general y de algunas otras formas en casos específicos. En el Ejemplo 7.7 se encontró evidencia en contra de la igualdad de contenido medio de vitamina B en las distintas variedades de té consideradas.

Resulta interesante cuestionarse sobre cuál o cuáles son las medias que difieren.

La próxima sección está dedicada a responder esta pregunta.

7.2.3.8 Intervalo de confianza para la diferencia de dos medias

El propósito radica ahora en comparar las medias de dos grupos, digamos i e i^* . A estas comparaciones se las conoce como **comparaciones a posteriori** o **post-hoc**.

Comenzamos construyendo un intervalo de confianza para $\mu_i - \mu_{i^*}$. El estimador puntual a considerar es $\bar{X}_i - \bar{X}_{i^*}$.

¿Cuál es su varianza? ¿Cómo se estima la misma?

Ya hemos visto que en el caso de normalidad, homocedasticidad e independencia de las variables, el intervalo de confianza de nivel $1 - \alpha$ para la diferencia de medias es

$$\left[\bar{X}_i - \bar{X}_{i^*} - t_{n-k, 1-\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_i} + \frac{1}{n_{i^*}}}, \bar{X}_i - \bar{X}_{i^*} + t_{n-k, 1-\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_i} + \frac{1}{n_{i^*}}} \right]$$

A partir de este intervalo, podemos deducir un test para estudiar las siguientes hipótesis

$$\begin{cases} H_0 : \mu_i = \mu_{i^*} \\ H_1 : \mu_i \neq \mu_{i^*} \end{cases}$$

El problema de este intervalo o de este test es que tiene nivel $1 - \alpha$ para la comparación de un par pero deja de tener este nivel cuando se quiere comparar varios pares, como ya hemos observado en la introducción del análisis de la varianza. Es debido a ello que cuando uno planea de antemano utilizar uno o muy pocos intervalos o tests, se puede usar intervalos de a pares elevando adecuadamente el nivel de confianza de los mismos, aunque en caso contrario, conviene emplear un método para **intervalos de confianza simultáneos**.

En este sentido, se dispone de muchas alternativas de comparación a nivel global. Podemos citar por ejemplo las de Dunnet, Newman–Keuls, Tukey o LSD (*least significant difference*). Todos estos procedimientos involucran el cálculo de un valor crítico que es luego comparado con las diferencias entre pares de los promedios muestrales. Si el valor crítico es mayor que la diferencia

entre los promedios de dos subpoblaciones, significa que los valores medios de esos dos grupos no son significativamente distintos.

Tradicionalmente, las comparaciones múltiples se realizan al mismo nivel de significación que el ANOVA. Sin embargo, esto puede variar según la necesidad del estudio considerado.

Algunas alternativas de intervalos simultáneos se exhiben en la Tabla 7.11 donde SCE indica la suma de los cuadrados estimados.

Prueba	Fórmula	Tabla
LSD	$T \sqrt{2SCE/n}$	t -Student
Dunnet	$D \sqrt{2SCE/n}$	D -Dunnet
Tukey HSD	$Q \sqrt{SCE/n}$	Q -Tukey
Newman-Keuls	$Q_{max} \sqrt{SCE/n}$	Q_{max} -Newman-Keuls
Duncan	$Q_0 \sqrt{SCE/n}$	Q_0 -Duncan
Sheffe	$S \sqrt{SCE/n}$	$S = \sqrt{(n-1)F}$

Tabla 7.11: Comparaciones múltiples

Básicamente, todas las pruebas son mejoras de la prueba original de t de Student. La prueba de Dunnet se emplea cuando el interés es comparar todos los competidores contra uno original, como por ejemplo, todas las mejoras de tratamiento al tradicional. La prueba LSD debe emplearse sólo si se desean unas pocas comparaciones establecidas a priori antes de realizar el análisis de la varianza). Mientras que las pruebas de Scheffe y de Tukey están diseñadas para comparar todos los pares de medias.

Ejemplo 7.8. En [6], Badimon y colaboradores llevaron a cabo un estudio a fin de determinar el efecto de la fracción lipoproteica HDL-VHDL sobre lesiones ateroscleróticas en conejos. Para ello, en Nueva Zelanda se escogieron 24 conejos que fueron asignados aleatoriamente y, en forma balanceada, a una de las siguientes dietas aterogénicas que consisten de un conjunto de alteraciones con el fin de generar un depósito de lípidos en la pared de las arterias, que finalmente se transformará en una placa de calcificación y facilitará la pérdida de elasticidad arterial y otros trastornos vasculares.

- ✿ **Dieta 1:** 60 días de dieta rica en colesterol 0.5%.
- ✿ **Dieta 2:** 90 días de dieta rica en colesterol 0.5%.
- ✿ **Dieta 3:** 90 días de dieta rica en colesterol 0.5% y luego 30 días con 50 mg de fracción lipoproteica HDL-VHDL por semana.



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

Luego del experimento, los animales fueron sacrificados. En todos los casos se comprobaron lesiones aterogénicas en la arteria aorta. Se midió el contenido de colesterol en la aorta en mg/g obteniéndose los resultados de la Tabla 7.12.

Dieta 1	Dieta 2	Dieta 3
13.4	10.4	7.5
11.0	14.2	7.2
15.3	20.5	6.7
16.7	19.6	7.6
13.4	18.5	11.2
20.1	24.0	9.6
13.6	23.4	6.8
18.3	13.6	8.5

Tabla 7.12: Colesterol en conejos

El modelo que vamos a aplicar es

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

donde $\varepsilon_{ij} \sim N(0, \sigma^2)$ para $1 \leq i \leq 3$ y $j \leq 8$.

Calculamos la media y el desvío del contenido de colesterol (variable dependiente) en la aorta correspondiente a cada una de las dietas, ver Tabla 7.13 para los resultados y Código 7.8 para la generación de los mismos.

Realizamos un *boxplot* (ver Código 7.8) para apreciar gráficamente si existen diferencias entre los contenidos medios de colesterol en la aorta de las dietas y, también ver si hay presencia de *outliers* en las distribuciones o asimetrías y si tiene sentido pensar que las varianzas son iguales.

En la Figura 7.14 se aprecia que las varianzas no parecen ser similares. No se observan *outliers* en ninguno de los diagramas de caja. Ahora, para comprobar si estas sospechas tienen significación estadística, vamos a ensayar la prueba de Levene (ver Código 7.8). Se obtienen las siguientes salidas.

Tratamiento	Media	Desviación típica	<i>n</i>
Dieta 1	15.225	2.9894	8
Dieta 2	18.025	4.8664	8
Dieta 3	8.138	1.5620	8
Totales	13.796	5.3608	24

Tabla 7.13: Resúmenes para datos del colesterol en conejos

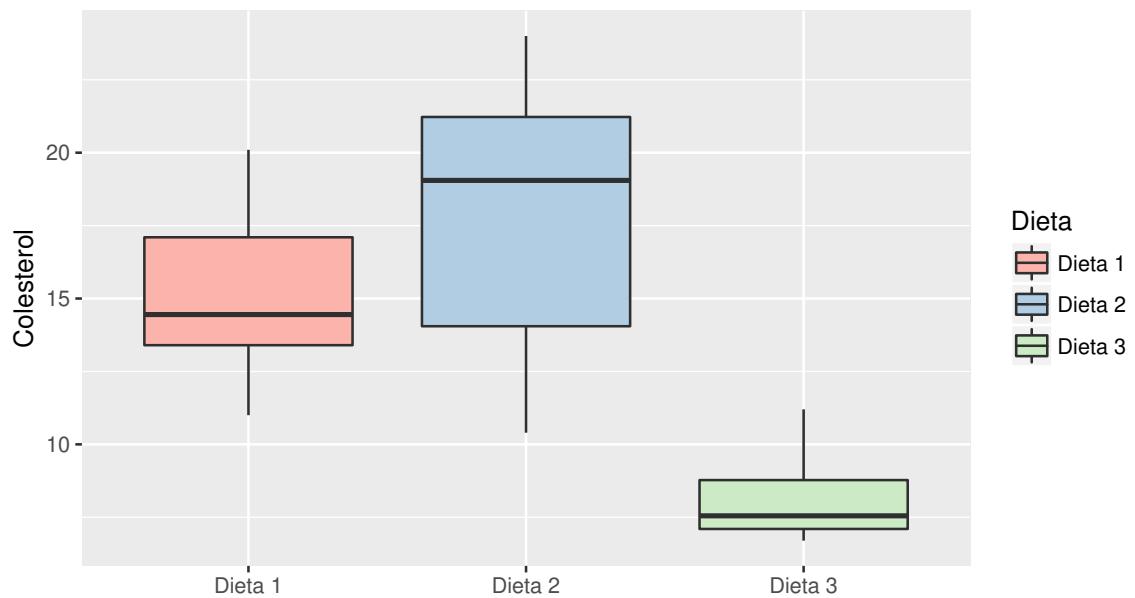


Figura 7.14: Boxplot comparativo para las distintas dietas

- ✿ El resumen es

```
Df   Sum Sq  Mean Sq  F value    Pr(>F)
Dieta      2    415.6   207.78    17.78  3.03e-05      ***
Residuals  21    245.4    11.69
- - -
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Esta salida indica que las diferencias entre los diámetros aórticos de los conejos sometidos a las distintas dietas son diferentes. Sin embargo, estos resultados sólo serán válidos si se satisfacen los supuestos del modelo de análisis de la varianza.

- ✿ La prueba de Shapiro-Wilk produce la siguiente salida

```
Shapiro-Wilk normality test
data: residuals(colesterol.anova)
W = 0.97939, p-value = 0.8843
```

Esta salida indica que puede sostenerse el supuesto de normalidad distribucional de los residuos.

- ✿ La prueba de Levene produce la siguiente salida

```
Levene's Test for Homogeneity of Variance (center = median)
Df   F value    Pr(>F)
group      2     3.639  0.04396
             21
- - -
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

De esta última salida se interpreta que no es posible suponer homocedasticidad en la distribución de los residuos.

Intentemos entonces una transformación para la variable respuesta. Para decidir el exponente de la transformación aplicamos el test de Box & Cox (ver Código 7.8). La salida de este test es la Figura 7.15, la cual sugiere una transformación de la variable respuesta con un exponente cercano a -0.5 .

Realizamos la transformación sugerida y un nuevo análisis de la varianza, que origina las siguientes salidas (nuevamente referimos al Código 7.8).

- ✿ El resumen es

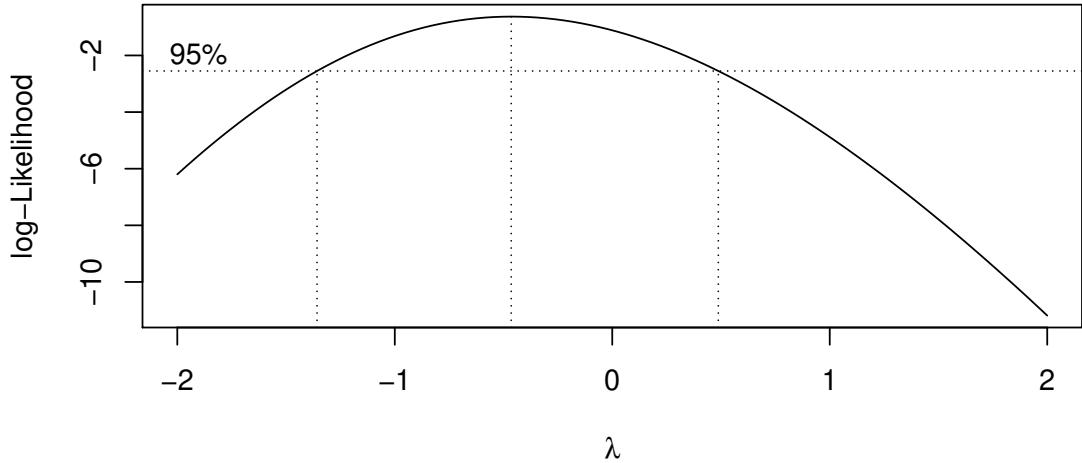


Figura 7.15: Salida del test de Box & Cox para el colesterol en conejos

	Df	Sum Sq	Mean Sq	F value	Pr(>F)		
Dieta	2	0.05841	0.029207	29.75	7.45e-07	***	
Residuals	21	0.02062	0.000982				

Signif. codes:	0	'***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

En esta salida se aprecia que las diferencias siguen siendo significativas aún con los datos transformados.

- * Hacemos el análisis diagnóstico del modelo para los nuevos datos.

Shapiro-Wilk normality test

```
data: residuals(tcolesterol.anova)
W = 0.97902, p-value = 0.8771
```

Esto indica que se cumple el supuesto de normalidad requerido para el modelo de análisis de la varianza para los datos transformados.

- * A partir de la siguiente salida, vemos que también se cumple el supuesto de homocedasticidad.

```

Levene's Test for Homogeneity of Variance (center = median)
    Df   F value           Pr(>F)
group    2   0.2626          0.7715
                21

```

Nos preguntamos por último, cuáles son las dietas que difieren entre sí. Utilizamos los intervalos de confianza simultáneos para las diferencias de medias de Tukey (ver Código 7.8 con datos extraídos de <https://goo.gl/sRR9yt>) y obtenemos la siguiente salida

```

Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = Colesterol~Dieta, data = conejos)
$Dieta
      diff      lwr      upr      p adj
Dieta 2-Dieta 1 -0.01743725 -0.05692446  0.02204997 0.5168243
Dieta 3-Dieta 1  0.09484116  0.05535394  0.13432838 0.00000151
Dieta 3-Dieta 2  0.11227841  0.07279119  0.15176563 0.0000013

```

De donde se puede apreciar que la Dieta 3 produce niveles de colesterol inferiores a los de las otras dos dietas. Más aún, los dos últimos intervalos no contienen al 0, lo cual indica que la Dieta 3 es diferente de las Dietas 1 y 2.

```

library(readxl) # Permite leer archivos xlsx
library(ggplot2) # Paquete para confeccionar dibujos
library(car) # Paquete con funciones que acompañan regresión aplicada
library(MASS)
# Paquete con funciones y bases de datos para la librería de Venables y Ripley

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

pordieta=split(conejos$Colesterol,conejos$Dieta) # Separa los datos según dieta
lapply(pordieta,mean) # Calcula las medias
lapply(pordieta,sd) # Calcula los desvíos estándar

ggplot(conejos, aes(x=Dieta, y=Colesterol, fill=Dieta)) +
  geom_boxplot() +
  xlab("") +
  scale_fill_brewer(palette="Pastell1")
# Produce boxplots

attach(conejos)

colesterol.anova=aov(Colesterol~Dieta, data=conejos) # Realiza el ANOVA
summary(colesterol.anova)

```

```

# Sirve para analizar si las diferencias son significativas
shapiro.test(residuals(colesterol.anova))
# Testea la normalidad de los residuos del modelo
leveneTest(Colesterol~Dieta, data=conejos)
# Testea el supuesto de homocedasticidad

boxcox(Colesterol~Dieta, plotit=T)
# Investiga qué transformación deja los datos más próximos a la normalidad

tcolesterol.anova=aov(Colesterol^(-0.5)~Dieta, data=conejos)
# Realiza el ANOVA para los datos transformados
summary(tcolesterol.anova)
# Analiza si las diferencias observadas siguen siendo significativas
shapiro.test(residuals(tcolesterol.anova))
# Testea la normalidad de los residuos
leveneTest(Colesterol^(-0.5)~Dieta, data=conejos)
# Testea el supuesto de homocedasticidad
TukeyHSD(tcolesterol.anova, conf.level=0.95)
# Realiza las comparaciones múltiples a posteriori entre los valores medios

```

Código 7.8: Código para el análisis del colesterol en conejos

¿Qué sucede si no se verifican los supuestos del análisis de la varianza ni para los datos originales ni para los datos transformados?

La alternativa en este caso son las pruebas no paramétricas. Siendo las más usadas para ANOVA, la prueba de la mediana y la prueba de Kruskal-Wallis, también conocida como **análisis de la varianza no paramétrico**. De estas dos pruebas, la más potente resulta ser la de Kruskal-Wallis siendo una generalización del test de Wilcoxon de rangos signados que ya hemos presentado.

7.2.3.9 Test de Kruskal-Wallis no paramétrico para muestras independientes

Esta prueba contrasta la hipótesis nula que establece que las k muestras independientes proceden de la misma población y, en particular, todas ellas tienen la misma posición central. La misma se basa en los rangos de las observaciones y no requiere el cumplimiento del supuesto de normalidad ni del supuesto de homocedasticidad.

El modelo que supone este test consiste en

$$\begin{aligned}
 \text{Población 1: } & Y_{11}, Y_{12}, \dots, Y_{1n_1} && \text{v.a.i.i.d. con escala al menos ordinal} \\
 \text{Población 2: } & Y_{21}, Y_{22}, \dots, Y_{2n_2} && \text{v.a.i.i.d. con escala al menos ordinal} \\
 & \vdots && \vdots \\
 \text{Población } k: & Y_{k1}, Y_{k2}, \dots, Y_{kn_k} && \text{v.a.i.i.d. con escala al menos ordinal}
 \end{aligned}$$

donde las variables de las k poblaciones también son independientes entre sí.

Las distribuciones de todas las subpoblaciones deben ser semejantes, de lo contrario, el rechazo de la hipótesis de nulidad implicaría que las distribuciones son distintas y no que sus medianas difieren, al igual que en la prueba de Wilcoxon-Mann-Whitney.

Las hipótesis a contrastar son

$$\begin{cases} H_0 : \theta_1 = \theta_2 = \cdots = \theta_k \\ H_1 : \exists(i,j) : \theta_i \neq \theta_j \end{cases}$$

El estadístico de contraste para esta prueba, cuando hay pocos o ningún empates, es

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_{i\cdot}^2}{n_i} - 3(N+1)$$

donde N es el total de observaciones y $R_{i\cdot}$ es la suma de los rangos de la muestra i , dados por R_{ij} , el rango en la distribución conjunta de la observación j del grupo i . Bajo H_0 , este estadístico tiene distribución aproximada Chi cuadrado con $k - 1$ grados de libertad. Por este motivo, la regla de decisión resulta:

- ✿ se rechaza H_0 cuando $H_{obs} > \chi_{k-1,1-\alpha}^2$.
- ✿ no se rechaza H_0 cuando $H_{obs} < \chi_{k-1,1-\alpha}^2$.

Podemos describir el procedimiento a partir de los siguientes pasos.

- ✿ Se ordenan todas las observaciones en sentido creciente y se reemplazan por su rango R_{ij} ($i = 1, \dots, k, j = 1, \dots, n_i$), en la muestra conjunta ordenada.
- ✿ En caso de empates, se asigna a cada una de las observaciones empatadas el rango promedio de ellas.
- ✿ Se calcula la suma de los rangos de cada grupo de observaciones. La suma de los rangos en la muestra combinada del i -ésimo grupo se designa con $R_{i\cdot}$ y el rango promedio del i -ésimo grupo se denota con $\bar{R}_{i\cdot}$.
- ✿ Se calcula el estadístico de contraste H .
- ✿ Se toma una decisión y se brinda la conclusión.

Cabe aclarar que decir “las poblaciones tienen la misma posición central” es equivalente a decir que tienen “el mismo valor esperado o media aritmética de los rangos”, o que “las poblaciones tienen igual mediana”.

Ejemplo 7.9. Se realizó una intervención educativa innovadora para mejorar el rendimiento de los estudiantes. Dentro de los grupos de clasificación, el A es el grupo de control y los restantes, B y C, son los grupos con distintas innovaciones.



<https://flic.kr/p/8tXgM4>

Se evaluó a los alumnos mediante una prueba objetivo sobre un total de 60 puntos. Las puntuaciones logradas por los alumnos se presentan en la Tabla 7.14.

Grupo	Puntajes						
A	13	27	26	22	28	27	
B	43	35	47	32	31	37	
C	33	33	33	26	44	33	54

Tabla 7.14: Calificaciones según los grupos

En la Figura 7.17 (ver Código 7.9 para su generación) se grafican los datos correspondientes a las tres distribuciones de los valores observados.

En la Tabla 7.15 se muestran los resultados de aplicar el test de Shapiro-Wilk a cada uno de los grupos. El código en R se muestra en 7.9.

Datos	W	p-valor
Grupo A	0.76163	0.02583
Grupo B	0.92057	0.5095
Grupo C	0.82769	0.07607

Tabla 7.15: Prueba de normalidad de los datos

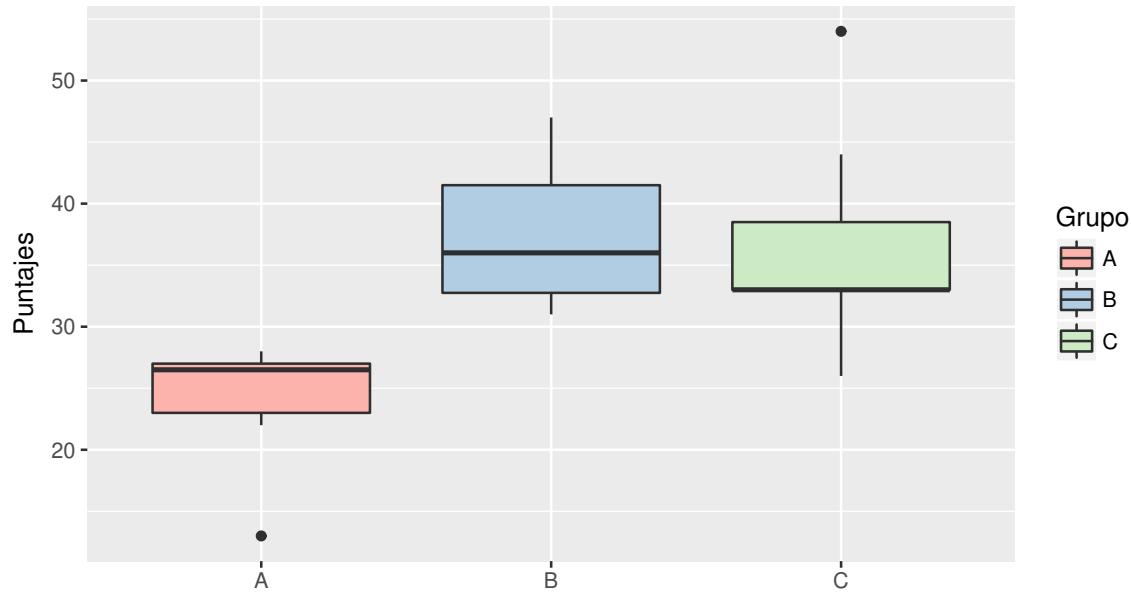


Figura 7.17: Distribución del rendimiento por grupo

Según los resultados de la Tabla 7.15, se rechaza la hipótesis de normalidad para la distribución de los recuentos en los puntajes correspondientes al grupo A. Por ende, aplicamos un análisis no paramétrico mediante el test de Kruskal-Wallis. Para ello planteamos las hipótesis

$$\begin{cases} H_0 : \text{los tres grupos tienen la misma posición para la variable de estudio dada por el puntaje} \\ H_1 : \text{al menos un grupo tiene diferente posición para la variable en estudio dada por el puntaje} \end{cases}$$

En la Tabla 7.16 se muestran los datos ordenados los datos y sus rankeamientos.

Establecemos la regla de decisión como se rechazamos H_0 si $H_{obs} > \chi^2_{2,0.95} = 5.99$, siendo el estadístico de contraste de nuestra prueba

$$H_{obs} = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{19(19+1)} \left(\frac{24.5^2}{6} + \frac{80^2}{6} + \frac{85.5^2}{7} \right) - 3(19+1) = 9.92$$

Por lo tanto, la decisión es rechazar H_0 debido a que $9.92 > 5.99$. Luego, no puede suponerse que la distribución de los rendimientos de las innovaciones sean iguales.

Con el análisis realizado hasta el momento, no se puede inferir acerca de la mediana ya que, como puede apreciarse en la Figura 7.17, las distribuciones en los distintos grupos no son similares.

Aplicando el test de Kruskal-Wallis de la suma de los rangos y, luego del mismo, el test de comparaciones múltiples, según el Código 7.9, obtenemos las siguientes salidas.

Puntaje	Grupo	Rango	Rango por grupo
13	A	1	
22	A	2	
26	A	3.5	
27	A	5.5	
27	A	5.5	
28	A	7	24.5
31	B	8	
32	B	9	
35	B	14	
37	B	15	
43	B	16	
47	B	18	80
26	C	3.5	
33	C	11.5	
44	C	17	
54	C	19	85.5

Tabla 7.16: Datos de los puntajes ordenados y rankeados

- ✿ Para la prueba de Kruskal-Wallis

```
Kruskal-Wallis rank sum test
data: Puntajes and Grupo
Kruskal-Wallis chi-squared = 9.9265, df = 2, p-value = 0.00699
```

- ✿ Para la prueba de comparaciones múltiples

Multiple comparison test after Kruskal-Wallis			
		p.value:	0.05
Comparisons	obs.dif	critical.dif	difference
A-B	9.250000	7.777876	TRUE
A-C	8.130952	7.494949	TRUE
B-C	1.119048	7.494949	FALSE

A partir de esta última salida, surge que las diferencias de las distribuciones son estadísticamente significativas y que el grupo A difiere significativamente de los grupos B y C, mientras que los grupos B y C no difieren significativamente entre sí.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(pgirmess)
# Paquete con herramientas para la lectura, escritura y transformación de datos

Puntajes=c(13,27,26,22,28,27,43,35,47,32,31,37,33,33,33,26,44,33,54)
Grupo=as.factor(c(rep("A",6), rep("B",6), rep("C",7)))
Rendimiento=data.frame(Grupo, Puntajes)
# Carga la base de datos

ggplot(Rendimiento, aes(x=Grupo, y=Puntajes, fill=Grupo)) +
  geom_boxplot() +
  xlab("") +
  scale_fill_brewer(palette="Pastell1")
# Produce boxplots

grupoA=Rendimiento[Rendimiento$Grupo=="A",2]
grupoB=Rendimiento[Rendimiento$Grupo=="B",2]
grupoC=Rendimiento[Rendimiento$Grupo=="C",2]
shapiro.test(grupoA)
shapiro.test(grupoB)
shapiro.test(grupoC)
# Aplica el test de Shapiro-Wilk a cada grupo

kruskal.test(Puntajes, Grupo)
# Realiza el test de Kruskal-Wallis
kruskalmc(Puntajes~Grupo)
```

```
# Realiza un test de comparación múltiple entre tratamientos luego del test de  
# Kruskal-Wallis
```

Código 7.9: Código para el análisis de innovaciones en la educación



7.3 Ejercitación

Ejercicio 1.

Se quiere comparar el tiempo que tardan en reparar computadoras dos conjuntos de técnicos. Para ello se seleccionan al azar los tiempos en fracción de jornada laboral de un grupo de 20 técnicos de dos sucursales distintas de cierta empresa. Los datos registrados se encuentran en la Tabla 7.17.

Tiempos 1	Tiempos 2		
0.17	0.21	0.18	0.20
0.26	0.22	0.33	0.30
0.19	0.28	0.23	0.32
0.34	0.25	0.16	0.20
0.52	0.90	0.19	0.19
0.33	0.33	0.30	0.22
0.23	0.22	0.21	0.27
0.20	0.17	0.20	0.24
0.18	0.39	0.16	0.29
0.22	0.27	0.21	0.27

Tabla 7.17: Tiempos de reparación según grupos de técnicos

1. ¿Satisfacen los datos el supuesto de normalidad? Si la respuesta es afirmativa, aplicar un test basado en este supuesto para analizar si los tiempos de los dos grupos son iguales o no. Si la respuesta es negativa, realizar una transformación de Box & Cox para normalizar los datos y aplicar la prueba a los datos transformados.
2. Aplicar una prueba no paramétrica y comparar los resultados con los obtenidos en el ítem anterior.

Ejercicio 2.

Un investigador estudió el contenido en sodio de marcas de cerveza comercializadas en Capital Federal y Gran Buenos Aires. Para ello, se seleccionaron las seis marcas más prestigiosas del mercado y se eligieron botellas o latas de 500 ml para cada marca seleccionada y se midió el contenido en sodio en miligramos de cada una de ellas. Los resultados de este muestreo son los de la Tabla 7.18.

1. Graficar la variable observada en cada grupo y analizar la presencia de *outliers*, la igualdad gráfica de las medias y las formas de las distribuciones.

Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6
24.4	10.2	19.2	17.4	13.4	21.3
22.6	12.1	19.4	18.1	15	20.2
23.8	10.3	19.8	16.7	14.1	20.7
22.0	10.2	19.0	18.3	13.1	20.8
24.5	9.9	19.6	17.6	14.9	20.1
22.3	11.2	18.3	17.5	15.0	18.8
25.0	12.0	20.0	18.0	13.4	21.1
24.5	9.5	19.4	16.4	14.8	20.3

Tabla 7.18: Cantidad de sodio por marca de cerveza

2. Calcular la media y el desvío de cada uno de los grupos. ¿Es posible que se satisfaga el supuesto de homogeneidad?
3. Establecer las hipótesis estadísticas de interés.
4. Contrastar las hipótesis con un nivel $\alpha = 0.05$.
5. Verificar el cumplimiento de los supuestos de normalidad y homocedasticidad. Si se verifican estos supuestos, concluir en el contexto del problema.

Ejercicio 3.

Para comparar cuatro suplementos de engorde en bovinos para carne, se seleccionaron al azar, cuarenta animales *Hereford* de iguales edades y sexos y de pesos homogéneos, para ser usados en un experimento. Los suplementos analizados poseen las siguientes características:

S1: suplemento constituido por grano partido y fuente A

S2: suplemento constituido por grano partido y fuente B

S3: suplemento constituido por grano entero y fuente A

S4: suplemento constituido por grano entero y fuente B.

Se asignaron aleatoriamente 10 animales por suplemento, los que fueron alimentados individualmente con una dieta estándar más el correspondiente suplemento durante 80 días. La variable en estudio o variable respuesta fue la eficiencia de conversión individual, en kilogramo de materia seca por kilogramo de ganancia de peso, y cuyos registros se presentan en la Tabla 7.19.

1. Realizar un análisis gráfico y descriptivo de la eficiencia de conversión lograda por los distintos suplementos.

	S1	S2	S3	S4
	3.3	4.6	6.7	6.3
	4.4	4.5	5.8	6
	4.9	5.0	5.0	6.7
	4.9	4.0	4.8	5.5
	3.9	4.5	5.3	6.6
	4.2	5.2	6.2	6.1
	4.7	4.9	5.0	5.3
	5.1	5.5	6.4	6.5
	4.6	4.8	5.9	6.3
	4.5	5.3	5.4	6.8

Tabla 7.19: Eficiencia de conversión según suplemento

2. Establecer las hipótesis de interés del problema y explicitar los supuestos necesarios.
3. Testear las hipótesis con nivel de significación del 5%.
4. Analizar el cumplimiento de los supuestos del modelo.
5. Concluir en términos del problema y en caso de rechazar H_0 , indicar cuáles medias son diferentes, utilizando para ello las comparaciones a posteriori de Tukey.

Ejercicio 4.

Se desea estudiar el efecto de una nueva droga analgésica para uso farmacéutico en pacientes con neuralgia crónica. Con tal fin, se la compara con la aspirina y con un placebo. En 30 pacientes elegidos al azar, se utiliza el método del doble ciego, asignando al azar 10 pacientes a cada tratamiento. La variable aleatoria observada está dada por el número de horas en que el paciente está libre de dolor después de haber sido medicado. Los resultados obtenidos se muestran en la Tabla 7.20.

	Media	Desvío
Placebo	2.50	0.13
Aspirina	2.82	0.20
Droga	3.20	0.17

Tabla 7.20: Comparación nuevo analgésico

Se tienen los p -valores de la Prueba de Levene ($p = 0.18$) y de la Prueba de Shapiro-Wilks ($p = 0.24$) de los residuos del modelo. Se pide lo siguiente.

1. Identificar la variable dependiente y el factor de interés.
2. Escribir el modelo en general y en términos del problema.
3. Analizar los resultados de las pruebas de hipótesis para los supuestos del modelo.
4. Plantear las hipótesis y construir la tabla de ANOVA sabiendo que $SC_{error} = \sum_{i=1}^k (n_i - 1)s_i^2$.
5. Comparara los tratamientos utilizando un test t con nivel global 0.05; es decir, como son 3 comparaciones, $\alpha = 0.05/3$ para cada una.
6. Adicionalmente, se indagó a los pacientes sobre efectos colaterales gástricos como respuesta al tratamiento. Los encuestados respondieron según una escala entre 0 y 5, donde 0 indica nunca y 5 siempre. Los resultados obtenidos se muestran en la Tabla 7.21.

Placebo	0	3	2	3	4	2	2	3	1	1
Aspirina	1	4	3	0	2	3	4	5	2	3
Droga	4	5	4	2	3	4	1	5	3	0

Tabla 7.21: Efectos gástricos colaterales

- (a) ¿Los investigadores deberían utilizar la misma prueba estadística que la empleada para comparar el tiempo libre de dolor? Justificar.
- (b) ¿Cuáles son las conclusiones de este estudio?

Ejercicio 5.

Se está estudiando el tiempo de cocción de un alimento antes de lanzarlo al mercado. Se han formado cuatro grupos y se les ha pedido que midan el tiempo transcurrido hasta que, según su juicio, el alimento quede a punto. Debido a que esta sensación es subjetiva, se usa un ANOVA para estimar la varianza que presenta el experimento. Todos los grupos usan fuentes de calor y utensilios similares. Si la Tabla 7.22 recoge los resultados redondeados en minutos, ¿qué estimación podría hacerse de la varianza de la población de estos alimentos? ¿Se observan diferencias entre los grupos?

1. Graficar los tiempos de cocción por tratamiento y calcular las medidas resumen de los mismos.
2. Establecer las hipótesis de interés y escribir el modelo detallando los supuestos.

Grupo A	Grupo B	Grupo C	Grupo D
25	121	81	25
36	36	81	25
36	36	36	36
25	64	9	25
36	36	25	36
16	81	36	25
25	49	9	25
36	25	49	25
49	64	169	25
36	49	1	25
25	121	81	25

Tabla 7.22: Tiempo de cocción por grupo

3. Realizar la prueba y el diagnóstico correspondiente. ¿Son válidos los resultados de la prueba? Si la respuesta es afirmativa, concluir en el contexto del problema. En caso contrario, intentar una transformación de potencia conveniente para normalizar y/o homocedastizar la variable respuesta.
4. Realizar nuevamente la prueba, en caso de ser necesario, y el diagnóstico del modelo correspondiente, concluyendo en términos del problema.
5. Comparar los resultados con los del test no paramétrico.

Ejercicio 6.

Se quiere comparar el trabajo de cuatro analistas de un laboratorio en el ensayo de determinación del porcentaje de alcohol metílico en muestras de un producto químico, mediante la técnica de cromatografía líquida de alta resolución (HPLC). Los analistas reportaron los resultados que se exhiben en la Tabla 7.23.

Analista	Porcentaje alcohol		
1	84.99	84.02	84.38
2	85.15	85.13	84.88
3	84.72	84.48	85.16
4	84.2	84.1	84.55

Tabla 7.23: Porcentajes de alcohol según analista

1. ¿Por qué no es adecuado aplicar el análisis de la varianza paramétrico en este caso?
2. Mediante la prueba no paramétrica de Kruskall-Wallis, determinar si el porcentaje de alcohol depende del analista que lo mide.

Capítulo 8

Comparación de medias en el caso multivariado

*Los números nunca mienten,
después de todo ellos implemente
cuentan diferentes historias
dependiendo de la Matemática de
los relatores.*

— Luis Alberto Urrea

La generalización a varias dimensiones de la densidad Normal univariada juega un papel fundamental en el análisis multivariado. Muchos de los fenómenos naturales del mundo real pueden ser estudiados por medio de la distribución Normal multivariada.

Incluso, a pesar de que el fenómeno estudiado no siga este modelo de distribución, las distribuciones de muchos de los estadísticos utilizados es aproximadamente Normal multivariada.

8.1 Distribución Normal univariada

Recordemos que una variable continua X tiene distribución Normal univariada con media μ y varianza σ^2 , simbólicamente $X \sim N(\mu, \sigma^2)$, cuando su función de densidad de probabilidad es

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \forall x \in \mathbb{R}$$

Algunas propiedades importantes de la distribución normal:

- ✿ Su gráfica es simétrica respecto de $x = \mu$.
- ✿ Su gráfica es asintótica respecto del eje de abscisas.

- ✿ Presenta un máximo en $x = \mu$ siendo el valor máximo de la función $\frac{1}{\sqrt{2\pi}\sigma}$.
- ✿ Presenta dos puntos de inflexión, en $x = \mu - \sigma$ y en $x = \mu + \sigma$.
- ✿ La combinación lineal de variables aleatorias Normales es otra variable aleatoria Normal.
- ✿ El área bajo la curva dentro del intervalo $(\mu + k\sigma; \mu + t\sigma)$ no depende de μ ni de σ sino de los valores reales que tomen k y t .

En las Figuras 8.1 y 8.2 se puede apreciar el efecto sobre la gráfica de la función de densidad cuando varían las medias y las varianzas respectivamente.

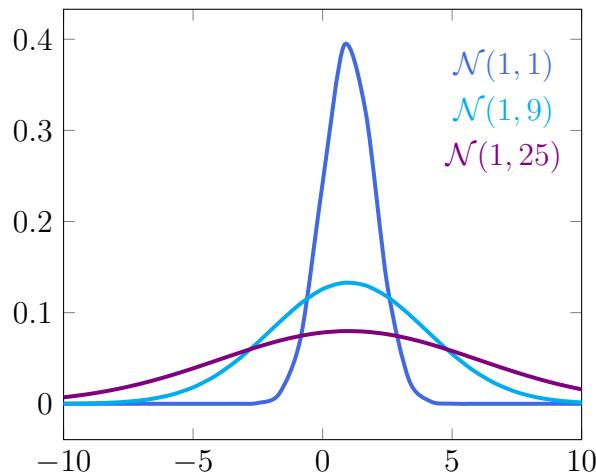


Figura 8.1: Distribución normal con varianzas distintas

8.2 Distribución Normal multivariada

Un vector aleatorio continuo $X = (X_1, \dots, X_n)^t$ tiene distribución Normal multivariada con vector de medias $\mu = (\mu_1, \dots, \mu_n)^t$ y matriz de covarianzas Σ , simbólicamente $X \sim N_n(\mu, \Sigma)$, cuando su función de densidad de probabilidad está dada por

$$f(X, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (X - \mu)^t \Sigma^{-1} (X - \mu) \right)$$

donde $-\infty < x_i < +\infty$ para todo $i = 1, \dots, n$.

Entre las propiedades importantes de esta distribución se pueden destacar las siguientes.

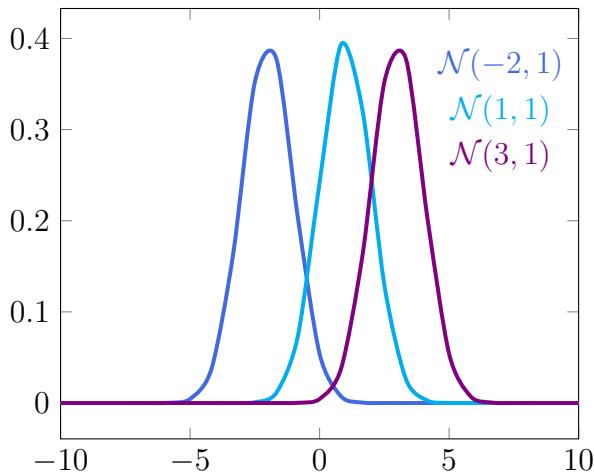


Figura 8.2: Distribución normal con medias distintas

- ✿ Es una generalización del caso univariado.
- ✿ Tiene propiedades matemáticas que la hacen muy manejable.
- ✿ Depende de un número relativamente reducido de parámetros: n para el vector de medias y para la matriz de covarianzas $\frac{n(n+1)}{2}$.
- ✿ En el caso de esta distribución, la ausencia de correlación es equivalente a independencia.
- ✿ Si bien los datos de los que disponemos rara vez siguen con exactitud una distribución Normal, esta distribución suele ser una aproximación útil.
- ✿ Al igual que en el caso univariado, esta distribución es el límite de la suma de vectores aleatorios independientes y con la misma distribución (Teorema Central del Límite).

La gráfica para el caso bivariado se puede apreciar en la Figura 8.3.

Si $X \sim N_p(\mu, \Sigma)$, entonces cualquier combinación lineal de sus coordenadas también tiene distribución Normal; es decir, si c es un vector de constantes conocidas, entonces

$$Y = cX \sim N(c\mu, c\Sigma c^t)$$

En particular para el caso de los vectores canónicos e_i , donde e_i tiene un 1 en la i -ésima coordenada y 0 en las restantes, se tiene que $e_i X$ corresponde a la coordenada X_i del vector aleatorio X ; es decir, en particular, esta coordenada también tiene distribución Normal. Por lo tanto, cada una

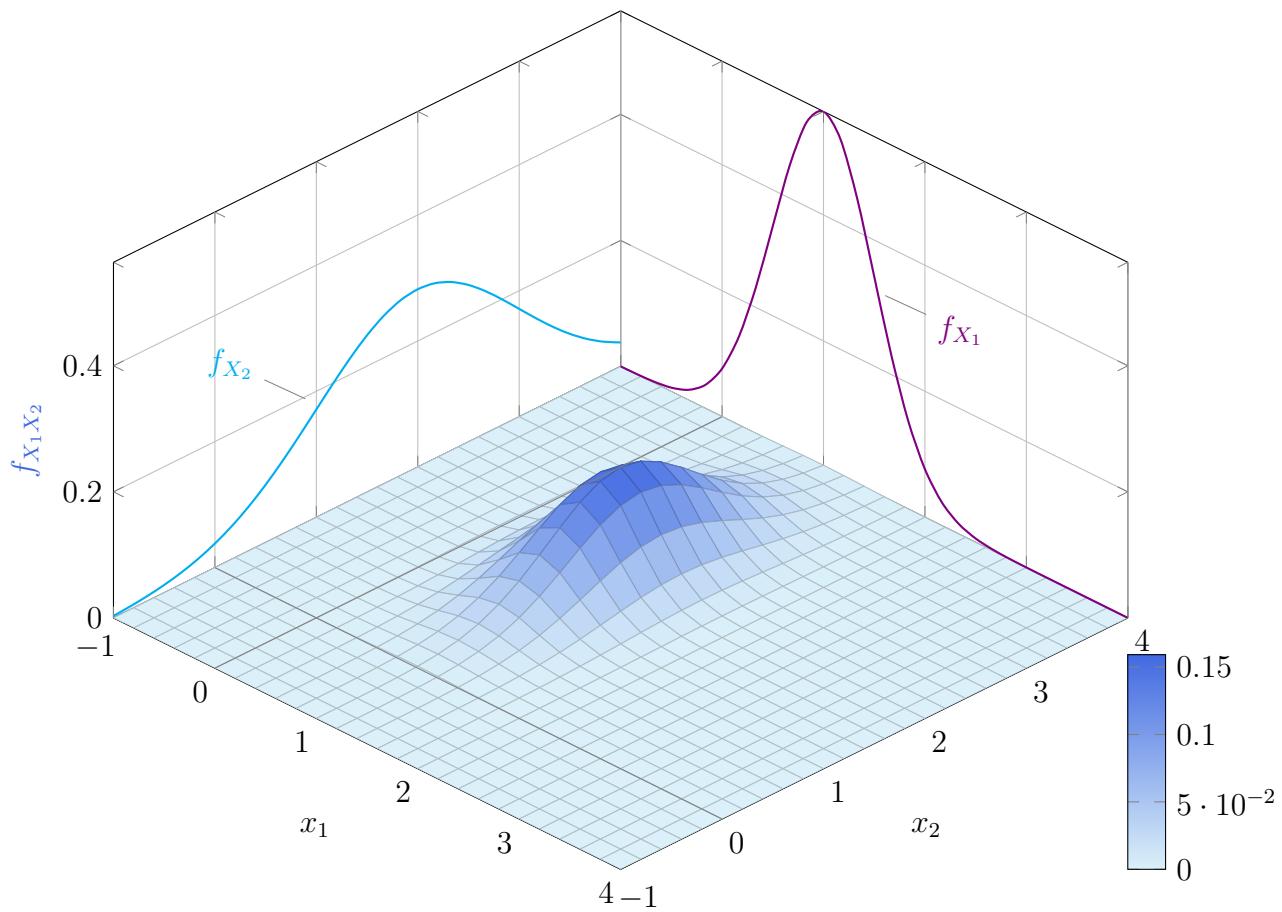


Figura 8.3: Ejemplo de distribución Normal bivariada

de las coordenadas de una distribución Normal multivariada tiene distribución Normal univariada. Equivalentemente, se puede afirmar que **las distribuciones marginales son normales**.

Esta característica permite definir a partir del vector aleatorio $X = (X_1, \dots, X_q, X_{q+1}, \dots, X_n)^t$ las siguientes variables Normales

$$Y_1 = (X_1, \dots, X_q)^t \quad \text{e} \quad Y_2 = (X_{q+1}, \dots, X_n)^t$$

Ejemplo 8.1. Consideramos el vector aleatorio $X \sim N_3(\mu, \Sigma)$ donde $\mu = (1, 2, 3)^t$ y Σ es la matriz identidad de tamaño 3×3 .

Sean las combinaciones lineales

$$\begin{cases} Y_1 &= 2X_1 + 3X_2 \\ Y_2 &= -1X_1 + 2X_3 \end{cases}$$

que pueden expresarse en notación matricial como

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Entonces, la distribución conjunta de la variable $Y = (Y_1, Y_2)^t$ es Normal multivariada con parámetros μ_Y y Σ_Y dados por:

$$\mu_Y = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 8 \\ 5 \\ 3 \end{pmatrix}$$

y

$$\Sigma_Y = \begin{pmatrix} 2 & 3 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 3 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 13 & -2 \\ -2 & 5 \end{pmatrix}$$



8.2.1 Estimadores de máxima verosimilitud

La **función de verosimilitud** de una distribución es la función de distribución conjunta considerada como función de los parámetros en lugar de como función de las variables de la muestra. La **estimación de máxima verosimilitud** es la que se obtiene maximizando la función de verosimilitud; es decir, calculando los valores de los parámetros que la hacen máxima para la muestra disponible.

En nuestro caso particular, los estimadores de máxima verosimilitud para la distribución Normal multivariada $N_p(\mu, \Sigma)$ para un vector de dimensión p , se obtienen maximizando la función de verosimilitud con respecto a los parámetros μ y Σ .

Se puede demostrar que estos estimadores para el caso Normal multivariado son

$$\hat{\mu} = \bar{X} \quad \text{y} \quad \hat{\Sigma} = S$$

Estos estimadores verifican las siguientes propiedades:

- ✿ $\bar{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$
- ✿ $nS \sim W$ donde $W_{p,n-1}(\Sigma)$ distribución de Wishart con $n - 1$ grados de libertad, siendo esta distribución una extensión al caso multivariado de la distribución χ^2 (ampliaremos este concepto en la próxima sección).
- ✿ $\hat{\mu}$ y $\hat{\Sigma}$ son independientes.

El comportamiento asintótico de los estimadores de máxima verosimilitud está dado por los siguientes resultados.

- **Ley débil de los grandes números para el caso multivariado:** Si X es una variable multivariada con esperanza $E(X) = \mu$ y matriz de covarianzas $V(X) = \Sigma$, entonces vales las siguientes convergencias en probabilidad

$$\bar{X} \xrightarrow{P} \mu \quad \text{y} \quad S \xrightarrow{P} \Sigma$$

- **Teorema central del límite para el caso multivariado:** Si X es una variable aleatoria p -variada con esperanza $E(X) = \mu$ y matriz de covarianzas $V(X) = \Sigma$, entonces se tiene la siguiente convergencia en distribución

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N_p(0, \Sigma)$$

8.2.2 Distribución de Wishart

Se dice que una matriz W cuadrada de orden p ; es decir, tiene p filas y p columnas, sigue una **distribución de Wishart** [49], $W \sim \mathcal{W}(\Sigma, p, n)$, si y sólo si W puede escribirse como

$$W = \sum_{i=1}^n X_i X_i^t$$

donde X_i son vectores aleatorios independientes e idénticamente distribuidos con $X_i \sim N_p(0, \Sigma)$. Dicho de otro modo, la distribución de Wishart obedece a la suma de los productos entre distribuciones Normales multivariadas independientes de media 0 y varianza común Σ , y sus respectivas traspuestas.

La distribución de Wishart generaliza la distribución χ_n^2 . De hecho, es fácil ver que si $p = 1$ entonces $\mathcal{W}(\sigma^2, 1, n) = \sigma^2 \chi_n^2$.

La suma de variables aleatorias independientes con distribución Wishart con iguales parámetros Σ y p , es una nueva variable aleatoria con distribución Wishart conservando estos mismos parámetros. Es decir, si $W_1 \sim \mathcal{W}(\Sigma, p, n_1)$ y $W_2 \sim \mathcal{W}(\Sigma, p, n_2)$ son independientes, entonces $W_1 + W_2 \sim \mathcal{W}(\Sigma, p, n_1 + n_2)$.

8.2.3 Distribuciones muestrales de la media y la varianza

Para el caso univariado de una muestra aleatoria normal con media μ y varianza σ^2 , las variables promedio muestral \bar{X} y varianza muestral S^2 satisfacen

$$\bar{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right) \quad \text{y} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

La última expresión equivale a decir

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

¿Cómo se puede generalizar esta propiedad para el caso de un vector Normal multivariado?

Sean X_1, X_2, \dots, X_n vectores aleatorios independientes idénticamente distribuidos tales que $X_i \sim N_p(\mu, \Sigma)$ para $i = 1, \dots, n$. Sean \bar{X} el vector de medias y V la matriz de varianzas-covarianzas muestrales, desconocida la media poblacional, vale decir centrada en la media muestral, tenemos respectivamente

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{y} \quad V = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^t$$

Entonces

$$\bar{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right) \quad \text{y} \quad (n-1)V \sim W(\Sigma, p, n-1)$$

Consideremos una muestra aleatoria simple de tamaño n de un vector aleatorio de p componentes con distribución $N_p(\mu, \Sigma)$, digamos $X = (X_1, X_2, \dots, X_n)^t$. Como $X_i \sim N_p(\mu, \Sigma)$, vale que $X_i - \mu \sim N_p(0, \Sigma)$ para todo $i = 1, \dots, n$.

Sea U la matriz definida como la suma de los productos entre las desviaciones de cada X_i respecto de la media poblacional y su traspuesta

$$U = X^t X = \sum_{i=1}^n (X_i - \mu)(X_i - \mu)^t$$

Vale decir la matriz de las observaciones centradas en las medias de la población. La matriz U tiene una distribución de Wishart $U \sim \mathcal{W}(\Sigma, p, n)$.

Para el caso particular en el que $p = 1$, $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ resulta ser una estimador insesgado de σ^2 , la varianza de la población cuando la media poblacional μ es conocida. Del mismo modo,

$$D = \frac{1}{n} U = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)(X_i - \mu)^t$$

es un estimador insesgado de Σ cuando el vector de medias poblacional μ es conocido. La matriz D es semi-definida positiva y resulta definida positiva cuando Σ es inversible.

8.2.4 Distribución de Hotelling

La distribución de Hotelling es una generalización de la distribución t -Student [25]. Recordemos que la variable aleatoria t de Student se define como el cociente entre una variable aleatoria Normal estándar y la raíz cuadrada de una variable aleatoria Chi cuadrado, independiente de la variable Normal del numerador, dividida por sus grados de libertad. Es decir, si $Z \sim N(0, 1)$ y $U \sim \chi_n^2$, entonces $T = \frac{Z}{\sqrt{U/n}} \sim t_n$. Como ya hemos visto, en esta variable se basa el estadístico de contraste para la media poblacional de una distribución Normal cuando la varianza poblacional es desconocida, obteniendo que

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}} \sim t_{n-1}$$

Consideremos el vector aleatorio $X \sim N_p(0, I_p)$, donde I_p indica la matriz identidad de tamaño $p \times p$. Sea además la matriz $V \sim \mathcal{W}(I_p, p, n)$ independiente de X , entonces $nX^t V^{-1} X$ sigue una distribución de Hotelling de parámetro n denotada $T_{p,n}^2$. Simbólicamente,

$$nX^t V^{-1} X \sim T_{p,n}^2$$

Existe una relación entre la distribución de Hotelling y la de Fisher-Snedecor. Si $Q \sim T_{p,n}^2$ es de Hotelling, entonces

$$\frac{n-p+1}{np} Q \sim F_{p, n-p+1}$$

donde $F_{p, n-p+1}$ denota la distribución de Fisher-Snedecor con p y $n-p+1$ grados de libertad.

Por otro lado, si $X \sim N_p(\mu, \Sigma)$ y $U \sim \mathcal{W}(\Sigma, p, n)$, siendo \bar{X} y U independientes, entonces

$$n(\bar{X} - \mu)^t U^{-1} (\bar{X} - \mu) \sim T_{p,n}^2$$

En particular, si X_1, X_2, \dots, X_n es una muestra aleatoria, con $X_i \sim N_p(\mu, \Sigma)$ ($i = 1, \dots, n$) y V es la matriz de varianzas-covarianzas muestral, se verifica que

$$(n-1)(\bar{X} - \mu)^t V^{-1} (\bar{X} - \mu) \sim T_{p,n-1}^2$$

8.2.5 Test del vector de medias para una población

Consideramos ahora una muestra aleatoria multivariada X_1, X_2, \dots, X_n donde para cada $i = 1, \dots, n$, $X_i \sim N_p(\mu, \Sigma)$ siendo Σ una matriz inversible. El objetivo de esta sección es construir un test para contrastar las siguientes hipótesis

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Esta prueba se basa en el estadístico

$$\frac{n-p}{p}(\bar{X} - \mu_0)^t V^{-1}(\bar{X} - \mu_0) \sim F_{p, n-p}$$

donde V denota la matriz de varianzas-covarianzas muestral.

Ejemplo 8.2. Nos interesa contrastar las hipótesis

$$\begin{cases} H_0 : \mu = (\sqrt{50}, 6)^t \\ H_1 : \mu \neq (\sqrt{50}, 6)^t \end{cases}$$

Se tiene una muestra de 40 individuos de una variable bidimensional con matriz de varianzas-covarianzas muestral dada por

$$V = \begin{pmatrix} 4.288 & 1.244 \\ 1.244 & 0.428 \end{pmatrix}$$

y vector de medias muestrales

$$\bar{X} = \begin{pmatrix} 6.9 \\ 6.2 \end{pmatrix}$$

En este caso, el estadístico construido resulta igual a

$$\frac{40-2}{2}[(6.9, 6.2) - (\sqrt{50}, 6)] \begin{pmatrix} 4.288 & 1.244 \\ 1.244 & 0.428 \end{pmatrix}^{-1} [(6.9, 6.2) - (\sqrt{50}, 6)]^t = 17.77$$

Como $P(F_{2,38} > 17.77) < 0.001$, se rechaza la hipótesis de nulidad. Se concluye que no se puede sostener que la media es igual a $\mu = (\sqrt{50}, 6)^t$.

8.2.6 Test para comparar medias de dos poblaciones

Sea X una variable aleatoria observada en dos poblaciones. Consideremos dos muestras multivariadas independientes con distribuciones correspondientes a las dos poblaciones. Suponiendo que

ambas poblaciones tienen distribución Normal con la misma matriz de varianzas-covarianzas, nos interesa comparar sus vectores medios para lo cual vamos a contrastar las siguientes hipótesis

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Sean \bar{X}_1 y \bar{X}_2 los vectores de medias muestrales, y V_1 y V_2 las matrices de varianzas-covarianzas muestrales. Un estimador insesgado de la diferencia $\mu_1 - \mu_2$ es $\bar{X}_1 - \bar{X}_2$. Como $\bar{X}_1 \sim N_p\left(\mu_1, \frac{1}{n_1}\Sigma\right)$ y $\bar{X}_2 \sim N_p\left(\mu_2, \frac{1}{n_2}\Sigma\right)$, tenemos que $\bar{X}_1 - \bar{X}_2 \sim N_p\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right)$.

Asumiendo la hipótesis nula como verdadera, $\mu_1 - \mu_2 = 0$ y entonces

$$\bar{X}_1 - \bar{X}_2 \sim N_p\left(0, \left(\frac{n_1 + n_2}{n_1 n_2}\right)\Sigma\right)$$

Por otro lado, un estimador insesgado para la matriz de varianzas-covarianzas poblacional común a las dos poblaciones, Σ , es

$$S = \frac{(n_1 - 1)V_1 + (n_2 - 1)V_2}{n_1 + n_2 - 2}$$

para el cual se verifica

$$(n_1 + n_2 - 2)S \sim \mathcal{W}(\Sigma, p, n_1 + n_2 - 2)$$

Debido a que \bar{X}_1 es independiente de \bar{X}_2 y ambas son independientes de S , bajo H_0 vale que

$$D^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^t S^{-1} (\bar{X}_1 - \bar{X}_2) \sim T_{p, n_1 + n_2 - 2}^2$$

donde D resulta la **distancia de Mahalanobis** entre los vectores de medias muestrales de ambas poblaciones. Además, por lo expuesto en la sección anterior,

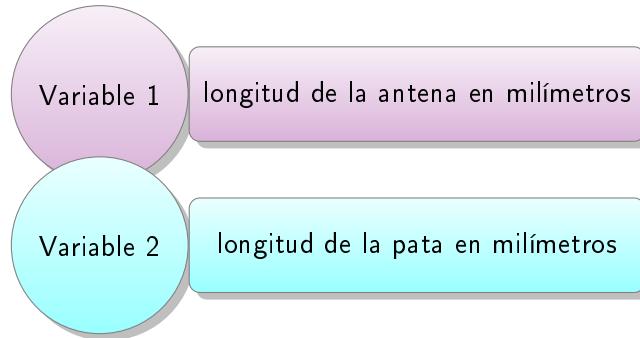
$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} D^2 \sim F_{p, n_1 + n_2 - p - 1}$$

Ejemplo 8.3. Se desea comparar dos especies de avispas, conocidas como *chaqueta amarilla* y *negra pequeña*.



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

Para ello, se consideran las siguientes variables



Para dos muestras independientes de tamaños $n_1 = 9$ y $n_2 = 6$, se han obtenido los datos que figuran en la Tabla 8.1.

Con el Código 8.1 se puede generar el diagrama de dispersión de la Figura 8.5.

Los vectores de medias \bar{X}_1 y \bar{X}_2 correspondientes a las especies ‘chaqueta amarilla’ y ‘negra pequeña’ respectivamente, se calculan con el Código 8.1 y están dados por

$$\bar{X}_1 = (1.41, 1.80) \quad \bar{X}_2 = (1.23, 1.93)$$

Las matrices de varianzas-covarianzas S_1 y S_2 correspondientes a las especies ‘chaqueta amarilla’ y ‘negra pequeña’ respectivamente, se calculan con el Código 8.1 y son

$$S_1 = \begin{pmatrix} 0.0103 & 0.0079 \\ 0.0079 & 0.0159 \end{pmatrix} \quad S_2 = \begin{pmatrix} 0.0036 & 0.0036 \\ 0.0036 & 0.0118 \end{pmatrix}$$

Nuevamente, aplicando el Código 8.1, se puede obtener la estimación (centrada) insesgada de la matriz de varianzas-covarianzas común

$$S = \frac{1}{13}(8S_1 + 5S_2) = \begin{pmatrix} 0.0077 & 0.0063 \\ 0.0063 & 0.0143 \end{pmatrix}$$

Antena	Pata	Especie
1.38	1.64	chaqueta amarilla
1.39	1.71	chaqueta amarilla
1.23	1.72	chaqueta amarilla
1.36	1.74	chaqueta amarilla
1.38	1.82	chaqueta amarilla
1.48	1.82	chaqueta amarilla
1.54	1.81	chaqueta amarilla
1.38	1.90	chaqueta amarilla
1.56	2.07	chaqueta amarilla
1.14	1.78	negra pequeña
1.21	1.86	negra pequeña
1.18	1.96	negra pequeña
1.28	1.96	negra pequeña
1.26	2.10	negra pequeña
1.29	1.90	negra pequeña

Tabla 8.1: Datos sobre las avispas

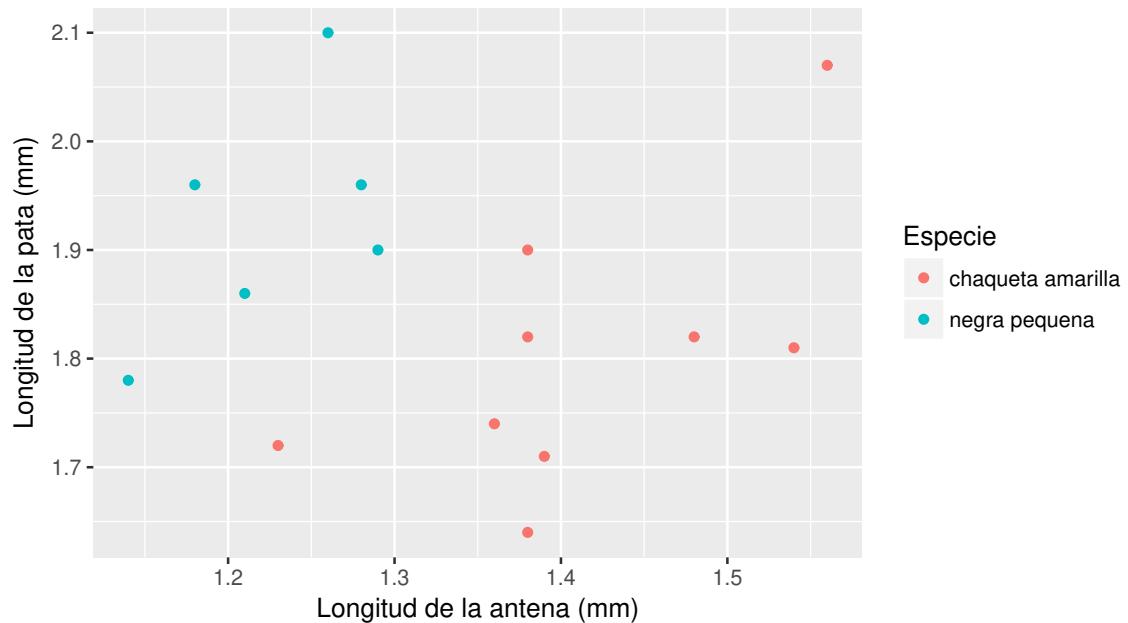


Figura 8.5: Diagrama de dispersión para las avispas

Finalmente, hallamos la distancia de Mahalanobis entre las dos medias muestrales (ver Código 8.1 con datos extraídos de <https://goo.gl/5pTjSS>)

$$D^2 = (\bar{X}_1 - \bar{X}_2)^t S^{-1} (\bar{X}_1 - \bar{X}_2) = 12.54575$$

Luego,

$$T_{obs}^2 = \frac{9 \cdot 6}{9 + 6} D_{obs}^2 = 45.1647 \quad \text{y} \quad F_{obs} = \frac{9 + 6 - 2 - 1}{2(9 + 6 - 2)} T_{obs}^2 = 20.8452$$

Recordemos que el último estimador sigue una distribución $F_{2,12}$. Considerando un nivel de significación del 5%, tenemos que $F_{obs} = 20.8452 > F_{2,12,0.95} = 3.885$. Por lo tanto, rechazamos la hipótesis de nulidad, vale decir que no puede sostenerse la hipótesis de igualdad de los vectores medios.

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

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

avispas$Especie=factor(avispas$Especie) # Declara las especies como factor

ggplot(avispas, aes(Antena, Pata)) +
  geom_point(aes(colour=Especie)) +
  xlab('Longitud_de_la_antena_(mm)') +
  ylab('Longitud_de_la_pata_(mm)')
# Realiza un diagrama de dispersión

especie.avispa=split(avispas, avispas$especie)
# Agrupa los datos según la especie

prom.esp1=apply(especie.avispa[[1]][,1:2], 2, mean)
prom.esp2=apply(especie.avispa[[2]][,1:2], 2, mean)
prom.total=apply(avispas[,1:2], 2, mean)
# Calcula los promedios para cada especie y del grupo general

S1=var(especie.avispa[[1]][,1:2])
round(S1,4)
S2=var(especie.avispa[[2]][,1:2])
round(S2,4)
# Calcula las matrices de varianzas-covarianzas para cada especie

S=(8*S1+5*S2)/13
round(S,4)
# Calcula las matrices de varianzas-covarianzas común
```

```
dist=t(prom.esp1-prom.esp2)*%solve(S)*%(prom.esp1-prom.esp2)
```

Código 8.1: Código para el análisis de las avispas

8.2.7 Análisis de perfiles

Se realiza un análisis de perfiles en las siguientes situaciones:

- ✿ El objetivo es comparar el comportamiento promedio de individuos de una o varias poblaciones y se dispone de mediciones repetidas sobre un conjunto de variables relacionadas.
- ✿ Las componentes del vector normal de interés no corresponden a diferentes variables sino a una misma variable repetida, por ejemplo en el tiempo o el espacio.
- ✿ Se quiere comparar transversalmente y longitudinalmente las medias de dos poblaciones.

Por ejemplo, consideramos dos grupos de personas, uno con n_1 individuos y el otro con n_2 individuos. En cada uno de estos grupos se aplica un tratamiento distinto y se mide el resultado del tratamiento en p instantes diferentes. En la Figura 8.6 se presenta un ejemplo de las medias en cada instante para cada uno de los grupos. En la abscisa se representan los instantes o repeticiones, mientras que la ordenada indica el valor de la media en este instante. El perfil se construye con las medias observadas en cada uno de los grupos, en este caso en tres instantes distintos. Notemos que, en este caso, las tres observaciones son realizadas sobre un mismo individuo o grupo, por lo cual no son independientes.

Más allá que estudiar la igualdad de las curvas o los perfiles, queremos dar respuesta a preguntas del estilo:

- ✿ ¿Los grupos se comportan de manera similar durante todo el proceso? Gráficamente, ¿las curvas son paralelas?
- ✿ ¿Los grupos tienen un nivel parecido? Es decir, ¿las curvas son del mismo nivel?
- ✿ ¿No hay cambio a lo largo del tiempo? Desde el dibujo, ¿la curva promedio es horizontal?

En la sección que sigue trataremos el caso de dos grupos.

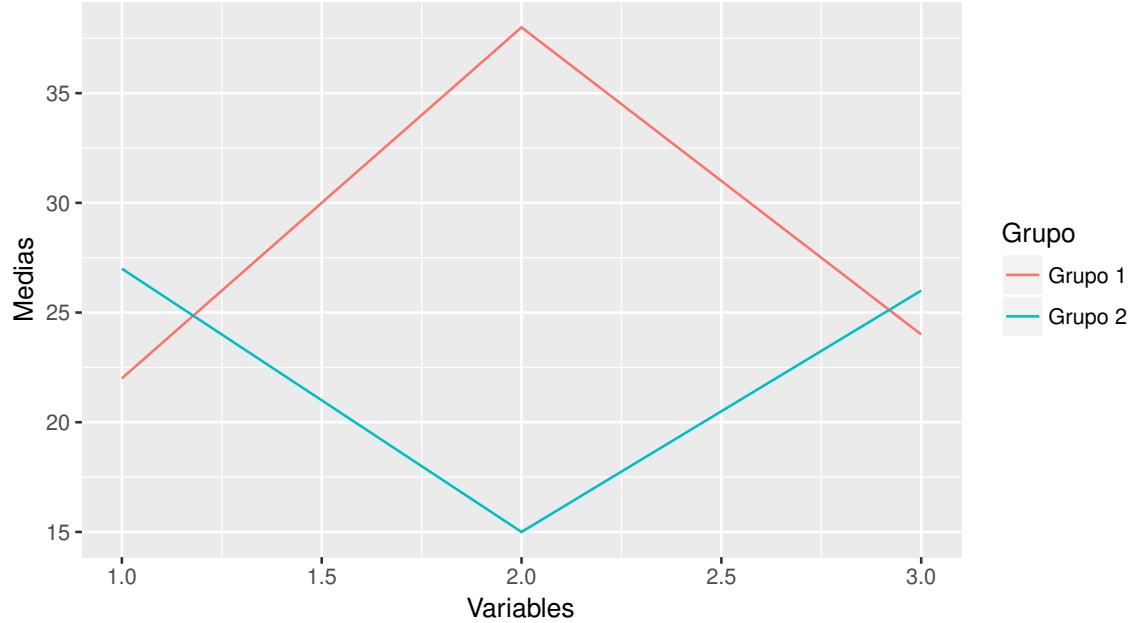


Figura 8.6: Ejemplo de comparación de perfiles

8.2.7.1 Caso de dos perfiles

Sean μ_1 y μ_2 los vectores de medias poblacionales correspondientes a las dos poblaciones consideradas. Es decir,

$$\mu_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix} \quad \text{y} \quad \mu_2 = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix}$$

Si estamos interesados en saber si los perfiles son idénticos para las dos poblaciones se debe realizar el test que plantea

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Sin embargo, si nuestro interés es probar que los perfiles son paralelos, conviene plantear las siguientes hipótesis

$$\begin{cases} H_0 : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \cdots = \mu_{1p} - \mu_{2p} \\ H_1 : \exists(i, j) : \mu_{1i} - \mu_{2i} \neq \mu_{1j} - \mu_{2j} \end{cases}$$

Equivalentemente, si el planteo fuera matricial, consideraremos la matriz $C \in \mathbb{R}^{(p-1) \times p}$ dada por $C_{ii} = 1$ y $C_{i,i+1} = -1$ para $i = 1, \dots, p-1$ y $C_{ij} = 0$ en el resto de los lugares. Entonces la prueba

consiste en

$$\begin{cases} H_0 : C(\mu_1 - \mu_2) = 0 \\ H_1 : C(\mu_1 - \mu_2) \neq 0 \end{cases}$$

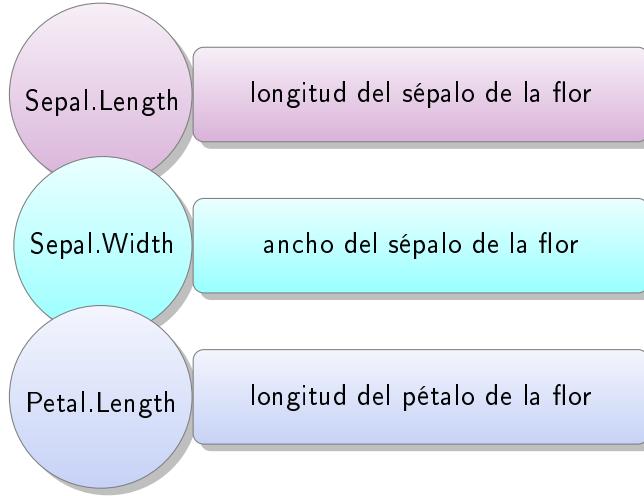
En este caso, $CX \sim N_{p-1}(C\mu, C\Sigma C^T)$ y el estadístico de contraste es

$$\frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^t C^t (CSC^t)^{-1} C (\bar{X}_1 - \bar{X}_2) \sim T_{p-1, n_1+n_2-2}^2$$

Siendo además,

$$\frac{n_1 + n_2 - p}{(n_1 + n_2 - 2)(p - 1)} T_{p-1, n_1+n_2-2}^2 \sim F_{p-1, n_1+n_2-p}$$

Ejemplo 8.4. Vamos a trabajar con el *data set iris* de R y nos referimos al Código 8.2 para todos los cálculos y gráficos que se van a realizar. Consideramos las dos primeras variedades de este archivo que son setosa y versicolor. Sobre estas dos variedades consideramos las primeras tres variables dadas por



Se dispone de 50 observaciones para estas variables en cada uno de los dos grupos considerados. El vector medio total estimando la misma media para los dos grupos es

$$\bar{X} = (5.471, 3.099, 2.861)$$

Estimamos los vectores medios de cada grupo para ver si los grupos presentan similitudes o diferencias en estas variables. Notamos respectivamente \bar{X}_s y \bar{X}_v a las medias de las especies setosa y versicolor,

$$\bar{X}_s = (5.006, 3.428, 1.462) \quad \text{y} \quad \bar{X}_v = (5.936, 2.770, 4.260)$$

En la Figura 8.7 se grafican ambos vectores medios muestrales.

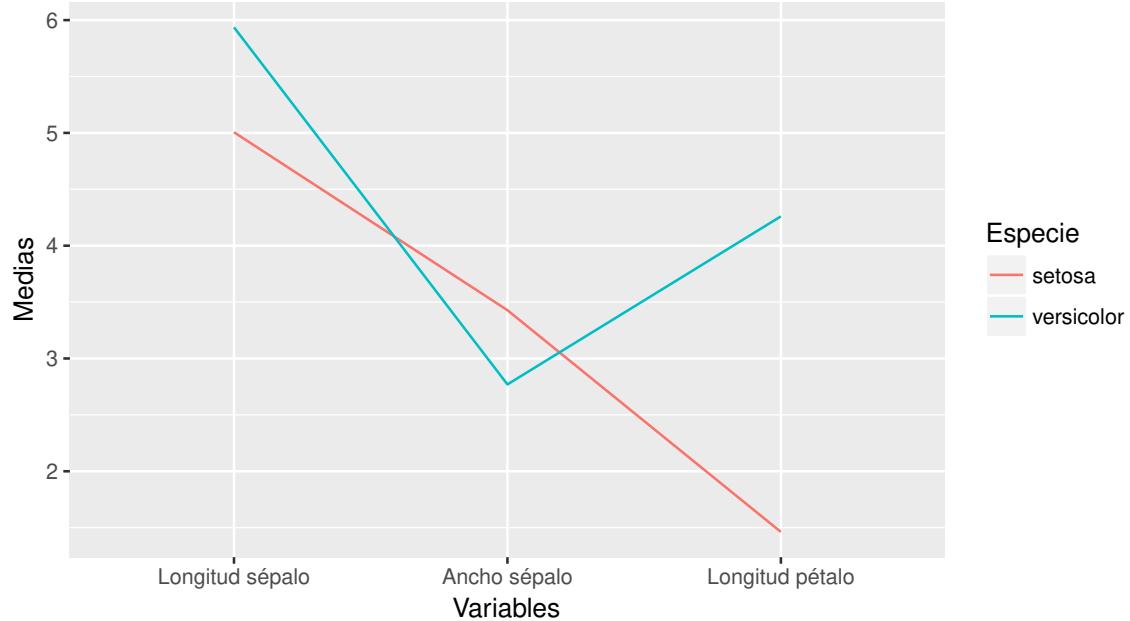


Figura 8.7: Perfiles según la especie

Estamos interesados en testear la hipótesis de que los perfiles son paralelos. Para ello, consideremos las matrices de varianzas-covarianzas muestrales de las especies setosa y versicolor, y la matriz de varianzas amalgamada, notadas respectivamente como S_s , S_v y S .

$$S_s = \begin{pmatrix} 0.124 & 0.099 & 0.016 \\ 0.099 & 0.144 & 0.012 \\ 0.016 & 0.012 & 0.030 \end{pmatrix} \quad S_v = \begin{pmatrix} 0.266 & 0.085 & 0.183 \\ 0.085 & 0.098 & 0.083 \\ 0.183 & 0.083 & 0.221 \end{pmatrix}$$

$$S = \begin{pmatrix} 0.195 & 0.092 & 0.100 \\ 0.092 & 0.121 & 0.048 \\ 0.100 & 0.048 & 0.126 \end{pmatrix}$$

Para realizar la prueba utilizaremos la matriz de varianzas-covarianzas común, S .

El estadístico de contraste del test de Hotelling para probar la igualdad de medias resultó $T_{obs}^2 = 2319.292$. Mientras que el valor crítico del test es $F_{3,96,0.95} = 2.7$. Como $F_{obs} = 757.3197 > F_{3,6,0.95} = 2.7$, se concluye que las diferencias entre los vectores medios de ambos grupos resultaron estadísticamente significativas con un nivel del 5%.

Nos preguntamos ahora si puede suponerse que los perfiles son paralelos. El estadístico de contraste del test de Hotelling aplicando la función de R, para contrastar paralelismo de los perfiles resultó $T_{obs}^2 = 1980.545$. El valor crítico del test es $F_{2,97,0.95} = 3.09$. Luego, debido a que $T_{obs}^2 > F_{2,97,0.95}$, se concluye que no hay evidencia a favor de la hipótesis de paralelismo con un nivel del 5%.

```

Sys.setenv(R_ZIPCMD= "C:/Rtools/bin/zip")
# Requerido para generar archivos xlsx
library(readxl) # Permite leer archivos xlsx
library(openxlsx) # Permite escribir archivos xlsx
library(ggplot2) # Paquete para confeccionar dibujos
library(corrplot)
#Paquete que incluye una estimación eficiente de covarianza y correlación
library(Hotelling) # Paquete que implementa el test de Hotelling

iris.especie=split(iris,iris$Species)
# Agrupa los datos por especie

setosa=data.frame(iris.especie[[1]][,-c(4,5)])
versicolor=data.frame(iris.especie[[2]][,-c(4,5)])
# Toma las tres primeras tres primeras para cada variedad

total=data.frame(rbind(iris.especie[[1]][,-c(4,5)],
iris.especie[[2]][,-c(4,5)]))
media.conjunta=apply(total,2,mean)
media.setosa=apply(setosa,2,mean)
media.versicolor=apply(versicolor,2,mean)
# Calcula la media conjunta y por especie

# Vamos a preparar los datos para el gráfico de perfiles de medias
ms=as.matrix(media.setosa)
mv=as.matrix(media.versicolor)
medias=rbind(ms,mv)
datos=cbind(rep(c(1,2,3),2),medias,c(rep("setosa",3),rep("versicolor",3)))
colnames(datos)=c("Variables","Medias","Especie")
data=data.frame(datos)
nombre=paste("C:/.../datosiris.xlsx")
write.xlsx(data, file=nombre)
datosiris=read_excel("C:/.../datosiris.xlsx")
datosiris[,1:2]=as.numeric(unlist(datosiris[,1:2]))
ggplot(datosiris, aes(x=Variables, y=Medias, colour=Especie)) +
geom_line() +
scale_x_discrete(limit=c("1", "2", "3"),
labels=c("Longitud_sépalo", "Ancho_sépalo",
"Longitud_pétalo"))

var.setosa=round(var(setosa),3)
var.versicolor=round(var(versicolor),3)
var.amalgamada=round(49*(var.setosa+var.versicolor)/98,3)
# Calcula la matriz de varianzas-covarianza por especie y amalgamada

dif.med=(media.setosa-media.versicolor)
# Calcula la diferencia entre los vectores medios

```

```

T2=(50*50/100)*t (dif.med) %*% solve (var.amalgamada) %*% dif.med
# Calcula el estadístico de Hotelling
F.obs=(96/(3*98))*T2
# Calcula el valor observado F de Fisher-Snedecor
pvalor=1-pf(F.obs,3,96)
# Estimamos el p-valor de la prueba
total.especie=data.frame(cbind(total,c(rep("setosa",50),rep("versicolor",50))))
colnames(total.especie)=c("Sepal.Length","Sepal.Width","Petal.Length",
"Especie")
fit=hotelling.test(.~Especie, data=total.especie)
# Aplica el test de Hotelling

# Replicamos lo anterior en el caso matricial
C=rbind(c(1,-1,0),c(0,1,-1))
transf.setosa=as.matrix(setosa)%*%t(C)
transf.versicolor=as.matrix(versicolor)%*%t(C)
transf.total=cbind(rbind(transf.setosa,transf.versicolor),
Especie=factor(c(rep("setosa",50),rep("versicolor",50))))
transf.difmed=C%*%(media.setosa-media.versicolor)
transf.var=C%*%var.amalgamada%*%t(C)
transf.T2=(50*50/100)*t(transf.difmed)%*%solve(transf.var)%*%transf.difmed
transf.F.obs=(96/(3*98))*transf.T2
transf.fit=hotelling.test(.~Especie, data=data.frame(transf.total))

```

Código 8.2: Código para el análisis de perfiles usando `iris`

Cuando las medias de dos grupos son significativamente distintas, puede ser de utilidad considerar estas variables para asignar un individuo a uno de los dos grupos.

Capítulo 9

Métodos de clasificación supervisada

Una vez realizado el análisis de perfiles y de haber determinado que los vectores medios de los dos grupos de estudio son diferentes, es probable que el investigador esté interesado en clasificar a un nuevo individuo dentro de alguno de estos grupos. Por ejemplo, si disponemos de un vector de información sobre pacientes que han respondido bien a cierto tratamiento y otros que no lo han tenido los mismos resultados positivos, podría resultar de interés decidir si a un nuevo paciente le conviene realizar este tratamiento o no. Del mismo modo, podría ocurrir aplicando este análisis para clientes que han cumplido con sus obligaciones y otros que no lo han hecho, con el objeto de decidir si resulta conveniente otorgar a un nuevo cliente una financiación o no.

En líneas generales, se dispone de un conjunto de observaciones que denominaremos **conjunto de entrenamiento**. Para este conjunto de individuos se saben al mismo tiempo los valores que asumen las variables de interés y el grupo al cual pertenece cada integrante. Por este motivo, esta técnica de clasificación se denomina **supervisada**.

9.1 Análisis discriminante

El **análisis discriminante** (AD) tiene por objetivo encontrar una función tal que, al aplicarla a un nuevo individuo, nos permita clasificarlo de acuerdo con el valor que éste presenta en un conjunto de variables que denominaremos **variables discriminantes** y asignarlo a uno de los grupos previamente conocidos o definidos.

Veremos luego que existen algoritmos de clasificación no supervisada, como el **análisis de conglomerado** o **cluster** (del inglés), donde se desconocen tanto la cantidad de grupos como la pertenencia de los individuos del conjunto de entrenamiento a cada uno de ellos.

Para el AD disponemos originalmente de una tabla de datos donde se registraron los valores de p variables observadas sobre N individuos y el grupo de pertenencia. De este modo, la tabla es de tamaño $N \times (p + 1)$.

Este análisis señala para cada una de las variables consideradas, su poder clasificatorio que está

asociado con su peso en la función discriminante.

Para comenzar, consideramos en primera instancia el caso más sencillo, que es el compuesto por dos muestras independientes que provienen de dos poblaciones Normales multivariadas,

✿ Población 1: $P_1 \sim N_p(\mu_1, \Sigma_1)$

✿ Población 2: $P_2 \sim N_p(\mu_2, \Sigma_2)$

Se supone conocido que un nuevo individuo I , con vector de observaciones X_I , proviene de alguna de estas dos poblaciones con probabilidades que denominaremos π_1 y π_2 respectivamente. Simbólicamente:

$$P(I \in P_1) = \pi_1 \quad \text{y} \quad P(I \in P_2) = \pi_2$$

Buscamos entonces una regla para predecir a cuál de estas dos poblaciones es más probable que pertenezca un nuevo individuo.

Se han considerado diferentes enfoques para dar respuesta a este problema y detallaremos los mismos a continuación.

Primer enfoque

Esta opción se define en función de la verosimilitud de I en cada población.

Por lo que, se asigna al sujeto a la Población 1 si $L(X_I, \mu_1, \Sigma_1) > L(X_I, \mu_2, \Sigma_2)$, siendo L la función de verosimilitud que consiste en la función de probabilidad considerada como función del vector de parámetros. Esto significa que se asignar al sujeto a la Población 1 si su función de probabilidad toma valor superior para el vector de parámetros de la Población 1 que para el vector de parámetros de la Población 2. Es decir, $I \in P_1$ cuando

$$|\Sigma_1|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X_I - \mu_1)^t \Sigma_1^{-1} (X_I - \mu_1)\right) > |\Sigma_2|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X_I - \mu_2)^t \Sigma_2^{-1} (X_I - \mu_2)\right)$$

Observar que en las probabilidades, ambos miembros de la desigualdad se multiplican por el número $\frac{1}{\sqrt{(2\pi)^p}}$ que es positivo.

En el caso particular en que $\Sigma_1 = \Sigma_2$, esta regla se simplifica diciendo que $I \in P_1$ cuando

$$\exp\left(-\frac{1}{2}(X_I - \mu_1)^t \Sigma^{-1} (X_I - \mu_1)\right) > \exp\left(-\frac{1}{2}(X_I - \mu_2)^t \Sigma^{-1} (X_I - \mu_2)\right)$$

Realizando cálculos algebraicos, se puede demostrar que esto es equivalente a

$$(\mu_1 - \mu_2)^t \Sigma^{-1} X_I > (\mu_1 - \mu_2)^t \Sigma^{-1} \left(\frac{\mu_1 + \mu_2}{2}\right)$$

Dicho de otra manera, si $b = (\mu_1 - \mu_2)^t \Sigma^{-1}$ y $k = (\mu_1 - \mu_2)^t \Sigma^{-1} \left(\frac{\mu_1 + \mu_2}{2}\right)$, asignamos I a la primera población cuando X_I verifica que $bX_I > k$.

Cabe observar que, en general, desconocemos el valor de μ_i para $i = 1, 2$, por lo que se estima con \bar{X}_i .

Este razonamiento puede extenderse a varios grupos.

Segundo enfoque

Este método se basa en la distancia de Mahalanobis.

Se asigna el sujeto a la Población i si el cuadrado de la distancia de Mahalanobis al vector medio del i -ésimo grupo es menor que el cuadrado de la distancia de Mahalanobis a los restantes grupos. La distancia de Mahalanobis al centro del i -ésimo grupo está dada por

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

El valor de D_i es una medida de lo lejos que está la observación x del vector de medias μ_i del i -ésimo grupo, considerando la matriz de varianzas-covarianzas de la población.

Tercer enfoque

En esta opción se usa la regla de la probabilidad a posteriori.

Se asigna el sujeto a la Población i si

$$P(I \in P_i / \{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}) > P(I \in P_j / \{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\})$$

para todo $j \neq i$.

La principal ventaja de la regla de la probabilidad a posteriori es que brinda una indicación sobre cuánta confianza se podría tener en que el investigador haya tomado la decisión correcta.

Observar que, si bien se le dice ‘probabilidad a posteriori’, en realidad no es una probabilidad. El hecho es que la observación pertenece a una población o a otra, mientras que la incertidumbre proviene de la capacidad de la regla creada por el investigador para elegir la población correcta. Por ejemplo, si contrastamos

$$\begin{cases} P(I \in P_i / \{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}) &= 0.53 \\ P(I \in P_j / \{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}) &= 0.47 \end{cases}$$

no estamos tan seguros de haber clasificado correctamente. Sin embargo, por el contrario si el contraste es

$$\begin{cases} P(I \in P_i / \{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}) &= 0.93 \\ P(I \in P_j / \{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}) &= 0.07 \end{cases}$$

estamos mucho más seguros de haber clasificado correctamente.

En el caso en que las matrices de varianzas-covarianza de las distintas poblaciones pueden suponerse iguales, las tres reglas presentadas coinciden.

9.1.1 Reglas basadas en estimaciones de los parámetros

Al disponer de un par de muestras de dos poblaciones, digamos P_1 y P_2 , en realidad se desconoce el verdadero valor de los parámetros μ_1 , μ_2 , Σ_1 y Σ_2 . Por tal razón, se debe trabajar con sus estimaciones puntuales.

En el caso de los vectores de medias poblacionales, utilizamos como estimadores a los vectores de medias muestrales

$$\hat{\mu}_1 = \bar{X}_1 \quad \text{y} \quad \hat{\mu}_2 = \bar{X}_2$$

Para la estimación de la matriz de varianzas-covarianzas, cuando puedan suponerse iguales, utilizamos la matriz de varianzas-covarianzas muestral amalgamada dada por

$$V = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

donde S_1 y S_2 son las matrices de varianzas-covarianzas muestrales de las Poblaciones 1 y 2 respectivamente, y n_i es la cantidad de individuos de la Población i para $i = 1, 2$.

Ejemplo 9.1. Retomemos nuevamente el Ejemplo 8.3 de las dos especies de avispas. En dicho ejemplo, hemos probado aplicando el test de Hotelling, que las medias de ambas poblaciones son significativamente distintas. Supongamos ahora que tenemos una nueva observación y queremos clasificarla.

Supongamos que la nueva observación es $x = (1.25, 1.8)$. En la Figura 9.1 se indica esta nueva observación en color negro. Si la nueva observación fuera la del punto azul o rojo, no tendríamos muchas dudas respecto de a qué especie asignar la nueva avispa. Sin embargo, en este caso no es tan claro a qué grupo debería asignarse la observación del punto negro debido a que se encuentra en una zona fronteriza entre ambos grupos.

Asimismo, cabe destacar que para este ejemplo las observaciones se indican en un color distinto para cada una de las poblaciones, lo cual resulta posible y sencillo dado que sólo hemos observado dos variables. Si hubiéramos observado más variables, esta visualización podría resultar compleja o incluso, imposible.

En la Figura 9.1 resulta evidente que las dos variables, dadas por la longitud de la antena y de la pata, permiten discriminar entre estas dos poblaciones de avispas.



La pregunta que deberíamos hacernos es la siguiente.

¿Cómo trazar una línea que discrimine las poblaciones de la mejor manera posible?

O bien,

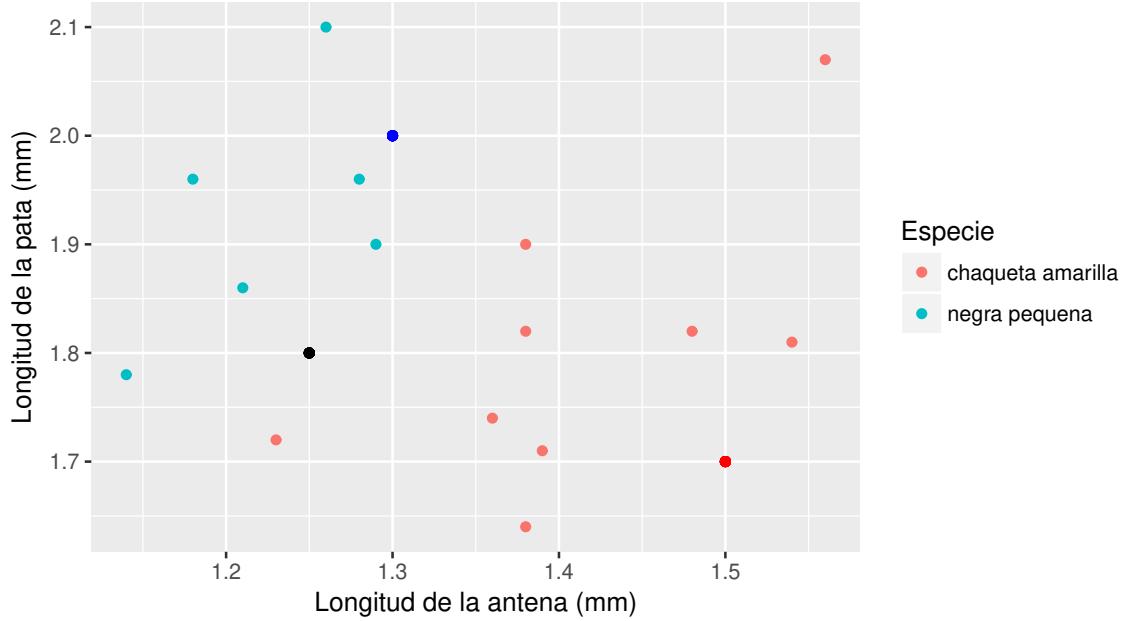


Figura 9.1: Perfiles según la especie con nuevas observaciones

¿En qué dirección proyectar de modo tal de lograr que las proyecciones aparezcan tan separadas como sea posible?

En la Figura 9.2 se muestra la línea que discrimina entre las dos especies de avispas consideradas en el Ejemplo 8.3.

Como ya hemos visto, la suma de cuadrados totales puede descomponerse en la suma de cuadrados entre y dentro de los grupos en el caso univariado. Extendamos esta idea para las matrices de varianzas-covarianza muestrales.

Se sabe que la covarianza total T es igual a la covarianza dentro de los grupos W , más la covarianza entre grupos B , que de manera matricial puede expresarse como

$$T = W + B$$

Buscamos proyectar las observaciones p -variadas $X = (X_1, X_2, \dots, X_p)$ sobre una dirección que maximice la separación entre las proyecciones de los grupos de interés. Es decir que la función discriminante para el caso de estudio, definirá las componentes discriminantes dadas por

$$Y_k = a_k^t X$$

donde a_k es un vector de coeficientes reales. De este modo,

$$\text{Var}(Y) = \text{Var}(a^t X) = a^t T a = a^t W a + a^t B a$$

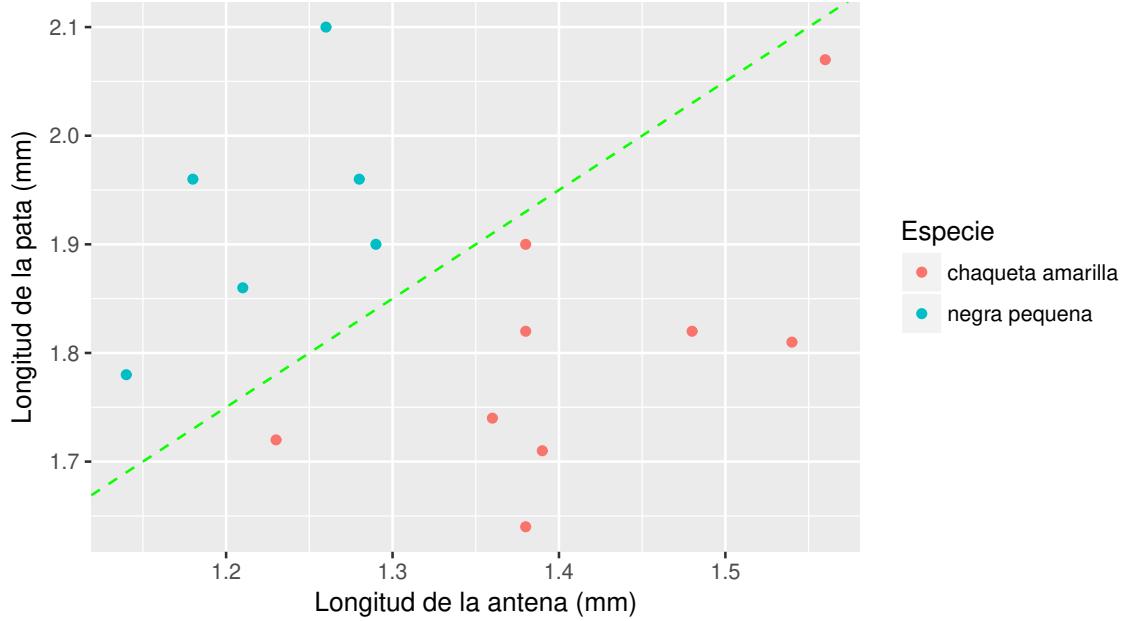


Figura 9.2: Línea discriminante entre las especies de avispas

Maximizar la variabilidad entre los grupos con el objetivo de discriminarlos mejor, implica maximizar la varianza entre grupos B , en relación con el total de la varianza T

$$\max_a \left\{ \frac{a^t B a}{a^t W a} \right\}$$

lo que equivale a maximizar

$$\max_a \{a^t W^{-1} B a\}$$

Notemos que este procedimiento coincide con el mismo esquema de maximización que hemos visto para el análisis de componentes principales. De este modo, la primera dirección corresponde al autovector asociado al mayor autovalor de la matriz $W^{-1}B$. Las siguientes coordenadas discriminantes se corresponden con los restantes autovectores de esta matriz y, naturalmente, serán independientes de la primera. Además, en forma análoga, se puede estimar la proporción de esta separación que logra explicar cada una de las coordenadas discriminantes, como el cociente entre su autovalor y la traza de la matriz $W^{-1}B$.

¿Cuál es la expresión para estas matrices?

Se tiene que

$$B = \sum_{i=1}^g (\bar{X}_{i\cdot} - \bar{X}_{..})(\bar{X}_{i\cdot} - \bar{X}_{..})^t \quad \text{y} \quad W = \sum_{i=1}^g (n_i - 1)S_i^2$$

donde g indica la cantidad de poblaciones de estudio.

Ejemplo 9.2. Nuevamente, nos referimos al Ejemplo 8.3 y para todos los cálculos y salidas del programa R que realizaremos, nos referimos al Código 9.1 con datos extraídos de <https://goo.gl/5pTjSS>.

Recordamos la estimación de los parámetros de interés que calculamos en el Ejemplo 8.3. La matriz de varianzas-covarianzas común, que resulta insesgada, es la matriz de varianza amalgamada dada por

$$\begin{pmatrix} 0.0077 & 0.0063 \\ 0.0063 & 0.0143 \end{pmatrix}$$

El vector medio total junto con los vectores medios por grupo se muestran en la Tabla 9.1.

	Antena	Pata
Media general	1.3373	1.8527
Media chaqueta amarilla	1.4111	1.8033
Media negra pequeña	1.2267	1.9267

Tabla 9.1: Salida medias

Las matrices de covarianzas dentro de los grupos y entre grupos son respectivamente

$$W = \begin{pmatrix} 0.1002 & 0.0817 \\ 0.0817 & 0.1863 \end{pmatrix} \quad \text{y} \quad B = \begin{pmatrix} 0.0177 & -0.0118 \\ -0.0118 & 0.0079 \end{pmatrix}$$

De este modo, la matriz discriminante es

$$W^{-1}B = \begin{pmatrix} 0.3552 & -0.2375 \\ -0.2192 & 0.1466 \end{pmatrix}$$

Los centroides con respecto a la especies chaqueta amarilla y negra pequeña son respectivamente 0.2537 y 0.0320. Además el punto de corte está en 0.1207. En realidad estamos clasificando sobre una proyección del tipo AC y, como se puede apreciar en la Figura 9.3, podemos equivocarnos en ambos sentidos.

A continuación procedemos con la verificación de los supuestos del Análisis Discriminante Lineal, que son el supuesto de normalidad multivariada distribucional y el de homocedasticidad para las matrices de varianzas-covarianzas.

La salida de R del test de Shapiro-Wilk para testear normalidad en el caso multivariado es

Shapiro-Wilk normality test

data: Z

W = 0.95885, p-value = 0.6725

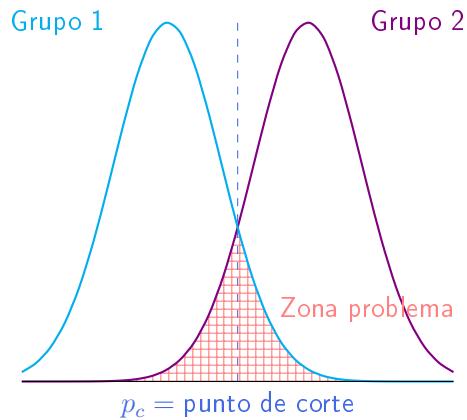


Figura 9.3: Zona problema en la clasificación

Debido a este resultado, no rechazamos la normalidad de la distribución multivariada.

Nos interesa ahora saber si se puede suponer que las matrices de varianzas-covarianzas de ambas especies son iguales. Para ello vamos a aplicar el **test M de Box**, cuya salida en R es

Box's M-test for Homogeneity of Covariance Matrices

data: avispas[, 1:2]

Chi-Sq (approx.) = 1.3654, df = 3, p-value = 0.7137

Como no se rechaza la hipótesis de nulidad que establece que las matrices de varianzas-covarianzas de los grupos son iguales, podemos suponer que no se rechaza el supuesto de homocedasticidad.

Entonces, vamos a calcular la función discriminante lineal. La salida del análisis discriminante lineal en R es

Prior probabilities of groups:

chaqueta amarilla	negra pequeña
0.6	0.4

Group means:

	avispas\$Antena	avispas\$Pata
chaqueta amarilla	1.411111	1.803333
negra pequeña	1.226667	1.926667

Coefficients of linear discriminants:

	LD1
avispas\$Antena	-14.601952
avispas\$Pata	9.011903

Esto debe interpretarse como la función discriminante lineal

$$f(X_1, X_2) = -14.6X_1 + 9.01X_2$$

donde X_1 y X_2 son las variables de interés dadas respectivamente por la longitud de la antena y de la pata de la avispa.

Los valores que toma la función discriminante, conjuntamente con los centroides según el paquete de R que se emplee, pueden diferir en una constante. Aunque, en cualquier caso la clasificación resulta ser la misma.

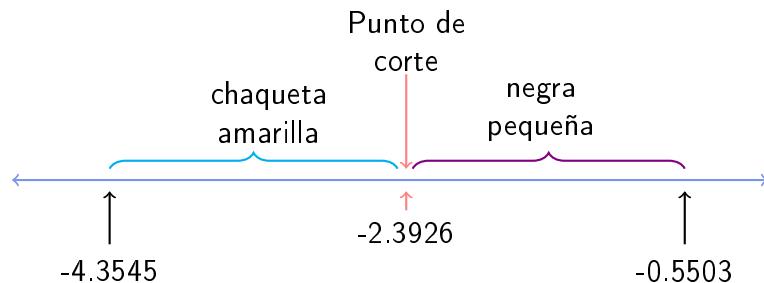
El método que estamos aplicando, proyecta los puntos de las observaciones sobre la dirección que mejor discrimina entre los grupos. Hallemos entonces la proyección de los dos centroides, indicando por proy_a y proy_n a las proyecciones de las especies chaqueta amarilla y negra pequeña respectivamente.

$$\text{proy}_a = -14.6 \cdot 1.4111 + 9.01 \cdot 1.8033 = -4.3545$$

$$\text{proy}_n = -14.6 \cdot 1.2267 + 9.01 \cdot 1.9267 = -0.5503$$

Ahora tenemos que encontrar un punto de corte que notamos p_c , para lo cual medimos la distancia entre las dos proyecciones anteriores, teniendo en cuenta la proporción de cada grupo. Esto nos permite establecer el valor a partir del cual se separan los grupos.

$$p_c = -4.36 + 0.6(\text{proy}_a - 0.4\text{proy}_n) = -2.3926$$



Esto significa que si la proyección de una observación resulta superior a -2.3926 , se clasificará al individuo como ‘negra pequeña’, mientras que en caso contrario se lo clasificará como ‘chaqueta amarilla’.

En la Tabla 9.2, se muestra la proyección para cada observación utilizando la función discriminante hallada. Luego, con la regla establecida del punto de corte en $p_c = -2.3926$ se hizo la asignación a uno de los dos grupos. Observar que la clasificación coincide con la especie observada originalmente.

¿Cómo clasificamos la nueva observación?

Para la nueva observación $(1.25, 1.8)$, calculamos

$$\text{proy}(1.25, 1.8) = -14.6 \cdot 1.25 + 9.01 \cdot 1.8 = -2.032 > p_c$$

Luego, clasificamos al nuevo individuo dentro del grupo ‘negra pequeña’.

Antena	Pata	Proyección	Clasificación
1.38	1.64	-5.3716	chaqueta amarilla
1.39	1.71	-4.8869	chaqueta amarilla
1.23	1.72	-2.4608	chaqueta amarilla
1.36	1.74	-4.1786	chaqueta amarilla
1.38	1.82	-3.7498	chaqueta amarilla
1.48	1.82	-5.2098	chaqueta amarilla
1.54	1.81	-6.1759	chaqueta amarilla
1.38	1.90	-3.0290	chaqueta amarilla
1.56	2.07	-4.1253	chaqueta amarilla
1.14	1.78	-0.6062	negra pequeña
1.21	1.86	-0.9074	negra pequeña
1.18	1.96	0.4316	negra pequeña
1.28	1.96	-1.0284	negra pequeña
1.26	2.10	0.5250	negra pequeña
1.29	1.90	-1.7150	negra pequeña

Tabla 9.2: Clasificación discriminante lineal de las avispas

```

library(readxl) # Permite leer archivos xlsx
library(mvnormtest)
# Paquete que generaliza el test de Shapiro-Wilk para el caso multivariado
library(biotools)
# Paquete con herramientas para análisis de conglomerados y de discriminante

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

avispas$Especie=factor(avispas$Especie) # Declara las especies como factor
especie.avispa=split(avispas, avispas$Especie)
# Agrupa los datos según la especie

prom.esp1=apply(especie.avispa[[1]][,1:2], 2, mean)
prom.esp2=apply(especie.avispa[[2]][,1:2], 2, mean)
prom.total=apply(avispas[,1:2], 2, mean)
# Calcula los promedios para cada especie y del grupo general

S1=var(especie.avispa[[1]][,1:2])
S2=var(especie.avispa[[2]][,1:2])
# Calcula las matrices de varianzas-covarianzas para cada especie

S=(8*S1+5*S2)/13

```

```

round(S,4)
# Calcula las matrices de varianzas-covarianzas común

W=13*S
round(W,4) # Calcula la matriz de covarianzas dentro de los grupos
B=(prom.esp1-prom.total)%*%t(prom.esp1-prom.total)+
(prom.esp2-prom.total)%*%t(prom.esp2-prom.total)
round(B,4) # Calcula la matriz de covarianzas entre grupos

mat.avispas=as.matrix(avispas[1:15,1:2])
# Convierte los datos en matriz

mat.disc=solve(W)%*%B
round(mat.disc,4)
# Calcula la matriz discriminante

avect1=eigen(mat.disc)$vectors[,1] # Calcula el primer autovector
coord.disc.1=mat.avispas%*%avect1 # Calcula las proyecciones

centroide.esp1=prom.esp1%*%avect1
centroide.esp2=prom.esp2%*%avect1
# Calcula los centroides por especie

corte=prom.esp1%*%avect1+(9/15)*(prom.esp2%*%avect1-prom.esp1%*%avect1)
# Calcula el punto de corte

clase=0
for(i in 1:15)
{ifelse(coord.disc.1[i,1]>corte,
clase[i]<-"chaqueta_amarilla", clase[i]<-"negra_pequeña")}
clase
# Clasifica los individuos con la función discriminante
table(clase,avispas$Especie)
# Compara la clasificación con la clase original

mshapiro.test(t(mat.avispas))
# Realiza el test de Shapiro de normalidad multivariada

boxM(avispas[,1:2], avispas$Especie)
# Realiza el test para comparar matrices de varianzas-covarianzas

z=lda(avispas$Especie~avispas$Antena+avispas$Pata, prior=c(9/15,6/15),
method="mle")
# Realiza el análisis discriminante lineal
proyecciones=-14.6*avispas[,1]+9.01*avispas[,2]
# Calcula las proyecciones aplicando la función discriminante lineal

```

Código 9.1: Código para el análisis discriminante de las avispas



La solución aplicada al Ejemplo 9.2, tal como la hemos presentado, corresponde al **Análisis Discriminante Lineal de Fisher**, que resulta válido solamente si se verifican los siguientes dos supuestos:

- ✿ normalidad en la distribución de ambos grupos.
- ✿ homocedasticidad entre los grupos dada por la igualdad de matrices de varianzas-covarianzas.

Se llaman **puntuaciones discriminantes** a los valores que se obtienen al evaluar la función discriminante f para cualquier individuo. Esta función discriminante es de la forma

$$D = f(X_1, X_2, \dots, X_p) = a_1X_1 + a_2X_2 + \dots + a_pX_p$$

Pudiendo expresar las coordenadas discriminantes de manera matricial

$$\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{pn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

Ejemplo 9.3. En la Figura 9.4, generada con el Código 9.2 con datos extraídos de <https://goo.gl/5pTjSS>, se puede visualizar la información de los datos del Ejemplo 8.3. En este gráfico, los puntos llenos indican las observaciones correspondientes a la especie negra pequeña, los puntos vacíos las observaciones correspondientes a la especie chaqueta amarilla y los rombos azules las medias de cada especie.

```
library(readxl) # Permite leer archivos xlsx
library(klaR) # Paquete con funciones para clasificación y visualización

avispas=read_excel("C:/.../avispas.xlsx")
# Importa la base con la cual se va a trabajar
avispas$Especie=factor(avispas$Especie) # Declara las especies como factor

colores=c("cadetblue1","plum2")
partimat(Especie~Antena+Pata, data=avispas, method='lda',
image.colors=colores, col.mean="royalblue", pch=18,
main="Gráfico de partición", print.err=0,
gs=c(rep(1,9),rep(2,6)))
# Produce un gráfico de partición de clases
```

Código 9.2: Código para la visualización de clases de las avispas



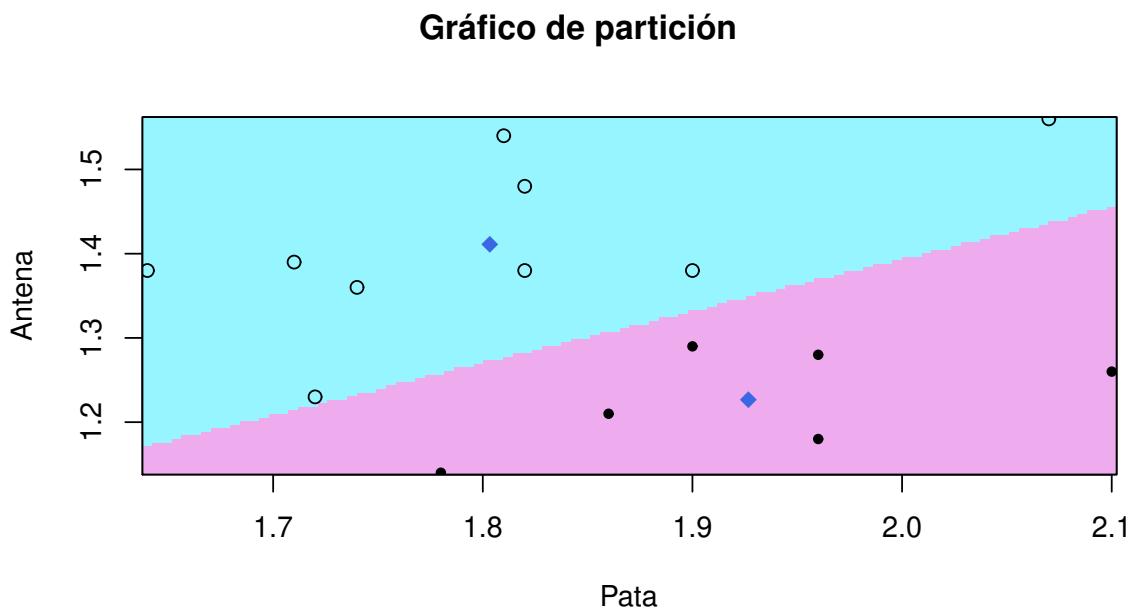


Figura 9.4: Diagrama de partición de las especies de las avispas

9.1.1.1 Estimación de las probabilidades de clasificación errónea

Después de construir la regla de discriminante, interesa conocer su capacidad para discriminar. En caso de no tener regla alguna, se “tira la moneda” y se asigna los sujetos a un grupo mediante el azar, con lo cual la capacidad de discriminar correctamente sería del 50%.

Resulta obvio la preferencia de una regla que no se equivoca en ningún caso, o bien, que clasifique correctamente en un 95% de los casos. Lamentablemente, esto no siempre es posible.

La **matriz de confusión** da una idea de la tasa de clasificaciones incorrectas por grupo y global.

Ejemplo 9.4. En el Ejemplo 8.3 todas las observaciones fueron bien clasificadas, situación no habitual. La matriz de confusión en este caso es diagonal y se exhibe en la Tabla 9.3 siendo generada por el Código 9.3 con datos extraídos de <https://goo.gl/5pTjSS>.

	chaqueta amarilla	negra pequeña
chaqueta amarilla	9	0
negra pequeña	0	6

Tabla 9.3: Matriz de confusión para las avispas

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

```

# Paquete con herramientas para análisis de conglomerados y de discriminante
avispas=read_excel("C:/.../avispas.xlsx")
# Importa la base con la cual se va a trabajar
avispas$Especie=factor(avispas$Especie) # Declara las especies como factor

z=lda(avispas$Especie~avispas$Antena+avispas$Pata, prior=c(9/15,6/15),
method="mle")
# Realiza el análisis discriminante lineal

prediccion=predict(z, avispas[,1:2])$class
# Calcula los valores predichos
table(avispas$Especie, prediccion)
# Tabula al grupo original comparando con los valores predichos

```

Código 9.3: Código para el cálculo de matriz de confusión de las avispas



En este ejemplo en particular, al conocerse el grupo de pertenencia de cada individuo, se puede comprobar la efectividad del método de clasificación observando el porcentaje de casos bien clasificados. Para estimar la probabilidad de clasificación correcta disponemos de las siguientes tres alternativas.

- ✿ **Clasificación ingenua:** cuando se utilizan los mismos datos para construir la regla y para estimar la probabilidad de clasificación correcta. En este caso se calcula la proporción de observaciones bien clasificadas con la regla construida a partir de ellas mismas.
- ✿ **Muestra de entrenamiento y de validación:** se parte el conjunto de datos disponibles en dos submuestras al azar con, aproximadamente, las dos terceras partes para construir o entrenar la regla y la tercera parte restante para validarla. Con la mayor de las submuestras, llamada *training sample* o **muestra de entrenamiento**), se construye la regla de clasificación y con la menor de las submuestras, denominada **muestra de validación**, se estima la probabilidad de buena clasificación.
- ✿ **Validación cruzada, *cross validation*, o *leave one out*:** se elimina la primera observación, se construye la regla sin ella y se la clasifica a esta observación con dicha regla. Luego se reincorpora la primera observación, se elimina la segunda y se procede de la misma forma que con la primera continuando de esta manera hasta la última observación. Finalmente, se estima la probabilidad de buena clasificación considerando la proporción de observaciones bien clasificadas de esta manera.

Podemos mencionar las siguientes observaciones

- ✿ Este análisis sólo tiene sentido cuando las medias de ambos grupos difieren significativamente.
- ✿ La ausencia de normalidad multivariante o la presencia de *outliers* conlleva a problemas en la estimación.
- ✿ Las matrices de varianzas-covarianzas distintas requieren el uso de técnicas de clasificación cuadráticas, conocidas como Análisis Discriminante Cuadrático de Fisher.
- ✿ En la validación cruzada es importante entender que para cada elemento que se elimina, se construye una regla distinta y por lo tanto, los coeficientes de la misma pueden variar de un caso a otro.
- ✿ La multicolinealidad genera problemas en la interpretación de los parámetros selección de variables.
- ✿ No todos los conjuntos son linealmente separables (situación que analizaremos con más detalle en lo que sigue).

9.1.1.2 Casos de más de dos grupos

Cuando el número de grupos es mayor a 2 y sosteniendo el supuesto de homocedasticidad, se presentarán las siguientes propuestas.

- ✿ **Primera propuesta:** se calcula la distancia de Mahalanobis al centroide (media) de cada grupo y un nuevo individuo se clasifica en el i -ésimo grupo si el valor de la distancia de Mahalanobis a ese grupo es la menor de todas.
- ✿ **Segunda propuesta:** se calcula la probabilidad a posteriori de que una nueva observación pertenezca a cada uno de los grupos. Se clasifica la observación en el grupo que maximiza dicha probabilidad o la función de verosimilitud.

9.1.2 Validación de los supuestos del análisis discriminante

Como ya hemos dicho en varias oportunidades, el análisis discriminante lineal sólo es válido cuando las variables originales tienen distribución Normal multivariada y las matrices de varianzas-covarianzas son iguales para todos los grupos. Es más, cuando la distribución conjunta es Normal multivariada, cualquier combinación lineal de sus componentes se distribuye normalmente. En particular, esto se verifica para cada una de las coordenadas del vector de observaciones. Es por ello que, si alguna de las variables originales no se distribuye de manera Normal, entonces es seguro que la distribución conjunta no es Normal multivariada. Luego, bastará con que una componente no tenga distribución Normal para asegurar que la distribución conjunta no es Normal multivariada.

Sin embargo, si todas las componentes tienen distribución Normal univariada, cabe preguntarse si este hecho es suficiente para probar que la distribución conjunta es Normal multivariada. La

respuesta es **no**. En este caso recomendamos realizar un test de bondad de ajuste multivariado para testear el supuesto de Normalidad distribucional.

Para comprobar el supuesto de homocedasticidad se puede aplicar la prueba M de Box. La hipótesis nula de esta prueba es que las matrices de varianzas-covarianzas de los grupos son iguales. Para comparar las matrices de varianzas-covarianzas se utiliza el determinante de la matriz de varianzas-covarianzas de cada uno de los grupos. Así como el test de Bartlet para el caso univariado, el test M de Box es sensible a la falta de normalidad multivariante. Es decir, matrices iguales pueden aparecer como significativamente diferentes por no cumplirse el supuesto de normalidad. Por otra parte, si las muestras son de tamaños grandes, este test pierde efectividad, resultando más fácil rechazar la hipótesis nula. Una alternativa robusta para esta prueba es el test de Levene para datos multivariados.

La prueba de Levene univariada realiza un análisis de la varianza sobre los valores absolutos de las diferencias entre los valores observados y el centro del grupo, que puede tomarse como la media o como la mediana en el caso de querer evitar la influencia de *outliers*. De forma análoga, la **prueba de Levene multivariada** propuesta por O'Brien [8], se basa en la razón F del ANOVA calculada sobre las distancias euclídeas de puntos individuales de cada grupo a su centroide c_i .

Denotando por x_{ij} al j -ésimo punto del i -ésimo grupo, y por x_{ijk} a la k -ésima coordenada de x_{ij} , se define

$$d_{ij}^c = \Delta(x_{ij}, c_i)$$

donde el **vector centroide** se define como el punto que minimiza la suma de cuadrados de las distancias a cada punto del grupo

$$c_i = \min_c \sum_{j=1}^{n_i} (d_{ij}^c)^2$$

siendo n_i la cantidad de elementos del i -ésimo grupo y

$$\Delta(x_{ij}, c_i) = \sqrt{\sum_{k=1}^p (x_{ijk} - c_i)^2}$$

Este vector centroide usualmente se asume como el vector de medias muestrales, definido como $\bar{x} = (\bar{x}_1, \dots, \bar{x}_p)$. Luego, la estadística F del ANOVA que se utiliza para probar H_0 es de la forma

$$F = \frac{(N - g) \sum_{i=1}^g n_i (D_{i\cdot}^c - D_{\cdot\cdot}^c)^2}{(g - 1) \sum_{i=1}^g \sum_{j=1}^{n_i} (d_{ij}^c - D_{i\cdot}^c)^2}$$

donde g es la cantidad de grupos, n_i el total de observaciones del i -ésimo grupo y $N = \sum_{i=1}^g n_i$ es

el total de observaciones. Además, $D_{i\cdot}^c = \sum_{j=1}^{n_i} d_{ij}^c$ es la suma de distancias de las observaciones del i -ésimo grupo a su centroide y $D_{..}^c = \sum_{i=1}^g D_{i\cdot}^c$ es la suma de todas las distancias.

Bajo la suposición de H_0 , el estadístico F de ANOVA sigue aproximadamente una distribución F de Fisher con $g - 1$ y $N - g$ grados de libertad. Anderson [4] propone el uso de medianas como una opción robusta para la prueba de Levene.

En este caso, la mediana p -dimensional no necesariamente está definida como el vector de medianas individuales para cada variable. En algunos programas estadísticos, como R, se encuentran rutinas implementadas para encontrar medianas espaciales de este tipo.

¿Qué corresponde hacer cuando se rechaza el supuesto de homocedasticidad?

Al rechazar la hipótesis nula que plantea $H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_g$, una alternativa es estandarizar por separado cada grupo con su respectiva matriz de varianzas-covarianzas estimada, de modo tal de obtener grupos con igual matriz de varianzas-covarianzas y, sobre este nuevo espacio de representación, calcular el discriminante lineal. En este caso, se proyectan todos los datos y se estima la varianza de las puntuaciones discriminantes dentro de cada grupo.

Destacamos algunas consideraciones a tener en cuenta sobre la selección de las variables para el modelo.

- ✿ Si se encuentra que una variable asume valores medios muy diferentes en los distintos grupos, es probable que resulte una buena variable para discriminar.
- ✿ En segunda instancia, con las variables para las cuales se notó diferencia, es conveniente testear la igualdad de vectores medios entre los grupos.
- ✿ Es conveniente estudiar si las matrices de varianzas-covarianzas de los distintos grupos y del grupo general son similares o no.
- ✿ Sólo en el caso de hallar diferencias significativas entre los vectores medios de los grupos y siendo que las matrices de varianzas-covarianzas resultaron similares, tendrá sentido utilizar la función discriminante lineal.

9.1.3 Interpretación de los coeficientes de la función discriminante

Los coeficientes estandarizados de la función discriminante son los que corresponden al cálculo de la función discriminante con todas las variables clasificadoras estandarizadas. Este recurso se utiliza para evitar ciertos problemas de escala que pudieran existir entre las variables.

Los coeficientes estandarizados a_{ij} pueden interpretarse como indicadores de la importancia relativa de cada una de las variables en cada función discriminante. De esta manera, si la variable x_j es importante en la función discriminante y_i , su respectivo coeficiente a_{ij} será grande en valor absoluto. Dicho de otro modo, hay una fuerte asociación entre la variable x_j y la proyección y_i .

Estos coeficientes son poco fiables si existen problemas de multicolinealidad entre las variables clasificadoras. Al estar correlacionadas las variables originales, y a veces en forma significativa, es conveniente ser cuidadoso a la hora de interpretar estos coeficientes.

9.1.4 Costos de clasificación

En algunas ocasiones es necesario ponderar los errores cometidos. Por ejemplo, es más costoso no indicar que un paciente tiene rechazo a un órgano transplantado, cuando efectivamente lo tiene, que indicar que sí lo tiene cuando en realidad esto no es así. En el primer caso el paciente puede agravarse, mientras que en el segundo caso es posible que se le de una medicación o se aumente la que está recibiendo en forma innecesaria.

Una manera de diferenciación en la regla discriminante entre los dos tipos de errores posibles es asignar un costo a cada error. También podría utilizarse información previa, como por ejemplo si uno supiera que el rechazo a un órgano transplantado ocurre en a lo sumo el 20% de los pacientes.

Con el fin de formalizar, supongamos que tenemos dos poblaciones:

- ✿ P_1 con función de densidad $f(x, \theta_1)$ donde θ_1 es un vector de parámetros que caracteriza a la densidad de la primera población.
- ✿ P_2 con función de densidad $f(x, \theta_2)$ donde θ_2 es un vector de parámetros que caracteriza a la densidad de la segunda población.

Una regla discriminante general, partirá al espacio p -dimensional en dos regiones R_1 y R_2 de modo tal que, si una observación pertenece a R_1 será clasificada en el primer grupo y en caso contrario, será clasificada en el segundo grupo.

Llamamos $C(i/j)$ al costo de clasificar a un individuo en la i -ésima población cuando en realidad el mismo pertenece a la j -ésima población y $P(i/j)$ a la probabilidad de que esto ocurra. También designamos con p_i a la probabilidad de que un individuo de la población general pertenezca al i -ésimo grupo.

¿Cuál será entonces el costo promedio de clasificación errónea de una observación seleccionada de la población general de forma aleatoria?

La respuesta es, el costo total está dado por

$$CT = p_1 C(2/1)P(2/1) + p_2 C(1/2)P(1/2)$$

Establecemos la regla de manejo que asignamos un individuo a la primera población cuando se verifica que

$$p_1 C(2/1) P(2/1) < p_2 C(1/2) P(1/2)$$

vale decir si

$$p_1 C(2/1) f(x, \theta_2) < p_2 C(1/2) f(x, \theta_1)$$

En el caso en que las probabilidades de pertenencia a ambos grupos fueran iguales, $X \in P_1$ si se verifica que $C(2/1)f(x, \theta_2) < C(1/2)f(x, \theta_1)$.

Más aún, si consideramos los costos de mala clasificación para ambos grupos iguales, $X \in P_1$ si vale que $f(x, \theta_2) < f(x, \theta_1)$.

Concluimos que, en el caso de costos iguales de clasificación y probabilidades de pertenencia a cada grupo idénticas, la regla se reduce a maximizar la función de verosimilitud.

9.2 Análisis discriminante cuadrático de Fisher

Cuando el supuesto de homocedasticidad no puede sostenerse, una opción es utilizar el Análisis Discriminante Cuadrático de Fisher que construye la regla: $x \in P_1$ cuando

$$\frac{1}{\sqrt{(2\pi)^p}} |\Sigma_1|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_1)^t \Sigma_1^{-1} (x - \mu_1)\right) > \frac{1}{\sqrt{(2\pi)^p}} |\Sigma_2|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_2)^t \Sigma_2^{-1} (x - \mu_2)\right)$$

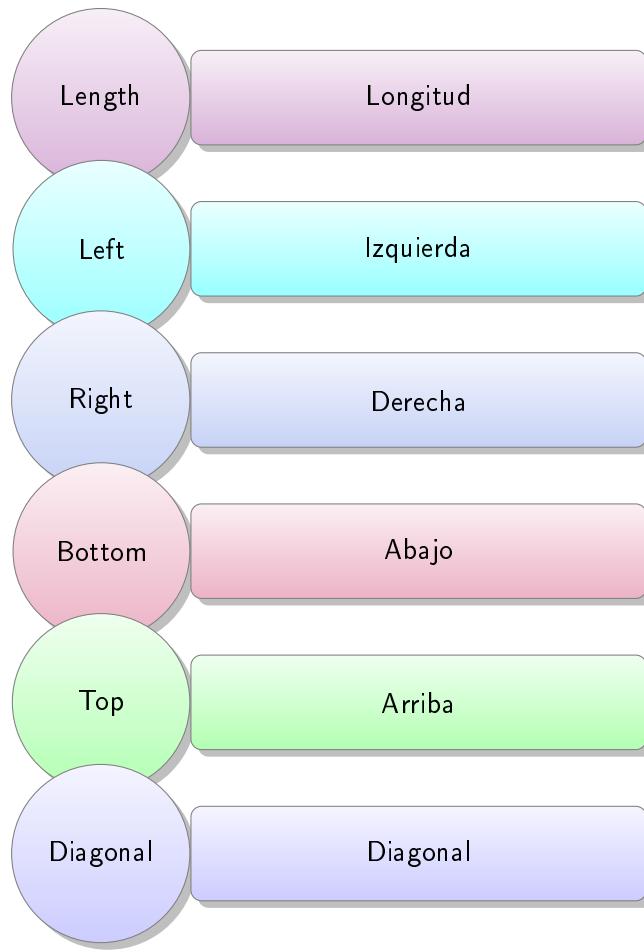
siendo $\Sigma_1 \neq \Sigma_2$.

Ejemplo 9.5. Vamos a utilizar la base de datos `banknote` de R. En la misma se dispone de medidas sobre 200 billetes de los cuales algunos son legítimos y otros falsos.



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

Sobre estos billetes se tomaron las siguientes medidas de longitud:



En la variable **Status** (Estado), se registró si cada billete es genuino o falso. Para el análisis computacional de lo que realizaremos nos referimos al Código 9.4.

En la Figura 9.6 se pueden apreciar los *boxplots* de cada una de estas variables en los grupos definidos por la variable categórica **Status**. En la misma, todas las variables analizadas parecen discriminar entre los billetes legítimos y los apócrifos.

Realicemos ahora la comparación de los vectores medios, para lo que podemos testear la igualdad de medias, obteniendo la siguiente salida en R:

```
$stats
$stats$statistic
[1] 2412.451
$stats$m
[1] 0.1624579
$stats$df
[1] 6 193
$stats$nx
```

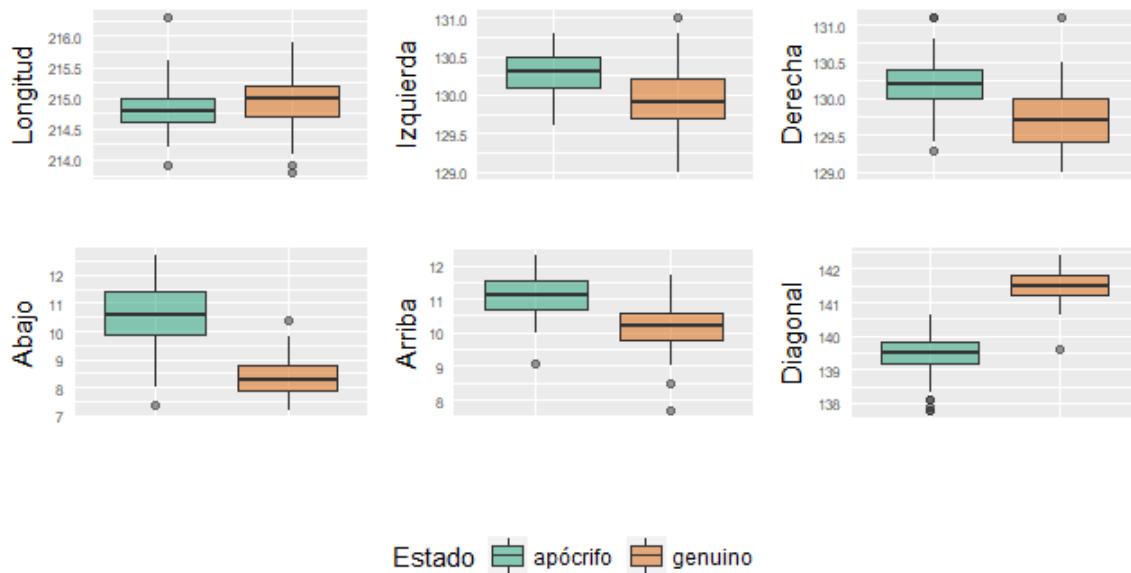


Figura 9.6: Análisis univariado por estado de billete

```
[1] 100
$stats$ny
[1] 100
$stats$p
[1] 6
$pval [1] 0
```

Por lo que se rechaza la hipótesis de igualdad de vectores medios.

Con el propósito de reahlizar un análisis gráfico del supuesto de normalidad univariada, mostramos en la Figura 9.7 los *QQ-plots* para cada variable de estudio. Se observa que algunos de los *QQ-plots* señalan severos apartamientos del supuesto de normalidad univariada lo que indica que no se cumple la normalidad multivariada. Sin embargo, vamos a aplicar un test para analizar este supuesto.

La salida de esta prueba en R es

```
Shapiro-Wilk normality test
data: Z
W = 0.95953, p-value = 1.758e-05
```

Por lo tanto, se rechaza el supuesto de normalidad multivariada.

Al efectuar el análisis del supuesto de homocedasticidad, la salida correspondiente en R es:

```
Box's M-test for Homogeneity of Covariance Matrices
data: banknote[, 2:7]
```

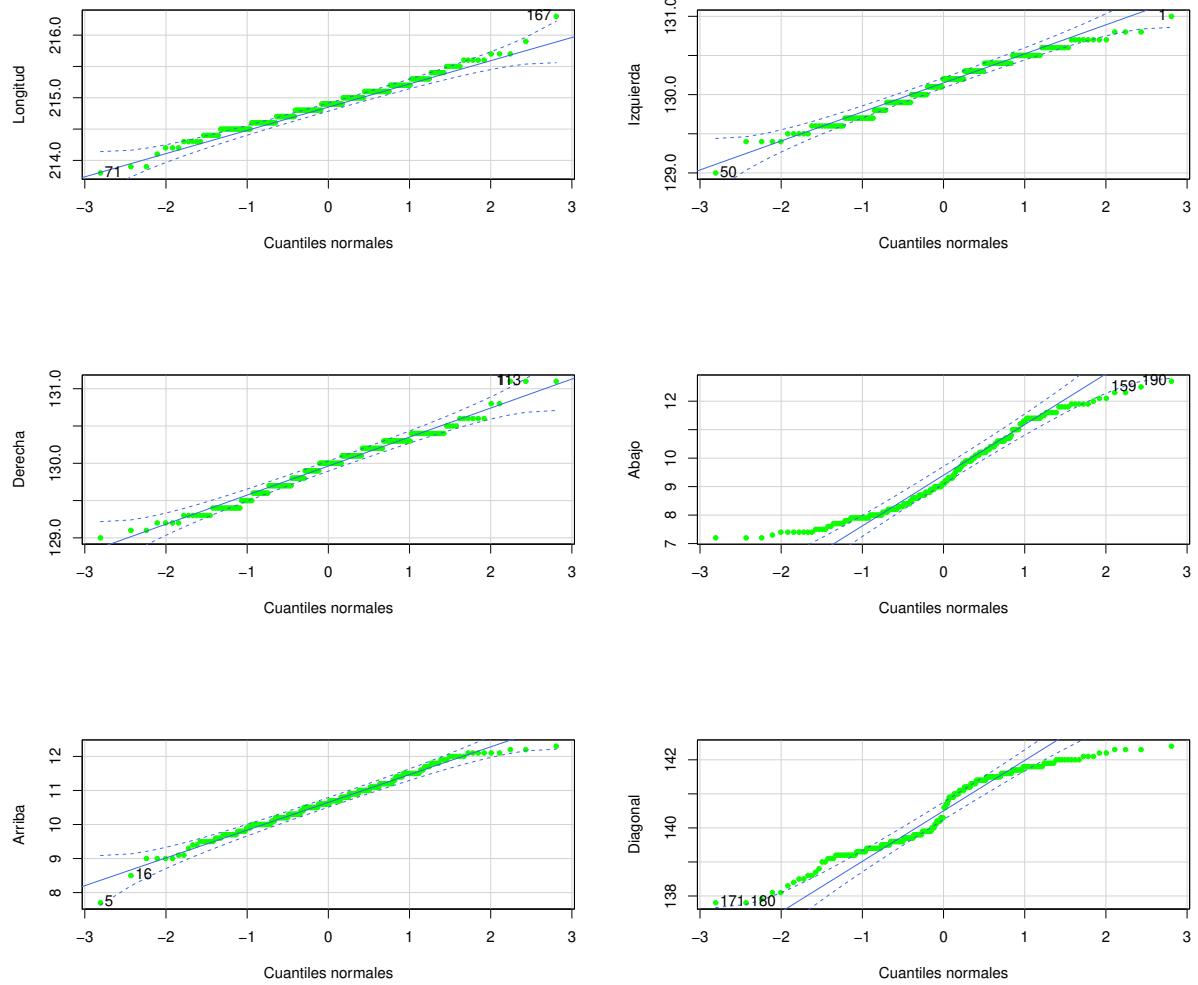


Figura 9.7: *QQ-plots* de las distintas medidas de billete

Chi-Sq (approx.) = 121.9, df = 21, p-value = 3.198e-16

Con lo cual también se rechaza el supuesto de homocedasticidad.

Analicemos si la matriz de correlación en ambos grupos es similar. En la Figura 9.8 se puede apreciar que las matrices de varianzas-covarianzas tienen formas diferentes en ambos grupos.

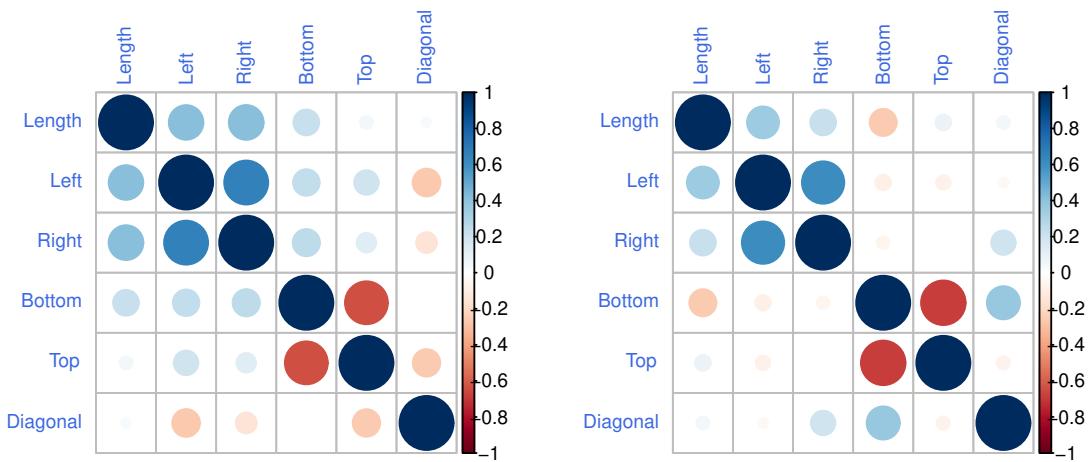


Figura 9.8: Correlogamas de los billetes según su estado

Aplicamos el análisis discriminante cuadrático, aún sin el supuesto de normalidad, para ver cómo clasifica. La salida en R resulta

Prior probabilities of groups:

counterfeit genuine

0.5 0.5

Group means:

	Length	Left	Right	Bottom	Top	Diagonal
counterfeit	214.823	130.300	130.193	10.530	11.13	139.450
genuine	214.969	129.943	129.720	8.305	10.168	141.517

La tasa de clasificación ingenua se observa en la Tabla 9.4.

La tasa de clasificación basada en muestra de entrenamiento de tamaño 120 se exhibe en la Tabla 9.5.

En las Figuras 9.9 y 9.10 se puede visualizar la clasificación de los billetes, donde los puntitos negros indican los que son apócrifos y los círculos vacíos los genuinos.

```
library(mclust) # Paquete con modelos Gaussianos para modelados basados en  
# conglomerados, clasificación y estimación de densidades
```

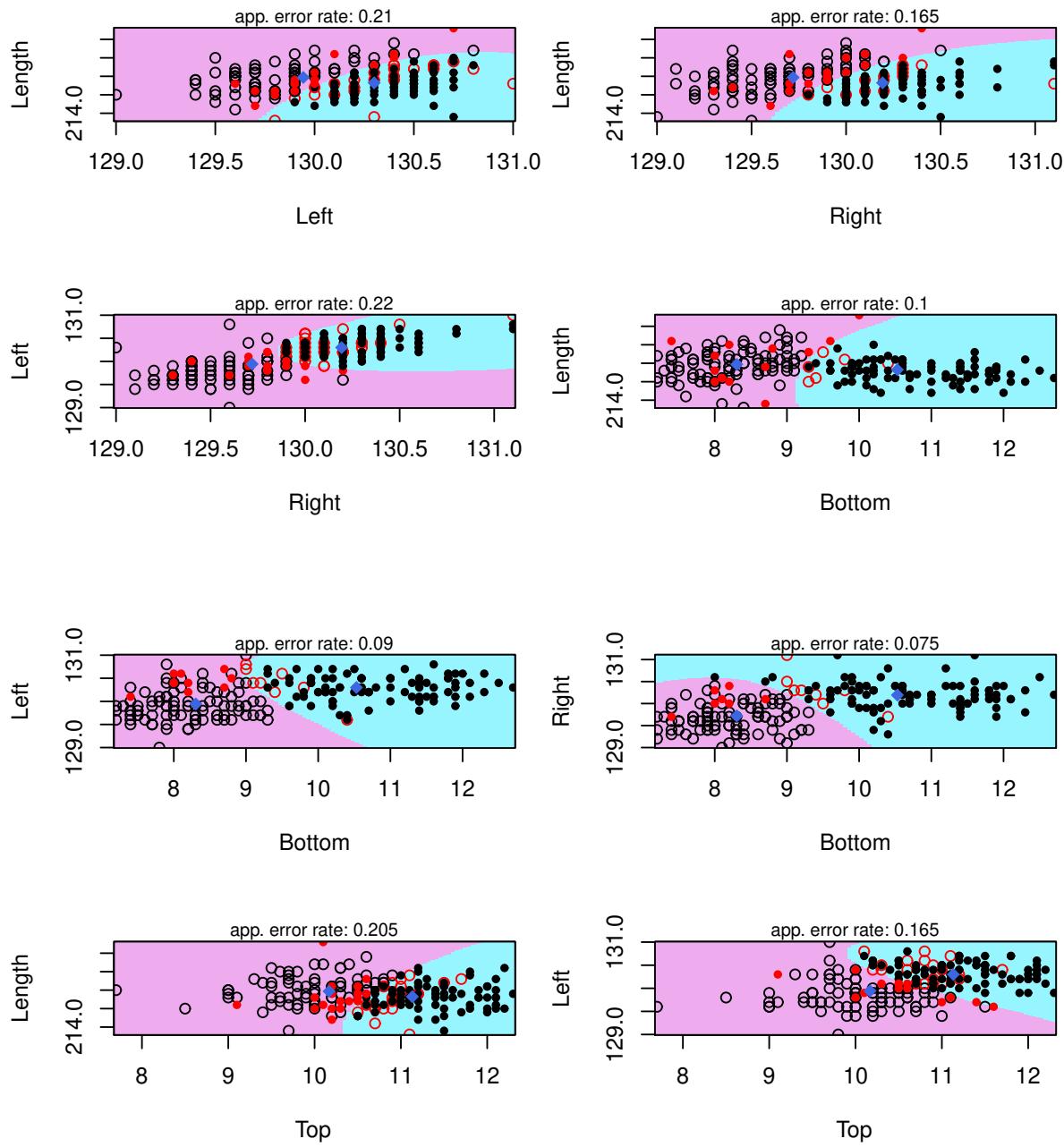


Figura 9.9: Partición por clases de billete

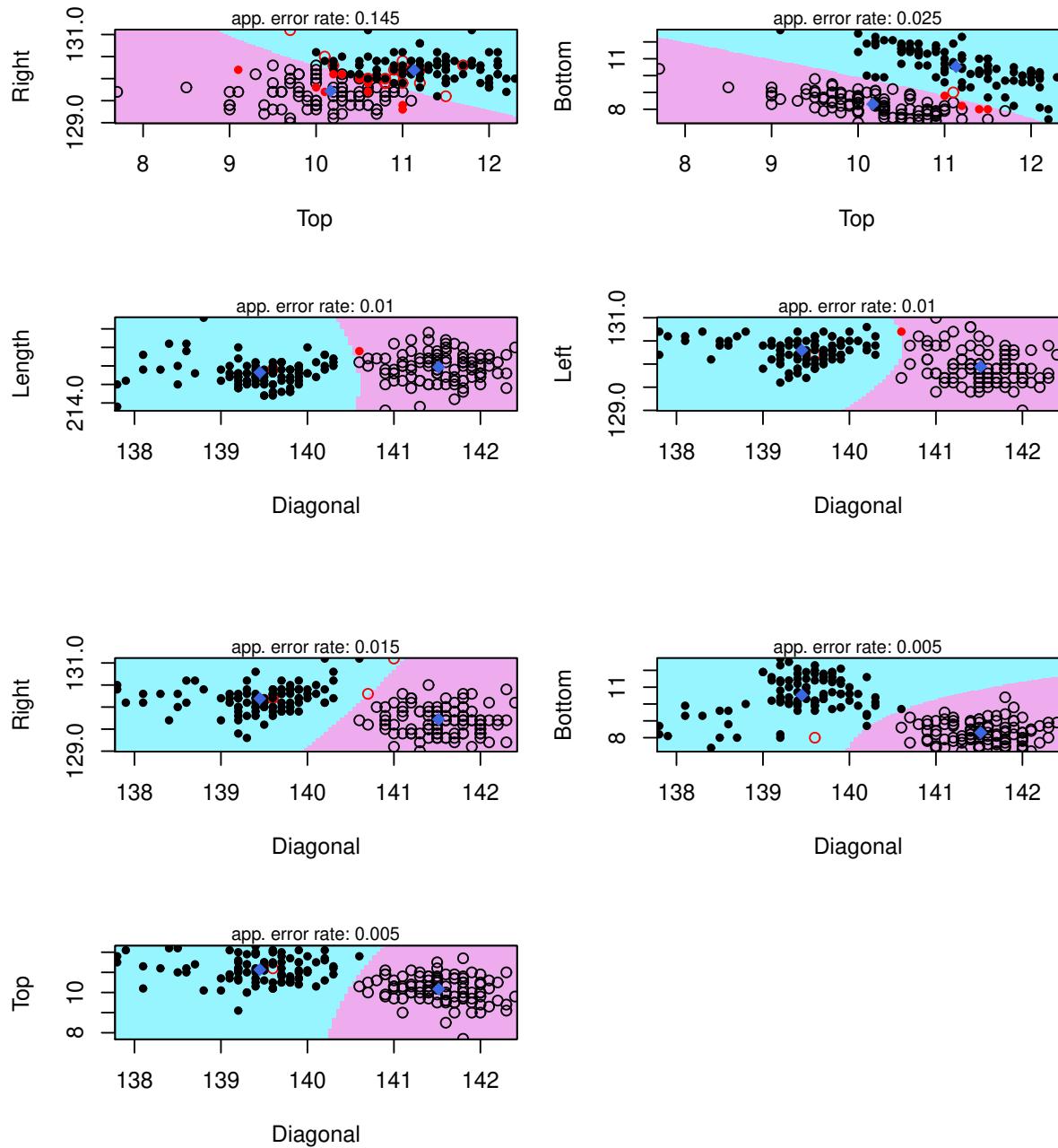


Figura 9.10: Partición por clases de billete (continuación)

		Clase predicha	
		Clase real	Apócrifo
Clase real	Apócrifo	100	0
	Genuino	1	99

Tabla 9.4: Matriz de confusión ingenua para los billetes

		Clase predicha	
		Clase real	Apócrifo
Clase real	Apócrifo	32	1
	Genuino	0	47

Tabla 9.5: Matriz de confusión con una muestra de entrenamiento

```

library(ggplot2) # Paquete para confeccionar dibujos
library(gridExtra) # Paquete para acomodar gráficos simultáneos
library(corrplot)
#Paquete que incluye una estimación eficiente de covarianza y correlación
library(Hotelling) # Paquete que implementa el test de Hotelling
library(car) # Paquete con funciones que acompañan regresión aplicada
library(mvnormtest)
# Paquete que generaliza el test de Shapiro-Wil para el caso multivariado
library(biotools)
# Paquete con herramientas para análisis de conglomerados y de discriminante
library(corrplot) # Paquete para la visualización gráfica de matrices
library(klaR) # Paquete con funciones para clasificación y visualización

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

data(banknote) # Base con la cual vamos a trabajar

bp.long=ggplot(data=banknote, aes(x>Status, y=Length, fill=status)) +
  geom_boxplot(position='identity', alpha=0.5) +
  xlab("") +
  ylab("Longitud") +
  scale_fill_brewer(palette="Dark2", name="Estado",
  breaks=c("counterfeit", "genuine"),

```

```

labels=c("apócrifo","genuino")) +
theme( axis.text.x=element_blank() , axis.ticks=element_blank() ,
axis.text.y=element_text(size=6)) +
theme(legend.position="bottom")
# Produce un boxplot para la longitud

bp.left=ggplot(data=banknote, aes(x>Status, y=Left, fill>Status)) +
geom_boxplot(position='identity', alpha=0.5) +
xlab("") +
ylab("Izquierda") +
scale_fill_brewer(palette="Dark2", name="Estado",
breaks=c("counterfeit", "genuine"),
labels=c("apócrifo","genuino")) +
theme( axis.text.x=element_blank() , axis.ticks=element_blank() ,
axis.text.y=element_text(size=6)) +
theme(legend.position="bottom")
# Produce un boxplot para la izquierda

bp.right=ggplot(data=banknote, aes(x>Status, y=Right, fill>Status)) +
geom_boxplot(position='identity', alpha=0.5) +
xlab("") +
ylab("Derecha") +
scale_fill_brewer(palette="Dark2", name="Estado",
breaks=c("counterfeit", "genuine"),
labels=c("apócrifo","genuino")) +
theme( axis.text.x=element_blank() , axis.ticks=element_blank() ,
axis.text.y=element_text(size=6)) +
theme(legend.position="bottom")
# Produce un boxplot para la derecha

bp.bot=ggplot(data=banknote, aes(x>Status, y=Bottom, fill>Status)) +
geom_boxplot(position='identity', alpha=0.5) +
xlab("") +
ylab("Abajo") +
scale_fill_brewer(palette="Dark2", name="Estado",
breaks=c("counterfeit", "genuine"),
labels=c("apócrifo","genuino")) +
theme( axis.text.x=element_blank() , axis.ticks=element_blank() ,
axis.text.y=element_text(size=6)) +
theme(legend.position="bottom")
# Produce un boxplot para abajo

bp.top=ggplot(data=banknote, aes(x>Status, y=Top, fill>Status)) +
geom_boxplot(position='identity', alpha=0.5) +
xlab("") +
ylab("Arriba") +
scale_fill_brewer(palette="Dark2", name="Estado",
breaks=c("counterfeit", "genuine"),

```

```

labels=c("apócrifo","genuino")) +
theme( axis .text .x=element _blank () , axis .ticks=element _blank () ,
axis .text .y=element _text (size=6)) +
theme(legend .position="bottom")
# Produce un boxplot para arriba

bp .diag=ggplot (data=banknote , aes(x=Status , y=Diagonal , fill=Status )) +
geom _boxplot (position='identity ' , alpha=0.5) +
xlab("") +
ylab("Diagonal") +
scale _fill _brewer (palette="Dark2" ,name="Estado" ,
breaks=c ("counterfeit" , "genuine") ,
labels=c("apócrifo","genuino")) +
theme( axis .text .x=element _blank () , axis .ticks=element _blank () ,
axis .text .y=element _text (size=6)) +
theme(legend .position="bottom")
# Produce un boxplot para la diagonal

mylegend1=g _legend (bp .diag)
# Guarda una leyenda

grid .arrange (arrangeGrob (bp .long + theme(legend .position="none") ,
bp .left + theme(legend .position="none") ,
bp .right + theme(legend .position="none") ,
bp .bot + theme(legend .position="none") ,
bp .top + theme(legend .position="none") ,
bp .diag + theme(legend .position="none") , nrow=2) ,
mylegend1 , nrow=2 , heights=c (10 ,3.5))
# Realiza un gráfico en simultáneo

fit=hotelling .test (.~Status , data=data .frame(banknote))
fit
# Realiza el test de Hotelling

qq .long=qqPlot (banknote$Length , xlab="Cuantiles_normales" , ylab="Longitud" ,
col="green" , pch=20, col .lines="royalblue" , lwd=1)
# Produce un qq-plot para la longitud

qq .left=qqPlot (banknote$Left , xlab="Cuantiles_normales" , ylab="Izquierda" ,
col="green" , pch=20, col .lines="royalblue" , lwd=1)
# Produce un qq-plot para la izquierda

qq .right=qqPlot (banknote$Right , xlab="Cuantiles_normales" , ylab="Derecha" ,
col="green" , pch=20, col .lines="royalblue" , lwd=1)
# Produce un qq-plot para la derecha

qq .bot=qqPlot (banknote$Bottom , xlab="Cuantiles_normales" , ylab="Abajo" ,

```

```

col="green", pch=20, col.lines="royalblue", lwd=1)
# Produce un qq-plot para abajo

qq.top=qqPlot(banknote$Top, xlab="Cuantiles_normales", ylab="Arriba",
col="green", pch=20, col.lines="royalblue", lwd=1)
# Produce un qq-plot para arriba

qq.diag=qqPlot(banknote$Diagonal, xlab="Cuantiles_normales", ylab="Diagonal",
col="green", pch=20, col.lines="royalblue", lwd=1)
# Produce un qq-plot para la diagonal

C=t(banknote[,2:7])
mshapiro.test(C)
# Realiza el test de Shapiro-Wilk

boxM(data=banknote[,2:7], grouping=banknote[, 1])
# Realiza el test M de Box

genuino=cor(banknote[banknote>Status=='genuine',2:7])
apocrifo=cor(banknote[banknote>Status=='counterfeit',2:7])
par(mfrow=c(1,2))
corrplot(genuino, tl.cex=0.7, cl.cex=0.7, tl.col="royalblue")
corrplot(apocrifo, tl.cex=0.7, cl.cex=0.7, tl.col="royalblue")
# Visualiza las matrices de correlación

ADC=qda(formula=Status~Length+Left+Right+Bottom+Top+Diagonal,
data=banknote)
# Realiza el análisis discriminante cuadrático
predicciones=predict(object=ADC, banknote)
# Clasifica y calcula las probabilidades a posteriori
table(banknote>Status, predicciones$class,
dnn=c('Clase_real','Clase_predicha'))
# Compara las clasificaciones

set.seed(12349) # Fija una semilla
entrenamiento=sample(1:200,120)
# Selecciona una muestra de entrenamiento de tamaño 120
modelo=qda(Status~Length+Left+Right+Bottom+Top +Diagonal,
data=banknote[entrenamiento,])
# Construye el modelo de predicción basados en la muestra de entrenamiento
pred=predict(modelo, newdata=banknote[-entrenamiento,])$class
# Clasifica con este modelo los datos de la muestra de validación
table(pred, banknote>Status[-entrenamiento])
# Comparamos las clasificaciones del conjunto de validación

colores=c("cadetblue1","plum2")
partimat(Status~, data=banknote, method='qda', image.colors=colores,
col.mean="royalblue", pch=18, gs=c(rep(1,100),rep(20,100)),

```

```
nplots.vert=2,nplots.hor=2, main="")
# Produce un gráfico de partición de clases
```

Código 9.4: Código para el análisis discriminante de los billetes

9.3 Alternativas robustas

Como ya hemos visto en secciones anteriores, la aplicación de la función discriminante requiere del cumplimiento del supuesto de normalidad multivariada. Sin embargo, este supuesto generalmente no se cumple y, en algunos casos, aún cumpliéndose, la función discriminante es afectada por la presencia de observaciones atípicas, más conocidas como *outliers*.

Cuando el supuesto de normalidad se satisface, pero existe una notoria presencia de *outliers*, se puede incluir la versión robusta. Por el contrario, cuando no se cumple el supuesto de normalidad y también existen observaciones atípicas, el modelo robusto ya no resulta adecuado y se debe recurrir a otras alternativas de clasificación.

Se define la **distanzia de Mahalanobis robusta** como

$$RD_{ij} = \sqrt{(x_{ij} - \hat{\mu}_{j,MCD})^t \hat{\Sigma}_{j,MCD}^{-1} (x_{ij} - \hat{\mu}_{j,MCD})}$$

donde $\hat{\mu}_{j,MCD}$ y $\hat{\Sigma}_{j,MCD}$ son respectivamente los estimadores de posición y de dispersión del j -ésimo grupo, basados en MCD (Minimun Covariance Determinant). En [26] se propone el estimador **FAST MCD** para garantizar la eficiencia computacional.

La **regla de discriminante cuadrático robusto** (RQDR) se define de la siguiente manera, si l denota la cantidad de poblaciones, π_k la k -ésima población y se quiere clasificar a un individuo X , entonces

$$X \in \pi_k \text{ si } \hat{d}_k^{RQ}(X) > \hat{d}_j^{RQ}(X) \text{ para todo } j = 1, \dots, l, j \neq k$$

donde

$$\hat{d}_j^{RQ}(X) = -\frac{1}{2} \ln |\hat{\Sigma}_{j,MCD}| - \frac{1}{2} (X - \hat{\mu}_{j,MCD})^t \hat{\Sigma}_{j,MCD}^{-1} (X - \hat{\mu}_{j,MCD}) + \ln(\hat{p}_j^R)$$

siendo la estimación robusta de la probabilidad de afiliación $\hat{p}_j^R = \frac{\tilde{n}_j}{\tilde{n}}$ con \tilde{n}_j la cantidad de *non-outliers* en el grupo j y $\tilde{n} = \sum_{j=1}^l \tilde{n}_j$

En el caso lineal, es suficiente estimar la varianza común, para lo cual se han propuesto los siguientes tres enfoques:

- * ponderar las matrices de covarianza robustas de cada grupo.

- ✿ ponderar las observaciones.
- ✿ basarse en un algoritmo para estimar el determinante común.

Ejemplo 9.6. Retomamos el Ejemplo 9.5, que clasifica billetes en apócrifos o genuinos, para aplicar la alternativa robusta. En la Tabla 9.6 se muestra el resultado de este análisis robusto que fue realizado mediante el Código 9.5.

		Clase predicha	
Clase real		Apócrifo	Genuino
	Apócrifo	51	49
	Genuino	50	50

Tabla 9.6: Matriz de confusión para la alternativa robusta

Nota: la clasificación del discriminante cuadrático en este caso resulta similar a la propuesta robusta.

```
library(MASS)
# Paquete con funciones y bases de datos para Estadística moderna aplicada

cov.gen=cov.rob(banknote[banknote$Status=="genuine",-1], method="mcd",
nsamp="best")
cov.apo=cov.rob(banknote[banknote$Status=="counterfeit",-1], method="mcd",
nsamp="best")
# Realiza las estimaciones robustas

prom.gen=rep(cov.gen$center,100)
prom.apo=rep(cov.apo$center,100)
var.gen=as.matrix(cov.gen$cov)
var.apo=as.matrix(cov.apo$cov)
# Guarda las estimaciones

DR.gen=as.matrix(banknote[,-1]-prom.gen) %*% solve(var.gen) %*%
t(as.matrix(banknote[,-1]-prom.gen))
DR.apo=as.matrix(banknote[,-1]-prom.apo) %*% solve(var.apo) %*%
t(as.matrix(banknote[,-1]-prom.apo))
# Calcula las distancias de Mahalanobis robustas

clase=0
for(i in 1:200) {
  ifelse(DR.gen[i]<DR.apo[i], clase[i]<- "Genuino", clase[i]<- "Apócrifo")
# Clasifica con las distancias
```

```
table(banknote$Status, clase)
# Compara las clasificaciones originales con las robustas
```

Código 9.5: Código análisis discriminante robusto de los billetes



9.4 Máquinas de soporte vectorial

Las **máquinas soporte vectorial** (SVM, del inglés *support vector machines*) tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico y fueron introducidas en los años 90 por Vapnik y sus colaboradores [10].

Desde el inicio, sus sólidos fundamentos teóricos han hecho que fueran aceptadas. La teoría de las SVM es una nueva técnica de clasificación y ha sido aplicada a múltiples disciplinas en los últimos años. Si bien originalmente fueron diseñadas para resolver problemas de clasificación binaria, en la actualidad se aplican para resolver problemas más complejos como los de regresión, agrupamiento y multiclasificación.

Entre los campos de aplicación más difundidos podemos mencionar los siguientes:

- ✿ visión artificial
- ✿ reconocimiento de caracteres
- ✿ clasificación de proteínas
- ✿ procesamiento de lenguaje natural
- ✿ análisis de series temporales

Dentro del conjunto de clasificadores, las SVM se pueden incluir en la **categoría de clasificadores lineales**, puesto que inducen a un hiperplano en el espacio original cuando los conjuntos son linealmente separables o bien en el espacio transformado, denominado espacio de características, cuando los conjuntos no son linealmente separables.

En muchas de estas aplicaciones, las SVM ha probado un desempeño superior al de las máquinas de aprendizaje tradicional como las redes neuronales, y se han convertido en herramientas poderosas para dar solución a los problemas de clasificación.

La definición de los vectores de soporte permite formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o incluso ningún conocimiento de los datos fuera de esta frontera.

Los datos son mapeados por medio de alguna transformación que denominaremos ***kernel*** o **núcleo**, a un espacio de características, que es un espacio de mayor dimensión, en el cual se logra

una mejor separación entre las clases. Esta función de frontera, en el espacio original, logra separar los datos en todas las clases distintas formando un agrupamiento con cada una de ellas.

Las SVM surgieron originalmente como clasificadores para dos clases. Sin embargo, es posible modificar la formulación del algoritmo para que sea posible aplicarlo para realizar una clasificación multiclase.

9.4.1 Separabilidad lineal

9.4.1.1 Linealmente separable

Supongamos que queremos encontrar una función lineal que separe objetos según su clase ubicados en un espacio bidimensional, como por ejemplo, las especies de la Figura 9.1.

En la mayoría de los casos la búsqueda de un hiperplano adecuado en un espacio de entrada es demasiado estricta para ser llevada a la práctica. Una opción posible para solucionar este problema es mapear el espacio de entrada en un espacio de dimensión mayor dentro del cual sea más sencillo buscar el hiperplano óptimo.

Cada punto de entrenamiento $x \in \mathbb{R}^n$ pertenece a una de dos clases, que podrían ser etiquetadas como 1 o -1 . Sea $z = \phi(x)$, notación correspondiente al mapeo en el espacio de características que llamaremos Z . Buscamos un hiperplano de la forma

$$wz + b = 0$$

donde w es un vector en \mathbb{R}^n y $b \in \mathbb{R}$, tal que separe los elementos en las dos clases definidas. Este hiperplano queda determinado por el par (w, b) y, dado un punto x_i , nos permitirá asignarlo a una de las clases definidas.

Sea (x_i, y_i) con $x_i \in \mathbb{R}^n$ y $y_i \in \{-1, 1\}$. La función de separación tiene la siguiente forma

$$f(z_i) = sg(w^t z_i + b) = \begin{cases} 1 & \text{si } y_i = 1 \\ -1 & \text{si } y_i = -1 \end{cases}$$

Hasta acá, el planteo es similar al realizado para el del análisis discriminante lineal.

Otra manera de pensar en la separación de los conjuntos es buscar el hiperplano que maximice el margen m entre los dos conjuntos, en la Figura 9.11 se muestra un ejemplo. Este problema suele resolverse mediante la aplicación del método de multiplicadores de Lagrange.

La idea entonces es extender la capacidad de discriminar de esta nueva metodología, a conjuntos que no estén separados linealmente, generalizando el criterio establecido. Desde un punto de vista algorítmico, el problema de optimización del margen geométrico se reduce a un problema de optimización cuadrático con restricciones lineales que puede ser resuelto mediante programación cuadrática o multiplicadores de Lagrange. La propiedad de convexidad garantiza la unicidad de la solución, en contraposición con otras técnicas como es el caso de redes neuronales.

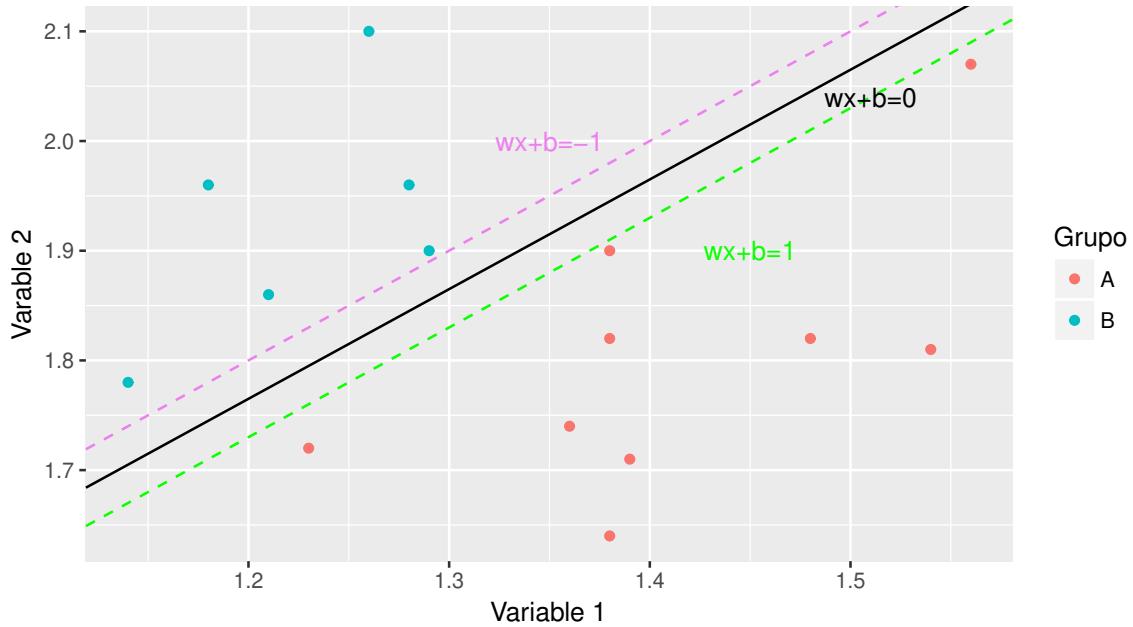


Figura 9.11: Margen entre conjuntos linealmente separables

Un conjunto S se dice **linealmente separable** cuando existe un par $(w, b) \in \mathbb{R}^n \times R$ tal que el sistema de inecuaciones

$$\begin{cases} wz_i + b \geq 1 & \text{si } y_i = 1 \\ wz_i + b \leq -1 & \text{si } y_i = -1 \end{cases}$$

resulta válido para todos los puntos $z_i \in S$.

Cuando se trata de un conjunto linealmente separable, una estrategia usual para encontrar w consiste en los siguientes pasos:

- ✿ Encontrar las envolventes convexas para los puntos de cada clase (ver Figura 9.12).
- ✿ Buscar los dos puntos más cercanos de cada envolvente.
- ✿ Encontrar el plano w que biseca a la recta que une ambos puntos.

Si se realiza un planteo matemático de ambos enfoques se puede ver que en realidad son equivalentes. Sin embargo, en el uso habitual de SVM se utiliza la terminología del segundo enfoque.

¿Cómo es el procedimiento en el caso que las clases no son linealmente separables?

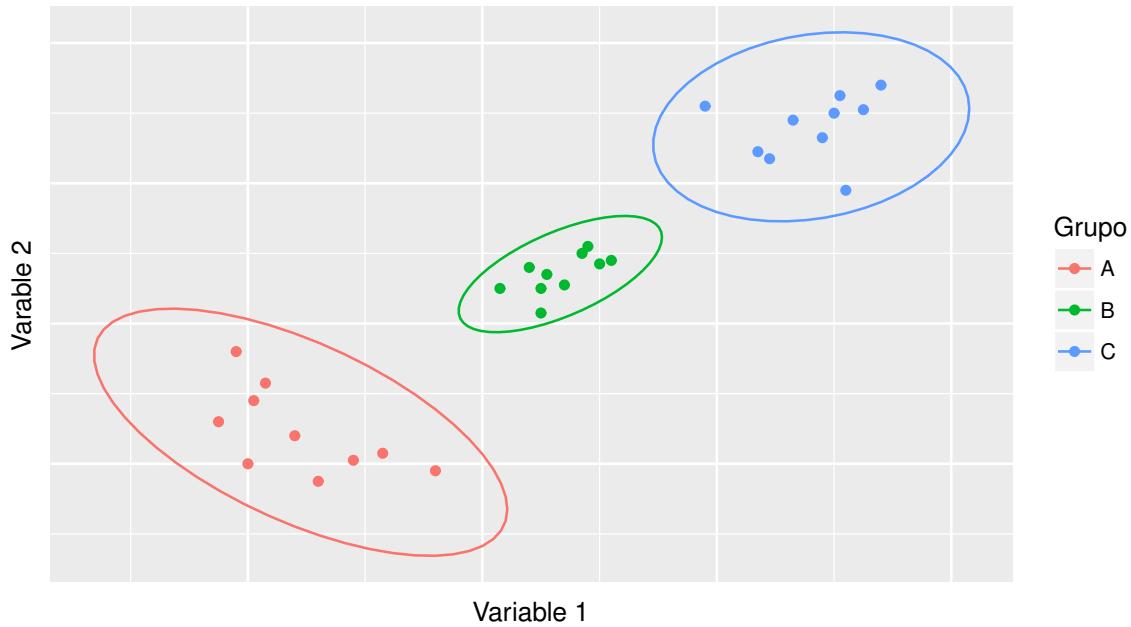


Figura 9.12: Ejemplo de envolventes conexas por clase

9.4.1.2 No linealmente separable

Cuando el supuesto de la definición de conjunto linealmente separable no se satisface, se dice que el conjunto **no es linealmente separable**. Si un conjunto S no es linealmente separable, se deberá admitir algunas violaciones a la formulación de la clasificación de las SVM.

Al trabajar con datos como los de la Figura 9.13, no hay posibilidad de separarlos linealmente. Sin embargo, pueden definirse nuevas variables, a partir de las originales de modo tal que en el nuevo espacio, los conjuntos resulten linealmente separables. Si las variables originales son x e y , las nuevas variables podrían ser x^2 , y^2 , xy , entre otras.

Realizamos las siguientes observaciones.

- ✿ El proceso de pasar de un espacio de entrada con m dimensiones a otro espacio de características de $m > m$ dimensiones, se llama **mapeo**.
- ✿ El aumento de la dimensión del problema conlleva el riesgo de sobreajuste, aunque no resulta tan grave con SVM.
- ✿ La disponibilidad inicial de contar con muchas variables conduce a un aumento exponencial de las variables en el mapeo.

Sintetizamos los pasos del procedimiento de las SVM de la siguiente manera:

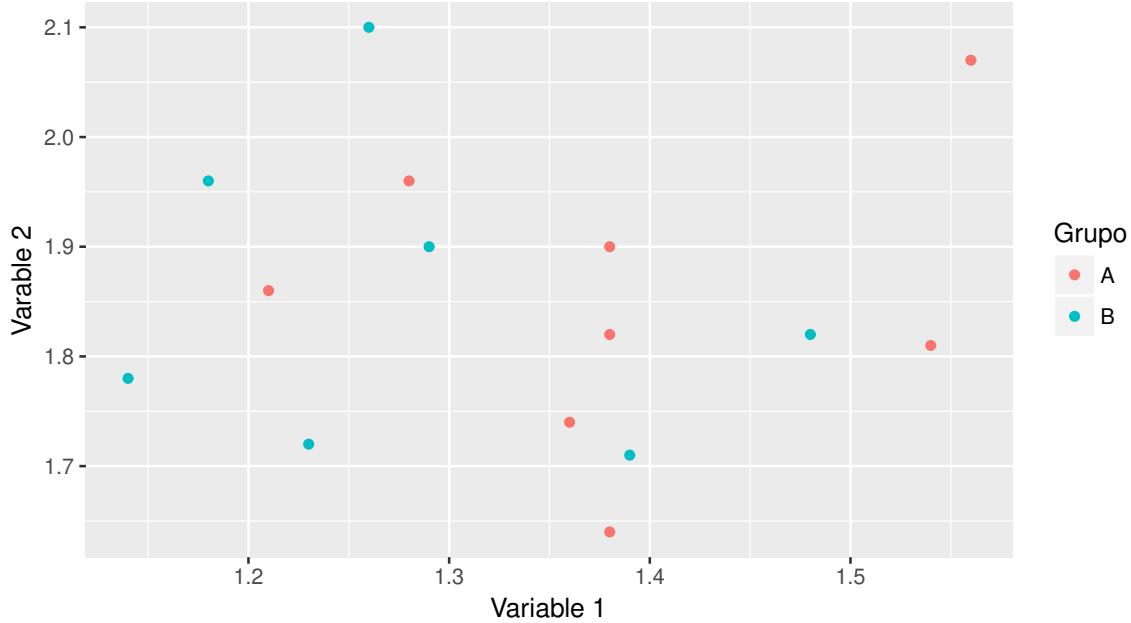


Figura 9.13: Ejemplo de conjunto no linealmente separable

- ✿ Se mapean los puntos de entrada a un espacio de características de una dimensión mayor; por ejemplo, si los puntos de entrada están en \mathbb{R}^2 pueden ser mapeados a \mathbb{R}^3 .
- ✿ Se busca en la imagen de este mapeo un hiperplano que los separe y que maximice el margen entre las clases.
- ✿ La solución del hiperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados **vectores soporte**.

9.4.2 *Kernels* y mapeo

Generalmente no se tiene ningún conocimiento sobre la función de mapeo más conveniente, que denotamos por ϕ , por ende el cálculo de la separación en el nuevo espacio parece imposible. Sin embargo, las SVM poseen una buena propiedad que no hace necesario ningún conocimiento acerca de ϕ .

Sólo es necesaria una función K que calcule el producto escalar de los puntos de entrada en el espacio de características Z ; vale decir

$$z_i \cdot z_j = \phi(x_i)\phi(x_j) = K(x_i, x_j)$$

Luego, si $\phi(x)$ denota el mapeo de la variable x , los ***kernels*** son objetos matemáticos que satisfacen

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)$$

El efecto de la aplicación del mapeo y los *kernels* se puede apreciar en la siguiente Figura 9.14 aplicada a la base de datos `mpg` de R relacionada con la economía de combustible según la marca de automóvil. El código de generación puede verse en 9.6

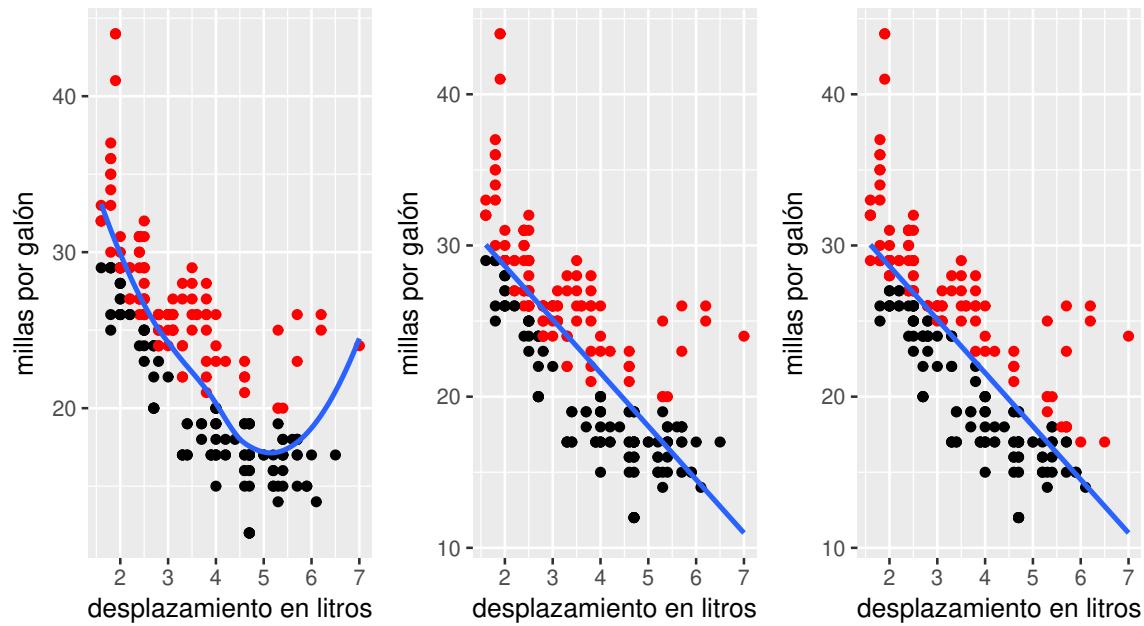


Figura 9.14: Efecto de mapeos

```
library(ggplot2) # Paquete para confeccionar dibujos
library(ggpubr) # Paquete con funciones que personalizan gráficos

plot.original=ggplot(mpg, aes(displ,hwy)) +
  geom_point(colour=class) +
  geom_smooth(method=loess, se=FALSE) +
  xlab('desplazamiento_en_litros') +
  ylab('millas_por_galón')
# Realiza una partición de clases con función de separación

class1=1*(mpg$hwy>(mpg$displ-5)^2+19)+1
class2=1*(mpg$hwy>(-20/7)*(mpg$displ)+33.5)+1
# Se definen nuevas clases

plot.1=ggplot(mpg, aes(displ,hwy)) +
  geom_point(colour=class1) +
  geom_smooth(method=lm, se=FALSE) +
  xlab('desplazamiento_en_litros') +
  ylab('millas_por_galón')
# Realiza una partición de clases con función de separación
```

```

plot.2=ggplot(mpg, aes(displ, hwy)) +
  geom_point(colour=class2) +
  geom_smooth(method = lm, se = FALSE) +
  xlab('desplazamiento_en_litros') +
  ylab('millas_por_galón')
# Realiza una partición de clases con función de separación

ggarrange(plot.original, plot.1, plot.2, nrow=1, ncol=3)
# Produce un gráfico en simultáneo

```

Código 9.6: Código para distintos mapeos para gasto de combustible

Formalizando lo expuesto hasta aquí, el algoritmo de las SVM a partir del producto escalar de dos vectores multidimensionales, busca una familia de hiperplanos que separan los grupos. La función que define este producto escalar es denominada ***kernel*** y la misma puede ser lineal, polinómica, radial o sigmoidal.

Para lograr la mejor clasificación, las SVM maximizan la distancia entre categorías sujeto asumiendo un costo y a un número óptimo de patrones de entrenamiento.

Entre los *kernels* conocidos podemos mencionar los siguientes

- ✿ **Kernel Lineal:** $K(x_i, x_j) = \langle x_i, x_j \rangle$
- ✿ **Kernel polinomial de grado h :** $K(x_i, x_j) = (\langle x_i, x_j \rangle + \tau)^h$
- ✿ **Kernel sigmoideo:** $K(x_i, x_j) = \tanh(\langle x_i, x_j \rangle + \tau)$
- ✿ **Kernel gaussiano:** $K(x_i, x_j) = \exp(\gamma|x_i - x_j|^2)$

Si se va a tratar con datos que no son linealmente separables, el análisis previo puede ser generalizado introduciendo algunas variables no negativas $\zeta_i \geq 0$, de tal modo que el sistema

$$\begin{cases} wz_i + b \geq 1 & \text{si } y_i = 1 \\ wz_i + b \leq -1 & \text{si } y_i = -1 \end{cases}$$

es modificado a

$$\left\{ y_i(wz_i + b) \geq 1 - \zeta_i \quad \text{para } 1 \leq i \leq n \right.$$

Los valores de ζ_i corresponden a los puntos que no satisfacen el sistema de inecuaciones original; es decir, el de la definición de conjunto separable linealmente. De este modo, $\sum_{i=1}^n \zeta_i$ se convierte en una medida de bondad de la clasificación.

De esta forma se convierte el problema de hallar el hiperplano en la minimización

$$\min \left\{ \frac{1}{2} w \cdot w + C \sum_{i=1}^n \zeta_i \right\}$$

La constante C es un parámetro de regularización y puede ser ajustada durante la formulación de las SVM.

Este problema de optimización cuadrática puede resolverse mediante su dual, que conduce a la ecuación

$$\max W(\alpha) = \max \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j z_i z_j \right\}$$

sujeto a $\sum_{i=1}^n \alpha_i y_i = 0$ y sabiendo que $0 \leq \alpha_i \leq C$ para todo $1 \leq i \leq n$.

Dentro de las ventajas de esta técnica podemos mencionar que

- ✿ El entrenamiento es relativamente sencillo.
- ✿ No existe un óptimo local.
- ✿ Se escala relativamente bien para datos en espacios de alta dimensión.
- ✿ El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente.
- ✿ Datos no tradicionales, como cadenas de caracteres o árboles, pueden ser ingresados como entrada de las SVM en lugar de vectores de características.

Por el contrario, la mayor debilidad radica en que es necesaria una buena función *kernel*; es decir, se necesitan metodologías eficientes para definir los parámetros de inicialización de las SVM.

Una observación interesante que puede hacerse es que una lección aprendida en las SVM se traduce a que un algoritmo lineal en el espacio de características es equivalente a un algoritmo no lineal en el espacio de entrada.

Ejemplo 9.7. Veamos cómo utilizar una máquina de soporte vectorial en R para lo cual nos referimos al Código 9.7. En la Figura 9.15 podemos ver la representación de los datos simulados, donde se aprecia que si bien los grupos están bastante separados, no son linealmente separables.

La Tabla 9.7 muestra la confusión del modelo. A partir de esta salida se sabe que en el grupo A, hay 34 individuos de los cuales 21 resultaron bien clasificados, hay un 38% de error. En el grupo B hay 34 individuos de los cuales 31 resultaron adecuadamente clasificados, el porcentaje de error es 9%. Finalmente, en el grupo C hay 31 individuos que fueron todos clasificados correctamente. La tasa de error global resulta de aproximadamente el 16%.

En la Figura 9.16 se muestra la representación de la clasificación por SVM donde se pueden identificar los datos mal clasificados de cada uno de los grupos mediante esta regla.

```
library(ggplot2) # Paquete para confeccionar dibujos
library(e1071) # Paquete que incluye análisis para las SVM
```

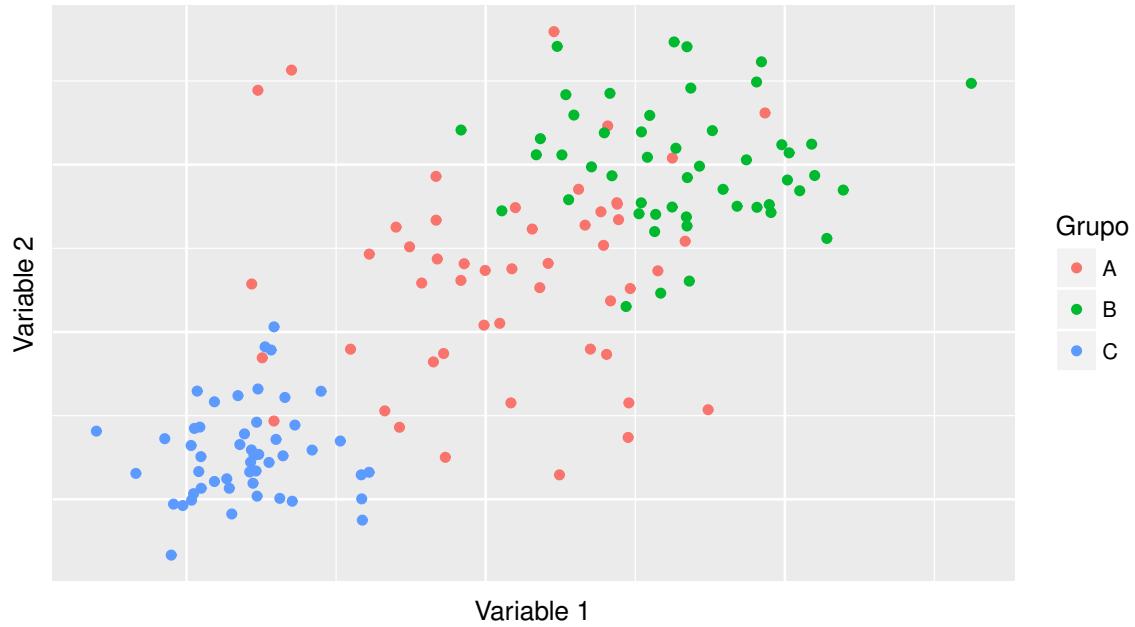


Figura 9.15: Representación gráfica de los datos simulados

		Clasificación por modelo		
		A	B	C
Grupo original	A	21	10	3
	B	3	31	0
C		0	0	31

Tabla 9.7: Tabla de confusión para el modelo por SVM

SVM classification plot

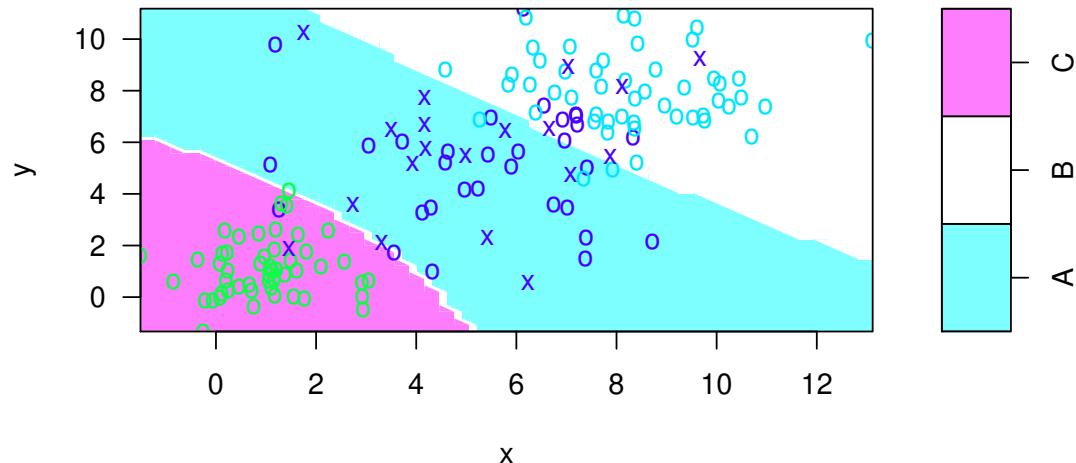


Figura 9.16: Representación gráfica de la clasificación por SVM

```
set.seed(12356) # Fija la semilla  
x=c(rnorm(50,5,2), rnorm(50,8,1.5), rnorm(50,1,1.2))  
y=c(abs(rnorm(50,5,2)),rnorm(50,8,1.5),rnorm(50,1,1.2))  
# Simula un conjunto de puntos en el plano  
  
Grupo=as.factor(c(rep("A",50),rep("B",50),rep("C",50)))  
# Agrupa los datos es tres grupos  
  
datos=data.frame(x,y, Grupo)  
# Arma la base de datos  
  
ggplot(datos, aes(x,y)) +  
  geom_point(aes(colour = factor(Grupo))) +  
  labs(colour="Grupo") +  
  xlab('Variable_1') +  
  ylab('Variable_2') +  
  theme(  
    axis.text.x=element_blank(),  
    axis.ticks.x=element_blank(),  
    axis.text.y=element_blank(),  
    axis.ticks.y=element_blank())  
# Produce una gráfico con los datos simulados  
  
eliminados=sample(1:nrow(datos),100)
```

```

# Elimina algunos datos de la muestra
validacion=datos[eliminados,]
# Arma la muestra de validación
entrenamiento=datos[-eliminados,]
# Armamos la muestra de entrenamiento
modelo.svm=svm(Grupo~y+x, data=entrenamiento, method="C-classification",
kernel="radial", cost=10, gamma=.1)
# Construye el modelo

predichos=data.frame(predict(modelo.svm, validacion))
clasificacion=cbind(validacion, predichos)
colnames(clasificacion)=c("x", "y", "Grupo", "Predichos")
table(validacion$Grupo, clasificacion$Predichos)
# Calcula la tabla de confusión del modelo

plot(modelo.svm, datos, symbolPalette=topo.colors(4), dataSymbol="o",
color.palette=cm.colors)
# visualiza la clasificación del modelo

```

Código 9.7: Código para la clasificación por SVM



9.5 Regresión logística

Los **modelos de regresión logística** se aplican cuando se desea conocer la relación entre una variable dependiente o respuesta dicotómica que toma sólo dos valores.

Las variables que se utilizan para estimar el grupo de pertenencia se denominan **variables predictoras o explicativas**. Estas variables pueden ser tanto cualitativas como cuantitativas y un mismo modelo puede incluir ambos tipos.

El modelo de regresión logística nos permite las siguientes acciones en simultáneo:

- ✿ Ponderar la importancia de la relación entre cada una de las variables predictoras y la variable dependiente (grupo de pertenencia).
- ✿ Estudiar la existencia de interacciones entre las variables predictoras, así como determinar la presencia de variables confusoras o modificadoras de efecto.
- ✿ Entrenar una metodología que nos permite clasificar a un nuevo individuo en una de las dos categorías, estimando la probabilidad de que el individuo pertenezca a cada una de ellas.

El modelo de regresión logística establece una estrategia general, cuyo principal objetivo es estudiar la influencia de ciertos factores sobre la probabilidad de ocurrencia de un cierto evento de interés.

Un modelo de regresión adecuado nos permitirá estimar la proporción de individuos en la población con la característica de interés, o bien la probabilidad de que un individuo tenga dicha característica, para cada posible combinación de valores de las variables explicativas.

En el caso de la regresión lineal clásica, la media de la variable respuesta $E(Y) = \mu$, se estima a partir de una combinación lineal de variables que denominaremos indistintamente explicativas, predictoras o covariables.

La expresión simbólica de este modelo es

$$\mu_Y(X_1, X_2, \dots, X_k) = E(Y/X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

En los modelos lineales generalizados, se modela una transformación de la media de la variable respuesta ($g(\mu)$), como una combinación lineal de las variables predictoras. Luego, la expresión simbólica para este modelo es

$$g(\mu_Y(X_1, X_2, \dots, X_n)) = g(E(Y/X_1, X_2, \dots, X_n)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Esta función g se conoce como **función de enlace** o **link**.

En el modelo de regresión logística, la media $p = E(Y)$ de una variable respuesta con distribución Binomial $Bi(1, p)$, se transforma mediante el enlace denominado **transformación logística** dado por

$$g(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Si la variable respuesta Y tiene distribución $Bi(1, p)$, el parámetro p es al mismo tiempo:

- ✿ la media de la variable Y .
- ✿ la probabilidad de que Y tome el valor 1.

Lo que puede expresarse simbólicamente como

$$E(Y) = P(Y = 1) = p$$

De esta manera, el modelo resulta de la forma

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{P(Y = 1/(X_1, X_2, \dots, X_k))}{1 - P(Y = 1/(X_1, X_2, \dots, X_k))}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

El modelo de regresión logística estima la probabilidad de que un individuo de la población, elegido al azar, pertenezca a cierto grupo, como puede ser afectado por una enfermedad, deudor de un crédito, víctima de un accidente, entre otros; como una función lineal de las variables predictoras valiéndose de una transformación que mapea \mathbb{R} en el intervalo $[0, 1]$. Con lo cual se tiene que

$$\ln\left(\frac{P(X_1, X_2, \dots, X_k)}{1 - P(X_1, X_2, \dots, X_k)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Esta transformación logra que todos los valores estimados por la función lineal resulten al transformarse valores de probabilidad. Despejando el valor de la probabilidad obtenemos

$$P(X_1, X_2, \dots, X_k) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}$$

Para entender de qué forma se transforman los valores estimados de $\alpha + \beta x$ en probabilidades, observemos el gráfico de la Figura 9.17.

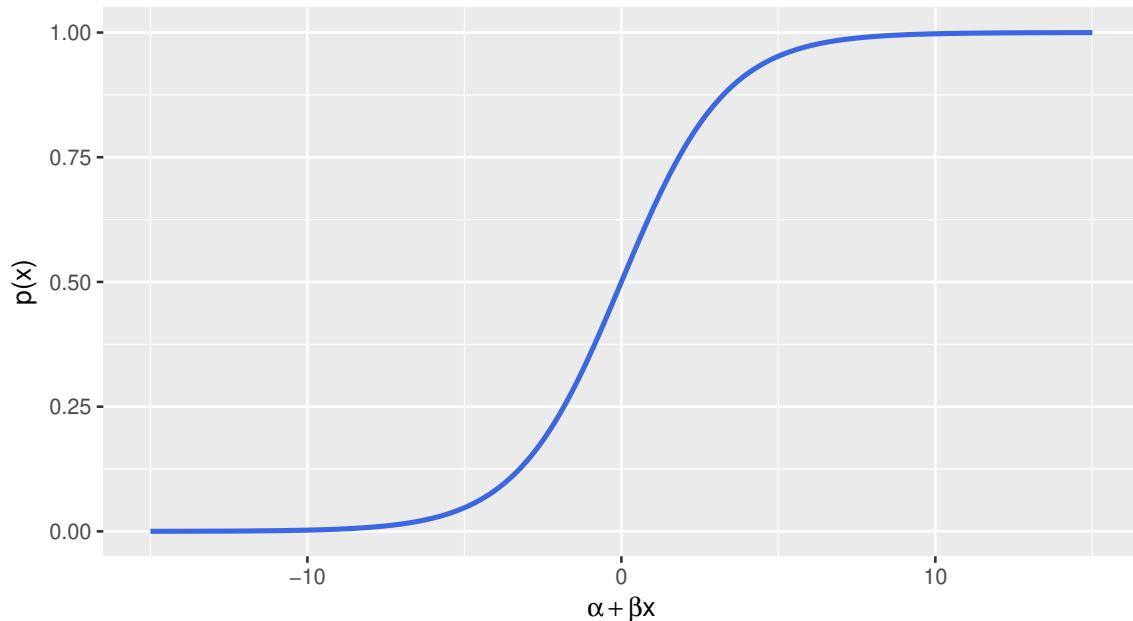


Figura 9.17: Curva logística

Si bien esta transformación, llamada **logit** es la más usada, dado que los coeficientes del modelo tienen una interpretación sencilla, existen otras transformaciones disponibles como **probit**.

En el modelo de regresión logística, los parámetros de mayor interés son los valores de β_i . Sin embargo, la interpretación de estos parámetros es un poco más delicada que en el caso de la regresión lineal.

La transformación logit, considera el cociente entre la probabilidad de que el individuo tenga la característica de interés y la probabilidad de que no la tenga. A este cociente se lo denomina **odds**, **oportunidad** o **chance**.

En la Tabla 9.8 podemos apreciar cómo se vinculan la probabilidad y el *odds*.

El coeficiente β es el cambio en la transformación logit cuando la variable X aumenta en una unidad, o bien, β es el cambio en el logaritmo del *odds ratio* (cociente de oportunidades) entre los

Probabilidad	<i>Odds</i>
0.01	0.010
0.10	0.111
0.20	0.250
0.30	0.429
0.40	0.667
0.50	1.000
0.60	1.500
0.70	2.333
0.80	4.000
0.90	9.000
0.99	99.000

Tabla 9.8: Comparación entre probabilidades y *odds*

grupos definidos por $X = x_0 + 1$ y por $X = x_0$. Simbólicamente, se tiene que

$$\ln(\text{odds}(x+1)) = \ln\left(\frac{p(x+1)}{1-p(x+1)}\right) = \alpha + \beta(x+1)$$

y

$$\ln(\text{odds}(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta x$$

Restando ambas ecuaciones, obtenemos

$$\beta = \ln(\text{odds}(x+1)) - \ln(\text{odds}(x))$$

Por lo tanto el *odds ratio* asociado a un cambio de una unidad para la variable X es e^β .

Cuando la variable es numérica, como puede ser por ejemplo la edad o los ingresos de una persona, este cociente es una medida que cuantifica el cambio en el riesgo cuando se pasa de un valor del factor a otro.

Ejemplo 9.8. En un estudio médico de cáncer de próstata, interesa predecir la ruptura capsular. Para ello se dispone de un conjunto de variables que se sospecha están asociadas con este evento.

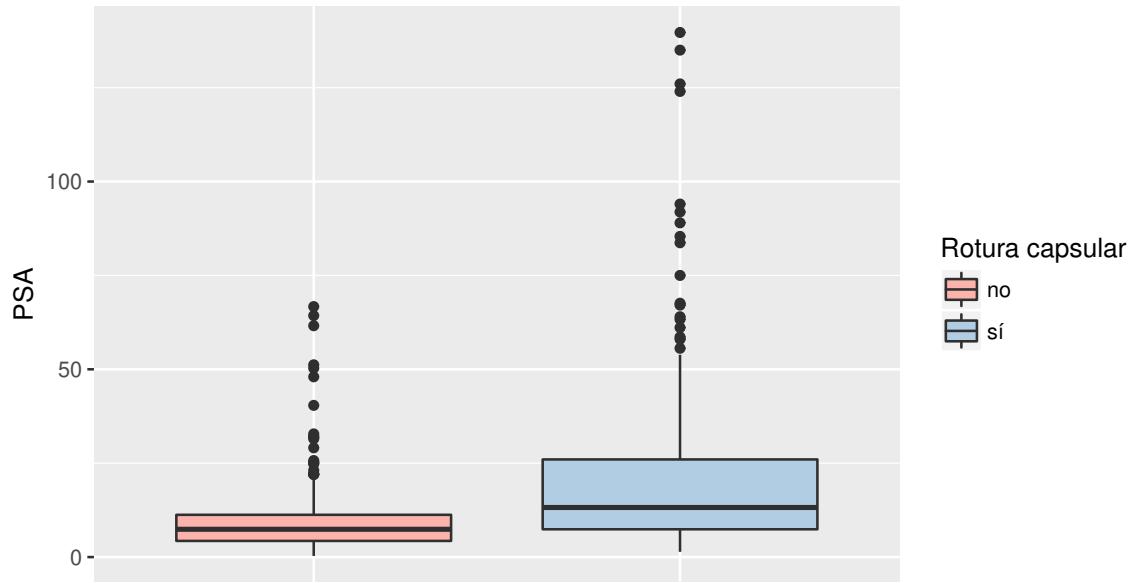


<https://flic.kr/p/m1UHzT>

Las variables de interés son



En todo este análisis, nos referimos al Código 9.8 con datos extraídos de <https://goo.gl/iewgmG>. Nos interesa saber si tiene sentido utilizar el valor del antígeno prostático (PSA) para predecir la ruptura capsular. Para ello graficamos en la Figura 9.19 un *boxplot* comparativo del PSA por los grupos definidos por la variable ruptura capsular.



[Figura 9.19:](#) PSA por ruptura de la cápsula

En la Figura 9.20 se aprecia por qué el modelo logístico es adecuado para modelar la probabilidad de ruptura capsular en función del antígeno prostático.

La salida de R para el modelo logístico resulta

Call:

```
glm(formula = Rotura ~ PSA, family = "binomial", data = prostata)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.1610	-0.9032	-0.7998	1.2528	1.6402

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.11370	0.16156	-6.893	5.45e-12	***
PSA	0.05018	0.00925	5.424	5.82e-08	***
<hr/>					
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

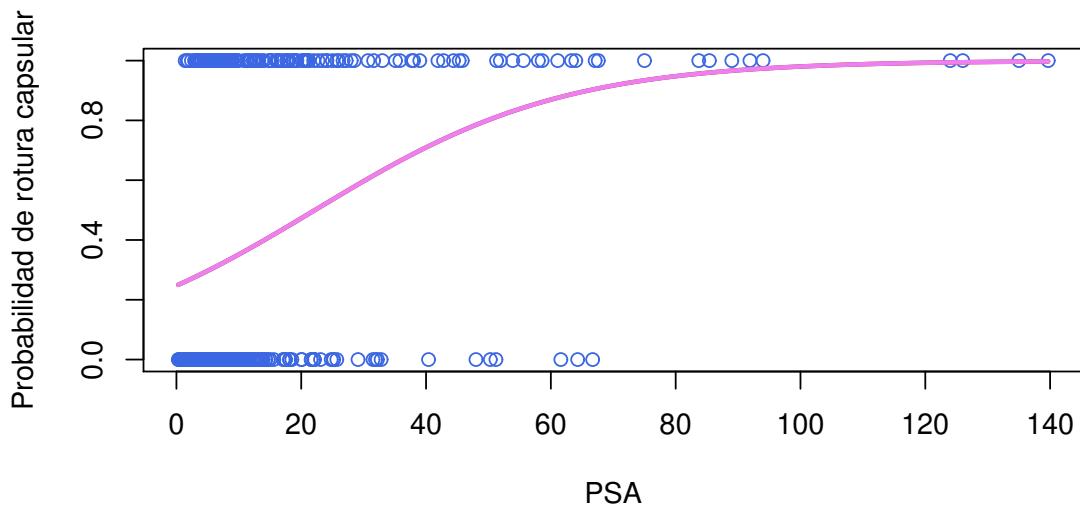


Figura 9.20: Probabilidad de rotura capsular en función de PSA

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 512.29 on 379 degrees of freedom

Residual deviance: 463.16 on 378 degrees of freedom

AIC: 467.16

Number of Fisher Scoring iterations: 5

El número 5.82e-08 de la tabla de coeficientes, indica que la variable PSA resulta significativa estadísticamente para predecir la ruptura capsular.

Con el fin de evaluar la calidad del modelo para clasificar, construimos la matriz de confusión de la Tabla 9.9.

		Predicciones	
Observaciones			
	no	sí	
no	210	17	
sí	108	45	

Tabla 9.9: Tabla de confusión para el modelo logístico

Si bien la tasa de buena clasificación, que resultó de 0.6710526, puede ser menor que en la clasificación utilizando SVM, este modelo nos informa cuánto impactan las variables predictoras

sobre la variable respuesta y en qué sentido. También nos permite evaluar el efecto de la interacción entre dos variables predictoras o incorporar una variable transformada.

Consideremos el modelo que usa como predictor a la variable índice de *gleason*. En la Figura 9.21 se aprecia que la distribución de este índice es diferente en los grupos definidos por la variable ruptura capsular.

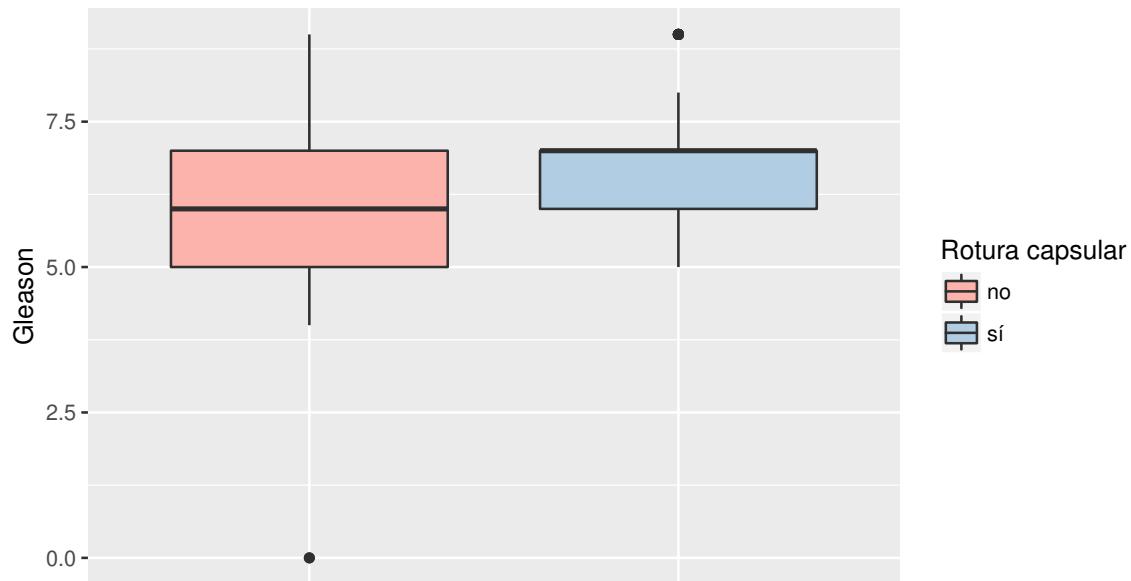


Figura 9.21: Gleason por ruptura de la cápsula

La salida de R para el modelo logístico resulta

Call:

```
glm(formula = Rotura ~ Gleason, family = "binomial", data = prostata)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.3633	-0.7960	-0.4529	1.0725	2.1579

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-8.4196	1.0023	-8.400	< 2e-16	***
PSA	1.2388	0.1525	8.121	4.61e-16	***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 512.29 on 379 degrees of freedom

Residual deviance: 415.70 on 378 degrees of freedom

AIC: 419.7

Number of Fisher Scoring iterations: 4

La variable resulta significativa y el porcentaje de error disminuye respecto del modelo anterior, con tasa de acierto de 0.7210526.

Parece razonable considerar conjuntamente ambas variables. De la salida de R, ambas resultan significativas, sin embargo, la clasificación no se ve beneficiada con la incorporación de las dos variables ya que la tasa de acierto es de 0.7052632. Concluimos entonces que la interacción de estas dos variables no resulta significativa.

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

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

prostata$Rotura=as.factor(prostata$Rotura)
ggplot(prostata, aes(x=Rotura, y=PSA, fill=Rotura)) +
  geom_boxplot() +
  xlab("") +
  scale_fill_brewer(palette="Pastel1", name="Rotura_capsular",
  breaks=c("0", "1"), labels=c("no", "sí")) +
  theme(axis.text.x=element_blank(),
  axis.ticks.x=element_blank())
# Produce boxplots

modelo_logistico=glm(Rotura~PSA, data=prostata, family="binomial")
# Construye el modelo logístico con respuesta binomial
plot(prostata$PSA, prostata$Rotura, col="royalblue", xlab="PSA",
ylab = "Probabilidad_de_rotura_capsular")
curve(predict(modelo_logistico, data.frame(PSA=x), type="response"),
add=TRUE, col="violet", lwd=2.5)
# Produce un gráfico con las variables de interés y la curva predictiva
summary(modelo_logistico)
# Realiza una síntesis del modelo

predicciones=ifelse(test=modelo_logistico$fitted.values>0.5, yes=1, no=0)
# Establece como punto de corte 0.5
matriz_confusion=table(prostata$Rotura, predicciones,
dnn = c("Observaciones", "Predicciones"))
# Calcula la matriz de confusión
mean(prostata$Rotura==predicciones)
# Calcula la tasa de buena clasificación
```

```

ggplot(prostata, aes(x=Rotura, y=Gleason, fill=Rotura)) +
  geom_boxplot() +
  xlab("") +
  scale_fill_brewer(palette="Pastel1", name="Rotura_capsular",
  breaks=c("0", "1"), labels=c("no", "sí")) +
  theme(axis.text.x=element_blank(),
  axis.ticks.x=element_blank())
# Produce boxplots

# Réplica para la variable Gleason
modelo_logist=glm(Rotura~Gleason, data=prostata, family="binomial")
summary(modelo_logist)

predic=ifelse(test= modelo_logist$fitted.values >0.5, yes=1, no=0)
matriz_confus=table(prostata$Rotura, predic,
dnn=c("Observaciones", "Predicciones"))
mean(prostata$Rotura==predic)

# Réplica para ambas variables en conjunto
modelo_logconjunto=glm(Rotura~PSA+Gleason, data=prostata, family = "binomial")
summary(modelo_logconjunto)
prediccionesconj=ifelse(test=modelo_logconjunto$fitted.values >0.5, yes=1,
no=0)
matriz_confconjunta=table(prostata$Rotura, prediccionesconj,
dnn = c("Observaciones", "Predicciones"))
mean(prostata$Rotura==prediccionesconj)

```

Código 9.8: Código para regresión logística

9.6 Ejercitación

Ejercicio 1.

La base de datos disponible en <https://goo.gl/FVqX22> consiste en 49 registros de gorriones sobre los que se han medido las variables: Largo, Alas, Cabeza, Pata, Cuerpo y Sobrevida explicadas en el Ejercicio 3 del Capítulo 2.

1. Comparar las medias de cada una de las variables entre los grupos. Realizar una exploración gráfica.
2. Comparar los vectores medios de ambos grupos. ¿Tiene sentido realizar un análisis discriminante?
3. Realizar el análisis discriminante a partir de las variables que se considere adecuado incluir.
4. ¿Se satisfacen los supuestos del modelo? ¿Resulta una buena clasificación?.

Ejercicio 2.

El archivo de datos disponible en <https://goo.gl/tej4gK> contiene 74 registros de pacientes sobre los cuales se han medido dos variables continuas dadas por Actividad AHF y Antígeno AHF. Con estas variables se pretende predecir el grupo de pertenencia respecto de la portación de hemofilia.

1. ¿Puede considerarse que ambas variables resultan ser de ayuda para esta clasificación?
2. Realizar un gráfico bivariado para ambos grupos de manera conjunta.
3. A partir del gráfico del ítem anterior, ¿es razonable que una función discriminante lineal sea adecuada?
4. Realizar un análisis discriminante con 50 registros elegidos al azar.
5. Utilizar los restantes registros para estimar la calidad de la regla discriminante.

Ejercicio 3.

Los datos del archivo disponible en <https://goo.gl/7bp93C> contienen las mediciones del pulso antes de realizar un ejercicio y del pulso después de realizarlo, para un conjunto de 40 individuos formado por hombres y mujeres, entre los cuales hay fumadores y no fumadores.

1. Interesa saber si la información del pulso antes y después de correr permite discriminar el sexo.
2. Idem inciso anterior pero para la categoría de fumador.

3. ¿En cuál de los dos casos se discrimina mejor?

Ejercicio 4.

La base de datos `iris` de R contiene 150 registros correspondientes a tres especies de la flor de iris. Estos datos pertenecen a un clásico ejemplo debido a Fisher (1936). El objetivo consiste en clasificar estas subespecies a partir de las 4 variables que incluyen medidas del sépalo y del pétalo de cada flor.

1. Analizar qué valores medios son diferentes en las especies.
2. Aplicar alguna regla de clasificación para discriminar los 3 grupos.
3. ¿Cuál es el porcentaje de bien clasificados? ¿Y los porcentajes de bien clasificados para cada especie?
4. ¿Qué registros no se clasificaron correctamente? ¿Puede darse una explicación sobre el porqué de esta situación?

Ejercicio 5.

La base de datos disponible en <https://goo.gl/zBTEwN> consiste de 37 registros geoposicionales que informan sobre las coordenadas en las cuales 37 tormentas se transformaron en huracanes, para 2 clasificaciones de huracanes conocidas como Baro y Trop. Estos datos son ficticios y pertenecen a Elsner, Lehmler, and Kimberlain [15].

1. Realizar un análisis discriminante teniendo como objetivo la clasificación de los huracanes.
2. Encontrar la expresión de la función discriminante.
3. ¿Qué cantidad de huracanes han sido bien clasificados? ¿Qué puede decirse sobre este resultado?
4. Apoyándose en el uso de gráficos, interpretar el resultado del inciso anterior.

Ejercicio 6.

Para regular la pesca de salmón, se desea identificar si el pescado es originario de Alaska o de Canadá. Cincuenta peces de cada lugar de origen fueron capturados y pesados cuando vivían en agua dulce y cuando vivían en agua salada. El objetivo es poder identificar si los nuevos pescados vienen de criaderos en Alaska o Canadá. Los datos correspondientes están disponibles en <https://goo.gl/HfeFPA>.

1. Graficar sobre un mismo diagrama de dispersión los pesos de los salmones de los dos orígenes en colores distintos, para visualizar si los vectores de medias de los grupos son similares o no.

2. Aplicar el test de Hotelling para testear si los vectores medios son iguales en ambos grupos.
3. Si se rechaza la hipótesis de nulidad del ítem anterior, construir una regla discriminante lineal o cuadrática, según corresponda, para poder clasificar a un nuevo pez según su origen.
4. Evaluar la capacidad discriminante de la regla construida.
5. Clasificar a un nuevo pez con la regla del ítem anterior, sabiendo que sus pesos en agua dulce y mar respectivamente fueron de 120 y 400 respectivamente.

Capítulo 10

Métodos de clasificación no supervisada

Como hemos visto en el Capítulo anterior, los métodos de clasificación supervisada consisten en hallar una función que permita asignar un valor objetivo a observaciones que el sistema no ha visto anteriormente a partir de un conjunto de datos de los que se conoce su valor objetivo; es decir, un conjunto de entrenamiento. Por el contrario, en los métodos de clasificación no supervisada no se tiene una salida esperada de asociación a los datos con los cuales se está trabajando, sino que partiendo de las propiedades de estos datos, se busca obtener una agrupación o caracterización, *clustering* en inglés, de los datos según la similaridad entre sus propiedades.

En los métodos no supervisados el algoritmo clasificador requiere simplemente de la información observada del grupo de estudio y ciertos parámetros que limiten el número de clases. Estos mecanismos de clasificación basan su efecto en la búsqueda de clases con suficiente separabilidad espectral como para conseguir diferenciar unos elementos de otros.

10.1 Distancias y medidas de proximidad

Con frecuencia, resulta de interés establecer alguna manera de medir la proximidad entre diferentes tipos de observaciones multivariadas. Una posibilidad para esta medida es una distancia; es decir, una función que aplicada a dos vectores de observaciones de cierto espacio, cuantifique la proximidad o similaridad entre ellos. Formalicemos ahora este concepto de distancia.

Sea X un conjunto, se define una **distancia** o **métrica** como una función que aplicada a un par de elementos de X da un valor numérico no negativo. Más precisamente, $X \times X \rightarrow \mathbb{R}$ y, para todo $x, y, z \in X$, se satisfacen las siguientes condiciones:

- ✿ **no negatividad:** $d(x, y) \geq 0$
- ✿ **simetría:** $d(x, y) = d(y, x)$
- ✿ **desigualdad triangular:** $d(x, z) \leq d(z, y) + d(y, z)$

A partir de la definición es sencillo probar que la distancia de un vector a sí mismo es nula; es decir, para todo $x \in X$, $d(x, x) = 0$. Más aún, si la distancia entre dos vectores es nula, quiere decir que los vectores son iguales, expresado simbólicamente como $d(x, y) = 0 \Rightarrow x = y$. En el caso de no exigirse el cumplimiento de esta última condición, la función d se denomina **pseudodistancia** o **pseudométrica**.

Se denomina **espacio métrico** al par (X, d) .

A continuación analizaremos diferentes medidas teniendo en cuenta el tipo de observaciones.

10.1.1 Medidas de distancia para vectores de observaciones continuas

Consideremos dos vectores x e y de dimensión k que contienen valores de variables continuas. Se define la **distancia de Minkowsky** o **norma L_p** como

$$d_p(x, y) = \left[\sum_{i=1}^k (x_i - y_i)^p \right]^{\frac{1}{p}}$$

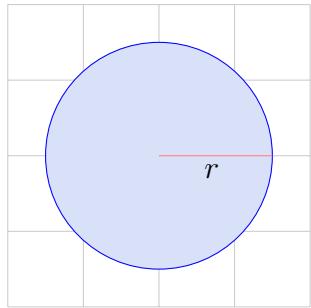
En la Tabla 10.1 se exhiben algunos casos particulares de esta medida.

p	Distancia	Fórmula
2	Euclídea	$d_2(x, y) = \left[\sum_{i=1}^k (x_i - y_i)^2 \right]^{\frac{1}{2}}$
1	Manhattan	$d_1(x, y) = \sum_{i=1}^k x_i - y_i $
∞	Chebyshev	$d_\infty(x, y) = \max_{1 \leq i \leq k} x_i - y_i $

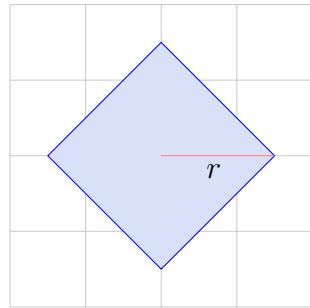
Tabla 10.1: Distintas medidas L_p

En el caso de la distancia Euclídea, los puntos que están a distancia r de un punto dado denominado centro, conforman una circunferencia de radio $r > 0$. El interior de esta circunferencia, está constituido por todos los puntos del espacio métrico cuya distancia al centro es menor que la longitud del radio r . El conjunto formado por la circunferencia y su interior se llama círculo. Para la distancia de Manhattan, o *city block*, los puntos que están a distancia r de un punto dado conforman un rombo cuyas diagonales miden $2r$. Mientras que para la distancia Chebyshev, los puntos que están a distancia r de un punto dado conforman un cuadrado cuyos lados miden $2r$. En la Figura 10.1 se muestran estas características geométricas.

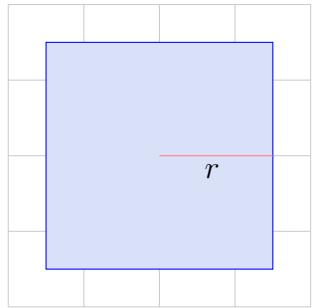
En la Figura 10.2 se puede apreciar que la distancia euclíadiana, representada en color rojo, no se corresponde con el camino real más corto para llegar de un punto a otro en el plano figurado de cierta ciudad, además de no ser éste único, como se señala en los colores verde, azul y violeta.



(a) Euclídea



(b) Manhattan



(c) Chebyshev

Figura 10.1: Interpretación geométrica de distancias L_p

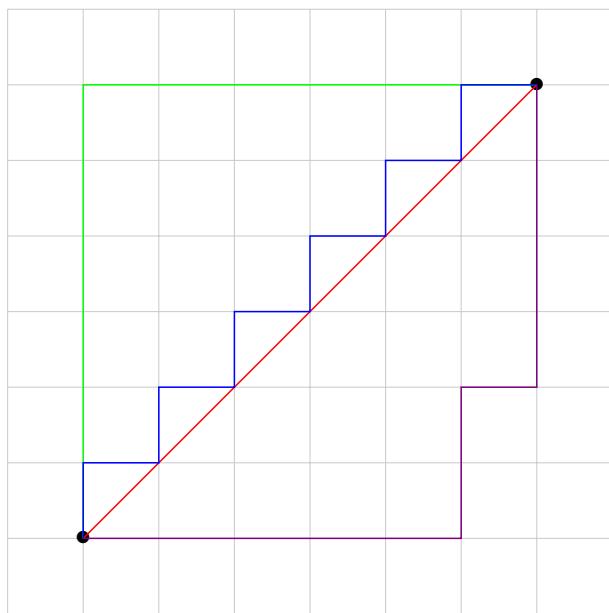


Figura 10.2: Ejemplo de la distancia *city blocks*

Ejemplo 10.1. En la Tabla 10.2 se consideran las observaciones correspondientes a tres nadadores para los cuales se han registrado los tiempos empleados para correr cada uno de los cuatro tramos en los que se dividió la carrera.

Nadador	x_1	x_2	x_3	x_4
A	10	10	13	12
B	12	12	14	15
C	11	10	14	13

Tabla 10.2: Tiempos para los nadadores

* Distancias euclídeas

$$d_2(A, B) = \sqrt{(10 - 12)^2 + (10 - 12)^2 + (13 - 14)^2 + (12 - 15)^2} = \sqrt{18}$$

$$d_2(A, C) = \sqrt{(10 - 11)^2 + (10 - 10)^2 + (13 - 14)^2 + (12 - 13)^2} = \sqrt{3}$$

$$d_2(B, C) = \sqrt{(12 - 11)^2 + (12 - 10)^2 + (14 - 14)^2 + (15 - 13)^2} = 3$$

* Distancias de Manhattan

$$d_1(A, B) = |10 - 12| + |10 - 12| + |13 - 14| + |12 - 15| = 8$$

$$d_1(A, C) = |10 - 11| + |10 - 10| + |13 - 14| + |12 - 13| = 3$$

$$d_1(B, C) = |12 - 11| + |12 - 10| + |14 - 14| + |15 - 13| = 5$$

* Distancias de Chebyshev

$$d_\infty(A, B) = \max\{|10 - 12| + |10 - 12| + |13 - 14| + |12 - 15|\} = 3$$

$$d_\infty(A, C) = \max\{|10 - 11| + |10 - 10| + |13 - 14| + |12 - 13|\} = 1$$

$$d_\infty(B, C) = \max\{|12 - 11| + |12 - 10| + |14 - 14| + |15 - 13|\} = 2$$



En la Tabla 10.3, se mencionan algunas versiones de medidas más generalizadas que incluyen pesos para cada variable $i = 1, \dots, k$. La distancia ponderada euclídea o de Minkowski se basa en asignar distintos pesos a las variables teniendo en cuenta las situaciones más y menos favorables. Se puede observar que la distancia de Mahalanobis es un caso particular de la distancia cuadrática usando como matriz de ponderación a la inversa de la matriz de covarianzas; es decir, $Q = V^{-1}$. La distancia de Canberra ha sido utilizada como métrica para comparar listas ordenadas o clasificadas [27] y para la detección de intrusos en seguridad informática [16].

Distancia	Fórmula	Comentarios
Ponderada Minkowski	$d_M(x, y) = \left[\sum_{i=1}^k w_i (x_i - y_i)^p \right]^{\frac{1}{p}}$	$w_i \in [0, 1], \sum_{i=1}^k w_i = 1$
Cuadrática	$d_Q(x, y) = (x - y)^t Q (x - y)$	$Q \in \mathbb{R}^{k \times k}$ definida positiva
Canberra	$d_C(x, y) = \sum_{i=1}^k \frac{ x_i - y_i }{ x_i + y_i }$	Si $x_i = y_i = 0$, el término i -ésimo se considera nulo

Tabla 10.3: Distintas medidas con peso

Ejemplo 10.2. Utilizando los datos de la Tabla 10.2, calculamos algunas distancias ponderadas.

- ✿ **Distancias ponderadas de Minkowski** utilizando los pesos $w_1 = 0.30$, $w_2 = 0.20$, $w_3 = 0.15$, $w_4 = 0.35$ y con $p = 2$.

$$d_M(A, B) = \sqrt{0.30 \cdot (-2)^2 + 0.20 \cdot (-2)^2 + 0.15 \cdot (-1)^2 + 0.35 \cdot (-3)^2} = 2.3022$$

$$d_M(A, C) = \sqrt{0.30 \cdot (-1)^2 + 0.20 \cdot 0^2 + 0.15 \cdot (-1)^2 + 0.35 \cdot (-1)^2} = 0.8944$$

$$d_M(B, C) = \sqrt{0.30 \cdot 1^2 + 0.20 \cdot 2^2 + 0.15 \cdot 0^2 + 0.35 \cdot 2^2} = 1.5811$$

- ✿ **Distancias cuadráticas** con $Q = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$.

$$d_Q(A, B) = (-2 \ -2 \ -1 \ -3) Q (-2 \ -2 \ -1 \ -3)^t = 109$$

$$d_Q(A, C) = (-1 \ 0 \ -1 \ -1) Q (-1 \ 0 \ -1 \ -1)^t = 15$$

$$d_Q(B, C) = (1 \ 2 \ 0 \ 2) Q (1 \ 2 \ 0 \ 2)^t = 44$$

- ✿ **Distancias de Canberra**

$$d_C(A, B) = \frac{|10 - 12|}{10 + 12} + \frac{|10 - 12|}{10 + 12} + \frac{|13 - 14|}{13 + 14} + \frac{|12 - 15|}{12 + 15} = 0.3299$$

$$d_C(A, C) = \frac{|10 - 11|}{10 + 11} + \frac{|10 - 10|}{10 + 10} + \frac{|13 - 14|}{13 + 14} + \frac{|12 - 13|}{12 + 13} = 0.1247$$

$$d_C(B, C) = \frac{|12 - 11|}{12 + 11} + \frac{|12 - 10|}{12 + 10} + \frac{|14 - 14|}{14 + 14} + \frac{|15 - 13|}{15 + 13} = 0.2058$$



10.1.2 Medidas de similaridad

Las medidas de similaridad constituyen una extensión de las distancias, satisfaciendo en general todas las condiciones de la definición de distancia excepto la propiedad de la desigualdad triangular.

En la Tabla 10.4 se presentan algunos ejemplos de medidas de similaridad considerando x e y dos vectores de k observaciones. La medida de correlación de Pearson también se denomina medida de separación angular o de producto interno normalizado.

Distancia	Fórmula
Correlación de Pearson	$s(x, y) = \frac{(x - \bar{x})^t(y - \bar{y})}{\ x - \bar{x}\ \ y - \bar{y}\ } = \frac{\sum_{i=1}^k (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^k (x_i - \bar{x})^2 \sum_{i=1}^k (y_i - \bar{y})^2}}$
Tanimoto	$s_T(x, y) = \frac{x^t y}{\ x\ + \ y\ - x^t y}$

Tabla 10.4: Distintas medidas de similaridad

La distancia de Tanimoto puede extenderse para el caso de datos nominales de la siguiente manera. Sean A y B dos conjuntos de cardinalidades n_A y n_B respectivamente. Denotamos el cardinal de la intersección entre A y B como $n_{A \cap B}$. Se define la medida de Tanimoto entre ellos como la razón del número de elementos que los conjuntos tienen en común entre el número de elementos distintos; es decir,

$$s_T(A, B) = \frac{n_{A \cap B}}{n_A + n_B - n_{A \cap B}}$$

Ejemplo 10.3. Para los conjuntos considerados en la Figura 10.3, se tiene que

$$d_T(A, B) = \frac{10}{40 + 50 - 10} = \frac{1}{8}$$

■

En el caso particular de variables binarias, las entradas de los dos vectores pueden resumirse en la Tabla 10.5. Donde a representa el número de posiciones donde los vectores X e Y coinciden en tomar el valor 0. Similarmente, d representa el número de coincidencias donde X e Y valen ambos 1, mientras que c y b representan el número de no coincidencias.

Observemos que, en este caso, la distancia de Tanimoto se reduce a

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

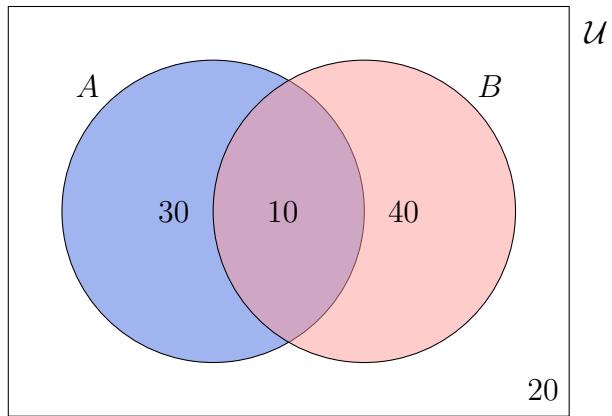


Figura 10.3: Número de características

		Y	
		0	1
X	0	a	b
	1	c	d

Tabla 10.5: Entradas de vectores binarios

Al momento de comparar similaridades y diversidades de una muestra, puede ser de utilidad contar con ciertos índices o coeficientes, siendo un concepto más amplio de medidas de distancia o similaridad para introducir una nueva técnica de análisis multivariado. Algunos ejemplos se presentan en la Tabla 10.6.

Ejemplo 10.4. Consideramos dos sujetos distintos, A y B, a los cuales se les ha observado presencia (1) o ausencia (0) de hipertensión arterial (HTA), obesidad (OBS), ataque cerebral (ACV) y diabetes (DBT). Queremos calcular la similitud entre dos pacientes en cuyas historias clínicas se han registrado los datos que figuran en la Tabla 10.7.

La distancia de Tanimoto es

$$d_T(A, B) = \frac{0 + 2}{0 + 2(1 + 1) + 2} = \frac{2}{6} = \frac{1}{3}$$

A partir de la Tabla 10.8, obtenemos los siguientes coeficientes:

- ✿ **Coeficiente de Coincidencias simples:** $\frac{0 + 2}{0 + 1 + 1 + 2} = \frac{1}{2}$
- ✿ **Coeficiente de Jaccard:** $\frac{2}{1 + 1 + 2} = \frac{1}{2}$

Coeficiente	Fórmula	Descripción
Coincidencias simples	$\frac{a + d}{a + b + c + d}$	Mide la proporción de emparejamiento entre dos observaciones
Jaccard	$\frac{d}{b + c + d}$	Mide la proporción de emparejamiento si al menos uno de los vectores admite un 1
Phi	$\frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$	Mide la correlación lineal entre los vectores
Anderberg	$\frac{t_1 - t_2}{2(a + b + c + d)}$ $t_1 = \max\{a, b\} + \max\{a, c\} + \max\{b, d\} + \max\{c, d\}$ $t_2 = \max\{a + c, b + d\} + \max\{a + b, c + d\}$	Mide la predicción del estado de una característica en un vector dado el estado de la misma en el otro vector
Dice	$\frac{2d}{b + c + 2d}$	Es similar a Jaccard asignando doble peso a la coincidencia de 1
Ochiai	$\frac{d}{\sqrt{(b + d)(c + d)}}$	Mide la cohesión entre agrupamientos
Russel-Rao	$\frac{a + d}{\sqrt{a + d + 2(b + c)}}$	Es el valor predeterminado para los datos de similitud binarios
Hamann	$\frac{(a + d) - (b + c)}{a + b + c + d}$	Mide la proporción de la diferencia entre coincidencias y discrepancias
Sneath-Sokal	$\frac{2(a + d)}{2(a + d) + b + c}$	Es similar a coincidencias simples asignando doble peso a los emparejamientos
Rogers-Tanimoto	$\frac{a + d}{a + d + 2(b + c)}$	Es opuesto a Sneath-Sokal asignando doble peso a las discrepancias
Yule	$\frac{ad - bc}{ad + bc}$	Mide la fracción de asociación perfecta de 1

Tabla 10.6: Coeficientes de similaridad

Paciente	HTA	OBS	ACV	DBT
A	1	1	1	0
B	1	1	0	1

Tabla 10.7: Registros para los pacientes

		B	
		0	1
A	0	0	1
	1	1	2

Tabla 10.8: Entradas binarias para los pacientes

- * **Coeficiente de Phi:** $\frac{0 \cdot 2 - 1 \cdot 1}{\sqrt{(0+1)(0+1)(1+2)(1+2)}} = -\frac{1}{3}$
- * **Coeficiente de Anderberg:** $\frac{(1+1+2+2) - (3+3)}{2(0+1+1+2)} = 0$
- * **Coeficiente de Dice:** $\frac{2 \cdot 2}{1+1+2+2} = \frac{2}{3}$
- * **Coeficiente de Ochiai:** $\frac{2}{\sqrt{(1+2)(1+2)}} = \frac{2}{3}$
- * **Coeficiente de Russel-Rao:** $\frac{0+2}{\sqrt{0+2+2(1+1)}} = \frac{2}{\sqrt{6}}$
- * **Coeficiente de Hamann:** $\frac{(0+2) - (1+1)}{0+1+1+2} = 0$
- * **Coeficiente de Sneath-Sokal:** $\frac{2(0+2)}{2(0+2)+1+1} = \frac{2}{3}$
- * **Coeficiente de Rogers-Tanimoto:** $\frac{0+2}{0+2+2(1+1)} = \frac{2}{3}$
- * **Coeficiente de Yule:** $\frac{0 \cdot 2 - 1 \cdot 1}{0 \cdot 2 + 1 \cdot 1} = -1$



10.2 Introducción al análisis de conglomerados o *clusters*

Los objetivos de esta sección consisten en:

- ✿ Obtener una representación “compacta” de los datos.
- ✿ Generar una clasificación de los datos.
- ✿ Alcanzar una mayor comprensión de la estructura de los datos a partir de esta clasificación.

Trabajamos con un conjunto de n observaciones que pueden ser animales, plantas, sucursales, hospitales, entre otros; y de los cuales disponemos de p variables que los caracterizan. En este contexto, queremos encontrar una división útil en un número de clases desconocido *a priori*. Una vez determinado el número de clases y realizada la división, resulta de interés estudiar las características distintivas de cada una de estas clases.

El **análisis de *clusters* o conglomerados** es una técnica de reducción de datos que pretende la subdivisión de la población en subgrupos más manejables. El mismo es un método propio del análisis exploratorio de datos, que permite descubrir asociaciones y estructuras en los datos que no son evidentes pero que pueden ser útiles una vez que se han detectado. Si, por ejemplo, un investigador ha recogido datos a partir de un cuestionario, puede enfrentarse a un número elevado de observaciones que no tendrá sentido a menos que clasifique en grupos manejables.

Los resultados de un análisis de *clusters* pueden contribuir a:

- ✿ la definición formal de un esquema de clasificación tal como una taxonomía para un conjunto de objetos.
- ✿ la determinación de modelos estadísticos para describir poblaciones y asignar individuos a éstas.
- ✿ el descubrimiento de rasgos característicos de ciertas subpoblaciones.

Este análisis se aplica con frecuencia en Psicología para la clasificación o descripción de tipologías personales, así como también en la segmentación del mercado.

Al igual que muchos procedimientos multivariados, el análisis de *clusters* conlleva inicialmente a una pérdida de información. Se recurre a técnicas de agrupamiento cuando no se conoce una estructura de asociación de los datos *a priori* y el objetivo operacional es identificar la agrupación natural de las observaciones.

Las técnicas de clasificación basadas en agrupamientos implican la distribución de las unidades de estudio en clases o categorías de manera tal que cada clase, o conglomerado, reúne unidades cuya similitud es máxima bajo cierto criterio. Es decir, los objetos en un mismo grupo comparten el mayor número permisible de características y los objetos en diferentes grupos tienden a ser distintos.

10.2.1 Análisis de *clusters* por individuos o variables

En líneas generales lo que se pretende agrupar son individuos. Sin embargo, existen algunas circunstancias en las cuales es interesante agrupar variables con el propósito de intentar buscar las que resulten de comportamiento similar. En este caso, la metodología es la misma que para el análisis de *clusters* por individuos y simplemente, se debe trasponer la matriz de datos para aplicar luego el método general.

Para agrupar objetos, casos o variables, es necesario seguir cierto algoritmo. Los algoritmos o métodos de agrupamiento permiten identificar clases existentes en relación a un conjunto dado de atributos o características.

El agrupamiento logrado dependerá de lo siguiente:

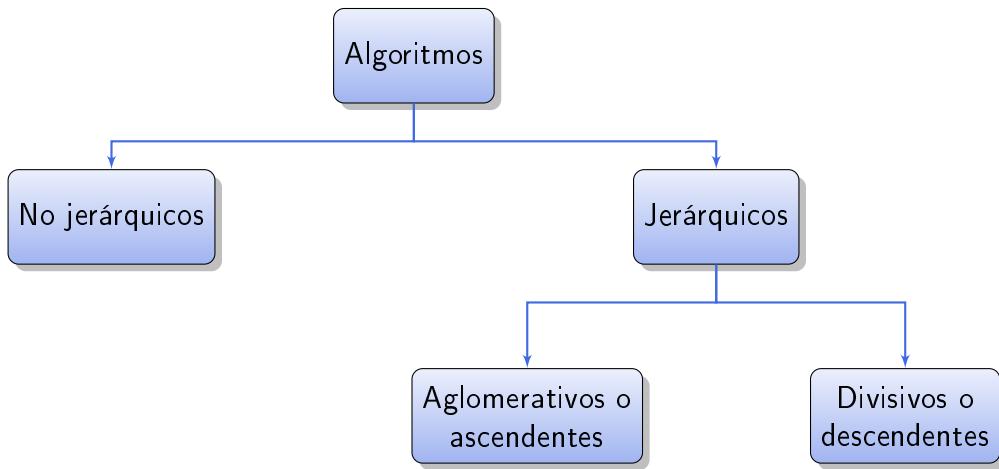
- ✿ el algoritmo de agrupamiento o división utilizado
- ✿ la distancia seleccionada
- ✿ la cantidad de grupos deseados (de existir esta información)
- ✿ las variables utilizadas para el método y las disponibles en la base
- ✿ la condición de estandarización o no de las variables seleccionadas

Existen diferentes alternativas para la validación de los *clusters*. Podemos mencionar los siguientes criterios.

- ✿ **Criterios externos:** comparan la clusterización con una segmentación previa de referencia.
- ✿ **Criterios internos:** analizan la significatividad de los *clusters* sólo considerando los datos usados en la clusterización.
- ✿ **Criterios relativos:** comparan la clusterización con otras resultantes de segmentaciones alternativas.

10.2.2 Métodos de agrupamiento

Los algoritmos de clasificación pueden dividirse de la siguiente manera



Entre los métodos no jerárquicos o de partición, se desea obtener una única descomposición o partición del conjunto original de objetos en base a la optimización de una función objetivo. El más conocido es el denominado K-medias o, en inglés, *K-means*.

Por otro lado, los algoritmos jerárquicos pretenden encontrar particiones jerarquizadas consecutivamente más (o menos) finas para que luego los objetos sean unidos (o separados) en grupos paso por paso. En este tipo de métodos, la clasificación resultante tiene un número creciente de clases anidadas mientras que en los no jerárquicos las clases no son anidadas.

Dentro de las ventajas de los métodos jerárquicos, podemos mencionar que:

- ✿ Sugieren el número de *clusters*.
- ✿ Establecen una jerarquía entre los *clusters*.
- ✿ Posibilitan visualizar el proceso mediante un gráfico denominado **dendograma**.

En contrapartida, algunas de sus desventajas son que los mismos resultan muy costosos en grandes bases de datos y que son lentos.

También existen los llamados métodos mixtos.

Muchas veces, informaciones preliminares disponibles o resultados de experimentos pilotos, pueden orientar al experimentador o usuario en la selección del número de clases. En otras ocasiones, se conoce algún valor máximo para el número de clases, y entonces el algoritmo se implementa especificando dicho valor y luego, en relación con los resultados obtenidos, se vuelven a realizar agrupamientos. El análisis de conglomerados es una técnica de clasificación no supervisada.

En el análisis de conglomerados de casos o registros individuales se parte de una matriz de datos de tamaño $n \times p$, siendo p el número de mediciones o variables en cada uno de los n objetos estudiados. La misma es luego transformada en una matriz de distancias de tamaño $n \times n$, donde el elemento ij -ésimo mide la distancia entre pares de objetos i y j para $1 \leq i, j \leq n$.

Cuando se dispone de numerosas variables para realizar el agrupamiento, es común utilizar, previo a la clusterización, **técnicas de reducción de dimensión** tales como el análisis de componentes principales, para obtener un número menor de variables capaces de expresar la variabilidad en los datos. Esta técnica puede facilitar la interpretación de los agrupamientos obtenidos.

En la práctica, se recomienda aplicar varios algoritmos de agrupamiento y de selección o combinación de variables para cada conjunto de datos seleccionando, finalmente, la interpretación más apropiada desde los agrupamientos realizados.

En el caso de comparación de varios agrupamientos alternativos, suele utilizarse el **coeficiente de correlación cofenética**, el cual indica la correlación de las distancias definidas por la métrica de árbol binario con las distancias originales entre objetos. Luego, se espera que el agrupamiento con mayor coeficiente sea el que mejor describe el agrupamiento natural de los datos.

Es importante destacar que los procedimientos de agrupamiento producen resultados exitosos cuando la matriz de datos tiene una estructura de posible interpretación desde el problema que originó la recolección de la información.

Debido a esto, logrados los grupos, es importante caracterizar los mismos a través de diversas medidas resumen para favorecer la interpretación del agrupamiento final.

En la Tabla 10.9 se comparan los métodos aglomerativos con los divisivos.

Aglomerativos	Divisivos
Parten de tantas clases como objetos haya	Parten de una única clase con todos los objetos
Se van obteniendo clases de objetos similares	Se va dividiendo en clases sucesivamente

Tabla 10.9: Métodos aglomerativos versus divisivos

10.2.3 Algoritmos jerárquicos

Los algoritmos jerárquicos producen agrupamientos de tal manera que un conglomerado puede estar completamente contenido dentro de otro, pero no está permitido otro tipo de superposición entre ellos.

Los resultados de agrupamientos jerárquicos se muestran en un diagramas de árboles en dos dimensiones, llamado **dendrograma**, en el que se pueden observar las uniones y/o divisiones que se van realizando en cada nivel del proceso de construcción de conglomerados; es decir, el historial del método.

El dendrograma es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de *clusters* y se utiliza para representar la clasificación jerárquica. Los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud o disimilitud entre los objetos. Las ramas en el árbol representan los conglomerados, y se unen en un nodo cuya posición a lo largo del eje de distancias indica el nivel en el cual la fusión ocurre. El nodo donde todas las entidades forman un único conglomerado, se denomina **nodo raíz**.

Una de las principales características de los procedimientos de agrupamiento jerárquicos aglomerativos es que la ubicación de un objeto en un grupo no cambia; o sea, una vez que un objeto se ubicó en un conglomerado, no se lo reubica en otro. Este objeto puede ser fusionado con otros pertenecientes a algún otro conglomerado para formar un tercero que incluye a ambos.

Los procedimientos jerárquicos descriptos anteriormente no realizan ninguna acción diferencial con observaciones aberrantes. Si una observación rara fue clasificada en etapas tempranas del procedimiento en algún grupo, ésta permanecerá ahí en la configuración final.

Algunos experimentadores, usan la técnica de la perturbación que consiste en la introducción de errores en los datos y reagrupamiento bajo la nueva situación, para probar la estabilidad de la clasificación jerárquica.

A continuación listamos los pasos de la clasificación jerárquica.

1. Decidir qué datos tomar para cada uno de los casos.

En primer lugar, se debe estudiar el tipo de variables con las cuales trabajar. Generalmente se toman varias variables todas del mismo tipo; es decir, todas continuas o todas categóricas. Esto se debe a que suele ser difícil considerar una distancia o medida de similaridad entre distintos tipos de variables.

Si bien sobre cada individuo es posible relevar un gran número de variables, esto no resultará necesariamente útil, dado que, por un lado la inclusión de variables irrelevantes no puede ser contrastada por el análisis de *clusters* y por otro, aumenta la posibilidad de errores de clasificación y genera ruido sobre la conclusión final.

En la selección de variables lo que implica la eliminación de información irrelevante, debe primar el objetivo de la investigación.

En caso en que las variables estén registradas en diferentes escalas y unidades de medición, se suele tipificar a las mismas de tal manera de lograr que todas las variables tengan media nula y desviación típica unitaria, con el fin de evitar la influencia de las unidades de medición de las mismas.

2. Elegir una medida de distancia entre los objetos a clasificar que son las clases iniciales.

Hemos visto que existen diferentes tipos de distancias y medidas de similaridad disponibles, por lo que la selección de la adecuada dependerá de las circunstancias y los casos de estudio.

3. Buscar los *clusters* que resultan ser los más similares.

En este paso se utiliza la minimización de la distancia seleccionada y el método de distancia al *cluster* elegido. Cabe observar que, una vez unidos dos objetos, los mismos no se separarán durante el resto del proceso.

4. Juntar estos dos *clusters* en uno nuevo.

El nuevo *cluster* tendrá a estos dos objetos. De esta manera, el número de *clusters* decrece en una unidad.

5. Seleccionar la técnica de clusterización y calcular la distancia entre este nuevo *cluster* y el resto.

Notar que no es necesario recalcular todas las distancias, solamente las del nuevo *cluster* con los anteriores.

6. Repetir desde el tercer paso hasta que todos los objetos estén reunidos en un único *cluster*.

Luego, se selecciona la cantidad de *clusters* adecuada para la respuesta al problema planteado; es decir, se decide en qué paso de la técnica hay que detenerse.

7. Interpretar los resultados.

Una vez determinados los grupos, corresponderá al investigador de cada campo, ya sea psicólogo, sociólogo, pedagogo u otro, analizar los grupos y el porqué de su formación para obtener las conclusiones relevantes de cada uno, así como las características en las que se diferencian cada conglomerado.

Es importante resaltar que los distintos algoritmos dependen del método utilizado en el paso 5 para calcular la distancia entre *clusters*. Por este motivo, los distintos métodos para el cálculo de las distancias entre *clusters* producen distintas clasificaciones, por lo que no existe una única clasificación correcta.

En la Tabla 10.10 se exhiben distintos algoritmos.

Aglomerativos	Divisivos
Enlace simple (vecino más próximo)	<i>Linkage</i> simple
Enlace completo (vecino más lejano)	<i>Linkage</i> completo
Promedio entre grupos	Promedio entre grupos
Método del centroide	Método del centroide
Método de la mediana	Método de la mediana
Método de Ward	Método de Ward
	Análisis de asociación

Tabla 10.10: Algoritmos jerárquicos

Con el propósito de entender la construcción de un dendograma y su significado utilizaremos el siguiente ejemplo sencillo que lo ilustra.

<i>Cluster</i>	X_1	X_2	X_3
1	1	1	2
2	1	2	5
3	2	1	4
4	5	5	2
5	5	5	6
6	7	5	4
7	8	8	5
8	7	6	5

Tabla 10.11: Observaciones originales

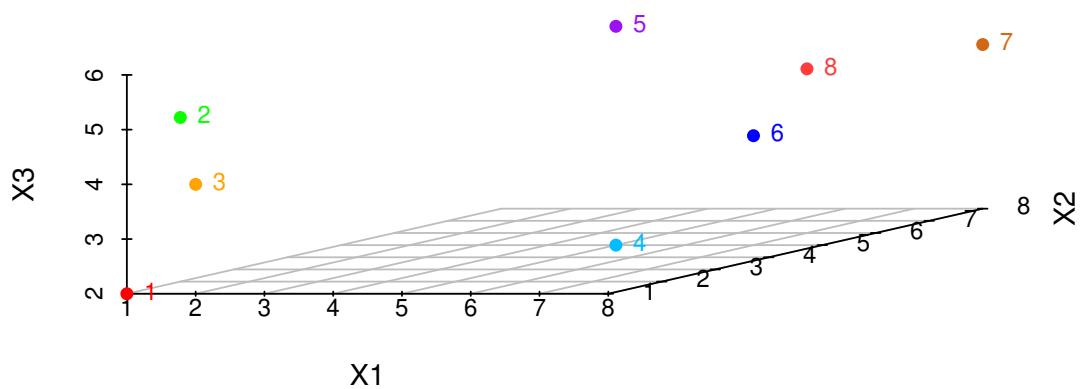


Figura 10.4: Representación de las observaciones originales

Ejemplo 10.5. Supongamos que disponemos de un conjunto de ocho objetos sobre los cuales se han observado las tres variables que figuran en la Tabla 10.11 y que se representan en la Figura 10.4.

A partir de estos datos consideraremos la matriz de distancias euclídeas entre las observaciones, que está dada por la Tabla 10.12.

	1	2	3	4	5	6	7
2	3.16						
3	2.24	1.73					
4	5.66	5.83	5.39				
5	6.93	5.10	5.39	4.00			
6	7.48	6.78	6.40	2.83	2.83		
7	10.34	9.22	9.27	5.20	4.36	3.32	
8	8.37	7.21	7.14	3.74	2.45	1.41	2.24

Tabla 10.12: Distancias euclídeas: primer paso

Realizamos las siguientes observaciones:

- ✿ Inicialmente tenemos 8 *clusters*; es decir, que cada uno de los objetos a clasificar se considera un *cluster* en primera instancia.
- ✿ De acuerdo con la matriz de distancias de la Tabla 10.12, los objetos o *clusters* más próximos son el 6 y el 8, con una distancia de 1.41, y por lo tanto los fusionamos construyendo un nuevo *cluster* que contiene los objetos 6 y 8.
- ✿ Llamaremos *A* al *cluster* formado por los elementos 6 y 8.
- ✿ Tenemos ahora 7 *clusters*.

Se nos plantea ahora un nuevo problema: cómo medir la distancia de este nuevo *cluster*, que se muestra en la Figura 10.5 al resto de los objetos.

Para este ejemplo lo que haremos es tomar como representante del grupo al centroide de los puntos que forman el *cluster*; es decir, el punto que tiene como coordenadas las medias de los valores de las variables para sus componentes. De este modo, las coordenadas de *A* resultan:

$$A = \left(\frac{5+7}{2}, \frac{5+6}{2}, \frac{6+5}{2} \right) = (6, 5.5, 5.5)$$

En la Tabla 10.13 se muestran los nuevos *clusters*. A partir de estas coordenadas calculamos la nueva matriz de distancias 10.14 entre los *clusters* que tenemos en este momento.

Ahora los *clusters* más similares son el *A* y el 5, con distancia 1.22, por lo que se deben fusionar en un nuevo *cluster*, al que llamaremos *B* y que se ilustra en la Figura 10.6. El centroide de este nuevo *cluster* es el punto (5.5, 5.25, 5.75). En la Tabla 10.15 se muestra el nuevo agrupamiento.

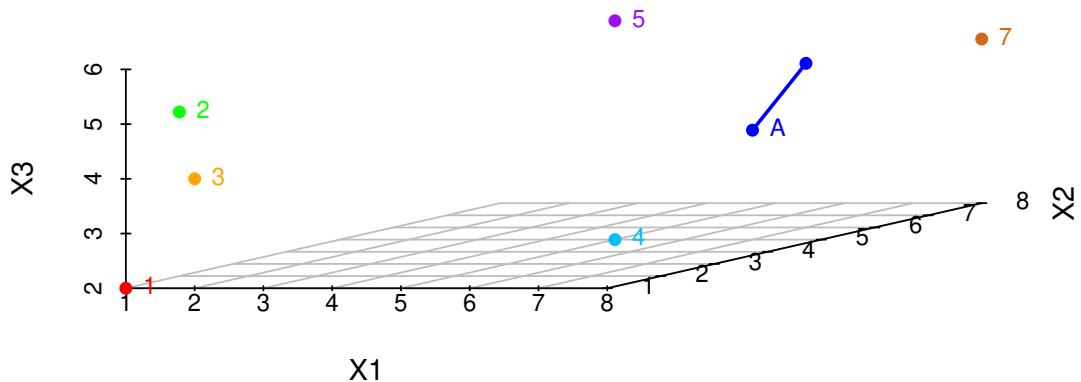


Figura 10.5: Representación del primer paso de clusterización

<i>Cluster</i>	X_1	X_2	X_3
<i>A</i>	6	5.5	5.5
1	1	1	2
2	1	2	5
3	2	1	4
4	5	5	2
5	5	5	6
7	8	8	5

Tabla 10.13: *Clusters* luego del primer paso

	<i>A</i>	1	2	3	4	5
1	7.58					
2	6.12	3.16				
3	6.20	2.24	1.73			
4	3.67	5.66	5.83	5.39		
5	1.22	6.93	5.10	5.39	4.00	
7	3.24	10.34	9.22	9.27	5.20	4.36

Tabla 10.14: Distancias euclídeas: segundo paso

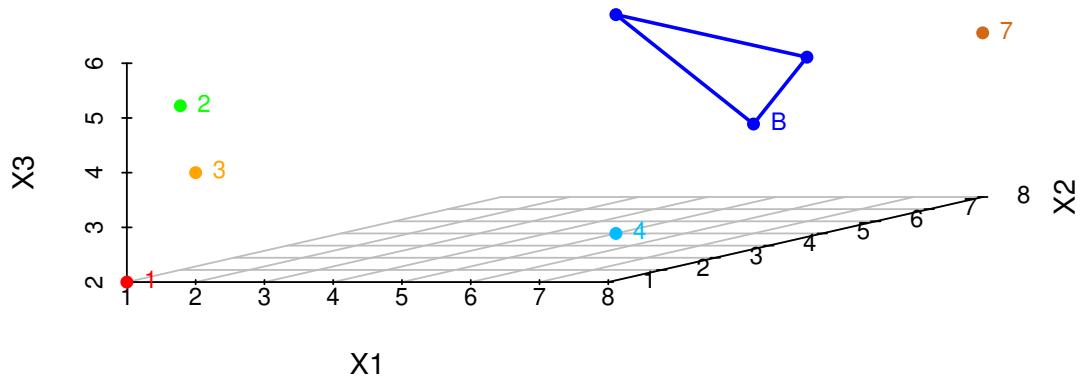


Figura 10.6: Representación del segundo paso de clusterización

<i>Cluster</i>	X_1	X_2	X_3
<i>B</i>	5.5	5.25	5.75
1	1	1	2
2	1	2	5
3	2	1	4
4	5	5	2
7	8	8	5

Tabla 10.15: *Clusters* luego del segundo paso

	<i>B</i>	1	2	3	4
1	7.24				
2	5.60	3.16			
3	5.78	2.24	1.73		
4	3.79	5.66	5.83	5.39	
7	3.79	10.34	9.22	9.27	5.20

Tabla 10.16: Distancias euclídeas: tercer paso

Recalculando la matriz de distancias euclídeas para los nuevos elementos, obtenemos la Tabla 10.16.

Repetimos el proceso agrupando en el *cluster C* los objetos más próximos 2 y 3, con distancia 1.73 y centroide (1.5, 1.5, 4.5), obteniendo la representación dada por la Figura 10.7, los *clusters* de la Tabla 10.17 y las distancias de la Tabla 10.18.

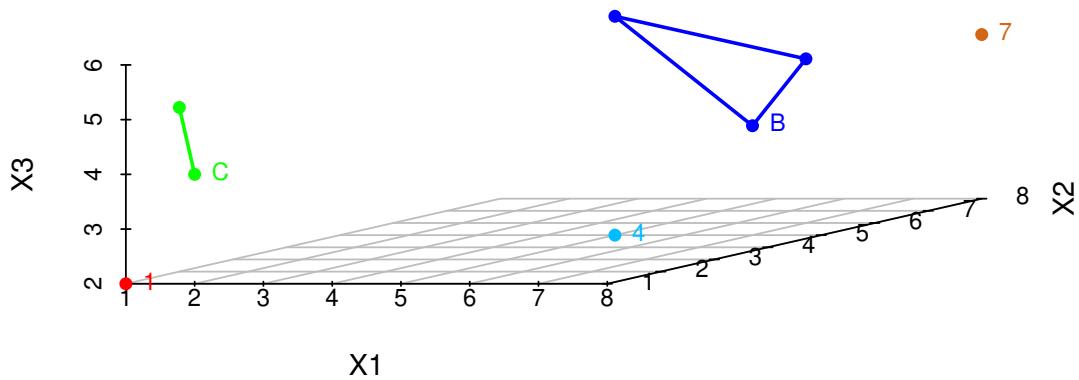


Figura 10.7: Representación del tercer paso de clusterización

<i>Cluster</i>	X_1	X_2	X_3
B	5.5	5.25	5.75
1	1	1	2
C	1.5	1.5	4.5
4	5	5	2
7	8	8	5

Tabla 10.17: *Clusters* luego del tercer paso

Ahora vemos que la distancia más pequeña es 2.60, por lo que agrupamos en el *cluster D* los objetos 1 y C, con centroide (1.25, 1.25, 3.25), obteniendo la representación dada por la Figura 10.8, los *clusters* de la Tabla 10.19 y las distancias de la Tabla 10.20.

Debido a que la mínima distancia es de 3.79 y se realiza tanto entre los objetos 4 y B, como entre 7 y B, agrupamos estos tres objetos en un nuevo *cluster* llamado E tal como se muestra en la Figura 10.9.

	B	1	C	4
1	7.24			
C	5.62	2.60		
4	3.79	5.66	5.55	
7	3.79	10.34	9.21	5.20

Tabla 10.18: Distancias euclídeas: cuarto paso

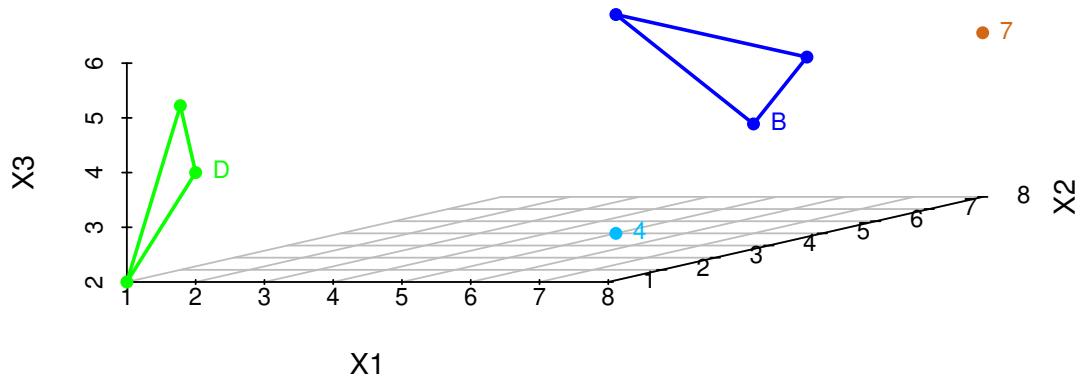


Figura 10.8: Representación del cuarto paso de clusterización

Cluster	X₁	X₂	X₃
<i>B</i>	5.5	5.25	5.75
<i>D</i>	1.25	1.25	3.25
<i>4</i>	5	5	2
<i>7</i>	8	8	5

Tabla 10.19: Clusters luego del cuarto paso

	<i>B</i>	<i>D</i>	4
<i>D</i>	6.35		
4	3.79	5.45	
7	3.79	9.71	5.20

Tabla 10.20: Distancias euclídeas: quinto paso

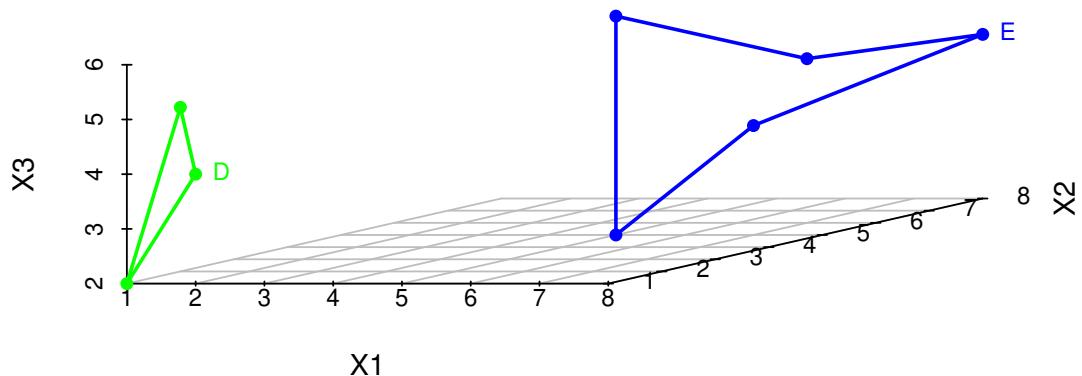


Figura 10.9: Representación del quinto paso de clusterización

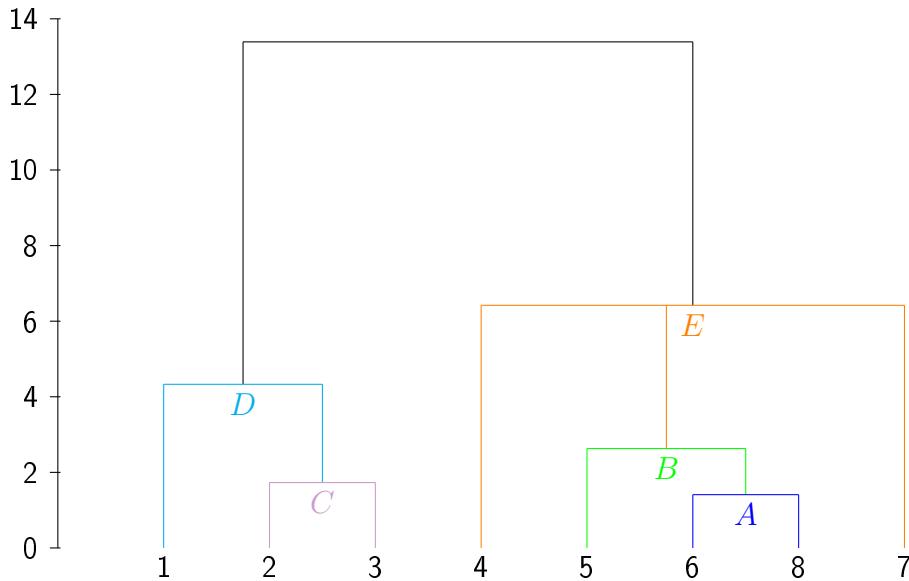


Figura 10.10: Dendograma

De esta manera nos han quedado solamente los dos *clusters* de la Tabla 10.21 con distancia 6.97 y que se fusionarán en el paso siguiente terminando de este modo el proceso.

<i>Cluster</i>	X_1	X_2	X_3
D	1.25	1.25	3.25
E	6.17	6.08	4.25

Tabla 10.21: *Clusters* luego del quinto paso

El proceso completo de fusiones se resume en el dendrograma de la Figura 10.10.

En el gráfico 10.4 parece evidente que tenemos dos *clusters*, el que denominamos *D* y el que denominamos *E*. En general, si cortamos el dendrograma mediante una línea horizontal como se muestra en la Figura 10.11, determinaremos el número de *clusters* en que dividimos el conjunto de objetos. Si, por ejemplo, decidimos cortar con la línea punteada roja, tendremos dos *clusters*, uno con tres elementos y otro con cinco elementos. Si en cambio elegimos la línea de corte punteada gris, tendremos cuatro *clusters* dos con un elemento cada uno y los otros dos con tres elementos cada uno.

La decisión sobre el número óptimo de *clusters* puede ser subjetiva, especialmente cuando se incrementa el número de objetos, ya que si se seleccionan demasiado pocos los *clusters* resultantes son heterogéneos y artificiales; mientras que si se seleccionan demasiados, la interpretación de los

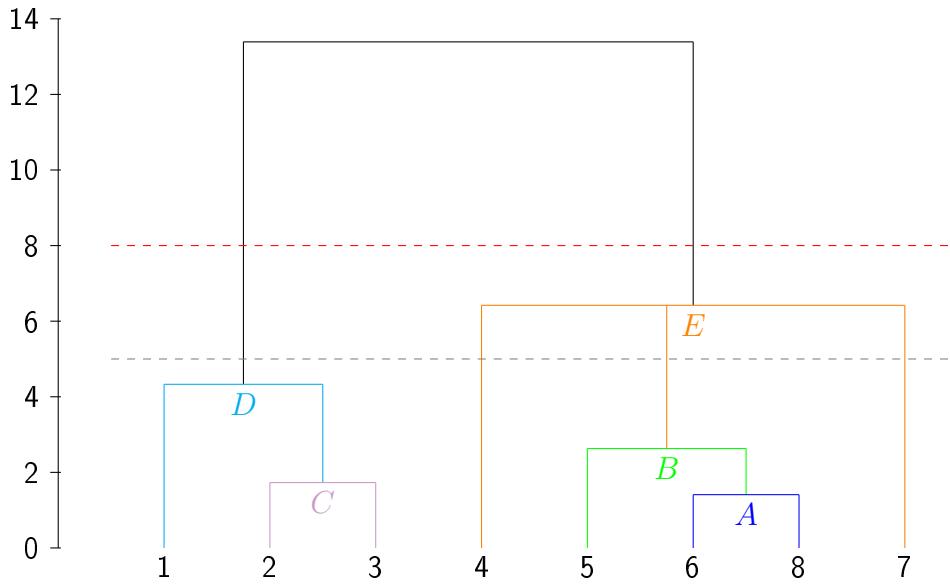


Figura 10.11: Dendograma con elección de cantidad de *clusters*

mismos suele ser complicada. Suele tomarse como ayuda en la decisión sobre el número de *clusters*, representaciones de los distintos pasos del algoritmo y la distancia a la que se produce la fusión. En los primeros pasos el salto en las distancias será pequeño, mientras que en los últimos el salto entre pasos será mayor. El punto de corte será aquel en el que comienzan a producirse saltos bruscos.

En el Ejemplo 10.5, el salto brusco se produce entre los pasos 5 y 6, luego el punto óptimo es el 5, en el que habría 2 *clusters*.

10.2.4 Algoritmos para medir distancia entre *clusters*

En este apartado vamos a presentar diferentes alternativas para medir la distancia entre *clusters*. Como ya indicamos, existen diversas formas de medir la distancia entre *clusters* las cuales producen diferentes agrupamientos y, por lo tanto, diferentes dendogramas.

Lamentablemente, no existe un criterio para decidir cuál de los algoritmos es el mejor. Esta decisión es usualmente subjetiva y depende del método que mejor refleje los propósitos de cada estudio particular.

10.2.4.1 Método de la media o *average linkage*

En el método de la media la distancia entre *clusters* se calcula como la distancia media, o promedio de las distancias, entre pares de observaciones considerando todos los pares formados por un elemento en cada *cluster*. Es decir, si $R = \{P_1, \dots, P_n\}$ y $S = \{Q_1, \dots, Q_m\}$ conforman dos *clusters*,

se mide la distancia entre ellos como

$$d(R, S) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m d(P_i, Q_j)$$

Ejemplo 10.6. Nos referimos al Ejemplo 10.5 a partir de la matriz de distancias 10.12. Luego de agrupar los elementos 6 y 8 en el *cluster A*, calculamos las distancias entre este nuevo *cluster* y el resto de los objetos obteniendo por resultado las distancias que se muestran en la Tabla 10.23. Los detalles de cálculo se presentan el la Tabla 10.22.

	6	8	Promedio
1	7.48	8.37	7.93
2	6.78	7.21	7.00
3	6.40	7.14	6.77
4	2.83	3.74	3.29
5	2.83	2.45	2.64
7	3.32	2.24	2.78

Tabla 10.22: Promedio distancias: primer paso

	A	1	2	3	4	5
1	7.93					
2	7.00	3.16				
3	6.77	2.24	1.73			
4	3.29	5.66	5.83	5.39		
5	2.64	6.93	5.10	5.39	4.00	
7	2.78	10.34	9.22	9.27	5.20	4.36

Tabla 10.23: Distancias con *average linkage*: primer paso

Ahora vemos que la distancia más pequeña se alcanza entre los elementos 2 y 3, por lo que los fusionamos en un *cluster* que denominamos *B*. Calculamos la distancia entre *B* y el resto de los elementos como se muestra en la Tabla 10.24 para luego obtener todas las distancias de la Tabla 10.25.

El menor valor de distancia es 2.64, luego juntamos *A* con 5 en *C* y calculamos la distancia entre *C* y el resto de los elementos como se muestra en la Tabla 10.26 para luego obtener todas las distancias de la Tabla 10.27.

		2	3	Promedio
A	6	6.78	6.40	
	8	7.21	7.14	6.88
	1	3.16	2.24	2.70
	4	5.83	5.39	5.61
	5	6.93	5.39	6.16
	7	9.22	9.27	9.25

Tabla 10.24: Promedio distancias: segundo paso

	A	1	B	4	5
1	7.93				
B	6.88	2.70			
4	3.29	5.66	5.61		
5	2.64	6.93	6.16	4.00	
7	2.78	10.34	9.25	5.20	4.36

Tabla 10.25: Distancias con *average linkage*: segundo paso

		5	6	8	Promedio
B	2	5.10	6.78	7.21	
	3	5.39	6.40	7.14	6.34
	1	6.93	7.48	8.37	7.59
	4	4.00	2.83	3.74	3.52
	7	4.36	3.32	2.25	3.31

Tabla 10.26: Promedio distancias: tercer paso

	1	B	4	C
B	2.70			
4	5.66	5.61		
C	7.59	6.34	3.52	
7	10.34	9.25	5.20	3.31

Tabla 10.27: Distancias con *average linkage*: tercer paso

El menor valor de distancia es ahora 2.70, por lo que agrupamos B con 1 en un nuevo *cluster* llamado D , para el cual calculamos su distancia al resto de los objetos como se muestra en la Tabla 10.28, obteniendo luego todas las distancias de la Tabla 10.29.

	1	2	3	Promedio
5	6.93	5.10	5.39	
C	6	7.48	6.78	6.40
	8	8.37	7.21	7.14
	4	5.66	5.83	5.39
	7	10.34	9.22	9.27
				9.61

Tabla 10.28: Promedio distancias: cuarto paso

	D	4	C
4	5.63		
C	6.76	3.52	
7	9.61	5.20	3.31

Tabla 10.29: Distancias con *average linkage*: cuarto paso

En este paso, la menor distancia es 3.31, siendo E el *cluster* que agrupa C con 7. Las distancias entre este *cluster* y los demás se muestran en la Tabla 10.30, resumiendo todas las distancias de la Tabla 10.31.

	5	6	7	8	Promedio
D	1	6.93	7.48	10.34	8.37
	2	5.10	6.78	9.22	7.21
	3	5.39	6.40	9.27	7.14
	4	4.00	2.83	5.20	3.74
					3.94

Tabla 10.30: Promedio distancias: quinto paso

Formamos un nuevo *cluster*, F , agrupando E y 4 y el proceso termina, siendo la distancia entre D y F igual a 7.10. El dendograma obtenido se muestra en la Figura 10.12.



Es importante observar que en el proceso se han utilizado solamente las distancias, de forma que para este procedimiento no es necesario disponer de los valores originales de las variables, basta con cualquiera de las matrices de distancias que utilizábamos en capítulos anteriores.

D		4
4		5.63
E	7.47	3.94

Tabla 10.31: Distancias con *average linkage*: quinto paso

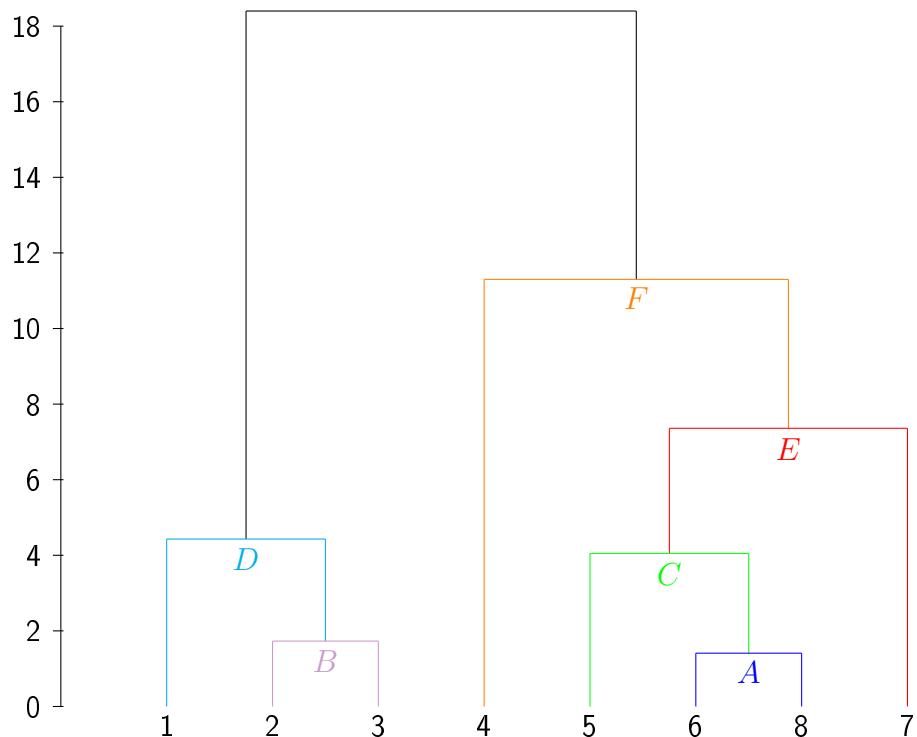


Figura 10.12: Dendograma con el método *average linkage*

Algunas de las características de los agrupamientos realizados con *average linkage* son las siguientes.

- ✿ Los *clusters* no resultan ni demasiado grandes ni demasiado pequeños.
- ✿ Se pueden utilizar medidas de similitud.
- ✿ Se tiende a fusionar *clusters* con varianzas pequeñas y a proporcionar *clusters* con la misma varianza.
- ✿ Se obtiene una buena representación gráfica de los resultados.

10.2.4.2 Método del vecino más próximo o *single linkage*

En el método del vecino más próximo la distancia entre dos *clusters* es el mínimo de las distancias entre un objeto cualquiera de uno de los *clusters* y un objeto cualquiera del otro. Es decir, si $R = \{P_1, \dots, P_n\}$ y $S = \{Q_1, \dots, Q_m\}$ conforman dos *clusters*, se mide la distancia entre ellos como

$$d(R, S) = \min\{d(P_i, Q_j) / i = 1, \dots, n; j = 1, \dots, m\}$$

Ejemplo 10.7. Nos referimos nuevamente al Ejemplo 10.5 a partir de la matriz de distancias 10.12. Luego de agrupar los elementos 6 y 8 en el *cluster* A , calculamos las distancias entre este nuevo *cluster* y el resto de los objetos obteniendo por resultado las distancias que se muestran en la Tabla 10.33. Los detalles de cálculo se presentan en la Tabla 10.32.

	6	8	Mínimo
1	7.48	8.37	7.48
2	6.78	7.21	6.78
3	6.40	7.14	6.40
4	2.83	3.74	2.83
5	2.83	2.45	2.45
7	3.32	2.24	2.24

Tabla 10.32: Mínimo de distancias: primer paso

Ahora vemos que la distancia más pequeña se alcanza entre los elementos 2 y 3, por lo que los fusionamos en un *cluster* que denominamos B . Calculamos la distancia entre B y el resto de los elementos como se muestra en la Tabla 10.34 para luego obtener todas las distancias de la Tabla 10.35.

El menor valor de distancia es 2.24, luego juntamos B con 1 en C y calculamos la distancia entre C y el resto de los elementos como se muestra en la Tabla 10.36 para luego obtener todas las distancias de la Tabla 10.37.

	A	1	2	3	4	5
1	7.48					
2	6.78	3.16				
3	6.40	2.24	1.73			
4	2.83	5.66	5.83	5.39		
5	2.45	6.93	5.10	5.39	4.00	
7	2.24	10.34	9.22	9.27	5.20	4.36

Tabla 10.33: Distancias con *single linkage*: primer paso

	2	3	Mínimo
A	6	6.78	6.40
	8	7.21	7.14
	1	3.16	2.24
	4	5.83	5.39
	5	6.93	5.39
	7	9.22	9.22

Tabla 10.34: Mínimo de distancias: segundo paso

	A	1	B	4	5
1	7.93				
B	6.40	2.24			
4	3.29	5.66	5.39		
5	2.64	6.93	5.39	4.00	
7	2.78	10.34	9.22	5.20	4.36

Tabla 10.35: Distancias con *single linkage*: segundo paso

	1	2	3	Minimo
A	6	7.48	6.78	6.40
	8	8.37	7.21	7.14
	4	5.66	5.83	5.39
	5	6.93	5.10	5.39
	7	10.34	9.22	9.22

Tabla 10.36: Mínimo de distancias: tercer paso

	A	C	4	5
C	6.40			
4	3.29	5.39		
5	2.64	5.39	4.00	
7	2.78	9.22	5.20	4.36

Tabla 10.37: Distancias con *single linkage*: tercer paso

El menor valor de distancia es ahora 2.64, por lo que agrupamos *A* con 5 en un nuevo *cluster* llamado *D*, para el cual calculamos su distancia al resto de los objetos como se muestra en la Tabla 10.38, obteniendo luego todas las distancias de la Tabla 10.39.

	5	6	8	Mínimo
C	1	6.93	7.48	8.37
	2	5.10	6.78	7.21
	3	5.39	6.40	7.14
	4	4.00	2.83	3.74
	7	4.36	3.32	2.24

Tabla 10.38: Mínimo de distancias: cuarto paso

	C	4	D
	4	5.39	
	D	5.10	2.83
	7	9.22	5.20

Tabla 10.39: Distancias con *single linkage*: cuarto paso

En este paso, la menor distancia es 2.24, siendo *E* el *cluster* que agrupa *D* con 7. Las distancias entre este *cluster* y los demás se muestran en la Tabla 10.40, resumiendo todas las distancias de la Tabla 10.41.

Formamos un nuevo *cluster*, *F*, agrupando *E* y 4 y el proceso termina, siendo la distancia entre *D* y *F* igual a 5.10. El dendograma obtenido resulta muy similar al de la Figura 10.12 con salvo que difiere ligeramente en las distancias en las que se fusionan los *clusters*. ■

Observemos que en el proceso, al igual que en el de *average linkage* se han utilizado solamente las distancias, de forma que para este procedimiento basta con cualquiera de las matrices de distancias

	5	6	7	8	Mínimo
C	1	6.93	7.48	10.34	8.37
	2	5.10	6.78	9.22	7.21
	3	5.39	6.40	9.27	7.14
	4	4.00	2.83	5.20	3.74
					2.83

Tabla 10.40: Mínimo de distancias: quinto paso

	C	4
4		5.39
E	5.10	2.83

Tabla 10.41: Distancias con *single linkage*: quinto paso

que utilizábamos en capítulos anteriores.

Mencionamos algunas características de los agrupamientos realizados con el método del vecino más próximo.

- ✿ Resulta útil para detectar *outliers*, estando entre los últimos pasos en unirse a la jerarquía.
- ✿ Se pueden usar medidas de similitud.
- ✿ Se tiende a construir *clusters* demasiado grandes.

10.2.4.3 Método del vecino más lejano o *complete linkage*

En el método del vecino más lejano la distancia entre dos *clusters* es el máximo de todas las distancias entre elementos de un *cluster* y elementos del otro. Es decir, si $R = \{P_1, \dots, P_n\}$ y $S = \{Q_1, \dots, Q_m\}$ conforman dos *clusters*, se mide la distancia entre ellos como

$$d(R, S) = \max\{d(P_i, Q_j) / i = 1, \dots, n; j = 1, \dots, m\}$$

Los pasos a seguir son similares a los realizados en el ejemplo para el método del vecino más próximo, tomando los valores máximos de las distancias en cuestión, en vez de los valores mínimos.

Entre las características de los agrupamientos realizados con el método del vecino más lejano podemos citar las siguientes.

- ✿ Resulta útil para detectar *outliers*.
- ✿ Se pueden usar medidas de similitud.
- ✿ Se tiende a construir *clusters* pequeños y compactos.

10.2.4.4 Método del centroide o *unweighted centroid*

El método del centroide es el que se utilizó en el Ejemplo 10.5 presentado con el fin de ilustrar la construcción de un dendograma.

Recordemos que en este caso, la distancia entre dos *clusters* se calcula como la distancia entre los centroides de los mismos, por tanto es necesario disponer de los valores originales de las variables.

Podemos mencionar como características de los agrupamientos realizados con el método del centroide que

- ✿ Las variables deben estar en escala de intervalo.
- ✿ Las distancias entre grupos se calculan como las distancias entre los vectores medios de cada uno de los grupos.
- ✿ Si los tamaños de los dos grupos a mezclar son muy diferentes, entonces el centroide del nuevo grupo será muy próximo al de mayor tamaño y probablemente estará dentro de ese grupo.

10.2.4.5 Método de Ward o varianza mínima

En esta técnica, la distancia entre dos *clusters* se calcula como la suma de cuadrados entre grupos en el ANOVA sumando para todas las variables.

En cada paso se minimiza la suma de cuadrados dentro de los *clusters* sobre todas las particiones posibles obtenidas fusionando dos *clusters* del paso anterior. Las sumas de cuadrados son más fáciles de entender cuando se expresan como porcentaje de la suma de cuadrados total.

Los cálculos son un poco más complejos que en los casos anteriores. Con el Código 10.1 se obtiene el dendograma de la Figura 10.13 producido al aplicar el método de Ward. Los datos para este código están disponibles en <https://goo.gl/DQegQk>.

```
library(readxl) # Permite leer archivos xlsx  
  
cluster=read_excel("C:/.../cluster.xlsx")  
# Importa la base con la cual se va a trabajar  
  
d=dist(cluster[,2:4], method="euclidean", p = 2)  
# Calcula la distancias euclídeas entre los datos originales  
dendoaux=hclust(d, method="ward.D")  
# Produce un dendograma  
dendo=as.dendrogram(dendoaux)  
# Se utiliza para personalizar el dendograma  
nodePar=list(lab.cex=1, pch=c(NA, 19), cex=0.7, col="royalblue")  
plot(dendo, xlab="", nodePar=nodePar, edgePar=list(col=2:3))  
# Grafica un dendograma
```

Código 10.1: Código para generar un dendograma

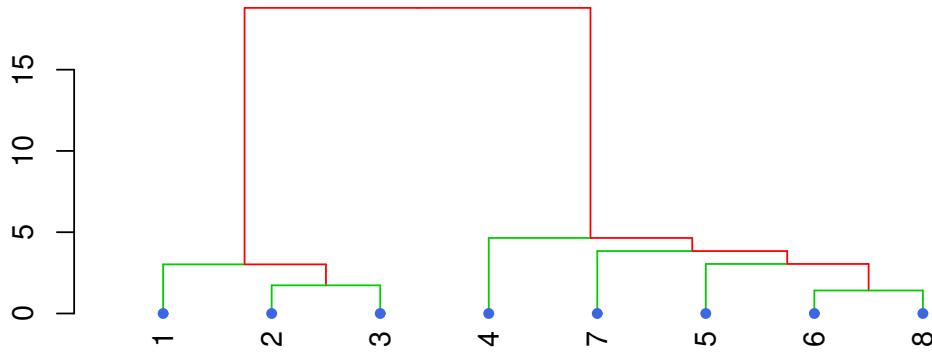


Figura 10.13: Dendrograma de Ward para los datos del Ejemplo 10.5

10.2.5 Métodos divisivos

Hasta el momento hemos tratado con métodos aglomerativos, considerando diferentes distancias y técnicas.

Veamos ahora cómo funcionan los métodos divisivos.

Los métodos divisivos trabajan en la dirección opuesta que la empleada en los métodos aglomerativos. En este método, se parte de un gran *cluster* que contiene a todos los elementos y se busca subdividirlo en *clusters* más pequeños. Si, por ejemplo, disponemos de n elementos, y los queremos partir en dos *subclusters*, disponemos de $2^{n-1} - 1$ posibles particiones. Ver la Tabla 10.42 a modo de ejemplo. Para cada una de estas posibles particiones, deberíamos calcular alguna medida que nos indique la eficiencia de la misma.

Si tenemos muchos objetos, este criterio puede resultar complejo. Sin embargo, para el caso de p variables binarias existen métodos computacionalmente simples y eficientes que se conocen con el nombre de **métodos monocategóricos**.

Los métodos monocategóricos dividen los *clusters* de acuerdo con la presencia o ausencia de cada una de las características. De esta forma, en cada subdivisión, los *clusters* contienen individuos con ciertos atributos, todos presentes o todos ausentes dentro del mismo.

<i>n</i>	<i>Cluster</i>	<i>Subclusters</i>
2	● ●	● ●
3	● ● ●	● ● ●

Tabla 10.42: Posibles particiones en dos *subclusters*

El término monocategórico se refiere al uso de una única variable para definir la partición en cada etapa del proceso. En contraposición, los **métodos policategóricos** se basan en más de una variable para la partición de cada etapa del proceso.

La elección de la variable en la cual se basará la siguiente partición depende de la optimización de un criterio que contemple al mismo tiempo la homogeneidad de los nuevos conglomerados y la asociación de las variables. Esto tiende a minimizar el número de particiones que deben realizarse.

Lance y Williams [31] propusieron como criterio de homogeneidad el índice *C* de caos dado por

$$C = pn \log(n) - \sum_{k=1}^p [f_k \log(f_k) - (n - f_k) \log(n - f_k)]$$

donde *n* es el número total de observaciones, *f_k* es el número de individuos que tienen el atributo *k* y *p* es la cantidad de variables consideradas sobre cada individuo.

A pesar de que los métodos divisivos son menos usados que los aglomerativos, tienen la ventaja de revelar la estructura principal de los datos.

10.2.6 Cantidad de *clusters*

Como ya hemos dicho en varias ocasiones, la decisión sobre el número óptimo de *clusters* es subjetiva, especialmente cuando se incrementa el número de objetos ya que si se seleccionan demasiado pocos, los *clusters* resultantes son heterogéneos y artificiales, mientras que si se seleccionan demasiados, la interpretación de los mismos suele resultar complicada.

Algunos programas, como *InfoStat*, proveen automáticamente el valor del **coeficiente de correlación cofenética**, el cual puede ser usado para seleccionar uno de varios agrupamientos alternativos. Este coeficiente indica la correlación de las distancias definidas por la métrica de árbol binario con las distancias originales entre objetos. Luego, se espera que el agrupamiento con mayor coeficiente sea el que mejor describe el agrupamiento natural de los datos.

Es importante destacar que los procedimientos de agrupamiento producen resultados exitosos cuando la matriz de datos tiene una estructura que es posible interpretar desde el problema que originó la recolección de la información. Debido a ello, logrados los grupos, es importante caracterizar los mismos a través de diversas medidas resumen para favorecer la interpretación del agrupamiento final.

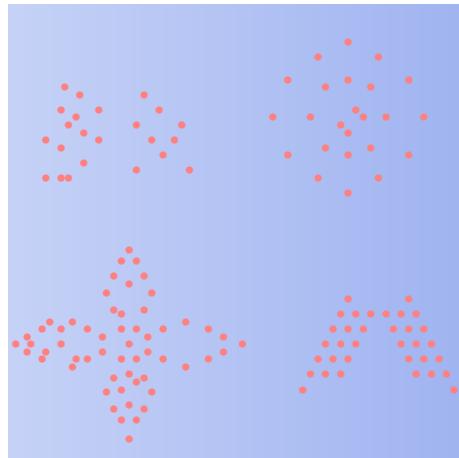


Figura 10.14: ¿Cuántos *clusters* se pueden ver?

Ya hemos mencionado que como ayuda a la decisión sobre el número de *clusters*, se suelen representar los distintos pasos del algoritmo y la distancia a la que se produce la fusión. En los primeros pasos el salto en las distancias será pequeño, mientras que en los últimos el salto entre pasos será mayor. El punto de corte puede ser aquel en el que comienzan a producirse saltos bruscos.

Algunos analistas recomiendan aplicar varios algoritmos de agrupamiento y de selección o combinación de variables para cada conjunto de datos. Seleccionando, finalmente, desde los agrupamientos realizados la interpretación más apropiada.

10.2.7 Métodos de partición no jerárquicos

Dentro de este tipo de métodos, el más popular es el algoritmo denominado **k-medias** o, en inglés, **k-means**.

¿En qué consiste?

Es importante destacar que un dato de entrada es el número de conglomerados deseados que denotaremos por k . Este algoritmo se implementa en los siguientes cuatro pasos.

1. Se divide a los n objetos en k subconjuntos no vacíos. La pregunta natural es: ¿Cómo elegir estos k subconjuntos? Existen las siguientes opciones:
 - * **Opción 1:** se asignan aleatoriamente.
 - * **Opción 2:** se toman los centros de los grupos como los puntos más alejados entre sí.

* **Opción 3:** se utiliza información previa disponible.

2. Se calcula el centroide o punto medio del *cluster*. Esta asignación es secuencial, con lo cual, cada vez que se reasigna un elemento se vuelve a calcular el centroide del nuevo *cluster*.
3. Se reasigna cada objeto al centroide de *cluster* más cercano.
4. Se vuelve al paso 2, hasta que no convenga realizar nuevas asignaciones. Esto sucede cuando no se puede mejorar el criterio de optimalidad establecido.

Algunas de las características de los agrupamientos realizados por el método de *k-means* se presentan en la Tabla 10.43.

Ventajas	Desventajas
Es relativamente eficiente	Sólo se aplica si la media está definida
Es computacionalmente rápido	Es sensible a la presencia de <i>outliers</i>
Se trabaja bien con datos faltantes (<i>missing values</i>)	No se satisface el criterio de optimización globalmente, en general se termina con un óptimo local Es necesario especificar el número de conglomerados <i>a priori</i>

Tabla 10.43: Ventajas y desventajas del método *k-means*

¿Cómo elegir la medida de proximidad o distancia?

La elección de esta medida depende en primera instancia de la naturaleza de los datos. Bajo ciertas circunstancias conviene discretizar el análisis de alguna variable continua o bien categorizarla. Por ejemplo, si la mayoría de las variables son categóricas y la variable “edad”, que es de escala continua, parece relevante para el análisis, se podría subdividir las categorías de esta variable numérica. Existen diferentes formas de hacer esta subdivisión]; por ejemplo, en un estudio médico podría considerarse la subdivisión en función del riesgo mientras que en un estudio de mercado podría estar en función de los niveles de disposición del dinero.

Hay que tener en cuenta que la selección de la distancia y de la técnica, pueden cambiar la disposición de los *clusters*.

¿Cómo tratar los valores perdidos o missing data?

La forma más simple, aunque no siempre implica ser la mejor, es utilizar únicamente los registros completos. Sin embargo, esta decisión puede reducir drásticamente la información disponible para el estudio.

Ya hemos visto que una matriz de distancias, digamos $D = (\delta_{ij})$, obtenida a partir de una matriz de datos multivariantes X , no cumple necesariamente la propiedad de la desigualdad triangular. Esta situación da lugar al problema de aproximar la matriz de distancias con una matriz que llamamos $U = (u_{ij})$ según algún criterio de proximidad adecuado.

La medida de proximidad que se utiliza es la correlación cofenética, que es el coeficiente de correlación lineal de Pearson entre los $n(n - 1)/2$ pares de distancias (δ_{ij}, u_{ij}) para $1 \leq i < j \leq n$. Este coeficiente vale 1 cuando las matrices D y U son iguales. Esto equivale a decir que la matriz D ya cumple la propiedad triangular y, por tanto, la clasificación es exacta.

Supongamos que los datos originales X_i han sido modelados usando un método *cluster* para producir un dendrograma T_i ; es decir, un modelo simplificado en el que los datos que son ‘cercanos’ se han asignado a un *cluster* y se representan en un dendrograma.

La distancia euclídea ordinaria entre las observaciones i y j está dada por $X_{ij} = |X_i - X_j|$. La distancia en el dendrograma de representación de los puntos de modelo T_i y T_j se nota por T_{ij} , siendo ésta la altura del nodo en el que estos dos puntos se unen primero.

El coeficiente de correlación cofenética se calcula como

$$C = \frac{\sum_{ij} (x_{ij} - \bar{x})(t_{ij} - \bar{t})}{\sqrt{\left[\sum_{i < j} (x_{ij} - \bar{x})^2 \right] \left[\sum_{i < j} (t_{ij} - \bar{t})^2 \right]}}$$

donde \bar{x} y \bar{t} son el promedio de los valore x_{ij} y t_{ij} , respectivamente.

En otras técnicas multivariadas existe la alternativa de estimar los datos faltantes mediante algún resumen estadístico de los datos disponibles en los restantes registros, como la media o la mediana por ejemplo. En esta estrategia no debe utilizarse en análisis de *clusters*.

¿Cómo establecer un criterio de optimalidad?

Un criterio de optimalidad muy difundido consiste en minimizar la suma de cuadrados dentro del grupo (SCD); es decir, minimizar las distancias al centroide del grupo de los elementos, siendo

$$\text{SCD} = \sum_{h=1}^k \sum_{j=1}^p \sum_{i=1}^{n_h} (x_{ijh} - \bar{x}_{jh})^2$$

Ejemplo 10.8. Consideremos las observaciones dadas en la Tabla 10.44

Los puntos iniciales se pueden visualizar en la Figura 10.15.

Nuestro objetivo es encontrar dos *clusters*. Para ello, elegimos como centroides a los puntos mas alejados que son $c_1 = (10, 19)$ y $c_2 = (60, 18)$.

<i>x</i>	<i>y</i>
15	17.6
10	19.0
20	18.0
48	19.0
60	18.0
50	18.2

Tabla 10.44: Datos para el algoritmo k-means

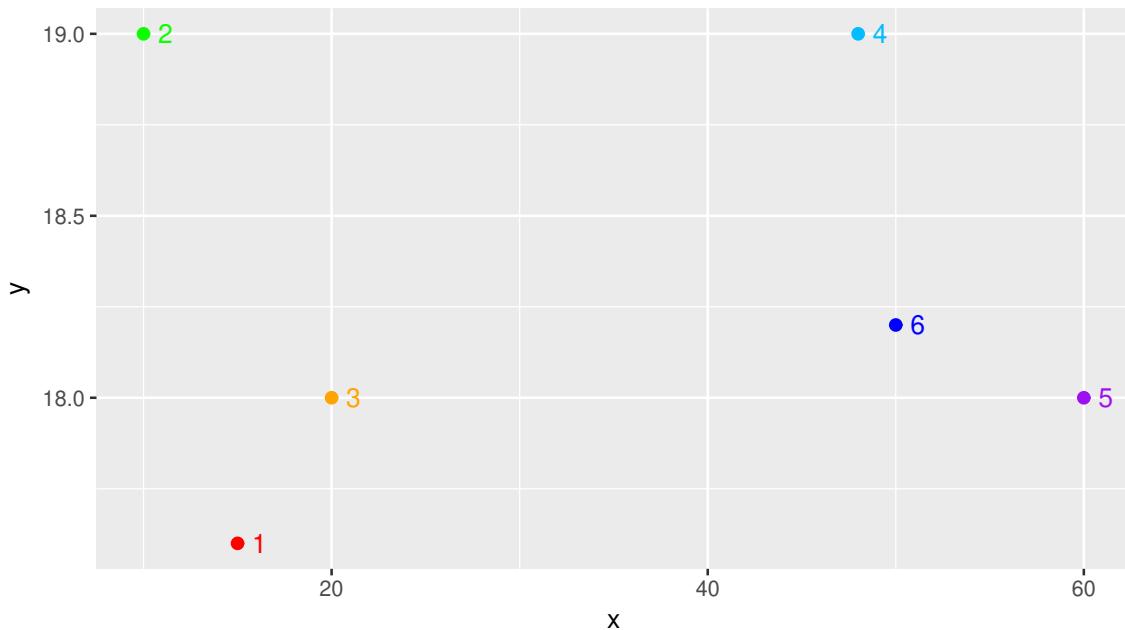


Figura 10.15: Puntos originales

Punto	Distancia a c_1	Distancia a c_2	Cluster asignado
1	5.19	45.00	1
2	0.00	50.01	1
3	10.05	40.00	1
4	38.00	12.04	2
5	50.01	0.00	2
6	40.01	10.00	2

Tabla 10.45: Distancia a los centroides en el primer paso

Las distancias de todos los puntos a los centroides elegidos se muestran en la Tabla 10.45

Calculamos los nuevos centroides para los dos *clusters* formados respectivamente por los tres primeros puntos y por los tres últimos. Se obtienen $c_3 = (15, 18.2)$ y $c_4 = (52.67, 18.4)$. En la Tabla 10.46 calculamos las distancias a los nuevos centroides, para ver si se reclasifican.

Punto	Distancia a c_3	Distancia a c_4	Cluster asignado
1	0.60	37.68	1
2	5.06	42.67	1
3	5.00	32.67	1
4	33.01	4.71	2
5	45.00	7.34	2
6	35.00	2.68	2

Tabla 10.46: Distancia a los centroides en el segundo paso

Si bien al cambiar los centroides cambiaron las distancias, no cambió la clasificación original. Como es un caso sencillo, no tiene sentido buscar el criterio de optimalidad.

Para los cálculos de distancias nos referimos al Código 10.2 con datos disponibles en <https://goo.gl/Ab9rp5>, <https://goo.gl/V2UAno> y <https://goo.gl/zZq3sz>.

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

# Lectura de las bases de datos
cluster=read_excel("C:/.../k-means1.xlsx")
cluster1=read_excel("C:/.../k-means1paso1.xlsx")
cluster2=read_excel("C:/.../k-means1paso2.xlsx")

# Cálculo de distancias
d0=round(dist(cluster, method="euclidean", p=2),2)
d1=round(dist(cluster1[,2:3], method="euclidean", p=2),2)
d2=round(dist(cluster2[,2:3], method="euclidean", p=2),2)
```

Código 10.2: Código para calcular distancias



Vamos a expresar el criterio de optimalidad de la siguiente manera

$$\min \sum_{h=1}^k \sum_{i=1}^{n_h} (x_i - \bar{x}_h)^t (x_i - \bar{x}_h) = \min \sum_{h=1}^k \sum_{i=1}^{n_h} S_{i,h}^2 = \min \sum_{h=1}^k \sum_{i=1}^{n_h} \text{tr}(W)$$

siendo

$$W = \sum_{h=1}^k \sum_{i=1}^{n_h} (x_i - \bar{x}_h)^t (x_i - \bar{x}_h)$$

Ejemplo 10.9. Ilustremos estas consideraciones para $p = 3$ y $k = 2$, con la distancia de Manhattan.

Tomemos los puntos del algoritmo jerárquico descripto en la Tabla 10.47 y representados en la Figura 10.16.

Objeto	x_1	x_2	x_3
1	1	1	2
2	2	2	1
3	2	2	3
4	3	3	5
5	4	5	4.7
6	5	5	5
7	8	8	7
8	7	8	5
9	5	7	6

Tabla 10.47: Datos para aplicar el método de k-means

En esta oportunidad, elegimos al azar entre los puntos los siguientes dos centroides : $c_{G_1} = (2, 2, 1)$ y $c_{G_2} = (8, 8, 7)$. Las distancias de los puntos a los mismos se muestra en la Tabla 10.48.

Objeto	Distancia a c_{G_1}	Distancia a c_{G_2}	Grupo de pertenencia
1	3	19	G_1
2	0	18	G_1
3	2	16	G_1
4	6	12	G_1
5	8.7	9.3	G_1
6	10	8	G_2
7	18	0	G_2
8	15	3	G_2
9	13	5	G_2

Tabla 10.48: Clasificación con $k = 2$ en el primer paso

Calculamos ahora los nuevos centroides de los grupos G_1 y G_2 . Estos centroides resultan de promediar las coordenadas de los puntos que integran cada uno de los clusters. Los nuevos centroides son $c_{G_1} = (2.4, 2.6, 3.14)$ y $c_{G_2} = (6.25, 7, 5.75)$ y los agrupamientos se muestran en la Tabla 10.49.

En este caso, vemos que no sólo cambiaron las distancias a los centroides, sino también la clasificación del objeto 5, que pasó de pertenecer a G_1 en el primer paso del algoritmo a pertenecer a G_2 en el segundo paso.

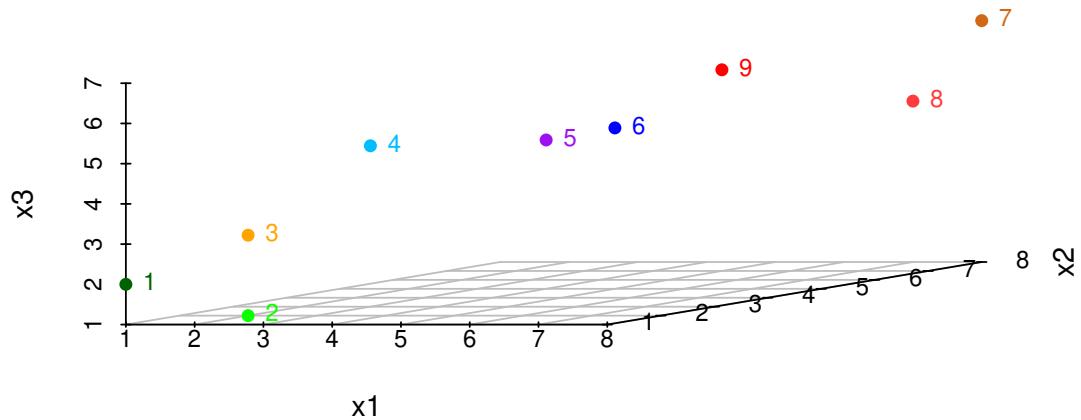


Figura 10.16: Representación de las observaciones originales tridimensionales

Objeto	Distancia a c_{G_1}	Distancia a c_{G_2}	Grupo de pertenencia
1	4.14	15.00	G_1
2	3.14	14.00	G_1
3	1.14	12.00	G_1
4	2.86	8.00	G_1
5	5.56	5.30	G_2
6	6.86	4.00	G_2
7	14.86	4.00	G_2
8	11.86	2.50	G_2
9	9.86	1.50	G_2

Tabla 10.49: Clasificación con $k = 2$ en el segundo paso

En este caso, los cálculos se obtienen con el Código 10.3 y los datos disponibles en <https://goo.gl/aU32Rk> y <https://goo.gl/R2hSXd>.

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

# Lectura de las bases de datos
cluster1=read_excel("C:/.../k-means2paso1.xlsx")
cluster2=read_excel("C:/.../k-means2paso2.xlsx")

# Cálculo de distancias
d1=round(dist(cluster1[,2:4], method="manhattan"),2)
d2=round(dist(cluster2[,2:4], method="manhattan"),2)
```

Código 10.3: Código para calcular distancias en el Ejemplo 10.9



Para determinar el número k de grupos suele utilizarse el siguiente contraste

$$F_{k,k+1} = \frac{\text{SCD}(k) - \text{SCD}(k+1)}{\text{SCD}(k+1)/(n-k-1)}$$

y se compara con el cuantil 0.95 de $F_{p,p(n-k-1)}$.

¿Qué sucedería si en Ejemplo 10.9 se propone que sean tres los conglomerados?

Ejemplo 10.10. A partir de los datos de la Tabla 10.47, proponemos estos tres centroides, elegidos aleatoriamente: $c_{G_1} = (1, 1, 2)$, $c_{G_2} = (3, 3, 5)$ y $c_{G_3} = (8, 8, 7)$. Calculamos las distancias de Manhattan de los puntos a los tres centroides elegidos y los clasificamos en el grupo del que están a menor distancia, como se exhibe en la Tabla 10.50.

Se obtienen los nuevos centroides $c_{G_1} = (1, 1, 2)$, $c_{G_2} = (3, 3, 5)$ y $c_{G_3} = (8, 8, 7)$, cuyas distancias se presentan en la Tabla 10.51.

Como no han cambiado las clasificaciones no se realizan movimientos. La pregunta que aún queda pendiente es si conviene realizar la partición en 2 o en 3 *clusters*. Mediante el Código 10.4 y datos disponibles en <https://goo.gl/Y9eMbb>, tenemos que $\text{SCD}(2) = 41.692$ y $\text{SCD}(3) = 15.393$. Utilizaremos para ello el criterio de F propuesto para $k = 2$,

$$F_{k,k+1} = \frac{\text{SCD}(k) - \text{SCD}(k+1)}{\text{SCD}(k+1)/(n-k-1)} = \frac{41.692 - 15.393}{15.393/(9-3)} = 10.25$$

Comparando este valor con el cuantil 0.95 de la distribución F de Snedecor $F_{0.95,3,18} = 3.15$. Puesto que $10.25 > 3.15$ resulta significativa la mejora de considerar 3 *clusters* en lugar de 2.

Objeto	Distancia a c_{G_1}	Distancia a c_{G_2}	Distancia a c_{G_3}	Grupo de pertenencia
1	0	7	19	G_1
2	3	6	18	G_1
3	3	4	16	G_1
4	7	0	12	G_2
5	9.7	3.3	9.3	G_2
6	11	4	8	G_2
7	19	12	0	G_3
8	16	9	3	G_3
9	14	7	5	G_3

Tabla 10.50: Clasificación con $k = 3$ en el primer paso

Objeto	Distancia a c_{G_1}	Distancia a c_{G_2}	Distancia a c_{G_3}	Grupo de pertenencia
1	1.34	9.23	16.34	G_1
2	1.66	8.23	15.34	G_1
3	1.66	6.23	13.34	G_1
4	5.66	2.43	9.34	G_2
5	8.36	0.87	6.64	G_2
6	9.66	1.77	5.34	G_2
7	17.66	9.77	2.66	G_3
8	14.66	6.77	1.66	G_3
9	12.66	4.77	2.34	G_3

Tabla 10.51: Clasificación con $k = 3$ en el segundo paso

```

library(readxl) # Permite leer archivos xlsx

cluster=read_excel("C:/.../k-means2.xlsx")
# Importa la base con la cual se va a trabajar
cluster=as.matrix(cluster[,2:4]) # Convierte en formato matriz

# Cálculo de promedios por grupo
xrayag1k2=as.matrix(apply(cluster[1:5,],2,mean))
xrayag2k2=as.matrix(apply(cluster[6:9,],2,mean))
xrayag1k3=as.matrix(apply(cluster[1:3,],2,mean))
xrayag2k3=as.matrix(apply(cluster[4:6,],2,mean))
xrayag3k3=as.matrix(apply(cluster[7:9,],2,mean))

# Inicialización
scdg1k2=matrix(0,nrow=5,ncol=3)
scdg2k2=matrix(0,nrow=4,ncol=3)
scdg1k3=matrix(0,nrow=3,ncol=3)
scdg2k3=matrix(0,nrow=3,ncol=3)
scdg3k3=matrix(0,nrow=3,ncol=3)

# Cálculo de diferencias de coordenadas
for (i in 1:5) {scdg1k2[i,]=cluster[i,]-t(xrayag1k2)}
for (i in 6:9) {scdg2k2[i-5,]=cluster[i,]-t(xrayag2k2)}
for (i in 1:3) {scdg1k3[i,]=cluster[i,]-t(xrayag1k3)}
for (i in 4:6) {scdg2k3[i-3,]=cluster[i,]-t(xrayag2k3)}
for (i in 7:9) {scdg3k3[i-6,]=cluster[i,]-t(xrayag3k3)}

# Cálculo de suma de cuadrados dentro del grupo
SCD2=sum(scdg1k2^2)+sum(scdg2k2^2)
SCD3=sum(scdg1k3^2)+sum(scdg2k3^2)+sum(scdg3k3^2)

```

Código 10.4: Código para calcular suma de cuadrados dentro del grupo

10.2.8 Otros métodos para elegir el número de *clusters*

En [43] se propone el estadístico de *gap* para determinar el número óptimo de conglomerados. Esta propuesta resulta independiente del algoritmo utilizado para la segmentación.

Supongamos que por algún algoritmo encontramos k *clusters* en un conjunto de n observaciones independientes p -variadas. Notamos las n_i observaciones del *cluster* i como $C_i = \{x_{i1}, x_{i2}, \dots, x_{in_i}\}$. En este *cluster* se definen las distancias euclídeas sobre los elementos y se suman para obtener

$$d_i = \sum_{j \neq k} (x_{ij} - x_{ik})^2$$

Cabe destacar que las distancias están contadas dos veces. Se define

$$W_k = \sum_{i=1}^k \frac{1}{2n_i} d_i$$

Es evidente que W_k está en función de la cantidad de *clusters* y, al crecer el valor de k , W_k decrece ya que la variabilidad dentro de los grupos también lo hace. La idea es comparar, para cada valor de k , el valor de $\log(W_k)$ con su valor esperado bajo la distribución de una hipótesis nula que supone que no hay *clusters*.

Se define ahora

$$\text{gap}_n^k = E_{H_0}[\log(W_k)] - \log(W_k)$$

Este estadístico requiere la suposición de una distribución de referencia para la hipótesis de nulidad.

El valor estimado para k , digamos \hat{k} es el que maximiza el estadístico de *gap* sobre la distribución de referencia elegida. Los autores del citado trabajo proponen dos alternativas para esta selección. La primera está basada en generar distribuciones uniformes dentro del rango de valores observados para cada una de las p variables, mientras que la segunda se basa en generar distribuciones uniformes a partir de las componentes principales de los datos originales.

Si se utiliza el programa R, se puede obtener el estadístico *gap* mediante la función `fviz_nbclust()` o con la función `clusGap()` incluida en el paquete `cluster`.

En [36] se describe una metodología para identificar departamentos con patrones de rendimientos similares (grupos) a través del uso de algoritmos de segmentación para datos anuales de cultivos de maíz, en el período 1969-2010, provistos por el Ministerio de Agricultura de la Nación Argentina, siendo un caso de segmentación de datos longitudinales.

10.3 Ejemplos

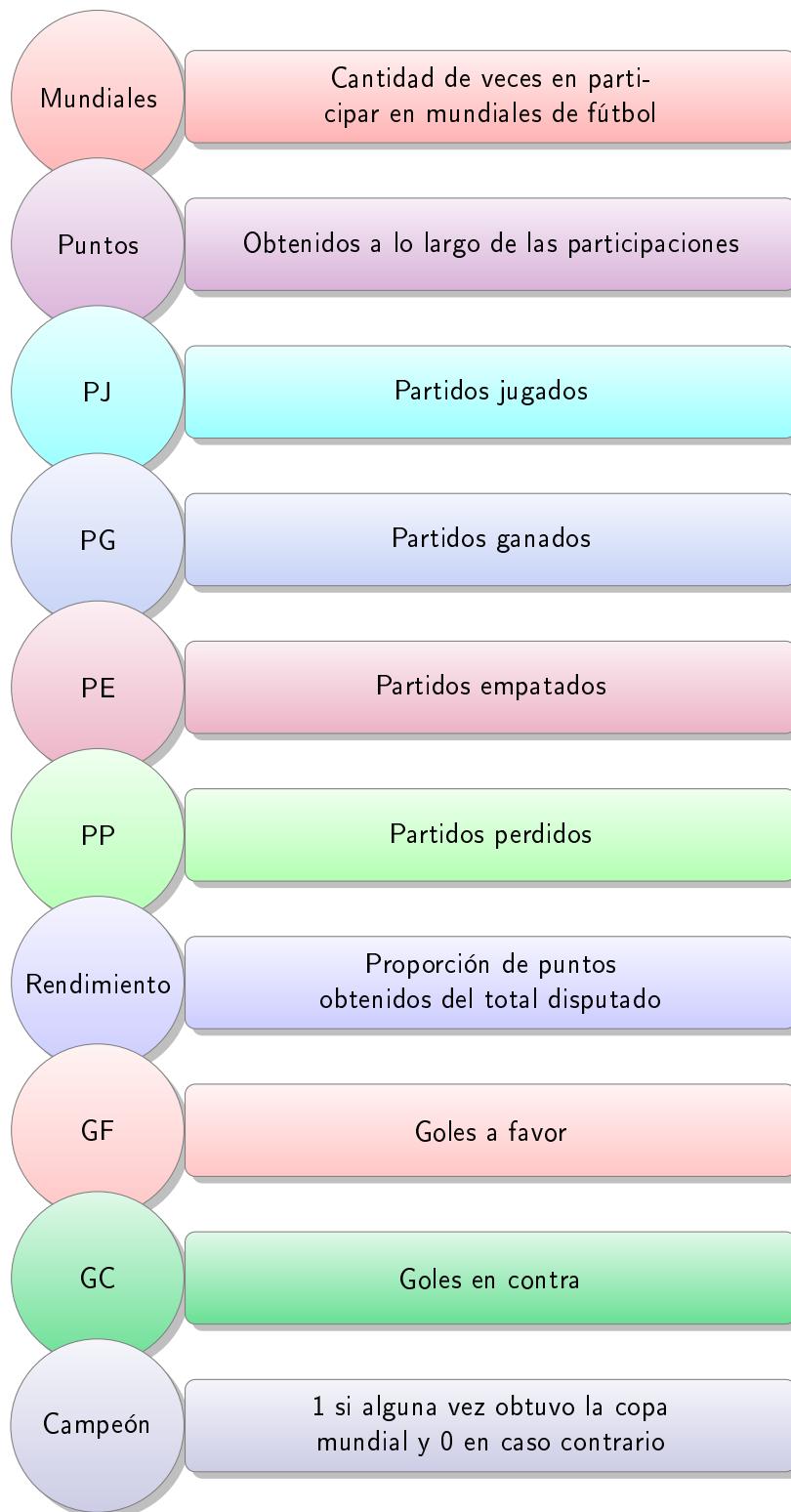
10.3.1 Ejemplo de agrupamiento jerárquico

Ejemplo 10.11. En el archivo disponible en <https://goo.gl/r9wyWE> se presentan 77 equipos representantes de países que participaron en mundiales de futbol.



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

Se han registrado las siguientes variables



Para el agrupamiento, nos referimos al Código 10.5 con el cual se produce el dendograma de la Figura 10.18 considerando la cantidad de cuatro *clusters*.

```
library(readxl) # Permite leer archivos xlsx
library(magrittr) # Proporciona un mecanismo para encadenar comandos
library(tibble) # Proporciona un mejor formato para los datos
library(dendextend) # Ofrece un conjunto de funciones para dendogramas

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

futnum=futbol[,2:11] # Selecciona las variables numéricas
fut=na.omit(futnum) # Elimina los registros con datos faltantes

fut %>% scale %>% dist() %>% hclust(method="ward.D") %>% as.dendrogram() -> dend
# Aplica el criterio de Ward a las variables estandarizadas

par(mar=c(6,1,0.1,0.1)) # Establece márgenes

dend %>%
set("branches_k_color", value=c("blue","red","purple","darkgreen"), k=4) %>%
# Personaliza las ramas
set("labels_col", value=c("blue","red","purple","darkgreen"), k=4) %>%
set("labels_cex", 0.65) %>% set("labels", futbol[,1]) %>%
# Personaliza las etiquetas
plot(axes=FALSE)
# Produce un dendograma personalizado
```

Código 10.5: Código para clasificar los países en el campeonato

10.3.2 Ejemplo de agrupamiento no jerárquico

Ejemplo 10.12. Utilizamos para este ejemplo la misma base de datos del Ejemplo 10.11. Con el Código 10.6 se producen distintas maneras de visualizar los *clusters*, tal cual se muestra en las Figuras 10.19, 10.20 y 10.21. Además, en las Figuras 10.22 y 10.23 se muestra un análisis de los goles a favor y en contra teniendo en cuenta el agrupamiento.

```
library(cluster) # Incluye métodos para el análisis de clusters
library(factoextra)
# Permite extraer y visualizar resultados de análisis de datos multivariados
library(ggplot2) # Paquete para confeccionar dibujos
library(readxl) # Permite leer archivos xlsx
```

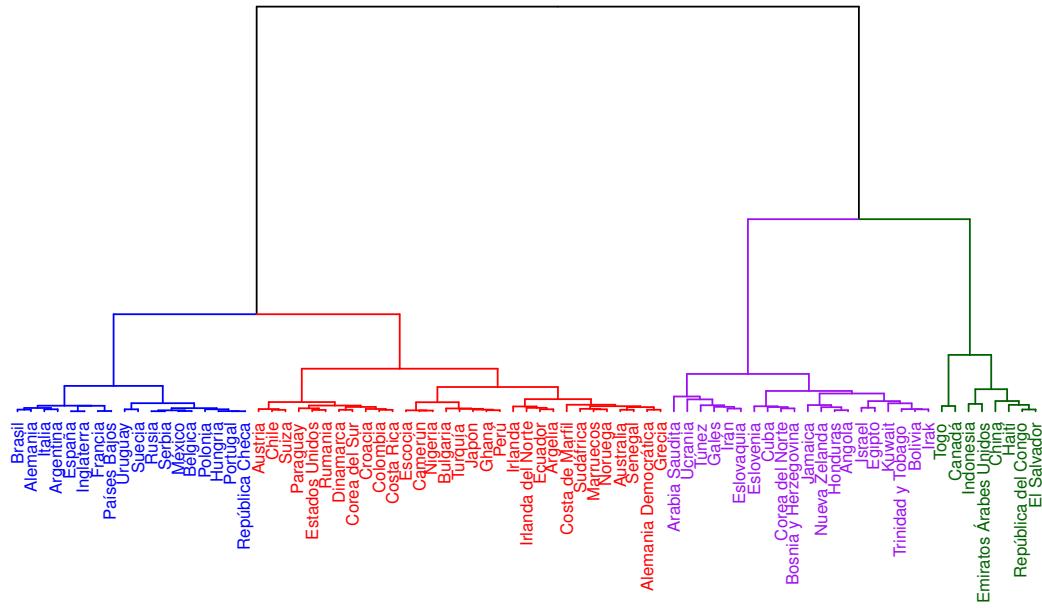


Figura 10.18: Agrupamiento de los países en el campeonato

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

futbol=data.frame(futbol) # Arregla los datos
df=na.omit(futbol) # Elimina los registros con datos faltantes
df=scale(futbol[, -1]) # Estandariza las variables
rownames(df)=futbol[, 1] # Pone nombre a las filas

set.seed(123) # Fija una semilla
kmfutbol=kmeans(df, centers=4, nstart=10)
# Aplica k-means con k=4

par(mar=c(5,5,1,2)) # Establece márgenes
clusplot(df, kmfutbol$cluster, main=NULL, color=TRUE, shade=TRUE, labels=4,
lines=0, plotchar=FALSE)
# Dibuja la clasificación
clusplot(df, kmfutbol$cluster, main=NULL, color=TRUE, shade=FALSE, labels=2,
lines=0, plotchar=FALSE, cex.txt=0.6, col.txt="black", cex=0)
# Dibuja la clasificación con etiquetas

fviz_cluster(kmfutbol, data=df, geom="text", labelsize=8, repel=TRUE) +
scale_color_brewer(palette = "Set1") +
ggtitle('')

# Dibuja la clasificación personalizada
```

```

Cluster=kmfutbol$cluster
datos=data.frame(cbind(futbol, Cluster))
# Guarda el grupo de pertenencia de cada país

ggplot(datos, aes(x=GC, y=GF)) +
  geom_point(aes(colour=factor(Cluster))) +
  scale_color_brewer(palette = "Set1") +
  xlab("Goles_en_contra") +
  ylab("Goles_a_favor") +
  labs(color='Cluster')
# Grafica relación de goles por grupo

ggplot(datos, aes(x=GC, y=GF)) +
  geom_point(aes(colour=factor(Cluster))) +
  geom_text(aes(label=País), size=3, hjust=0, vjust=0) +
  scale_color_brewer(palette = "Set1") +
  xlab("Goles_en_contra") +
  ylab("Goles_a_favor") +
  labs(color='Cluster')
# Grafica relación de goles por país

```

Código 10.6: Código para el agrupamiento no jerárquico de los países en el campeonato

10.3.3 Ejemplo de aplicación a *text mining*

La **minería de texto**, en inglés *text mining*, es equivalente al análisis de variables medidas por textos. El objetivo principal de este proceso consiste en obtener información de calidad a partir de un texto, como por ejemplo el descubrimiento de patrones o tendencias. A diferencia de las variables numéricas, los datos en este caso pueden provenir de diferentes fuentes y, por lo tanto, tener distintas estructuras. La técnica de *text mining* se basa en estructurar los datos de entrada para derivar un patrón dentro de los datos ya estructurados para su interpretación y evaluación.

Ejemplo 10.13. El objetivo de este ejemplo consiste en analizar la frecuencia con la que aparecen ciertas palabras en poemas del escrito Pablo Neruda. A tal fin, nos referimos a los datos disponibles en <https://goo.gl/vzjxEX>.

El análisis se produce mediante el Código 10.7, mediante el cual se obtienen el histograma de la Figura 10.24 y las nubes de palabras de las Figuras 10.25, 10.26, 10.28 y 10.27. Otra de las cosas que podemos obtener de este código son las siguientes:

- ✿ Las palabras que aparecen en los textos analizados con una frecuencia igual o superior a 15 son: “noche”, “ojos”, “alma”, “viento” y “más”. La última de estas palabras, si bien tiene una

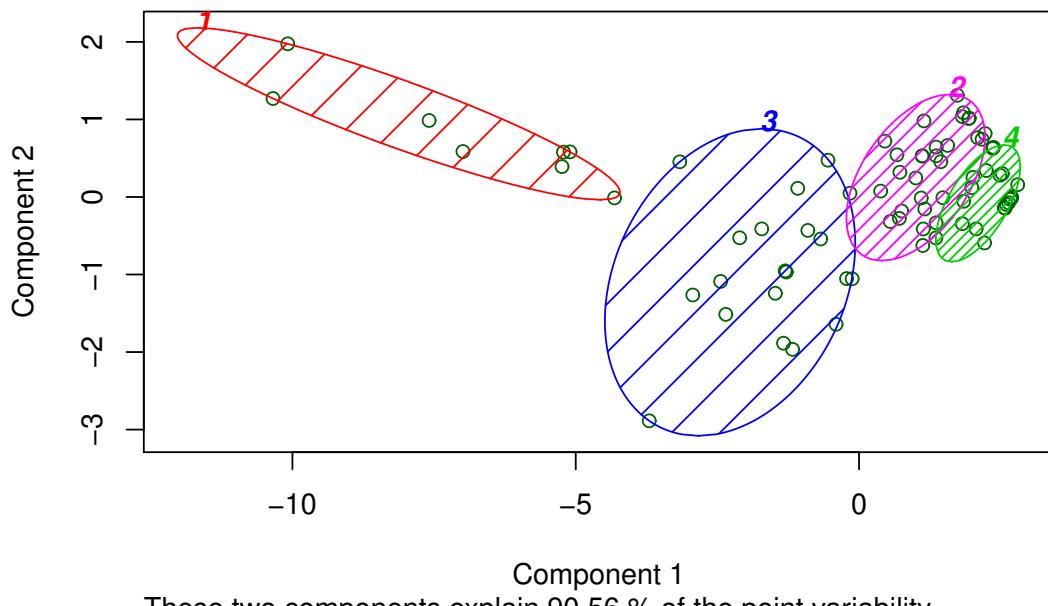


Figura 10.19: Grupos de países con el algoritmo k-means

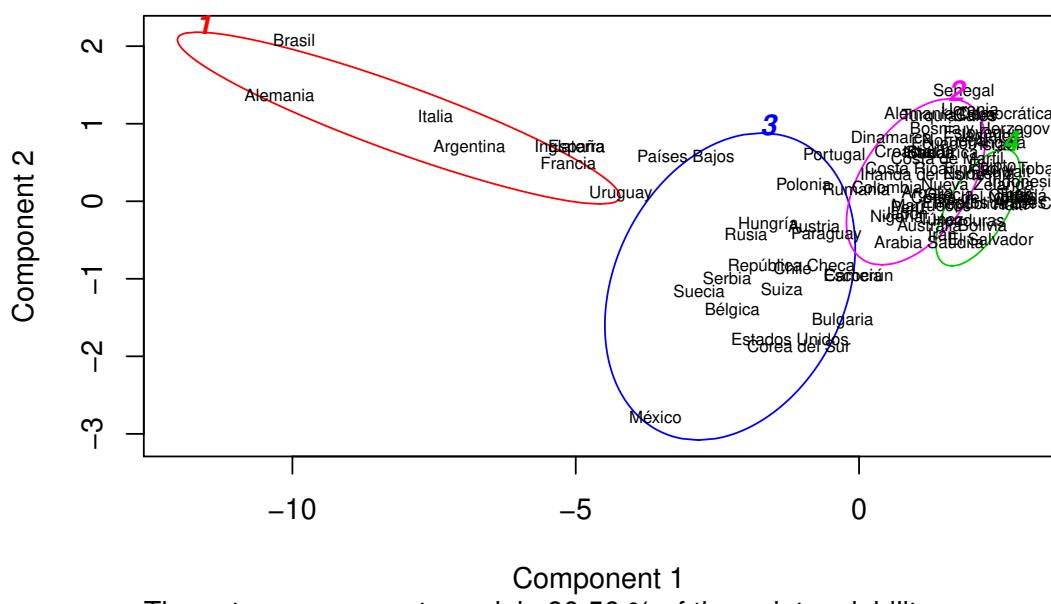


Figure 10.20. Gráfica de los pasos con el algoritmo de una rectificación de la pendiente.

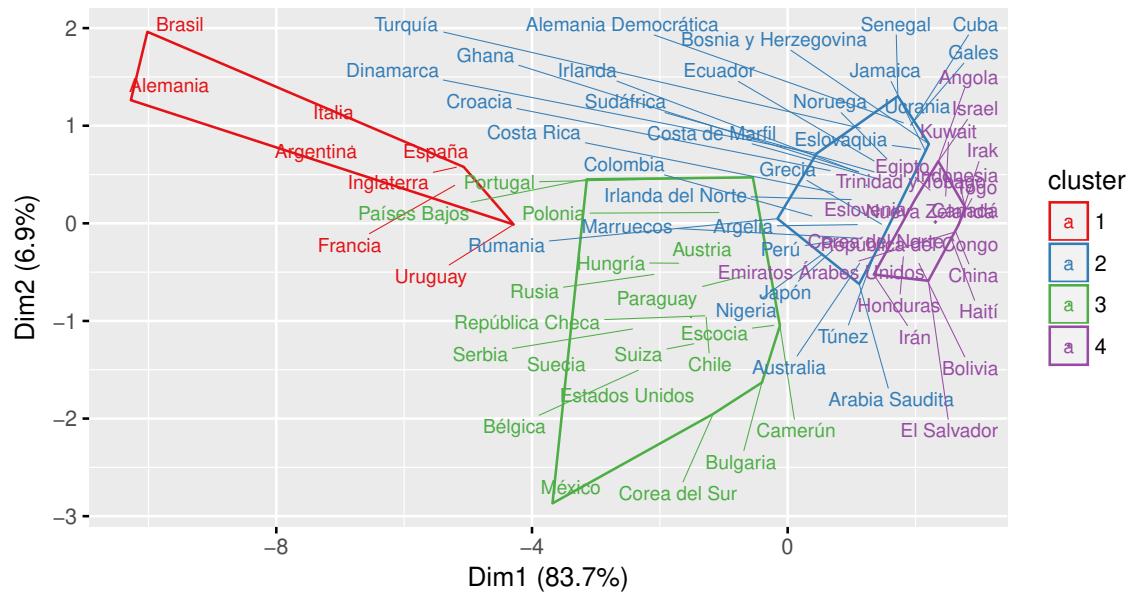


Figura 10.21: Otra manera de visualizar los grupos de países según k-means

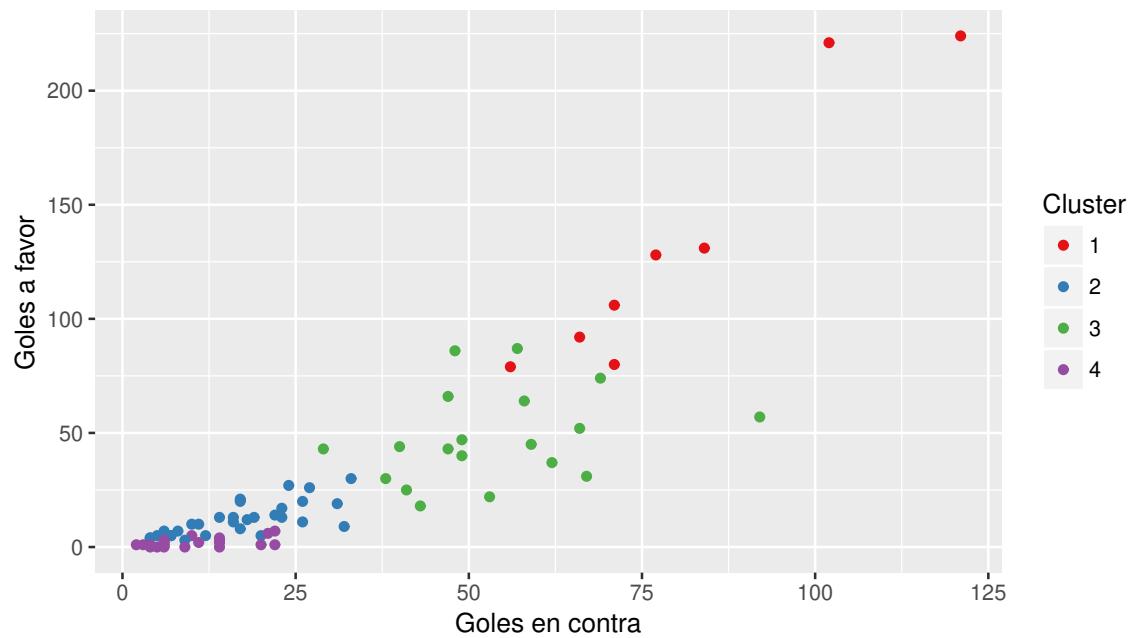


Figura 10.22: Relación entre goles a favor y en contra según agrupamiento

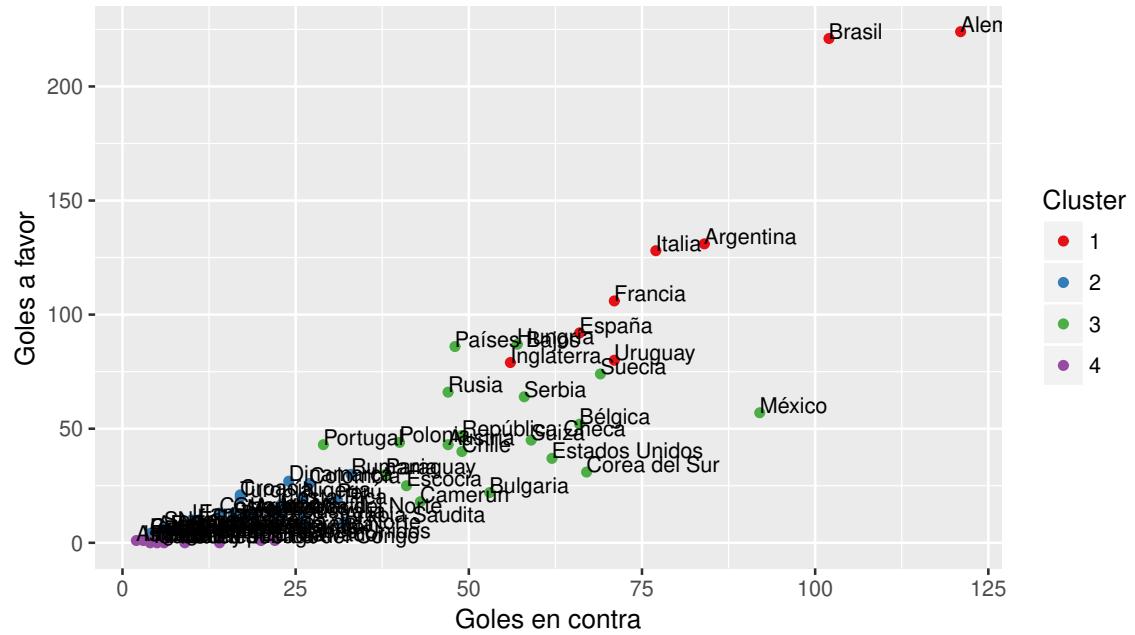


Figura 10.23: Relación entre goles a favor y en contra según agrupamiento con etiquetas

frecuencia alta, no parece resultar relevante lo que implica que tal vez de debiera hacer una mayor depuración en el análisis del documento.

- ✿ Se puede identificar, por ejemplo, qué palabras están asociadas con la palabra “noche”: “escribir” (0.44), “versos” (0.32) y “estrellada” (0.31).

```

library(tm) # Entorno para análisis de text mining
library(wordcloud) # Permite visualizar nube de palabras
library(wordcloud2) # Permite más opciones para nubes de palabras
library(devtools) # Colección de herramientas de desarrollo para paquetes
library(magrittr) # Proporciona un mecanismo para encadenar comandos
library(ggplot2) # Paquete para confeccionar dibujos
library(dendextend) # Ofrece un conjunto de funciones para dendogramas

texto=readLines('C:/.../poemas.txt', skip=0, n=-1)
# Lee el texto completo pues n=-1
str(texto)
# Muestra la estructura interna

## Eliminamos caracteres especiales en español
texto=gsub("á", "a", texto)
texto=gsub("é", "e", texto)

```

```

texto=gsub("í", "i", texto)
texto=gsub("ó", "o", texto)
texto=gsub("ú", "u", texto)
texto=gsub("ñ", "ni", texto)

docs=Corpus(VectorSource(texto))
# Crea el corpus; es decir, el acervo de documentos a analizar
inspect(docs)
# Muestra el documento

tab=content_transformer(function (x, pattern) gsub(pattern, " ", x))
# Función definida para reemplazar caracteres especiales por espacios

## Limpiamos el documento
docs=tm_map(docs, tab, "í")
docs=tm_map(docs, tab, "ó")
docs=tm_map(docs, content_transformer(tolower))
# Convierte todo a minúscula
docs=tm_map(docs, removeNumbers)
# Elimina números
docs=tm_map(docs, removeWords, stopwords("spanish"))
# Elimina palabras con poco valor para el análisis (preposiciones o interjecciones)
docs=tm_map(docs, removePunctuation)
# Elimina signos de puntuación
docs=tm_map(docs, stripWhitespace)
# Elimina los espacios vacíos excesivos

## Calculamos valores de los términos del documento
dtm=TermDocumentMatrix(docs)
m=as.matrix(dtm)
v=sort(rowSums(m), decreasing=TRUE)
d=data.frame(word=names(v), freq=v)
head(d, 10)

## Armamos la nube de palabras
set.seed(1234) # Fija una semilla
wordcloud(words=d$word, freq=d$freq, min.freq=1, max.words=200,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Set1"))

## Mejoramos la nube removiendo palabras que no son relevantes y
## unificando palabras con la misma relevancia

docs.nuevo=tm_map(docs, removeWords, c("alla", "alli", "aqui", "asi", "aun",
"hace", "hacia", "tan", "todas", "veces",
"vez", "voy"))
docs.nuevo=tm_map(docs.nuevo, content_transformer(function(x)
gsub(x, pattern="olas", replacement="ola")))
docs.nuevo=tm_map(docs.nuevo, content_transformer(function(x)

```

```

gsub(x, pattern="triste", replacement="tristeza")))
docs.nuevo=tm_map(docs.nuevo, content_transformer(function(x)
gsub(x, pattern="tristes", replacement="tristeza")))
docs.nuevo=tm_map(docs.nuevo, content_transformer(function(x)
gsub(x, pattern="solo", replacement="solo")))

## Generamos la nueva nube de palabras
dtm.nuevo=TermDocumentMatrix(docs.nuevo)
m.nuevo=as.matrix(dtm.nuevo)
v.nuevo=sort(rowSums(m.nuevo), decreasing=TRUE)
d.nuevo=data.frame(word=names(v.nuevo), freq=v.nuevo)
set.seed(4321) # Fija una semilla
wordcloud(words=d.nuevo$word, freq=d.nuevo$freq, min.freq=1, max.words=80,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Set1"))

wordcloud2(data=d.nuevo, color="random-light", backgroundColor="black")
# Produce una nube de palabras animada

## Construimos una nueva nube con argumentos distintos
m.nuevo <- m.nuevo %>% rowSums(m.nuevo) %>% sort(decreasing=TRUE)
# Proporciona las sumas de renglones ordenadas de mayor a menor
m.nuevo=data.frame(palabra=names(m.nuevo), freq=m.nuevo)
# Arreglo con la frecuencia de cada palabra
m.nuevo=m.nuevo[m.nuevo$frec>4]
# Selecciona las palabras con frecuencia superior a 4

## Armamos un histograma de frecuencias
m.nuevo[1:20, ] %>%
ggplot(aes(palabra, freq)) +
geom_bar(stat='identity', color='royalblue', fill='lightblue1') +
coord_flip() +
geom_text(aes(hjust=1.3, label=freq), size=3, color='royalblue') +
labs(x='Palabra', y='Número de usos')

## Generamos un dendograma
dtm.clus=removeSparseTerms(dtm.nuevo, sparse=.98)
# Remueve palabras dispersas
dtm.clus %>% scale %>% dist() %>% hclust(method="complete") %>% as.dendrogram()
-> dend
# Aplica el criterio completo a las variables estandarizadas
dend %>%
set("branches_k_color", k=12) %>%
# Personaliza las ramas
set("labels_col", k=12) %>% set("labels_cex", 0.8) %>%
# Personaliza las etiquetas
plot(axes=TRUE)
# Produce un dendograma personalizado

```

```

## Otras funciones
findFreqTerms(dtm.nuevo, lowfreq=15)
# Muestra las palabras con frecuencia superior a 15
findAssocs(dtm.nuevo, terms=c("alma","ojos","cuerpo","amo","corazon",
"noche", "viento"), corlimit=.3)
# Calcula la asociación entre términos frecuentes

```

Código 10.7: Código para el análisis de *text mining*

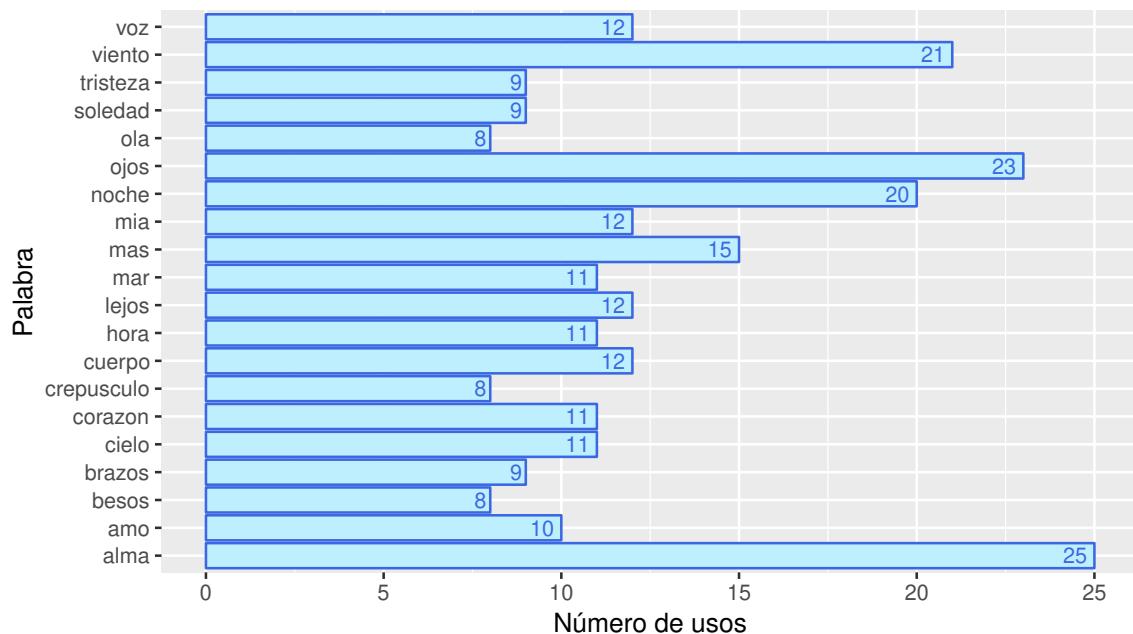


Figura 10.24: Palabras más frecuentes en poemas de Neruda

10.3.4 Ejemplo de aplicación a imágenes

Al realizar la captura de una imagen real a través de una computadora, la continuidad del tamaño, de la intensidad y de los colores se ve truncada. La combinación de características físicas continuas que nuestra mente está habituada a manejar debe ser convertida en números finitos con el objeto de ser procesados.

La visión continua debe ser discretizada para obtener una imagen digital. Esta conversión demanda la determinación de la resolución espacial y de la profundidad de color. La representación de imágenes color se basa en modelos matemáticos denominados **espacios de color**. Una imagen se puede definir como una función de dos dimensiones $f(x, y)$, donde x e y son coordenadas espaciales en el plano y f es un campo escalar que se denomina **intensidad** de la imagen en ese punto.

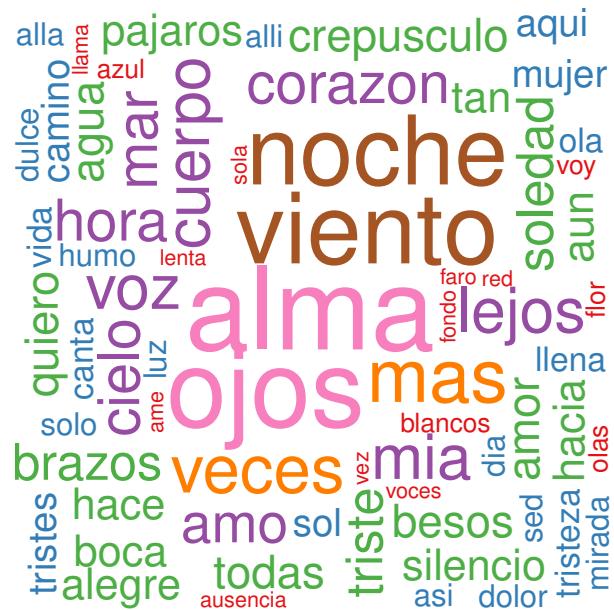


Figura 10.25: Nube de palabras más frecuentes en poemas de Neruda



Figura 10.26: Refinamiento de nube de palabras más frecuentes en poemas de Neruda



Figura 10.27: Nube de palabras con frecuencia mayor a 4 en poemas de Neruda

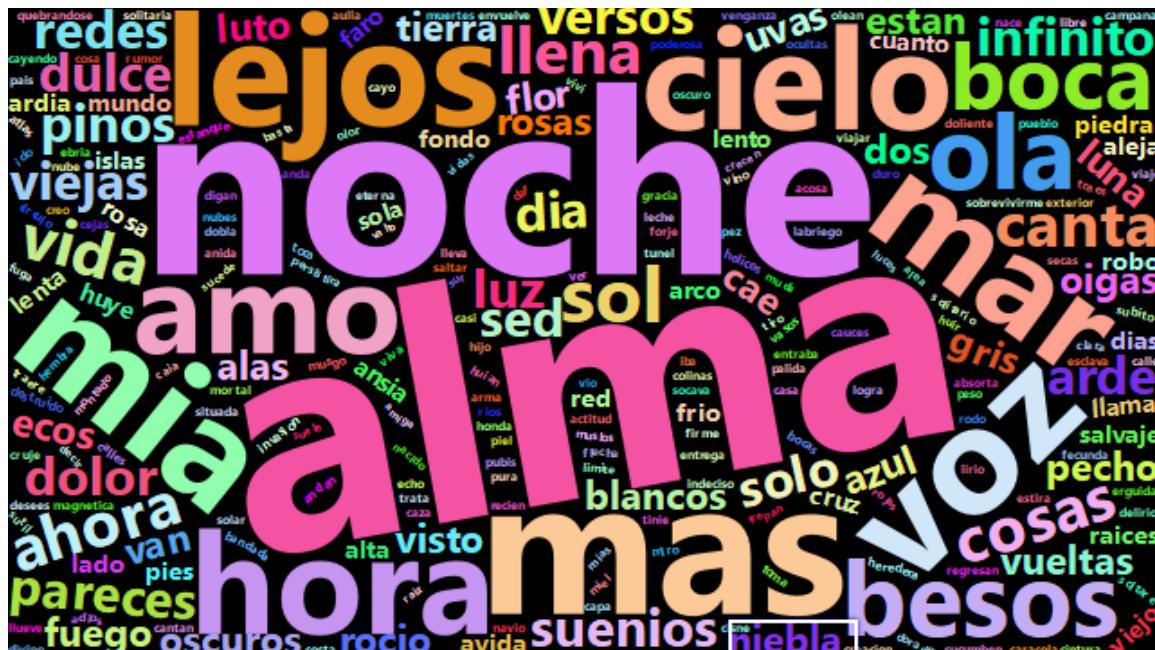


Figura 10.28: Otra forma para nube de palabras más frecuentes en poemas de Neruda

La denominación **escala de grises** se refiere a la intensidad de imágenes monocromáticas. Por otro lado, las imágenes en color están formadas por la combinación de imágenes 2-D (de dos dimensiones), como por ejemplo en el sistema *RGB*: *red-green-blue*, una imagen consiste en tres imágenes componentes individuales. Por este motivo, muchas de las técnicas desarrolladas para imágenes monocromáticas se pueden extender a imágenes de color, simplemente procesando cada una de sus componentes.

Convertir una imagen en formato digital requiere que tanto las coordenadas como la intensidad de la misma sean digitalizadas.

El uso de herramientas de procesamiento digital de imágenes se ha ido extendiendo a diversas áreas y ha dejado de ser una actividad exclusiva de los científicos, ganando cada vez más espacio en nuestra vida cotidiana. Podemos mencionar que esta disciplina tiene, entre otras, aplicaciones específicas a cuestiones relacionadas con:

- ✿ la Astronomía y la exploración del espacio,
- ✿ la Medicina,
- ✿ ciertas actividades de entretenimiento,
- ✿ el procesamiento de documentos,
- ✿ la industria y las máquinas,
- ✿ el hogar.

Los **algoritmos de clasificación** contribuyen a facilitar la interpretación de los datos provistos por una imagen. La **clasificación** se puede definir como el procedimiento para categorizar en forma automática los *pixels* de una imagen en clases.

En el siguiente ejemplo vamos a implementar un algoritmo denominado **segmentación** o *clustering* para clasificar los *pixels* de una imagen.

Ejemplo 10.14. Consideremos la imagen aérea de la siguiente figura. En la misma pueden apreciarse diferentes colores que corresponden a vegetación, agua, cielo y rocas. Si bien el ojo humano es capaz de distinguir estos colores con facilidad, no es tan sencillo para un algoritmo [3].



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

Aplicamos a la imagen aérea el algoritmo *k-means* con $k = 3, 4, 5, 6, 7, 8$ (ver el Código 10.8). En la Figura 10.30 se aprecian los *clusters* correspondientes a cada una de las particiones de *k-means*.

Para decidir cuál es la cantidad de *clusters* más apropiada, utilizamos el criterio del test explicado. En la Tabla 10.52 se aprecian las sumas de cuadrados dentro de los grupos y entre los grupos de cada clusterización.

<i>k</i>	Suma de cuadrados dentro	Suma de cuadrados entre
3	7005.233	11861.37
4	4457.626	14408.97
5	3430.948	15435.65
6	2626.265	16240.34
7	2275.888	16590.71
8	1906.423	16960.18

Tabla 10.52: Suma de cuadrados

El estadístico alcanza su valor mínimo para $k = 6$. Este algoritmo es una alternativa para la segmentación de bordes; es decir, la localización de *pixels* donde cambia el conglomerado.

```
library(jpeg) # Paquete para trabajar con archivos de imagen JPEG
library(ggplot2) # Paquete para confeccionar dibujos

mapa=readJPEG("C:/.../aerea.jpg")
# Lee la imagen de un archivo jpg

imgDm=dim(mapa) # Obtiene la dimensión de la imagen
imgRGB=data.frame(
x=rep(1:imgDm[2], each=imgDm[1]),
```



(a) $k = 3$



(b) $k = 4$



(c) $k = 5$



(d) $k = 6$



(e) $k = 7$



(f) $k = 8$

Figura 10.30: Reconstrucción de la imagen

```

y=rep(imgDm[1]:1 , imgDm[2]) ,
R=as.vector(mapa[, ,1]),
G=as.vector(mapa[, ,2]),
B=as.vector(mapa[, ,3])
) # Asigna los canales RGB a los datos

clusters=3
# Variar con 4,5,6,7,8
kmimg=kmeans(imgRGB[, c("R","G","B")], centers=clusters)
colores=rgb(kmimg$centers[kmimg$cluster ,])
# Aplica la clusterización por el algoritmo de k-means

ggplot(data=imgRGB, aes(x=x, y=y)) +
geom_point(colour=colores) +
theme_void()
# Recrea la imagen

## Calculamos sumas de cuadrados

kmimg3=kmeans(imgRGB[, c("R","G","B")], centers=3)
kmimg4=kmeans(imgRGB[, c("R","G","B")], centers=4)
kmimg5=kmeans(imgRGB[, c("R","G","B")], centers=5)
kmimg6=kmeans(imgRGB[, c("R","G","B")], centers=6)
kmimg7=kmeans(imgRGB[, c("R","G","B")], centers=7)
kmimg8=kmeans(imgRGB[, c("R","G","B")], centers=8)

ssd3=sum(kmimg3$withinss)
ssd4=sum(kmimg4$withinss)
ssd5=sum(kmimg5$withinss)
ssd6=sum(kmimg6$withinss)
ssd7=sum(kmimg7$withinss)
ssd8=sum(kmimg8$withinss)

sse3=kmimg3$betweens
sse4=kmimg4$betweens
sse5=kmimg5$betweens
sse6=kmimg6$betweens
sse7=kmimg7$betweens
sse8=kmimg8$betweens

ssd=c(ssd3,ssd4,ssd5,ssd6,ssd7,ssd8) # Suma de cuadrados dentro del grupo
sse=c(sse3,sse4,sse5,sse6,sse7,sse8) # Suma de cuadrados entre grupos

n=242*640 # Cantidad de objetos
est=0
for (k in 1:5) {est[k]=(ssd[k]-ssd[k+1])/(ssd[k+1]/(n-k-1))}
# Calcula los estadísticos
which.min(est) # Dice que el valor mínimo del estadístico es para k=6

```

[Código 10.8](#): Código para el análisis de clusterización de una imagen



10.4 Ejercitación

Ejercicio 1.

Consideremos el conjunto de datos representado por la matriz

$$M = \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 5 & 0 \\ 6 & 1 \\ 3 & 2 \\ 4 & 1 \end{pmatrix}$$

1. Graficar en \mathbb{R}^2 y construir el dendrograma correspondiente utilizando el criterio del vecino más lejano, utilizando la distancia euclídea.
2. Igual que el ítem anterior pero utilizando el criterio de vecino más cercano y promedio.
3. Repetir los mismos ejercicios utilizando las variables estandarizadas y comparar los resultados obtenidos.

Ejercicio 2.

Dada la siguiente matriz de distancias D , realizar los dendrogramas correspondientes a los métodos vecino más cercano, vecino más lejano y promedio, utilizando la distancia euclídea.

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 18 & 10 & 0 & 0 & 0 \\ 20 & 15 & 24 & 0 & 0 \\ 18 & 20 & 8 & 6 & 0 \end{pmatrix}$$

¿Se encuentran diferencias entre los resultados obtenidos por los diferentes algoritmos?

Ejercicio 3.

El objetivo es obtener cinco agrupamientos de los datos correspondientes al archivo disponible en <https://goo.gl/WNGHuH>.

1. Realizar un Análisis en Componentes Principales. ¿Qué proporción de la variabilidad total en las variables medidas explican las dos primeras componentes? Utilizando un gráfico de individuos, determinar grupos en los datos. ¿Cuántos grupos hay? ¿Qué tipo de pizzas pertenecen a cuáles agrupamientos? Comparar con el ítem anterior.
2. Aplicar un método de agrupamiento a los resultados del ítem anterior (valores de los casos sobre las componentes).

3. Aplicar el método de *K*-Medias a los datos de manera de obtener 5 grupos y comparar con los resultados anteriores.
4. Resumir los resultados. ¿Tienen los datos una estructura como para agruparlos? En el caso afirmativo, ¿en cuántos grupos sería más conveniente agrupar? Justificar.

Ejercicio 4.

Un Museo encuesta a un grupo de niños visitantes al final el recorrido. Dicha encuesta está diseñada con distintas preguntas generales y algunas que pueden ayudar a identificar grupos y diseñar estrategias que vayan acorde con los niños que están más interesados en asistir a un museo. Los datos recabados se encuentran disponibles en <https://goo.gl/Ljy9Sr>. Algunos de los interrogantes que encontramos en esta encuesta están dados por las siguientes variables.

Sexo: varón o mujer

Edad: en años

Diversión: ¿Es divertido ir al museo?*

Compras: Cuando voy al museo, ¿le pido a mis papás que me compren algo de lo que venden adentro?*

Aprendizaje ¿Puedo aprender en la escuela lo mismo que en el museo?*

Excursión: ¿Prefiero ir al museo en excursiones con la escuela?*

Juego: ¿Ir al museo en mi tiempo libre me quita tiempo para jugar?*

Interés: ¿No me interesa en lo más mínimo asistir al museo?*

Gusto: ¿Te gustó tu visita al museo? Con niveles Sí y No.

*Con niveles del 1 al 7 que recorren desde el desacuerdo total al acuerdo total.

El grupo de investigadores desea clasificar las opiniones generales que se tienen en relación al Museo y pretende agrupar a los 25 niños que respondieron la encuesta. Justificar la aplicación del análisis de *cluster*, demostrando que existe fuerte asociación entre las variables que van a configurarlo. Analizar el número de *clusters* y las características de los mismos.

Ejercicio 5. Se quiere agrupar a 7 alumnos de primer año de Psicología en base a sus notas en las asignaturas de las áreas de

X_1 : Básica

X_1 : Metodología

X_1 : Evolutiva

X_1 : Social

X_1 : Clínica

Para ello se calcula la media por área obteniendo la Tabla 10.53.

Estudiante	X_1	X_2	X_3	X_4	X_5
E_1	8	9	7	8	6
E_2	7	8	7	8	8
E_3	2	3	8	7	2
E_4	1	2	6	7	1
E_5	1	1	1	9	8
E_6	2	3	1	8	9
E_7	7	9	4	7	9

Tabla 10.53: Estudiantes del Psicología

Con los datos de la Tabla 10.53,

1. realizar los dendogramas, utilizando el método de Ward, para los datos crudos.
2. realizar los dendogramas, utilizando el método de Ward, para los datos estandarizados por variable.
3. ¿A qué se deben las diferencias observadas en los dendrogramas?
4. ¿Cuál de las alternativas debería seleccionarse teniendo en cuenta este coeficiente y la interpretabilidad de los resultados?

Ejercicio 6.

En el archivo disponible en <https://goo.gl/XcKCQN> se encuentran los consumos medios por tipo de proteína de varios países europeos. Interesa estudiar las categorías diferenciales de este consumo.

1. Utilizando el método de Ward y la distancia euclídea, particionar en dos *clusters*. ¿Como podría llamarse a cada uno de ellos?
2. Idem anterior pero en cuatro *clusters*. Utilizando el dendograma, ¿con cuál de las clasificaciones habría que quedarse?
3. Realizar una clusterización de las variables.
4. Comparar los resultados obtenidos con el análisis de componentes principales.

Apéndice A

Nociones elementales de Álgebra Lineal

Sea \mathbb{K} un cuerpo de escalares. Se define un **espacio vectorial** como una cuaterna $(\mathbb{V}, \mathbb{K}, +, \cdot)$ donde \mathbb{V} es un conjunto cuyos elementos se llaman **vectores** y $+ : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{V}$ y $\cdot : \mathbb{K} \times \mathbb{V} \rightarrow \mathbb{V}$ son dos operaciones que satisfacen lo siguiente:

- * $v + w = w + v$ para todo $v, w \in \mathbb{V}$
- * $(u + v) + w = u + (v + w)$ para todo $u, v, w \in \mathbb{V}$
- * existe $0 \in \mathbb{V}$ tal que $0 + v = v + 0 = v$ para todo $v \in \mathbb{V}$
- * para todo $v \in \mathbb{V}$ existe $-v \in \mathbb{V}$ tal que $v + (-v) = -v + v = 0$
- * $\alpha(v + w) = \alpha v + \alpha w$ para todo $\alpha \in \mathbb{K}$ y $v, w \in \mathbb{V}$
- * $(\alpha + \beta)v = \alpha v + \beta v$ para todo $\alpha, \beta \in \mathbb{K}$ y $v \in \mathbb{V}$
- * $(\alpha\beta)v = \alpha(\beta v) = \beta(\alpha v)$ para todo $\alpha, \beta \in \mathbb{K}$ y $v \in \mathbb{V}$
- * $1v = v$ para todo $v \in \mathbb{V}$

Por abuso de notación, se suele decir simplemente el espacio vectorial \mathbb{V} .

En análisis multivariado, utilizaremos como cuerpo de escalares a los números reales; es decir, trabajaremos con espacios vectoriales reales.

Un **subespacio** vectorial W es un subconjunto de un espacio vectorial V , que satisface por sí mismo la definición de espacio vectorial con las mismas operaciones que V .

Alcanza con que el subconjunto contenga al vector nulo y sea cerrado para las dos operaciones para garantizar que W es subespacio de V .

Una **transformación lineal** $T : \mathbb{V} \rightarrow \mathbb{W}$ es una función entre dos espacios vectoriales tal que para todo $v, w \in \mathbb{V}$ y $\alpha \in \mathbb{K}$ se satisface

- * $T(v + w) = T(v) + T(w)$

* $T(\alpha v) = \alpha T(v)$

Recordemos que trabajaremos con $\mathbb{K} = \mathbb{K}$.

Una manera cómoda para realizar operaciones es trabajar con la matriz asociada a una transformación lineal. Para esta definición es necesario introducir el concepto de **coordenadas en una base**. Sea $B = \{v_1, v_2, \dots, v_n\}$ una base para el espacio vectorial \mathbb{V} . Sea $v \in \mathbb{V}$, entonces existen únicos escalares $\alpha_1, \alpha_2, \dots, \alpha_n$ tales que $v = \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n$. Las coordenadas de v en la base B son $[v]_B = (\alpha_1, \alpha_2, \dots, \alpha_n)^t$.

Sea $T : \mathbb{V} \rightarrow \mathbb{W}$ una transformación lineal y sean $B = \{v_1, v_2, \dots, v_n\}$ y $B' = \{w_1, w_2, \dots, w_m\}$ bases de \mathbb{V} y \mathbb{W} respectivamente. Se define la **matriz asociada** a T en las bases B y B' como $M_{BB'}(T) \in \mathbb{R}^{m \times n}$ cuya columna j -ésima es $[T(v_j)]_{B'}$. Resulta que $M_{BB'}(T)[v]_B = [T(v)]_{B'}$.

Apéndice B

Nociones de Estadística

Consideramos la siguiente notación:

- ✿ $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ la matriz diagonal cuyos elementos son los autovalores de la matriz de covarianzas poblacional Σ .
- ✿ $\widehat{\Lambda} = \text{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \dots, \widehat{\lambda}_p)$ la matriz diagonal cuyos elementos son los autovalores de la matriz de covarianzas muestral $\widehat{\Sigma}$.
- ✿ $\Gamma = \{\gamma_1, \gamma_2, \dots, \gamma_p\}$ el conjunto de autovectores de la matriz de covarianzas poblacional Σ .
- ✿ $\widehat{\Gamma} = \{\widehat{\gamma}_1, \widehat{\gamma}_2, \dots, \widehat{\gamma}_p\}$ el conjunto de autovectores de la matriz de covarianzas muestral $\widehat{\Sigma}$.
- ✿ $\widehat{L}_j = \sqrt{n}(\widehat{\lambda}_j - \lambda_j)$.

Se puede demostrar que valen las siguientes propiedades:

- ✿ Los elementos de $\widehat{\Lambda}$ son los estimadores de máxima verosimilitud de los de Λ .
- ✿ Los elementos de $\widehat{\Gamma}$ son los estimadores de máxima verosimilitud de los de Γ .
- ✿ \widehat{L}_j converge en distribución a una $\mathcal{N}(0, 2\lambda_j^2)$.
- ✿ Si la distribución original de las observaciones es Normal, entonces \widehat{L}_j son asintóticamente independientes. Mientras que si la distribución original no es Normal, esta propiedad no se cumple necesariamente.

Test para porcentajes

Cuando queremos decidir cuál es la cantidad de componentes principales q a seleccionar de modo tal que las mismas cubran cierta proporción, $0 < p_0 < 1$, de la variabilidad total del conjunto, disponemos de una prueba para darle validez estadística a la decisión. La misma se conoce como prueba para porcentajes y se define de la siguiente manera. Si $X \in \mathbb{R}^{n \times p}$ y $q < p$ es la cantidad de componentes a seleccionar, las hipótesis de interés en esta prueba pueden plantearse como

$$\begin{cases} H_0 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} \leq p_0 \\ H_1 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} > p_0 \end{cases}$$

Este test es equivalente al test dado por

$$\begin{cases} H_0 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} = p_0 \\ H_1 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} > p_0 \end{cases}$$

La hipótesis nula se puede expresar en función del elemento θ_{p_0} definido como

$$\theta_{p_0} = (1 - p_0) \sum_{i=1}^q \lambda_i - p_0 \sum_{i=q+1}^p \lambda_i$$

siendo el estimador para este estadístico

$$\hat{\theta}_{p_0} = (1 - p_0) \sum_{i=1}^q \hat{\lambda}_i - p_0 \sum_{i=q+1}^p \hat{\lambda}_i$$

Por la propiedad de independencia asintótica de los autovalores bajo normalidad, la varianza de este estimador resulta

$$\sigma_{p_0}^2 = 2 \left[(1 - p_0)^2 \sum_{i=1}^q \lambda_i^2 - p_0^2 \sum_{i=q+1}^p \lambda_i^2 \right]$$

y la estimación para esta varianza es

$$\hat{\sigma}_{p_0}^2 = 2 \left[(1 - p_0)^2 \sum_{i=1}^q \hat{\lambda}_i^2 - p_0^2 \sum_{i=q+1}^p \hat{\lambda}_i^2 \right]$$

Esto nos permite reescribir las hipótesis de la siguiente manera

$$\begin{cases} H_0 : \theta_{p_0} \leq 0 \\ H_1 : \theta_{p_0} > 0 \end{cases}$$

Puesto que tenemos una combinación lineal de variables asintóticamente normales, θ_{p_0} también tiene distribución asintótica normal y rechazaremos H_0 cuando $\frac{\sqrt{n}\widehat{\theta}_{p_0}}{|\widehat{\sigma}_{p_0}|} \geq z_\alpha$.

Prueba de esfericidad

Consideramos la variable aleatoria $X \sim \mathcal{N}_p(\mu, \Sigma)$, donde p indica la cantidad de variables, y sean $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ los autovalores de la matriz Σ con respectivos autovectores $\gamma_1, \gamma_2, \dots, \gamma_p$. Notamos a sus estimadores de máxima verosimilitud por $(\widehat{\lambda}_1, \widehat{\lambda}_2, \dots, \widehat{\lambda}_p)$ y $\{\widehat{\gamma}_1, \widehat{\gamma}_2, \dots, \widehat{\gamma}_p\}$.

Nos interesa probar si a partir de la componente $r+1$, no hay direcciones de mayor variabilidad que las otras. Planteamos para ello las siguientes hipótesis

$$\begin{cases} H_0 : \lambda_{r+1} = \lambda_{r+2} = \dots = \lambda_p \\ H_1 : \lambda_{r+1} \geq \lambda_{r+2} \geq \dots \geq \lambda_p \text{ con al menos una desigualdad estricta} \end{cases}$$

Utilizamos el estadístico de contraste

$$M_r = \frac{\prod_{j=r+1}^p \widehat{\lambda}_j}{\left((p-r)^{-1} \sum_{j=r+1}^p \widehat{\lambda}_j \right)^{p-r}}$$

La distribución del estadístico, bajo H_0 (es decir, cuando H_0 es verdadera) es

$$-n \log M_r \xrightarrow{d} \chi^2_{0.5(p-r)(p+1-r)-1}$$

Se rechaza para valores chicos del estadístico.

Apéndice A

Nociones elementales de Álgebra Lineal

Sea \mathbb{K} un cuerpo de escalares. Se define un **espacio vectorial** como una cuaterna $(\mathbb{V}, \mathbb{K}, +, \cdot)$ donde \mathbb{V} es un conjunto cuyos elementos se llaman **vectores** y $+ : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{V}$ y $\cdot : \mathbb{K} \times \mathbb{V} \rightarrow \mathbb{V}$ son dos operaciones que satisfacen lo siguiente:

- * $v + w = w + v$ para todo $v, w \in \mathbb{V}$
- * $(u + v) + w = u + (v + w)$ para todo $u, v, w \in \mathbb{V}$
- * existe $0 \in \mathbb{V}$ tal que $0 + v = v + 0 = v$ para todo $v \in \mathbb{V}$
- * para todo $v \in \mathbb{V}$ existe $-v \in \mathbb{V}$ tal que $v + (-v) = -v + v = 0$
- * $\alpha(v + w) = \alpha v + \alpha w$ para todo $\alpha \in \mathbb{K}$ y $v, w \in \mathbb{V}$
- * $(\alpha + \beta)v = \alpha v + \beta v$ para todo $\alpha, \beta \in \mathbb{K}$ y $v \in \mathbb{V}$
- * $(\alpha\beta)v = \alpha(\beta v) = \beta(\alpha v)$ para todo $\alpha, \beta \in \mathbb{K}$ y $v \in \mathbb{V}$
- * $1v = v$ para todo $v \in \mathbb{V}$

Por abuso de notación, se suele decir simplemente el espacio vectorial \mathbb{V} .

En análisis multivariado, utilizaremos como cuerpo de escalares a los números reales; es decir, trabajaremos con espacios vectoriales reales.

Un **subespacio** vectorial W es un subconjunto de un espacio vectorial V , que satisface por sí mismo la definición de espacio vectorial con las mismas operaciones que V .

Alcanza con que el subconjunto contenga al vector nulo y sea cerrado para las dos operaciones para garantizar que W es subespacio de V .

Una **transformación lineal** $T : \mathbb{V} \rightarrow \mathbb{W}$ es una función entre dos espacios vectoriales tal que para todo $v, w \in \mathbb{V}$ y $\alpha \in \mathbb{K}$ se satisface

- * $T(v + w) = T(v) + T(w)$

* $T(\alpha v) = \alpha T(v)$

Recordemos que trabajaremos con $\mathbb{K} = \mathbb{K}$.

Una manera cómoda para realizar operaciones es trabajar con la matriz asociada a una transformación lineal. Para esta definición es necesario introducir el concepto de **coordenadas en una base**. Sea $B = \{v_1, v_2, \dots, v_n\}$ una base para el espacio vectorial \mathbb{V} . Sea $v \in \mathbb{V}$, entonces existen únicos escalares $\alpha_1, \alpha_2, \dots, \alpha_n$ tales que $v = \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n$. Las coordenadas de v en la base B son $[v]_B = (\alpha_1, \alpha_2, \dots, \alpha_n)^t$.

Sea $T : \mathbb{V} \rightarrow \mathbb{W}$ una transformación lineal y sean $B = \{v_1, v_2, \dots, v_n\}$ y $B' = \{w_1, w_2, \dots, w_m\}$ bases de \mathbb{V} y \mathbb{W} respectivamente. Se define la **matriz asociada** a T en las bases B y B' como $M_{BB'}(T) \in \mathbb{R}^{m \times n}$ cuya columna j -ésima es $[T(v_j)]_{B'}$. Resulta que $M_{BB'}(T)[v]_B = [T(v)]_{B'}$.

Apéndice B

Nociones de Estadística

Consideramos la siguiente notación:

- ✿ $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ la matriz diagonal cuyos elementos son los autovalores de la matriz de covarianzas poblacional Σ .
- ✿ $\widehat{\Lambda} = \text{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \dots, \widehat{\lambda}_p)$ la matriz diagonal cuyos elementos son los autovalores de la matriz de covarianzas muestral $\widehat{\Sigma}$.
- ✿ $\Gamma = \{\gamma_1, \gamma_2, \dots, \gamma_p\}$ el conjunto de autovectores de la matriz de covarianzas poblacional Σ .
- ✿ $\widehat{\Gamma} = \{\widehat{\gamma}_1, \widehat{\gamma}_2, \dots, \widehat{\gamma}_p\}$ el conjunto de autovectores de la matriz de covarianzas muestral $\widehat{\Sigma}$.
- ✿ $\widehat{L}_j = \sqrt{n}(\widehat{\lambda}_j - \lambda_j)$.

Se puede demostrar que valen las siguientes propiedades:

- ✿ Los elementos de $\widehat{\Lambda}$ son los estimadores de máxima verosimilitud de los de Λ .
- ✿ Los elementos de $\widehat{\Gamma}$ son los estimadores de máxima verosimilitud de los de Γ .
- ✿ \widehat{L}_j converge en distribución a una $\mathcal{N}(0, 2\lambda_j^2)$.
- ✿ Si la distribución original de las observaciones es Normal, entonces \widehat{L}_j son asintóticamente independientes. Mientras que si la distribución original no es Normal, esta propiedad no se cumple necesariamente.

Test para porcentajes

Cuando queremos decidir cuál es la cantidad de componentes principales q a seleccionar de modo tal que las mismas cubran cierta proporción, $0 < p_0 < 1$, de la variabilidad total del conjunto, disponemos de una prueba para darle validez estadística a la decisión. La misma se conoce como prueba para porcentajes y se define de la siguiente manera. Si $X \in \mathbb{R}^{n \times p}$ y $q < p$ es la cantidad de componentes a seleccionar, las hipótesis de interés en esta prueba pueden plantearse como

$$\begin{cases} H_0 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} \leq p_0 \\ H_1 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} > p_0 \end{cases}$$

Este test es equivalente al test dado por

$$\begin{cases} H_0 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} = p_0 \\ H_1 : \frac{\sum_{j=1}^q \lambda_j}{\sum_{j=1}^p \lambda_j} > p_0 \end{cases}$$

La hipótesis nula se puede expresar en función del elemento θ_{p_0} definido como

$$\theta_{p_0} = (1 - p_0) \sum_{i=1}^q \lambda_i - p_0 \sum_{i=q+1}^p \lambda_i$$

siendo el estimador para este estadístico

$$\hat{\theta}_{p_0} = (1 - p_0) \sum_{i=1}^q \hat{\lambda}_i - p_0 \sum_{i=q+1}^p \hat{\lambda}_i$$

Por la propiedad de independencia asintótica de los autovalores bajo normalidad, la varianza de este estimador resulta

$$\sigma_{p_0}^2 = 2 \left[(1 - p_0)^2 \sum_{i=1}^q \lambda_i^2 - p_0^2 \sum_{i=q+1}^p \lambda_i^2 \right]$$

y la estimación para esta varianza es

$$\hat{\sigma}_{p_0}^2 = 2 \left[(1 - p_0)^2 \sum_{i=1}^q \hat{\lambda}_i^2 - p_0^2 \sum_{i=q+1}^p \hat{\lambda}_i^2 \right]$$

Esto nos permite reescribir las hipótesis de la siguiente manera

$$\begin{cases} H_0 : \theta_{p_0} \leq 0 \\ H_1 : \theta_{p_0} > 0 \end{cases}$$

Puesto que tenemos una combinación lineal de variables asintóticamente normales, θ_{p_0} también tiene distribución asintótica normal y rechazaremos H_0 cuando $\frac{\sqrt{n}\widehat{\theta}_{p_0}}{|\widehat{\sigma}_{p_0}|} \geq z_\alpha$.

Prueba de esfericidad

Consideramos la variable aleatoria $X \sim \mathcal{N}_p(\mu, \Sigma)$, donde p indica la cantidad de variables, y sean $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ los autovalores de la matriz Σ con respectivos autovectores $\gamma_1, \gamma_2, \dots, \gamma_p$. Notamos a sus estimadores de máxima verosimilitud por $(\widehat{\lambda}_1, \widehat{\lambda}_2, \dots, \widehat{\lambda}_p)$ y $\{\widehat{\gamma}_1, \widehat{\gamma}_2, \dots, \widehat{\gamma}_p\}$.

Nos interesa probar si a partir de la componente $r+1$, no hay direcciones de mayor variabilidad que las otras. Planteamos para ello las siguientes hipótesis

$$\begin{cases} H_0 : \lambda_{r+1} = \lambda_{r+2} = \dots = \lambda_p \\ H_1 : \lambda_{r+1} \geq \lambda_{r+2} \geq \dots \geq \lambda_p \text{ con al menos una desigualdad estricta} \end{cases}$$

Utilizamos el estadístico de contraste

$$M_r = \frac{\prod_{j=r+1}^p \widehat{\lambda}_j}{\left((p-r)^{-1} \sum_{j=r+1}^p \widehat{\lambda}_j \right)^{p-r}}$$

La distribución del estadístico, bajo H_0 (es decir, cuando H_0 es verdadera) es

$$-n \log M_r \xrightarrow{d} \chi^2_{0.5(p-r)(p+1-r)-1}$$

Se rechaza para valores chicos del estadístico.

Referencias

- [1] Sobre la eficiencia asintótica de ciertas pruebas no paramétricas de dos muestras, Los anales de la estadística matemática.
- [2] Alan Agresti et al., *A survey of exact inference for contingency tables*, Statistical science **7** (1992), no. 1, 131–153.
- [3] Michael R Anderberg, *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, vol. 19, Academic press, 2014.
- [4] Marti J Anderson, *Distance-based tests for homogeneity of multivariate dispersions*, Biometrics **62** (2006), no. 1, 245–253.
- [5] Kevin Ashton et al., *That ‘internet of things’ thing*, RFID journal **22** (2009), no. 7, 97–114.
- [6] Juan José Badimón, Lina Badimón, and Valentín Fuster, *Regression of atherosclerotic lesions by high density lipoprotein plasma fraction in the cholesterol-fed rabbit.*, The Journal of clinical investigation **85** (1990), no. 4, 1234–1241.
- [7] Robert G Bland, Donald Goldfarb, and Michael J Todd, *The ellipsoid method: A survey*, Operations research **29** (1981), no. 6, 1039–1091.
- [8] Morton B Brown and Alan B Forsythe, *Robust tests for the equality of variances*, Journal of the American Statistical Association **69** (1974), no. 346, 364–367.
- [9] Juan de Burgos Román, *Álgebra lineal y geometría cartesiana*, McGraw-Hill,, 2006.
- [10] B Buser, *A training algorithm for optimal margin classifier*, Proc. 5th Annual ACM Workshop on Computational Learning Theory, 1992, 1992, pp. 144–152.
- [11] Herman Chernoff, *Chernoff faces*, International Encyclopedia of Statistical Science, Springer, 2011, pp. 243–244.
- [12] WJ Conover, *Practical nonparametric statics*, John Wiley & Sons, Inc., New York (1999), 130–133.

- [13] Jan De Leeuw and Patrick Mair, *Multidimensional scaling using majorization: Smacof in r*, (2011).
- [14] 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.
- [15] JB Elsner, GS Lehmler, and TB Kimberlain, *Objective classification of atlantic hurricanes*, Journal of Climate **9** (1996), no. 11, 2880–2889.
- [16] Syed Masum Emran and Nong Ye, *Robustness of chi-square and canberra distance metrics for computer intrusion detection*, Quality and Reliability Engineering International **18** (2002), no. 1, 19–28.
- [17] Ronald Aylmer Fisher, *The design of experiments*, Oliver And Boyd; Edinburgh; London, 1937.
- [18] John C Gower, *Some distance properties of latent root and vector methods used in multivariate analysis*, Biometrika **53** (1966), no. 3-4, 325–338.
- [19] Paul E Green, *Marketing applications of mds: Assessment and outlook: After a decade of development, what have we learned from mds in marketing?*, Journal of Marketing **39** (1975), no. 1, 24–31.
- [20] Michael J Greenacre, *Correspondence analysis*, London: Academic Press, 1984.
- [21] 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.
- [22] Kenneth Hoffman, Ray Kunze, and Hugo E Finsterbusch, *Álgebra lineal*, Prentice-Hall Hispanoamericana, 1973.
- [23] Jan Holler, Vlasios Tsiatsis, Catherine Mulligan, Stamatis Karnouskos, Stefan Avesand, and David Boyle, *Internet of things*, Academic Press, 2014.
- [24] Harold Hotelling, *Analysis of a complex of statistical variables into principal components.*, Journal of educational psychology **24** (1933), no. 6, 417.
- [25] _____, *The generalization of student's ratio*, Breakthroughs in statistics, Springer, 1992, pp. 54–65.
- [26] Mia Hubert and Katrien Van Driessen, *Fast and robust discriminant analysis*, Computational Statistics & Data Analysis **45** (2004), no. 2, 301–320.

- [27] Giuseppe Jurman, Samantha Riccadonna, Roberto Visintainer, and Cesare Furlanello, *Canberra distance on ranked lists*, Proceedings of Advances in Ranking NIPS 09 Workshop, Citeseer, 2009, pp. 22–27.
- [28] Joseph B Kruskal, *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*, Psychometrika **29** (1964), no. 1, 1–27.
- [29] William H Kruskal and W Allen Wallis, *Use of ranks in one-criterion variance analysis*, Journal of the American statistical Association **47** (1952), no. 260, 583–621.
- [30] Ingrid Juliana Lagos and Jose Alberto Vargas, *Sistema de familias de distribuciones de johnson, una alternativa para el manejo de datos no normales en cartas de control*, Revista Colombiana de Estadística **26** (2003), no. 1, 25–40.
- [31] GN Lance and WT Williams, *Note on a new information-statistic classificatory program*, Comput. J **11** (1968), 195.
- [32] Prasanta Chandra Mahalanobis, *On the generalized distance in statistics*, ., National Institute of Science of India, 1936.
- [33] RARD Maronna, R Douglas Martin, and Victor Yohai, *Robust statistics*, vol. 1, John Wiley & Sons, Chichester. ISBN, 2006.
- [34] 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.
- [35] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman, *Applied linear regression models*, Mc McGraw Hill, 1996.
- [36] Cecilia Oliva, *Métodos para la segmentación de datos longitudinales. aplicación a datos de rendimientos de cultivos en argentina*, Ph.D. thesis, Tesis Lic. Buenos Aires, Argentina, UBA. 72p, 2015.
- [37] Karl Pearson, *Principal components analysis*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **6** (1901), no. 2, 559.
- [38] 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.
- [39] Robin L Plackett, *Karl pearson and the chi-squared test*, International Statistical Review/Revue Internationale de Statistique (1983), 59–72.

- [40] Alain Li Wan Po, *Statistics for pharmacists*, McGraw Hill Professional, 1998.
- [41] Peter J Rousseeuw and Katrien Van Driessen, *A fast algorithm for the minimum covariance determinant estimator*, *Technometrics* **41** (1999), no. 3, 212–223.
- [42] 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.
- [43] Robert Tibshirani, Guenther Walther, and Trevor Hastie, *Estimating the number of clusters in a data set via the gap statistic*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** (2001), no. 2, 411–423.
- [44] Valentin Todorov, Peter Filzmoser, et al., *An object-oriented framework for robust multivariate analysis*, Citeseer, 2009.
- [45] Warren S Torgerson, *Multidimensional scaling of similarity*, *Psychometrika* **30** (1965), no. 4, 379–393.
- [46] Stefan Van Aelst and Peter Rousseeuw, *Minimum volume ellipsoid*, Wiley Interdisciplinary Reviews: Computational Statistics **1** (2009), no. 1, 71–82.
- [47] 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.
- [48] Frank Wilcoxon, *Individual comparisons by ranking methods*, *Biometrics bulletin* **1** (1945), no. 6, 80–83.
- [49] John Wishart, *The generalised product moment distribution in samples from a normal multivariate population*, *Biometrika* **20** (1928), no. 1/2, 32–52.