Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

Carrera: Licenciatura En Sistemas De Información

Asignatura: Base de Datos II

Tema 18: Fase de Exploración en Minería de Datos

Grupo Nº5

Alumnos:	L.U.:
Alcalá Miguel Fernando	54.018
Alegre Gastón Nahuel	54.479
Benitez Elio Gastón	54.229
Castillo Claudio M. F.	46.817
Dabove Grbavac Tyago Daniel	54.619

INDICE

INDICE DE FIGURAS

INDICE DE TABLAS

Introducción

En esta fase de exploración se refiere al análisis previo o preliminar de los datos, nos pueden servir para tener una idea de los datos con los que estamos trabajando. Hacer esta fase exploratoria antes de la fase de minería de datos es importante para los métodos ya sean predictivos o descriptivos de minería de datos, hay muchos de esos métodos o algoritmos que solo son aplicables cuando las variables con las que vamos a trabajar cumplen determinadas condiciones desde el punto de vista matemático.

Por ejemplo: hay algoritmos o métodos de minería de datos que nos dicen que es necesario que las variables con las que vamos a trabajar cumplan con el requisito de **normalidad**, es decir que los datos tengan una **distribución normal**. También se puede requerir que se cumplan con la simetría es decir que los datos tengan una distribución simétrica.

Que tenga una distribución normal significa que los valores de esos datos se ajustan a una campana de gauss, es decir que vamos a tener muchos datos muy parecidos entre sí. También tendremos un rango de datos donde va a estar la mayor cantidad de datos de esa variable(moda), con pocos datos que son inferiores a la mayoría y otros pocos datos que son superiores a la mayoría(Esto también se conoce como la simetría de los datos , una distribución no siempre es simétrica).

En la Figura 1 podemos observar de forma gráficamente cómo sería así.

Figura [1] Distribución normal.

Esto se aplica por ejemplo si tuviéramos que aplicar un método de minería de datos que tiene como prerrequisito que determinadas variables cumplan con el test de normalidad, entonces tenemos que hacer previamente el test de normalidad y asegurarnos que esas variables lo cumplan antes de hacer el proceso de minería. A esto se le llama la **fase exploratoria o exploración.**

CAPÍTULO 1: Desarrollo

1.1 Exploración en el proceso de extracción del conocimiento

Una vez los datos están recopilados, integrados y limpios, todavía no estamos listos (en muchos casos) para realizar una tarea de minería de datos. Luego de esto tendremos que realizar un reconocimiento o análisis exploratorio de los datos con el objetivo de conocerlos mejor de cara a la tarea de minería de datos.

Después de la fase de selección, el proceso de extracción del conocimiento contempla la fase de exploración. Dado que los datos provienen de diferentes fuentes, es necesaria su **exploración** mediante técnicas formales de análisis exploratorio de datos, buscando entre otras cosas la distribución de los datos, su simetría , normalidad y las correlaciones existentes en la información (que es de lo que se habla en la introducción).

Antes de aplicar cualquier técnica de minería de datos o incluso de análisis multivariante en general, es preciso realizar un análisis previo de los datos de que se dispone. Es necesario examinar las variables individuales y las relaciones entre ellas, así como evaluar y solucionar problemas en el diseño de la investigación y en la recogida de datos. La primera tarea que se suele abordar es el *análisis exploratorio y gráfico de los datos*. La mayoría del *software* estadístico dispone de herramientas que aportan técnicas gráficas preparadas para el examen de los datos que se ven mejoradas con medidas estadísticas más detalladas para su descripción. Estas técnicas permiten el examen de las características de la distribución de las variables implicadas en el análisis, las relación bivariantes (y multivariantes) entre ellas y el análisis de las diferencias entre grupos. Hay que tener presente que las representaciones gráficas nunca sustituyen a las medidas de diagnóstico forma estadístico, pero proporcionan una forma alternativa de desarrollar una perspectiva de carácter de los datos y las interrelaciones que existen. Incluso si son multivariantes.

Se busca conocer la **distribución** de los datos, su **simetría** y **normalidad** y las **correlaciones** existentes en la información.

Se utiliza:

- **Análisis exploratorio y gráfico** de los datos.
- Medidas de **diagnóstico formal** estadístico.

- EJ: contrastes de ajustes de los datos a una distribución, contrastes de asimetría, contrastes de aleatoriedad, etc.

Estos supuestos dependen de la técnica particular que se aplique y suelen ser el contraste de la normalidad de todas y cada una de las variables que forman parte del estudio, el testeo de la linealidad de las relaciones entre las variables que intervienen en el estudio (la relación entre la posible variable dependiente y las variables independientes que la explican ha de ser una ecuación lineal), la comprobación de la homocedasticidad de los datos que consiste en ver que la variación de la variable dependiente que se intenta explicar a través de las variables independientes no se concentra en un pequeño grupo de valores independientes (se tratará por tanto de ver la igualdad de varianzas para los datos agrupados según valores similares de la variable dependiente) y la comprobación de la multicolinealidad o existencia de relaciones entre las variables independientes. A veces también es necesario contrastar la ausencia de correlación serial de los residuos o auto correlación, que consiste en asegurar que cualquiera de los errores de predicción no está correlacionado con el resto.

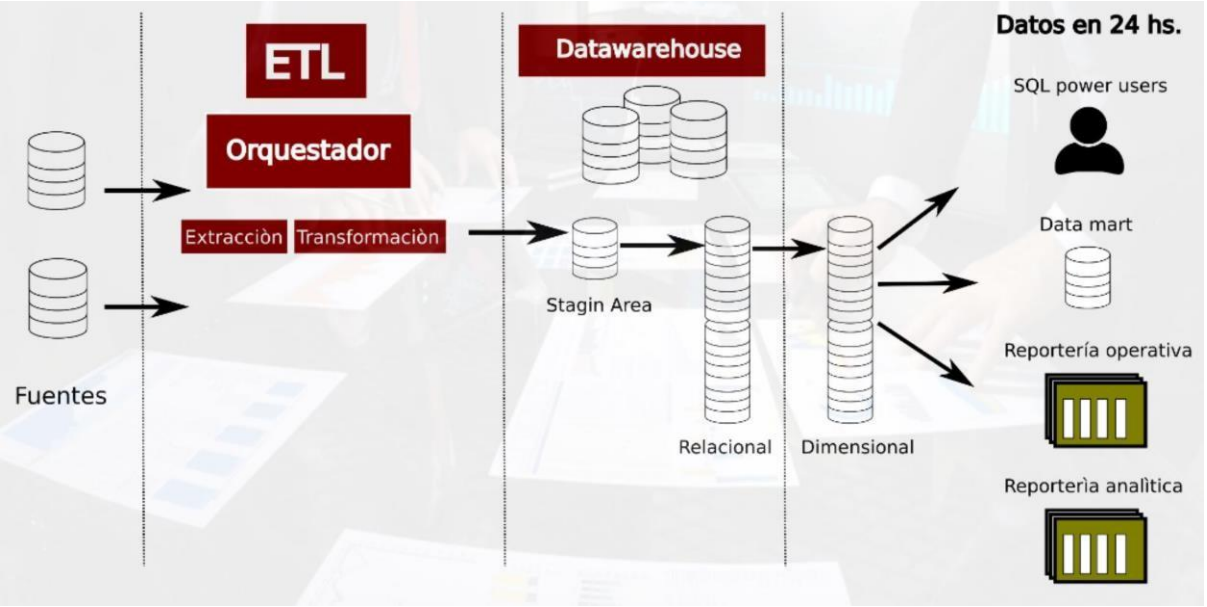


Figura [2]. Arquitectura dw para bi

Fase de selección en una arquitectura típica datawarehouse en bi. En dicha estructura el análisis exploratorio se puede realizar sobre la base de datos relacional (dw) y/o dimensional (dmt).

Proceso de extraccion del conocimiento

CAPÍTULO 2: Análisis exploratorio

Las técnicas del análisis exploratorio de datos permiten analizar la información exhaustivamente y detectar las posibles anomalías que presentan las observaciones. J. W. Tuckey ha sido uno de los pioneros en la introducción de este tipo de análisis. Los estadísticos descriptivos más habitualmente utilizados han sido la media y la desviación típica. Sin embargo, el uso automático de estos índices no es muy aconsejable. La media y la desviación típica son índices convenientes sólo cuando la distribución de datos es aproximadamente normal o, al menos, simétrica y unimodal. Pero las variables objeto de estudio no siempre cumplen estos requisitos. Por lo tanto es necesario un examen a fondo de la estructura de los datos.

Se recomienda iniciar un análisis exploratorio de datos con gráficos que permitan visualizar su estructura. Estamos ante las herramientas de exploración visual. Sin embargo, para la exploración formal, el uso de estadísticos robustos (o resistentes) es muy aconsejable cuando los datos no se ajustan a una distribución normal. Estos estadísticos son los que se ven poco afectados por valores atípicos. Suelen estar basados en la mediana y en los cuartiles y son de fácil cálculo. Fruto del análisis exploratorio, a veces es necesario realizar transformación de variables.

2.1 El contexto de la vista minable

Imagínese que le cae del cielo una base o almacén de datos con una nota: “extraiga usted conocimiento de aquí”. Usted se preguntará, entre otras cosas, lo siguiente:

- ¿Qué parte de los datos es pertinente analizar?
- ¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?
- ¿Qué conocimiento puede ser válido, novedoso e interesante?
- ¿Qué conocimiento previo me hace falta para realizar esta tarea?

No será capaz de extraer conocimiento si no se les responde a dichas preguntas. Del mismo modo, una herramienta de minería de datos, no puede digerir un conjunto de datos y

producir algo razonable, si no se le *orienta*. La razón fundamental del porqué esto es así radica no sólo en la incapacidad actual de las herramientas de realizar algunas tareas de una manera completamente automática, sino, fundamentalmente, en que la extracción de conocimiento viene a cubrir unas necesidades y expectativas, que deben indicarse, en cierto modo, de forma interactiva.Básicamente tratamos de establecer específicamente cuales son los parámetros de búsqueda.

Por tanto, es necesario expresar y proporcionar las respuestas a las cuatro preguntas anteriores, ya sea mediante lenguajes de minería de datos, ya sea interactivamente en herramientas especializadas o seleccionando aquellas herramientas necesarias.

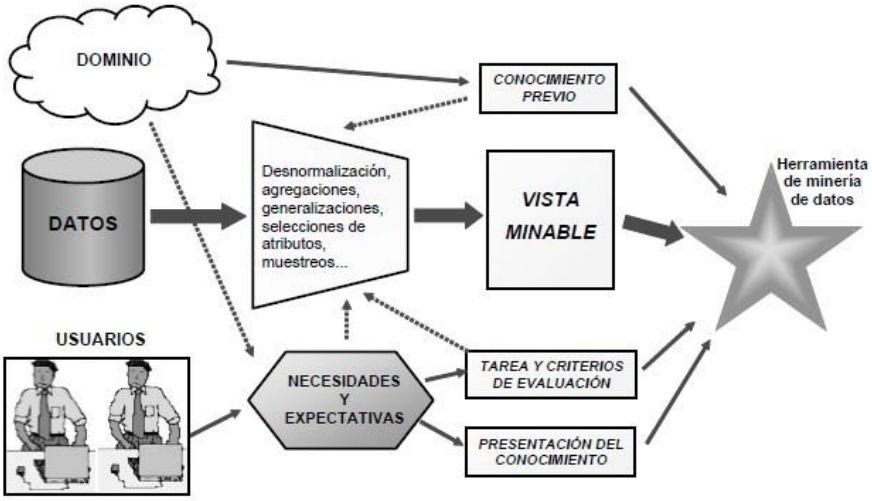


Figura [3]. De los datos, dominio y usuarios a la vista minable y elementos asociados.

- **Vista minable:** ¿qué parte de los datos es pertinente analizar? Una vista minable [Ng et al. 1998] consiste en una vista en el sentido más clásico de base de datos: una tabla. La mayoría de métodos de minería de datos, como veremos, son sólo capaces de tratar una tabla en cada tarea. Por tanto, la vista minable ha de recoger toda (y sólo) la información necesaria para realizar la tarea de minería de datos.

CAPÍTULO 3: Herramientas de exploración visual

En el siguiente paso se examina la posible presencia de normalidad, simetría y valores atípicos (outliers) en el conjunto de datos. Para ello suelen utilizarse los gráficos de caja y bigote. No obstante los gráficos de caja siempre deben ir acompañados de los histogramas digitales (o gráficos de tallo y hojas), ya que los primeros no detectan la presencia de distribuciones multimodales

3.1 Histogramas de frecuencias

De todas formas, siempre es conveniente iniciar el análisis exploratorio de datos con la construcción del histograma de frecuencias asociado, para poder así influir la distribución de probabilidad de los datos, su normalidad, su simetría y otras propiedades interesantes en el análisis de datos.

Como ejemplo considerar la variable *X* definida como el consumo de combustible en litros a los 1000 kilómetros de los automóviles de una determinada marca. Los valores para *X* son los siguientes.

43,1	36,1	32,8	39,4	36,1	19,9	19,4	20,2	19,2	20,5	20,2	25,1	20,5	19,4	20,6
20,8	18,6	18,1	19,2	17,7	18,1	17,5	30	27,5	27,2	30,9	21,1	23,2	23,8	23,9
20,3	17	21,6	16,2	31,5	29,5	21,5	19,8	22,3	20,2	20,6	17	17,6	16,5	18,2
16,9	15,5	19,2	18,5	31,9	34,1	35,7	27,4	25,4	23	27,2	23,9	34,2	34,5	31,8
37,3	28,4	28,8	26,8	33,5	41,5	38,1	32,1	37,2	28	26,4	24,3	19,1	34,3	29,8
31,3	37	32,2	46,6	27,9	40,8	44,3	43,4	36,4	30,4	44,6	40,9	33,8	29,8	32,7
23,7	35	23,6	32,4	27,2	26,6	25,8	23,5	30	39,1	39	35,1	32,3	37	37,7
34,1	34,7	34,4	29,9	33	34,5	33,7	32,4	32,9	31,6	28,1	30,7	25,4	24,2	22,4
26,6	20,2	17,6	28	27	34	31	29	27	24	23	36	37	31	38
36	36	36	34	38	32	38	25	38	26	22	32	36	27	27
44	32	28	31											

Para explorar esta información elaboramos la tabla de frecuencias asociada a los datos y estudiamos la posible normalidad y simetría de la distribución del consumo de combustible.

Como se trata de una variable cualitativa con 154 valores comprendidos entre 13 y 49, será necesario agruparlos en intervalos o cases. Para ello tomamos 12 intervalos de igual anchura (12 es un entero que aproxima a la raíz cuadrada de N = 154). La anchura de los intervalos será (49 - 13)/12 = 3. Se obtiene a tabla de frecuencias de la figura.

Intervalo	Límite inferior	Límite superior	Marca de clase	n_i	$f_i = n_i/N$	N_i	$F_i = n_i/N$
1	13,0	16,0	14,5	1	0,0065	1	0,0065
2	16,0	19,0	17,5	14	0,0909	15	0,0974
3	19,0	22,0	20,5	22	0,1429	37	0,2403
4	22,0	25,0	23,5	15	0,0974	52	0,3377
5	25,0	28,0	26,5	22	0,1429	74	0,4805
6	28,0	31,0	29,5	16	0,1039	90	0,5844
7	31,0	34,0	32,5	22	0,1429	112	0,7273
8	34,0	37,0	35,5	22	0,1429	134	0,8701
9	37,0	40,0	38,5	11	0,0714	145	0,9416
10	40,0	43,0	41,5	3	0,0195	148	0,9610
11	43,0	46,0	44,5	5	0,0325	153	0,9935
12	46,0	49,0	47,5	1	0,0065	154	1,0000

Tabla [1] Tabla de frecuencias.

Hemos observado los 154 valores sobre el consumo de los automóviles que inicialmente no aportan mucha información. Evidentemente hay una variabilidad en el consumo de los automóviles; sin embargo, es muy difícil detectar qué patrón sigue dicha variabilidad para determinar mejor la estructura de los datos. Por ello, en primer lugar, ha sido conveniente realizar una ordenación de los datos según su magnitud, es decir, una tabla de frecuencias, que aportará algo de luz sobre la distribución de frecuencias subyacentes. La siguiente es la construcción del histograma de frecuencias, gráfico adecuado para una variable cuantitativa con sus valores agrupados en intervalos. Su representación se presenta en la Figura [4].

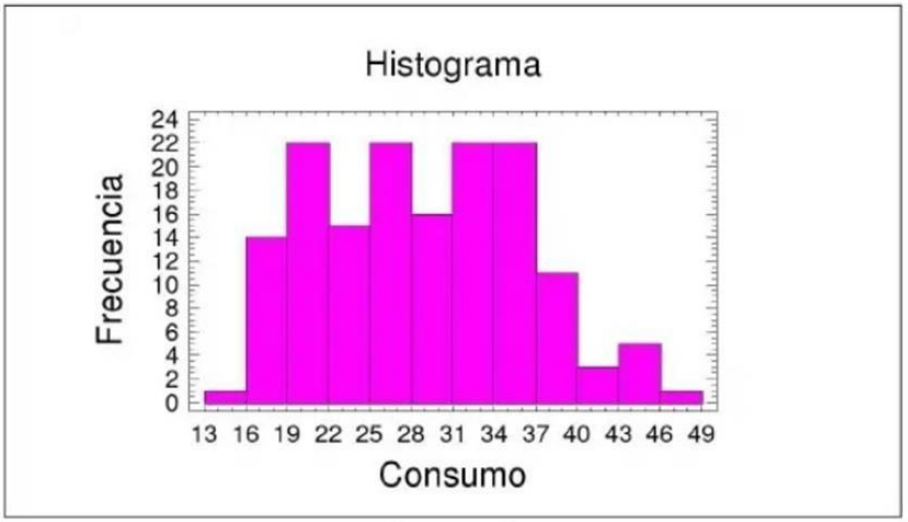


Figura [4] Histograma de frecuencias.

Se observa que la distribución subyacente que modela los datos sobre la variable consumo de los automóviles es aproximadamente simétrica y ajustable a forma de campana, lo que permite pensar en la existencia de normalidad y simetría en la distribución de *X*.

Vemos así que el histograma da una idea clara de la distribución de la variable, incluyendo un modelo probabilístico para su modelación, en este caso la distribución normal. El simple examen de los datos tabulados inicialmente no aportaba información alguna, sin embargo su graficación da luz al proceso.

3.2 Diagramas de tallo y hojas

El diagrama de tallo y hojas es un procedimiento semigrafico para presentar la información para variables cuantitativas para presentar la información para variables cuantitativas que es especialmente útil cuando el número total de datos es pequeño (menor que 50). Los principios para la realización del diagrama (debido a Tukey) son los siguientes:

- Redondear los datos a dos o tres cifras significativas.

- Disponerlos en dos columnas separadas por una línea vertical de tal forma que para los datos con dos dígitos la cifra de las decenas se encuentre a la izquierda de la línea vertical (tallo del diagrama). Por ejemplo, 87 se escribirá 8 7. Para datos con tres dígitos el tallo estará formado por los dígitos de las centenas y las decenas que se escribirán a la izquierda de la línea vertical, y las hojas están formadas por el dígito de las unidades, que se escribirá a la derecha de la línea vertical.

- Cada tallo define una clase, y se escribe solo una vez. A su derecha se van escribiendo por orden las sucesivas hojas correspondientes a ese tallo. El número de hojas para tallo representa la frecuencia de cada clase.

El diagrama de tallo y hojas también llamado *histograma digital*, es una combinación entre un histograma de barras y una tabla de frecuencias. Al mantener los valores de la variable, el diagrama de tallo y hojas resulta más informativo que el clásico histograma de barras, ya que conserva los datos originales y, al mismo tiempo, compone un perfil que ayuda a estudiar la forma y simetría de la distribución. Se trata pues de una herramienta de análisis exploratorio de datos que muestra el rango de los datos, donde están más concentrados, su simetría yla presencia de datos atípicos. Este procedimiento no es muy aconsejable para conjuntos muy grandes.

A continuación se presenta el diagrama de tallo y hojas (Figura [4]) para la variable *X* relativa a la variable consumo de los automóviles definida en el apartado anterior.

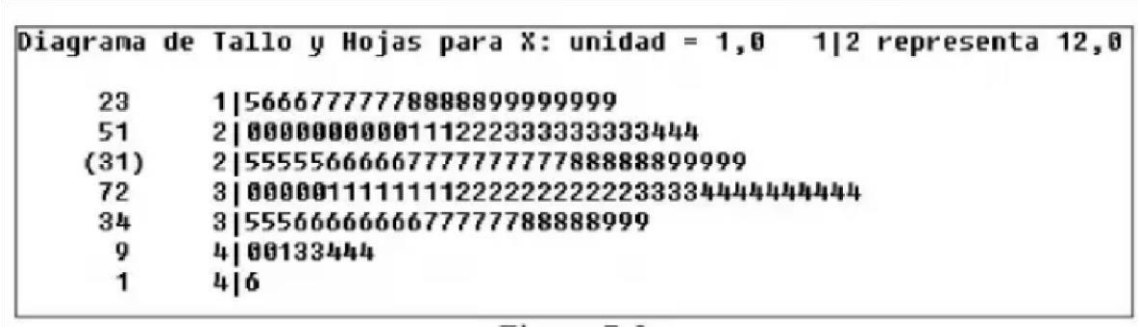


Figura [5] Diagrama de tallo y hojas.

El rango de *X* ha sido dividido en 7 clases o intervalos llamados tallos, cada uno de ellos representado por una fila del diagrama. El primer número de cada fila (separado de los demás) presenta la frecuencia absoluta de la clase correspondiente. El segundo número de cada fila presenta la cifra de las decenas de cada valor de *X* en su correspondiente clase. El resto de los números de cada fila (llamados *hojas*) son las cifras de las unidades de todos los elementos de la clase definida por la fila. De esta forma, además de presentar la distribución de los en forma de histograma horizontal, en el diagrama se observan los propios elementos. Las hojas permiten analizar la simetría, la normalidad y otras características de la distribución de igual forma que un histograma.

3.3 Gráfico de cajas y bigotes

El gráfico de cajas y bigotes permite analizar y resumir un conjunto de datos univariante dado. Esta herramienta de análisis exploratorio de datos va a permitir estudiar la simetría de los datos, detectar valores típicos y vislumbrar un ajuste de los datos a una distribución de frecuencias determinada.

El gráfico de cajas y bigotes divide los datos en cuatro áreas de igual frecuencia, una caja central dividida en dos áreas por una línea vertical y otras dos áreas representadas por dos segmentos horizontales (bigotes) que parten del centro de cada lado vertical de a caja, la caja central encierra el 50 por ciento de los datos. El sistema dibuja a mediana como una línea vertical en el interior de la caja. Si esta línea en el centro de la caja está situada en los cuartiles inferior y superior de la variable. Partiendo del centro de cada lado vertical de la caja se dibujan los dos bigotes, uno hacia la izquierda y el otro hacia la derecha. El bigote de la izquierda tiene un extremo en el primer cuartil Q1, y el otro en el valor dado por el primer cuartil menos de 0,5 veces el rango intercuartílico, esto es, Q1 – 1,5 * (Q3 - Q1).

El bigote de la derecha tiene un extremo en el tercer cuartil Q3 y el otro en el valor dado por el tercer cuartil más 1,5 veces el rango intercuartílico, esto es, Q3 + 1,5 * (Q3 – Q1). El sistema considera valores atípicos (outliers) los que se encuentren a la izquierda del bigote izquierdo y a la derecha del bigote derecho. El sistema separa estos datos del resto y los representa mediante puntos alineados con la línea horizontal central para que sean fáciles de detectar. En el interior de la caja se representa la media con un signo más.

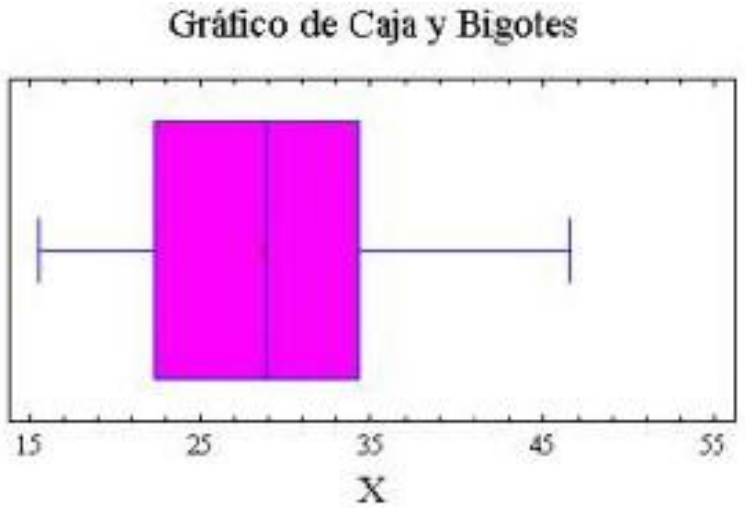


Figura [6] Gráfico de cajas y bigotes 1.

El gráfico permite afirmar que la variable *X* (consumo de los automóviles cada 1000 kilómetros) varía entre 15,5 y 46,6 y que el 50% central de los coches consume entre 22 (primer cuartil) y 34,5 (tercer cuartil) litros a los 1000 kilómetros.

3.4 Gráfico múltiple de caja y bigotes

En estadística es típico dividir el conjunto de datos de una variable en subgrupos racionales, que pueden ser por ejemplo estratos definidos según una determinada variable de estratificación. El gráfico múltiple de caja y bigotes va a permitir analizar, resumir y comparar simultáneamente varios conjuntos de datos univariantes dados, correspondientes a los diferentes grupos en que se pueden subdividir los valores de una variable. Esta herramienta de análisis exploratorio de datos va a permitir estudiar la simetría de los datos, detectar valores atípicos y representar medias, medianas, rangos y valores extremos para todos los grupos. Al ser la representación simultánea para todos los conjuntos de datos, se podrán comparar medias, medianas, rangos, valores extremos, simetrías y valores atípicos de todos los grupos. El gráfico múltiple representará horizontalmente un gráfico de caja y bigotes para cada grupo de valores de la variable en estudio.

Si clasificamos el consumo de los automóviles (variable X) según su cilindrada (variable Y), estamos haciendo subgrupos con los valores de X según la variable de estratificación Y.

Los posibles valores de Y son 8, 6, 5, 4 y 3 cilindros. Los valores de X para cada valor de Y vienen dados a continuación:

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
43,1	4	36,1	4	32,8	4	39,4	4	36,1	4	19,9	8	19,4	8	20,2	8	19,2	6	20,5	6
20,2	6	25,1	4	20,5	6	19,4	6	20,6	6	20,8	6	18,6	6	18,1	6	19,2	8	17,7	6
18,1	8	17,5	8	30	4	27,5	4	27,2	4	30,9	4	21,1	4	23,2	4	23,8	4	23,9	4
20,3	5	17	6	21,6	4	16,2	6	31,5	4	29,5	4	21,5	6	19,8	6	22,3	4	20,2	6
20,6	6	17	8	17,6	8	16,5	8	18,2	8	16,9	8	15,5	8	19,2	8	18,5	8	31,9	4
34,1	4	35,7	4	27,4	4	25,4	5	23	8	27,2	4	23,9	8	34,2	4	34,5	4	31,8	4
37,3	4	28,4	4	28,8	6	26,8	6	33,5	4	41,5	4	38,1	4	32,1	4	37,2	4	28	4
26,4	4	24,3	4	19,1	6	34,3	4	29,8	4	31,3	4	37	4	32,2	4	46,6	4	27,9	4
40,8	4	44,3	4	43,4	4	36,4	5	30,4	4	44,6	4	40,9	4	33,8	4	29,8	4	32,7	6
23,7	3	35	4	23,6	4	32,4	4	27,2	4	26,6	4	25,8	4	23,5	6	30	4	39,1	4
39	4	35,1	4	32,3	4	37	4	37,7	4	34,1	4	34,7	4	34,4	4	29,9	4	33	4
34,5	4	33,7	4	32,4	4	32,9	4	31,6	4	28,1	4	30,7	6	25,4	6	24,2	6	22,4	6
26,6	8	20,2	6	17,6	6	28	4	27	4	34	4	31	4	29	4	27	4	24	4
23	4	36	4	37	4	31	4	38	4	36	4	36	4	36	4	34	4	38	4
32	4	38	4	25	6	38	6	26	4	22	6	32	4	36	4	27	4	27	4
44	4	32	4	28	4	31	4												

Tabla [2] Consumo de los automóviles según su cilindrada**.**

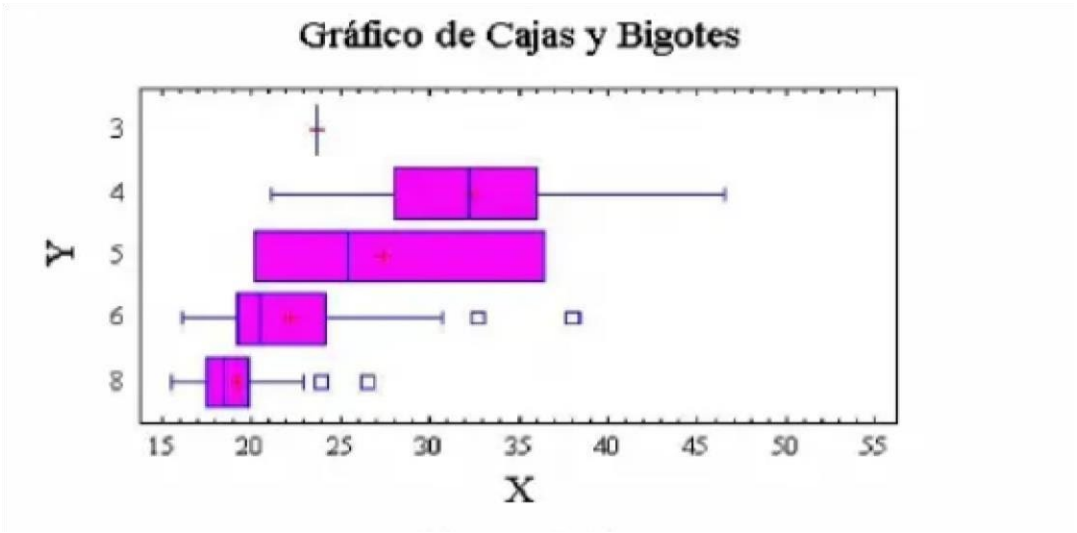


Figura [7] Gráfico múltiple de cajas y bigotes de consumo de automóviles (x) según su cilindrada (y)

El Gráfico permite afirmar que la variable X' (litros consumidos a los 1000 kilómetros) para los coches de 8 cilindros varía entre 15,5 y 23, y que el 50% central de estos coches consume entre 17,5 (primer cuartil) y 20 (tercer cuartil) litros a los 1000 kilómetros, existiendo 2 valores de X anormalmente grandes (outliers), ya que en la Figura aparecen dos puntos alineados con el bigote de la parte derecha. La distribución de X para los coches de 8 cilindros es ligeramente asimétrica hacia la derecha, ya que la zona de la derecha en el área central de la Figura es mayor que la de la izquierda, y la mediana corresponde aproximadamente al valor 18,5 de X, siendo la media 19,5 aproximadamente.

Para los coches de 6 cilindros los litros consumidos cada 1000 kilómetros

(variable X) varían entre 16 y 31, concentrándose el 50% central de los valores de X entre 19 (primer cuartil) y 24 (tercer cuartil), existiendo 2 valores atípicos de X anormalmente grandes (outliers), ya que en la Figura aparecen dos puntos alineados con el bigote de la parte derecha. La distribución de X para los coches de 6 cilindros es asimétrica hacia la derecha, la mediana de X' se aproxima a 20,5 y la media a 22,5.

Para los coches de 5 cilindros los litros consumidos cada 1000 kilómetros

(variable X) varían entre 20,2 y 36,5, concentrándose el 50% central de los valores de X entre los mismos valores, no existiendo bigotes ni outliers. La distribución de X para los coches de 5 cilindros es asimétrica hacia la derecha, la mediana de X se aproxima a 25,5 y la media a 27,5.

Para los coches de 4 cilindros los litros consumidos cada 1000 kilómetros

(variable X) varían entre 21 y 47, concentrándose el 50% central de los valores de X entre 28 (primer cuartil) y 36 (tercer cuartil), no existiendo outliers. La distribución de X para los coches de 4 cilindros es prácticamente simétrica con valores de mediana y media aproximados a 32. Para los coches de 3 cilindros hay un único valor de X, lo que no permite construir el gráfico de caja y bigotes.

Si comparamos los distintos gráficos, vemos que la asimetría de X es más fuerte para los coches de 5 y 6 cilindros, para los de 8 es menor y para los de 4 no existe. Valores de X anormalmente grandes sólo aparecen para los coches de 6 y 8 cilindros. Las medias y las medianas varían bastante para los diferentes grupos de valores de X determinados por los valores de Y. [1]

3.5 Gráfico de simetría

El gráfico de simetría es una herramienta que permite analizar visualmente el grado de simetría de una variable. En el eje de abscisas se representan las distancias de los valores de la variable a la mediana que quedan por debajo de ella, y en el eje de ordenadas se representan las distancias de los valores de la variable a la mediana que quedan por encima de ella. Si la simetría fuese perfecta, el conjunto de puntos resultante sería la diagonal principal. Mientras más se aproxime la gráfica a la diagonal más simetría existirá en la distribución de la variable.

Para el ejemplo de la variable *X*, variable definida por el número de litros consumidos por los automóviles cada 1000 kilómetros que venimos considerando durante todo el capítulo, tenemos el Gráfico de simetría de la Figura 7-8. [1]

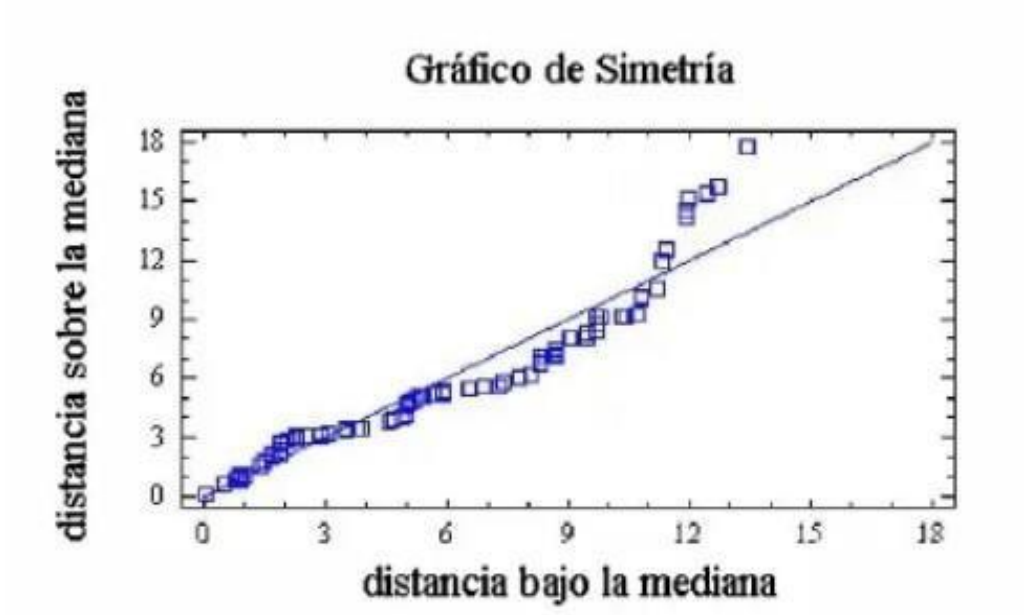


Figura [8] Gráfico de simetría.

Para la variable *X,* se observa un buen grado de simetría, ya que los puntos de la gráfica se ajustan bien a la diagonal.

Los pasos prácticos para elaborar el gráfico de simetría son los siguientes:

1. Se calcula la mediana de la variable (en nuestro caso 28,9).
2. Se ordenan los valores de la variable de mayor a menor (en orden descendente).
3. Se calcula las diferencias d_i entre los valores de la variable ordenados y la mediana.
4. Se toman los valores positivos d_i ordenados de menor a mayor y se les denomina p_i. Estos valores serán las distancias sobre la mediana.
5. Se toman los valores negativos d_i ordenados de menor a mayor y se les denomina n_i.

   Estos valores cambiados de signo serán las distancias bajo la mediana.
6. Se grafican los puntos de coordenadas (-n, p).

3.6 Gráficos para variables cualitativas

La exploración visual de variables cualitativas suele llevarse a cabo mediante diagramas de rectángulos, diagramas de sectores y pictogramas.

Los *diagramas de rectángulos* se construyen asignando a cada modalidad de la variable cualitativa un rectángulo con altura igual o proporcional a su frecuencia absoluta ni t con base constante.

Como ejemplo en la Figura [9] se presenta un diagrama de rectángulos que representa los activos según las distintas modalidades de la variable rama de actividad. Sobre cada rectángulo se presenta la frecuencia absoluta ni en miles de activos de la correspondiente rama de actividad. Sobre el eje de abscisas se presentan las propias ramas de actividad y sobre el eje de ordenadas se presentan diferentes valores de las frecuencias absolutas por intervalos que sirven como referencia para situar la altura de cada rectángulo.

Figura [9] Activos por ramas de actividad.

Los diagramas de sectores constituyen el tipo de gráfico más utilizado para representar distribuciones de frecuencias de variables cualitativas. La variable se representa en un círculo cuyas porciones (sectores circulares) tienen un área proporcional a las frecuencias absolutas de las modalidades de la variable. Para realizar el gráfico de sectores basta con asignar a cada modalidad de la variable un sector circular cuyo ángulo central sea proporcional a la frecuencia absoluta de la modalidad.

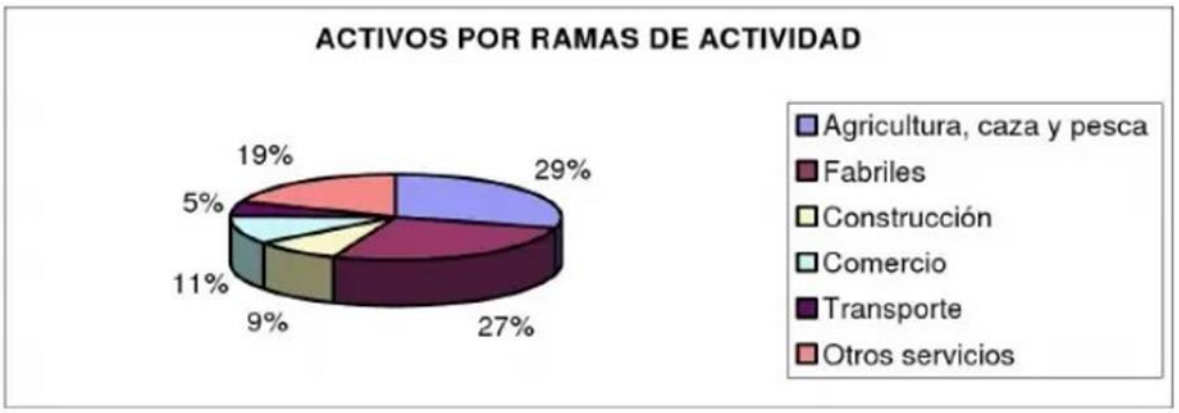
RAMA	ACTIVOS (n_i)	$f_i = n_i/N$	$\alpha_i = 360 f_i$
Agricultura, caza y pesca	3706,3	0,29	104,79
Fabriles	3437,8	0,27	97,20
Construcción	1096,3	0,09	31,00
Comercio	1388,3	0,11	39,25
Transporte	648,7	0,05	18,34
Otros servicios	2454,8	0,19	69,41
$N=$	12732,2		

Tabla [3] Activos por rama de actividad.

Figura [10] gráfico de sectores.

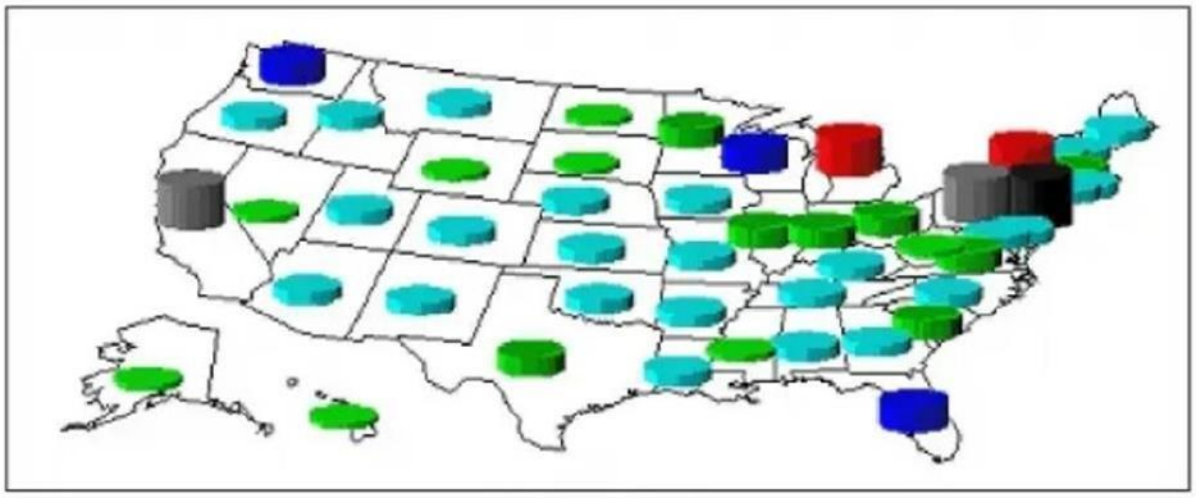
Otra forma habitual de construir gráficos de sectores consiste en asignar al sector circular relativo a la modalidad i-esima un porcentaje igual al tanto porciento que representa su frecuencia absoluta ni, sobre la frecuencia total N = Σni. Matemáticamente, la expresión del porcentaje pi relativo a la modalidad i-esima se expresa como sigue:

$$p_i = 100\frac{n_i}{N} = 100 f_i$$



Figura[11] Gráfico de sectores anterior con porcentajes.

Los pictogramas se construyen representando de manera pictórica cada modalidad de la variable cualitativa indicando por una silueta sugestiva el significado de cada unidad de carácter. Por ejemplo, en la Figura 7.15 se presenta sobre cada estado del mapa de Estados Unidos un cilindro cuya altura es proporcional a la cantidad de instalaciones con residuos peligrosos en 1997.



Figura[12] Mapa de cilindros.

En los pictogramas suele usarse cualquier silueta para representar la frecuencia de cada modalidad de la variable cualitativa. En la Figura [13] se presenta el pictograma anterior con un cono sobre cada estado cuya altura indica la frecuencia de instalaciones con residuos peligrosos.

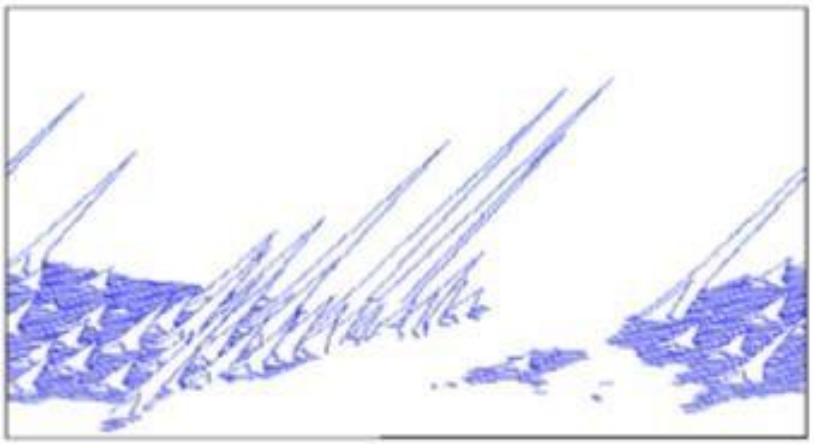


Figura [13] pictograma de frecuencias.

CAPÍTULO 4: Herramientas de exploración formal

Es necesario tener muy en cuenta que las representaciones gráficas, aunque proporcionar una forma alternativa de desarrollar una perspectiva del carácter de los datos y de las interrelaciones que existen entre ellos incluso si son multivariantes, nunca sustituyen a las medidas de diagnóstico formal estadístico como los contrastes de ajuste de los datos a una distribución, los contrastes de asimetría, los contrastes de aleatoriedad, el uso de estadísticos robustos, etc. La exploración gráfica de los datos siempre debe ir acompañada de los contrastes de exploración formal.

Algunas de las herramientas de exploración formal son las siguientes:

- Contrastes de la bondad de ajuste de una distribución.
- Contraste de Kolmogorov-Smirnov Lilliefors de la bondad de ajuste a una distribución.
- Estadísticos robustos de centralización.
- Estadísticos robustos de dispersión
- Estadísticos robustos de asimetría y curtosis.

CAPÍTULO 5: Transformaciones de las variables

Cuando el análisis exploratorio lo indique, los datos originales (no los estandarizados ni los previamente modificados) pueden necesitar ser transformados. Suelen considerarse cuatro tipos de transformaciones:

Transformaciones lógicas: se unen categorías del campo de definición de las variables para reducir su amplitud. De esta forma pueden eliminarse categorías sin respuestas. También pueden convertirse variables de intervalo en ordinales o nominales y crear variables ficticias (dummy).

Transformaciones lineales: se obtienen al sumar, restar, multiplicar o dividir las observaciones originales por una constante para mejorar su interpretación. Estas transformaciones no cambian la forma de la distribución, ni las distancias entre los valores ni el orden, y por tanto no provocan cambios considerables en las variables.

Transformaciones algebraicas: se obtienen al aplicar transformaciones no lineales monótonas a las observaciones originales (raíz cuadrada, logaritmos, etc.) por una constante para mejorar su interpretación. Estas transformaciones cambian la forma de la distribución al cambiar las distancias entre los valores, pero mantienen el orden.

Transformaciones no lineales no monotónicas: cambian las distancias y el orden entre los valores. Pueden cambiar demasiado la información original.

Con estas transformaciones se arreglan problemas en los datos. Por ejemplo: una asimetría negativa puede minorarse con una transformación parabólica o cubica, una asimetría positiva fuerte puede suavizarse mediante una transformación hiperbólica o hiperbólica cuadrática (con signo negativo) y una asimetría positiva débil puede suavizarse mediante una transformación de raíz cuadrada, logarítmica o recíproca de la raíz cuadrada (con signo negativo). La transformación logarítmica puede conseguir estacionalidad en media y en varianza para los datos. Suele elegir como transformación aquella que arregla mejor el problema, una vez realizada. Si ninguna arregla el problema, realizamos el análisis sobre los datos originales sin transformar. Combinado transformaciones lineales y algebraicas pueden modificarse los valores extremos de la distribución. [1]

CAPÍTULO 6: Supuestos subyacentes en las técnicas de minería de datos

Una etapa importante en las técnicas de minería de datos es la comprobación de los supuestos estadísticos subyacentes a las variables que intervienen en los modelos. La presencia de múltiples variables provoca complejidad que llevan a distorsiones y sesgos cuando no se cumplen determinados supuestos que se estudiarán a continuación (normalidad, homocedasticidad, linealidad, ausencia de autocorrelación o correlación serial y ausencia de multicolinealidad). [1]

Se debe realizar la **comprobación de los supuestos subyacentes** en los métodos multivariantes para la minería de datos; Estos **supuestos** suelen ser:

- El **contraste de la normalidad** de todas t c/u de las variables que forman parte del estudio. ➢ El **testeo de la linealidad** de las relaciones entre las variables ➢ La **comprobación de la homocedasticidad** de los datos:

    - Consiste en ver que la **variación de la variable dependiente** que se intenta explicar a través de las variables independientes **no se concentra** en un pequeño grupo de valores independientes.

La **comprobación de la multicolinealidad** o existencia de relaciones entre las variables independientes.

- La **contrastación de la ausencia de correlación seria de los residuos** o autocorrelación.

    - Consiste en asegurar que cualquiera de los errores de predicción no está correlacionado con el resto.

6.1 Normalidad

Tanto los métodos estadísticos univariantes como los multivariantes se basan en los supuestos de normalidad univariante y multivariante respectivamente. Todas las variables que intervienen en un método de análisis multivariante deben ser normales, aunque ello no garantiza la normalidad multivariante. El recíproco siempre es cierto, es decir, la normalidad multivariante implica la normalidad de cada variable. No obstante suele bastar con la normalidad de cada variable, aunque en procesos críticos puede extinguirse la normalidad multivariante.

Existen métodos gráficos como contrastes estadísticos formales, para comprobar la normalidad de las variables que intervienen en un método multivariante. [1]

6.2 Gráfico normal de probabilidad

Los gráficos normales de probabilidad sirven para determinar si un conjunto de datos dado se ajusta razonablemente a una distribución normal. El gráfico normal de probabilidad (Figura 717) presenta en el eje de abscisas los valores de la variable (X,, y en el eje de ordenadas las frecuencias relativas acumuladas de dichos valores (F). La normalidad de los datos será perfecta cuando el gráfico de los puntos (X, F') resulte ser una línea recta situada sobre la diagonal del primer cuadrante. Las diferencias que existan entre el gráfico de probabilidad y la línea recta marcarán la regla de decisión para aceptar o no la normalidad del conjunto de datos dado.

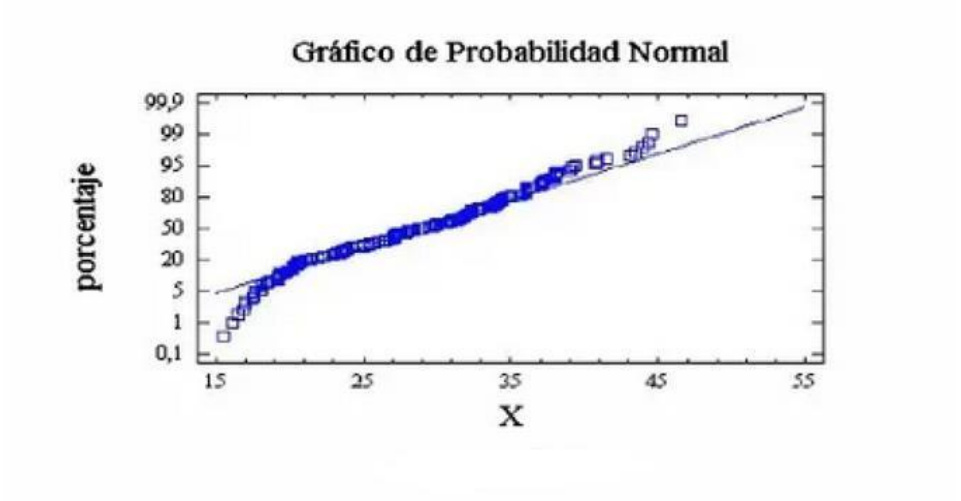


Figura [14] Gráfico de probabilidad Normal.

6.3 Heteroscedasticidad

En cualquier modelo multivariante suele suponerse que la variable u (término de error) es una variable aleatoria con esperanza nula (E (u) = 0) y matriz de covarianzas constante y diagonal (Var (u) = σ^2)

En cualquier modelo multivariante suele suponerse que la variable u (término de error) es una variable aleatoria con esperanza nula (E(u) = 0) y matriz de

covarianzas constante y diagonal (*Var* (u) = σ^2), matriz escalar). Es decir, que para todo 1, la variable u, tiene media cero y varianza o? no dependiente de 1, y además Cov(us11) = 0 para todo ¿ y para todo j distintos entre sí, pudiendo escribir Var(u)= 0" 1.

El hecho de que la varianza de u, sea constante para todo £ (que no dependa de 1), se denomina hipótesis de homocedasticidad. Si se relaja esta hipótesis y la varianza de 4, no es constante estamos ante la presencia de heteroscedasticidad. La importancia del incumplimiento de la hipótesis de homocedasticidad radica, entre otras cosas, en que los estimadores obtenidos por MCO no son de varianza mínima aunque sigan siendo insesgados. Además, para cada variable del modelo se estimará una varianza del error.

Para analizar la heteroscedasticidad de un modelo suele comenzarse por el análisis gráfico de los residuos, siendo esenciales las gráficas de los residuos (a poder ser estudentizados) respecto de la variable endógena y respecto de las exógenas, que deben de

presentar una estructura aleatoria libre de tendencia. El gráfico de los residuos contra cada variable exógena permite detectar como variable más culpable de heteroscedasticidad aquella cuyo gráfico se separa más de la aleatoriedad. También es un instrumento gráfico útil la gráfica de valores observados contra valores predichos, cuyos puntos han de ser lo más ajustados posible a la diagonal del primer cuadrante. para resolver problemas de heteroscedasticidad es conveniente tomar logaritmos.

6.4 Multicolinealidad

En un modelo multivariante suele suponerse como hipótesis que sus variables (sobre todo las variables exógenas) Xi, X;,...X; son linealmente independientes, es decir,

no existe relación lineal exacta entre ellas. Esta hipótesis se denomina hipótesis de ja, y cuando le, decimos que el modelo presenta multi

La matriz de correlaciones es un instrumento que ayuda a detectar la presencia de multicolinealidad. Valores altos en esta matriz son síntoma de posible dependencia entre las variables implicadas.

Entre las soluciones más comunes para la multicolinealidad se tiene: ampliar la muestra, transformar las variables adecuadamente, suprimir algunas variables con de las variables por sus componentes principales más significativas (puntuaciones) o utilizar métodos específicos de ajuste como la regresión en cadena.

6.5 Autocorrelación

En cualquier modelo multivariante suele suponerse que la variable u (término de error) es una variable aleatoria con esperanza nula (E(u)=0) y matriz de covarianzas constante y diagonal (Var(u)= σ^2 Ik matriz escalar . Es decir, que para todo 1, la variable 1 tiene media cero y varianza O? no dependiente de £, y además Cov(usu)=0 para todo ¡ y para todo j distintos entre sí, pudiendo escribir Var(u)= 0? I..

El hecho de que Cov(u,u,) = 0 para todo ¡ distinto de j se denomina hipótesis de no autocorrelación. En este apartado estudiaremos el modelo lineal cuando esta hipótesis no se cumple, es decir, cuando existe autocorrelación o correlación serial.

Por tanto, en presencia de autocorrelación será necesario estimar los elementos de la matriz de varianzas covarianzas residual V. Esta tarea suele simplificarse suponiendo que las perturbaciones aleatorias del modelo siguen un determinado esquema de comportamiento que reduce el número de parámetros a estimar. Los esquemas más típicos son:

Modelo autorregresivo de orden 1 AR(1) $\rightarrow u_t = \rho u_{t-1} + e_t$
Modelo autorregresivo de orden 2 AR(2) $\rightarrow u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t$
Modelo de medias móviles de orden 1 MA(1) $\rightarrow u_t = e_t + \rho e_{t-1}$

6.6 Linealidad

La linealidad es un supuesto implícito en todas las técnicas multivariantes basadas en medidas de correlación (regresión múltiple, regresión logística, análisis factorial, etc.). Como los efectos no lineales nunca están representados en el valor de la correlación, su presencia tendría efectos nocivos en el modelo multivariante.

La no linealidad se resuelve tomando como modelo multivariante el modelo no lineal que se detecte que ajusta mejor las variables en estudio. El análisis gráfico permite detectar qué tipo de no linealidad puede estar presente en nuestros datos.

Los gráficos de dispersión de las variables con secuencias no lineales y los gráficos residuales con falta de aleatoriedad permiten detectar la falta de linealidad, simplemente observando su forma. Si aparecen secuencias no lineales de puntos en los gráficos de dispersión, tendremos problemas de falta de linealidad. Lo mismo ocurre si aparecen secuencias no aleatorias en los gráficos residuales

Conclusión

La fase de selección comprende la extracción e ingesta con información extraída generalmente desde bases de datos transaccionales y luego su transformación y carga (ETL), luego esta información será volcada a bases de datos especialmente dispuestas para ese fin organizadas en bases de datos de stagin , datawarehouse y/o datamarts (según la variante utilizada) , esta última se pondrá a disposición del usuario utilizando alguna interfaz en forma de planillas de cálculo o herramientas diseñadas para tal fin como por ejemplo power bi. Estos datos ya extraídos y organizados en bases de datos, así dispuestos generalmente son utilizados para devolver respuestas a preguntas de negocio del usuario utilizando alguno de los métodos para analizar y evaluar los datos, por ej. OLAP. Pero también pueden ser utilizados para descubrir patrones ocultos en los datos que no pueden ser vistos de forma simple. Para lograr sus objetivos, en la fase de exploración de la minería de datos se trata de conocer los datos con los que estamos trabajando, para ello se vale de algoritmos matemáticos con los cuales se realiza un análisis para determinar distintos parámetros tales como su distribución, simetría y normalidad, entonces se realiza un análisis exploratorio de los datos mediante herramientas de exploración visual o de exploración formal, ambas técnicas son complementarias entre sí. Al realizar distintos modelos estadísticos y tener muchas variables, es probable obtener cierta distorsión y sesgo en los datos analizados. Entonces es necesario realizar una comprobación de supuestos subyacentes para: contrastar la normalidad de los datos, testear la linealidad de las variables y comprobar la homocedasticidad de los datos. La fase de exploración es previo a la realización de la minería de datos propiamente dicha y es importante para asegurarnos que los datos cumples distintas condiciones desde el punto de vista matemático por ejemplo si un método de minería de datos nos pide que determinadas variables cumplan con el test de normalidad, entonces previamente realizamos dicho test para asegurar que las variables lo cumplan.

Bibliografía

[1] Pérez López, César – Santín González, Daniel, «Capítulo 7: Fase de exploración en

Minería de Datos», en *"Minería de datos. Técnicas y herramientas"*, 1° Edición – 2° Reimpresión 2008. Madrid, Editorial Thomson: pp. 497-525

[2] Introducción a la minería de datos: José Hernández Orallo, María José Ramírez Quintana,

César Ferri Ramírez – Editor Pearson Educación, 2004 – ISBN 8420540919

[3] **Material Teórico de clase**: Mineria_de_Datos_Introduccion.pdf