Descriptiva

Teresa Villagarcía

1 Estadística y Calidad

El interés por la mejora continua de la Calidad y la productividad ha generado una demanda importante de métodos estadísticos que permitan obtener información sobre procesos, productos o servicios. Cada vez es más fácil tener acceso a enormes cantidades de información, pero si no tenemos técnicas que permitan procesar estos datos y extraer de ellos las ideas y conclusiones importantes, no nos sirven de nada.

Cuando se pretende mejorar cualquier aspecto de la gestión de una empresa: producción, venta, gestión..., es preciso conocer bien ese aspecto, estudiarlo y detectar los posibles puntos de mejora. En de…nitiva, es preciso saber gestionar bien y e…cientemente la información. La estadística nos proporciona numerosos métodos para resumir información y produce un tipo de análisis de datos que ha demostrado su fortaleza en todas las especialidades cientí…cas y de gestión.

Las técnicas estadísticas de análisis de datos actuales, son muy intensivas en el uso de ordenadores, y utilizan métodos numéricos y grá…cos. Por ello para poder aplicar estas técnicas y ser operativo con ellas es preciso estudiar la estadística con un fuerte apoyo informático. Actualmente existen en el mercado numerosos programas estadísticos que pueden ser utilizados por usuarios con distintos grados de preparación.. La utilización de un paquete u otro es indiferente, ya que lo importante es saber qué técnica emplear. Una vez decida la técnica que se va a utilizar, la mayoría del software disponible puede darnos resultados satisfactorios.

2 Estadística Descriptiva.

El objetivo de la estadística descriptiva es extraer la información que contiene un conjunto de datos. Para lograr esto, es preciso resumir la información y la estadística es la técnica más e…ciente de lograr resumir la información.

Distinguiremos dos formas básicas de extracción de la información: Analítica (Es decir utilizando valores numéricos) y Grá…ca. Ambos procedimientos son complementarios, y la utilización conjunta de ellos permitirá lograr altas cotas de e…ciencia.

3 Tipos de Datos.

El análisis que se aplique a un conjunto de datos dependerá en gran medida del tipo de datos (Variables) que se quiera analizar. Distinguiremos varios tipos de datos (Variables):

- 1. Datos cualitativos
- 2. Datos cuantitativos
  - 2.1 Datos Transversales
  - 2.2 Datos temporales.

1. Datos cualitativos:

Son datos cualitativos aquellos que recogen alguna característica no numérica. Ejemplos de variables cualitativas son: el sexo de un individuo, su provincia o nacionalidad de origen, su estado civil. Si se están estudiando hoteles de una determinada cadena ([1]), una variable cualitativa puede ser su situación, que se clasi…caría en Céntrico o Extrarradio.

2. Datos cuantitativos:

Son datos que se representan de una forma natural con números. Por ejemplo Altura de una persona, Peso, Ingresos. Si se tratase de hoteles, podríamos pensar en Número de Habitaciones, Número de trabajadores o Valoración Global que le dan los clientes.

Dentro de los datos cuantitativos distinguimos a su vez datos transversales y datos temporales.

2.1 Datos Transversales:

[1] Vamos a utilizar el ejemplo de la Calidad en una cadena de hoteles durante todo el texto.

Datos transversales son aquellos que se obtienen de muchos individuos en un determinado instante de tiempo. Ejemplos de este tipo de datos son la altura de 200 personas, o el número de trabajadores de 45 hoteles.

2.2 Datos temporales:

Se denomina serie temporal a la sucesión de observaciones de una variable a lo largo del tiempo. Ejemplos de serie temporal son la evolución de la in‡ación en España desde 1980, las temperaturas medias mensuales en Madrid, el Número de clientes mensuales en un hotel desde Junio de 1992 hasta Diciembre de 1988.

> Las técnicas que se aplican al estudio de los datos, van a variar en función del tipo de datos que tengamos.

4 Técnicas Grá…cas

La forma más rápida y e…ciente de captar información en los datos es mediante diversos grá…cos que tienen por objetivo destacar las estructuras internas que pudieran tener los datos. Vamos a estudiar tres técnicas grá…cas de gran utilidad. El histograma de frecuencias, el Diagrama de Tallo y Hojas y, …nalmente, el Diagrama de Caja.

4.1 Histograma

El histograma proporciona información sobre la frecuencia con que se obtienen observaciones de cada valor. Su interpretación es simple y es sencillo de realizar. Vamos a estudiarlo con un ejemplo.

Los datos que se presentan son las alturas de 117 estudiantes de ingeniería industrial.

Altura de 117 alumnos de la escuela de Ingenieros Industriales
180 178 192 180 162 183 168 160 182 172 163 175
163 182 179 174 182 178 159 157 175 175 178 179
189 180 182 165 178 155 178 182 178 180 183 179
170 165 185 162 170 174 190 178 163 170 180 189
180 175 167 167 173 172 175 175 165 180 173 165
163 169 162 169 178 163 184 172 169 176 164 178
187 181 199 190 169 179 184 187 175 176 179 161
178 178 169 179 175 177 169 175 178 177 184 180
175 175 184 156 173 192 186 180 169 171 172 180
193 182 185 177 170 173 192 166 173

La observación de los datos puede llevarnos algunas conclusiones, pero evidentemente no parece una forma e…ciente de obtener información. Nos gustaría poder ”entender” cómo es la altura de una forma mucho más sencilla. También nos gustaría saber si, por ejemplo un chico de 1.60 es bajo, alto o normal.

La primera técnica grá…ca que vamos a introducir es el histograma. Un histograma es una representación de las frecuencias con que aparecen los distintos valores en la muestra. Para realizar un histograma es preciso obtener la tabla de frecuencias de la variable.

La tabla de frecuencias de la variable se obtiene contando el número de alturas que se han encontrado en cada intervalo. En nuestro caso vamos a dividir el intervalo de las alturas en 8 clases o intervalos. El valor mínimo observado es 155 y el máximo 199. Entonces vamos a dividir el intervalo 155 a 199 en 8 clases. Cada intervalo tendrá un tamaño de:

$$(199-155)/8=5.5\text{cm}$$

Los intervalos o clases serán por tanto:

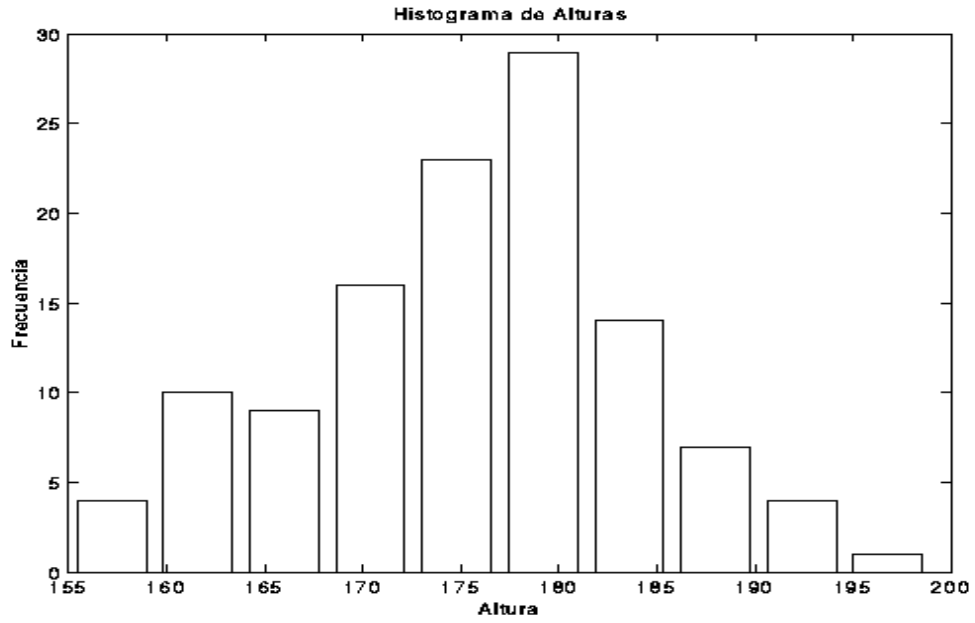
155-160.5, 160.5-166, 166-171.5... etc

Construimos una tabla de frecuencias en la que se recoge el número de individuos que se observan en cada una de las clases. A esta tabla se le llama Tabla de Frecuencias.

Tabla de Frecuencias

Intervalo	Min	Máx	Punto Medio	Frec Abs	Frec Rel
1	155	160.5	157.75	1	0.0427
2	160.5	166	163.25	15	0.1282
3	166	171.5	168.75	15	0.1282
4	171.5	177	174.25	27	0.2308
5	177	182.5	179.75	35	0.2991
6	182.5	188	185.25	11	0.0940
7	188	193.5	190.75	8	0.0684
8	193.5	199	196.25	1	0.0085

Con las frecuencias obtenidas construimos un Histograma que consiste en dibujar una barra de altura la frecuencia relativa, sobre cada uno de los intervalos. La …gura muestra el histograma



El histograma muestra las frecuencias obtenidas en cada uno de los intervalos. Es un grá..co muy útil pues permite resumir la información rápidamente. Así por ejemplo, podemos decir que la altura habitual de los estudiantes oscila entre 155 y 199, pero la gran mayoría tiene una altura entre 1.70 y 1.85. Con una simple ojeada a un histograma se obtiene mucha más información

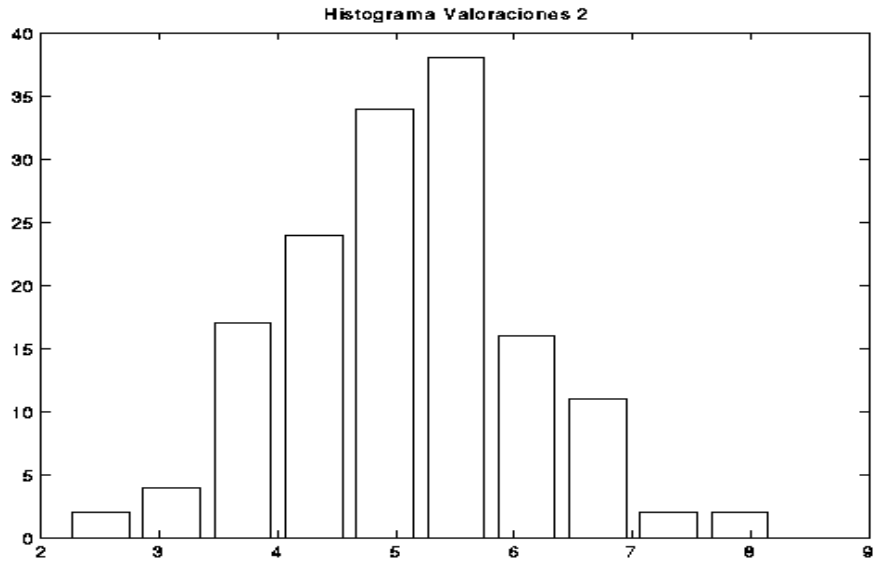
que con la horrible tabla de datos del ejemplo.

Ejemplo

Los histogramas de las …guras adjuntas representan la valoración obtenida por los servicios de una empresa en una encuesta realizada entre sus clientes:

Como puede observarse los clientes de esta empresa tienen opiniones confrontadas respecto a la calidad del servicio. Existe una minoría que está satisfecha y una mayoría que suspende la calidad de la empresa. ¿Ante esta situación que debemos hacer?

La Empresa 2, cuyo histograma de valoraciones se presenta a continuación tiene una situación completamente diferente.



Sus clientes están medianamente satisfechos pero no se detecta una bimodalidad. La Empresa 2 deberá mejorar en su conjunto. La Empresa 1 debe tratar de averiguar cómo están distribuidos sus clientes y aprender del subgrupo que está satisfecho para tratar de contentar a los demás.

Además del histograma existen otros grá...cos interesantes. Vamos a estudiar el diagrama de Tallo y Hojas y el diagrama de Caja.

4.2 Diagrama de Tallo y hojas.

El diagrama de tallo y hojas ofrece una información análoga a la del histograma pero es mucho más sencillo de realizar si no se dispone de ordenador.

Vamos a introducirlo con un ejemplo:

Las notas obtenidas por un grupo de alumnos de Estadística ha sido:

Notas de Estadística:
4.3 5.2 6 7.2 6.5 5 4 6.2 7.5 9 4.4 5 6 8 3.4 2
8.9 10 5 5.5 3 5.5

Diagrama de tallo y Hojas

0 j
1 j
2 j0
3 j45
4 j304
5 j200055
6 j0520
7 j25
8 j09
9 j0
10j0

El diagrama de tallo y hojas se construye separando para cada dato el último dígito de la derecha. Por ejemplo en la columna 4j304 se están representando los datos 4.3 4.0 y 4.4 que son todos los datos que empiezan por 4.

La visión de un diagrama de tallo y hojas, permite detectar rápidamente pautas en los datos. Así, en nuestro caso el profesor que ha puesto estas notas parece equilibrado. Hay muchos aprobados, y también hay buenas notas. Si hubiésemos obtenido un diagrama como el siguiente:

Diagrama de tallo y Hojas

0 j00000000333050305
1 j246342000255676
2 j09356623245
3 j45
4 j304
5 j20
6 j0520
7 j
8 j
9 j0
10 j0

¿Qué tipo de profesor tendríamos?

4.3 Diagrama de Caja.

El diagrama de caja es un grá…co muy interesante que proporciona información sobre la existencia de datos atípicos. Se explicará tras la sección dedicada a medidas analíticas.

5 Variables cualitativas.

Decimos que una variable es cualitativa cuando no tiene una representación numérica clara.

Por ejemplo, entre los coches vendidos en Estados Unidos en 1982 había coches norteamericanos, Japoneses y Europeos.

Se sabe que había un 54.84% de coches fabricados en EE.UU. un 16.77% de coches fabricados en Europa y un 28.39% de coches japoneses. Una buena representación grá…ca de estos datos es mediante una Tarta. En ella queda meridianamente re‡ejada la proporción entre los datos de los fabricantes de automóviles.

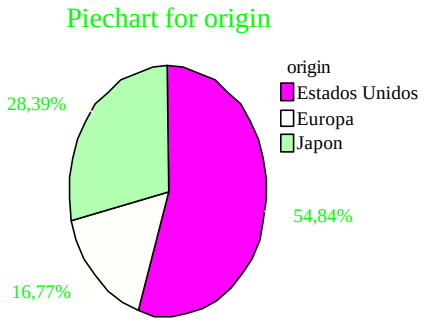


Diagrama de Tarta

Una representación alternativa es el diagrama de barras que ofrece una información similar.:

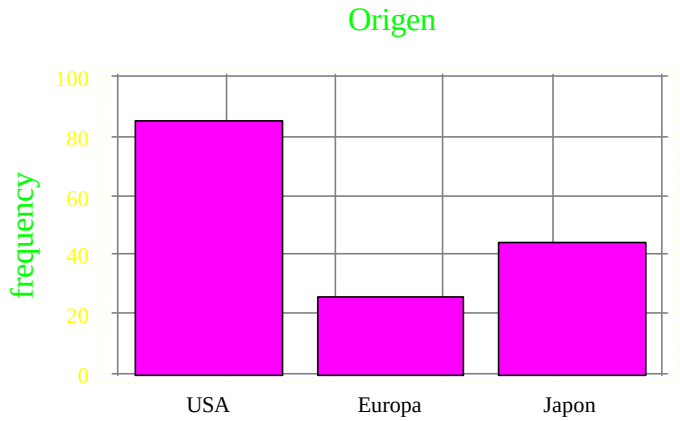


Diagrama de Barras

5.1 Diagrama de Pareto.

El diagrama de Pareto es una consecuencia de que cuando se analizan las causas de un problema, en general son relativamente pocas. Esencialmente

es un diagrama de barras tal que las barras están ordenadas de mayor a menor.

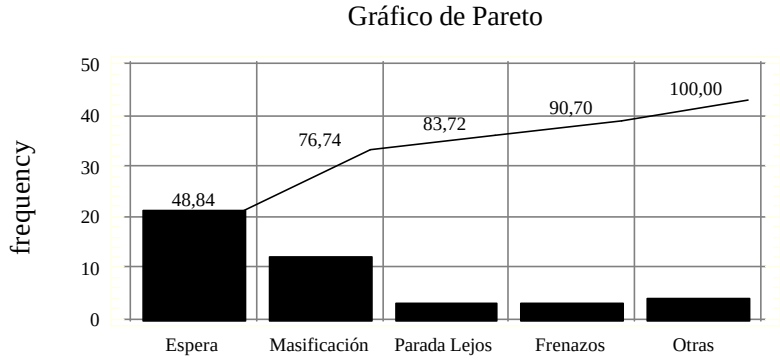
El ejemplo que se presenta a continuación ilustra el análisis de Pareto.

Ejemplo:

Se ha tomado nota durante dos meses de las reclamaciones de los clientes de un servicio de autobuses. Las causas de las reclamaciones se han clasi…cado y se han obtenido los datos siguientes:

Causa	Número de Quejas
Retrasos	21
Masi…cación	12
Parada lejos	3
Frenazos	3
Otras	4

Si se realiza el diagrama de barras obtenemos:



Como puede observarse el diagrama de Pareto ofrece también los porcentajes acumulados de las diversas causas ordenadas. En este ejemplo de autobuses podemos observar que el 76% de las quejas se re…eren a retrasos y masi…cación, que generalmente van asociados, ya que cuando un autobús se retrasa, se acumulan los viajeros que lo están esperando con la consiguiente masi…cación.

6 Series Temporales.

Las series temporales surgen constantemente cuando se quiere estudiar la evolución de una variable a través del tiempo. El grá…co natural de una serie temporal es su representación a lo largo del tiempo.

Cuando estudiamos series temporales es muy importante conocer una serie de conceptos:

- Periodicidad de la serie: Es la frecuencia con que se toman los datos. Las series pueden ser de periodicidad anual (Se tienen un dato por año) mensual (un dato al mes) trimestral (un dato al trimestre) u otras.
- Tendencia: Decimos que una serie tiene tendencia cuando su grá…co aumenta o disminuye de una forma sistemática con el tiempo.
- Estacionalidad: Decimos que una serie tiene estacionalidad si se observa un ciclo que está ligado al mes del año en que se ha tomado el dato.

A continuación se presenta un conjunto de grá…cos de series temporales.

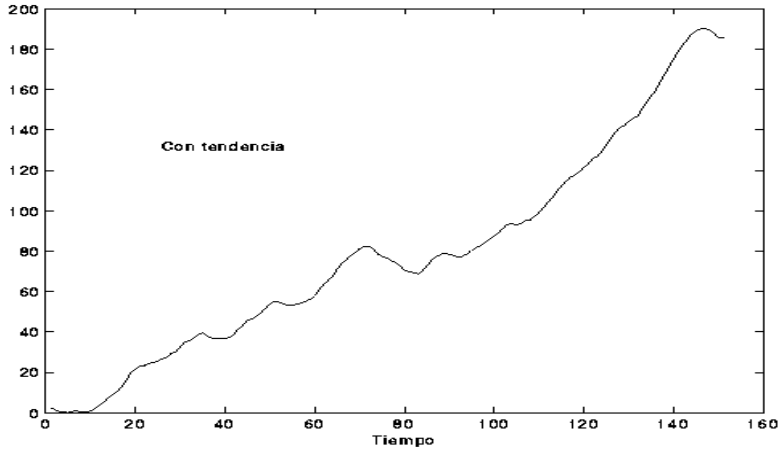


Figura 1: Serie con tendencia

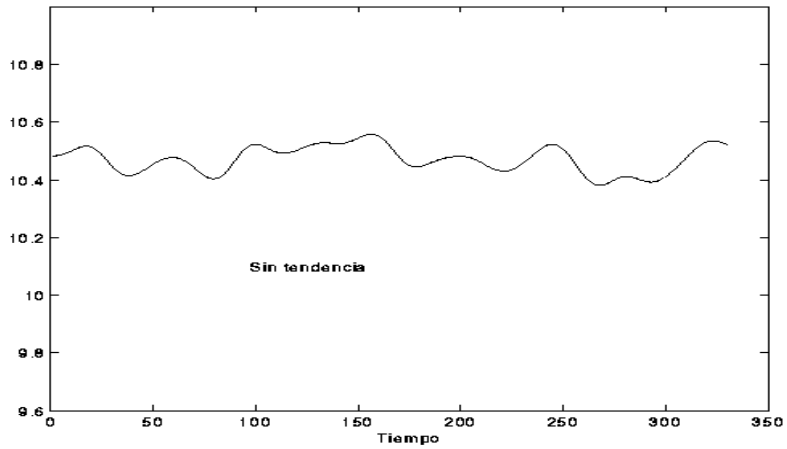


Figura 2: Serie sin tendencia

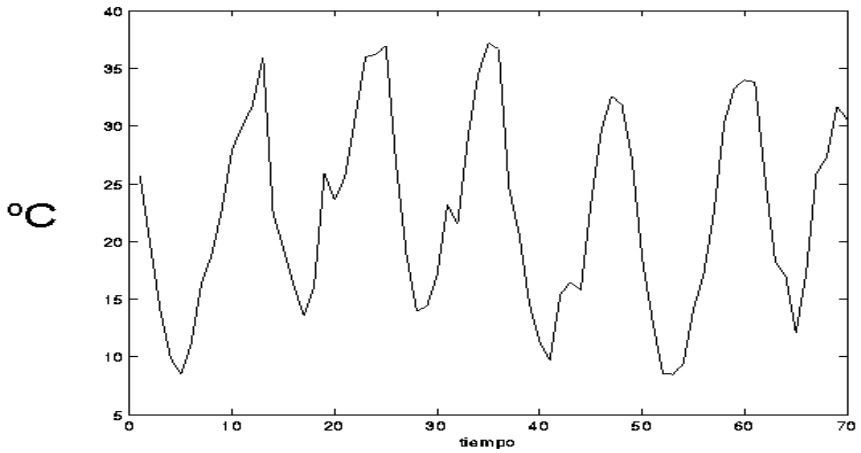
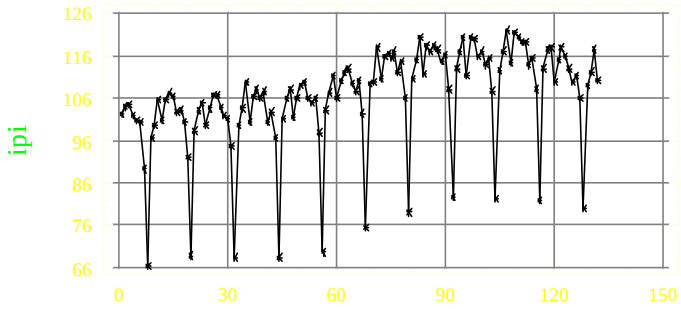


Figura 5: Temperatura en Madrid. Meses.

Figura 30: Serie IPI en Francia



7 Medidas Analíticas para la descripción de datos.

Además los grá…cos, se pueden usar medidas numéricas para describir conjuntos de datos. Distinguimos dos tipos de medidas:

- Posición o Centralización
- Dispersión

7.1 Medidas de Posición o Centralización.

Las medidas de centralización proporcionan información sobre dónde está localizada la muestra. La medida más conocida es la media de los datos que tiene una expresión matemática:

$$\text{Media} = \overline{x} = \frac{x_1 + x_2 + ¢¢¢ + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Así, por ejemplo la media de las alturas de los alumnos de ingeniería es: 175,4 centímetros.

Además de la media, existen otras medidas de centralización, como es la mediana. Para calcular la mediana de una serie de datos se procede de la siguiente manera:

- 1. Se ordenan los datos de menor a mayor
- 2. Se obtiene el valor central de los datos ordenados. Es valor es la mediana.

- 3. Si en número de datos es par y, consecuentemente, hay dos datos centrales, se calcula el valor medio de esos dos datos centrales. Ese valor es la mediana.

La mediana tiene una serie de ventajas respecto de la media. Vamos a verlas con un ejemplo.

Supongamos que tenemos 11 datos de la in‡ación en los países del euro. Esos datos son:

In‡ación en Países del Euro
1.2 2.0 0.8 2.4 1.3 0.3 2.0 1.7 1.6 0.9 2.9

La media de las in‡aciones es: 1.55
La mediana será:

- 1. Ordenamos de menor a mayor:

0.3 0.8 0.9 1.2 1.3 1.6 1.7 2.0 2.0 2.4 2.9

El valor central es 1.6, que es la mediana, que representa muy bien la zona en que se sitúan los datos.

Supongamos ahora que ha habido un error al teclear los datos y en lugar de introducir el valor 2.9 correspondiente a un país europeo, hemos introducido 2900, que no corresponde a ningún país.
Los datos serán:

1.2 2.0 0.8 2.4 1.3 0.3 2.0 1.7 1.6 0.9 2900

y la media será 264,9 que es un valor absurdo.

Sin embargo la mediana, será:

² 1. Ordenamos de menor a mayor:

0.3 0.8 0.9 1.2 1.3 1.6 1.7 2.0 2.0 2.4 2900

El valor central sigue siendo 1.6 que representa muy bien la zona en que se sitúan los datos.

La mediana apenas cambia aunque en la muestra haya algún dato erróneo.

La media es muy sensible a la existencia de algún dato erróneo.

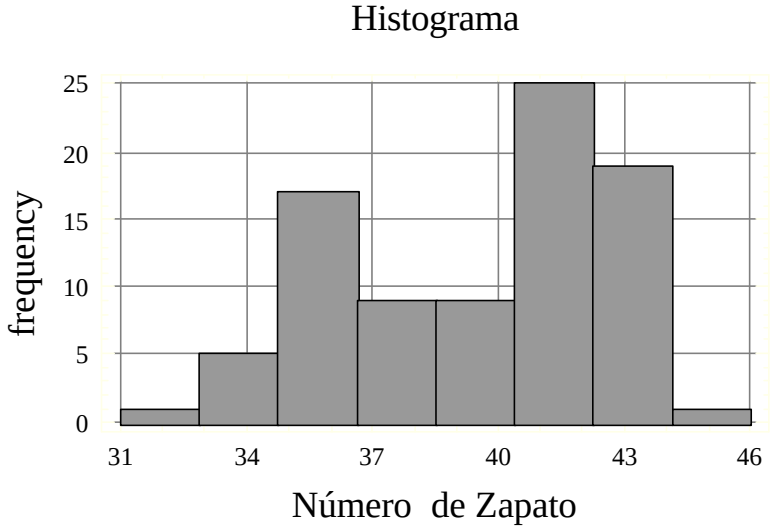
Finalmente otra medida de centralización muy utilizada es la Moda. La moda dice cual es valor que aparece más en la muestra. En el caso de las alturas de los estudiantes la moda es 178, que también es un valor que representa la zona central de datos.

7.2 Distribuciones Bimodales.

Cuando una distribución de frecuencias (Un histograma) presenta dos modas, es decir dos "montañas" se denomina bimodal. Estas situaciones suelen indicar que existen dos poblaciones diferentes para el fenómeno que se está estudiando y es peligroso utilizar las medidas analíticas usuales si no se estudian bien los datos mediante técnicas grá...cas, ya que el valor medio o la mediana pueden quedar en zonas numéricas en las que apenas si hay datos.

Ejemplo:

Los datos representan el número de zapato de una muestra de 80 personas.



¿Que tipo de datos tenemos?

7.3 Media Ponderada.

En ocasiones no nos interesa calcular la media de una muestra, en la que todos los valores tienen el mismo peso o importancia. Por ejemplo, supongamos que estamos estudiando la valoración que los clientes de un hotel hacen de la calidad del mismo. Entre las preguntas que se les realizan a los clientes se encuentran:

Limpieza de la habitación
Rapidez en los trámites de llegada
Iluminación adecuada del BAR

Supongamos que el valor medio obtenido en estas cuatro preguntas haya sido:

Variable/ Atributo de Calidad	Valor Medio Obtenido Caso 1	Valor Medio Obtenido Caso 2
Limpieza de la habitación	3	8
Rapidez en los trámites de llegada	5	7
Iluminación adecuada en el BAR	10	3
Media	6	6

Obsérvese que en ambos casos la media obtenida es idéntica. Sin embargo, si el Hotel del Caso 1, toma como índice de Calidad la nota media de 6, estará cometiendo un grave error, porque posiblemente para los clientes la iluminación del Bar es mucho menos importante que la limpieza de la habitación. En el caso 1, el hotel tiene un gravísimo problema, y la medida de calidad que obtiene no lo detecta.

En el caso 2, el hotel tiene un problema de calidad pequeño: Cambiar la iluminación del Bar, pero los clientes están satisfechos de la limpieza y la recepción. Sin embargo, la nota o indicador ...nal es igual en ambos casos.

Este problema lo resuelve la media ponderada. Supongamos que sabemos que los clientes valoran mucho la limpieza de la habitación. Entonces podemos dar más peso al calcular la media a la limpieza de la habitación.

Esto es muy sencillo. Si decidimos que la limpieza de la habitación debe tener un peso del 50% en la cali...cación ...nal, la recepción un 40% y la iluminación del Bar un 10% (Estos coe...cientes deben sumar 100), entonces la media ponderada por los coe...cientes será en el caso 1:

$$0,5x3+0,4x5+0,1x10=4.5$$

mientras que en el segundo caso,

$$0,5x8+0,4x7+0,1x3=7.1$$

La media ponderada ofrece una visión mucho más realista de los problemas de ambos hoteles.

La expresión para calcular una media ponderada es:

$$\text{Media Ponderada} = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

donde

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$$

Existen diversas técnicas para calcular los ponderadores $\alpha_1; \alpha_2; \cdots; \alpha_n$. Un posibilidad es realizar regresiones como se ilustrará en la sección correspondiente.

7.4 Medidas de Dispersión

Además de obtener una medida numérica de la ubicación de los datos, es importante obtener datos sobre si la muestra está muy concentrada en torno a la media o no.

Las medidas de dispersión más frecuentes son la desviación típica y los rangos y percentiles.

La desviación típica da una medida de la distancia de media de los datos a la media de la muestra. Si unos datos tienen mucha desviación quiere decir que su histograma será muy ancho y habrá mucha variabilidad.

Cuando la desviación típica sea pequeña tendremos datos muy centrados en torno al media y consecuentemente habrá poca variabilidad.

Su expresión matemática es:

$$s = \sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots (x_n - \bar{x})^2)/n - 1}$$

siendo $x_1; x_2$ etc los valores de la muestra y $\bar{x}$ el valor medio de la muestra.

Ejemplo: Supongamos que 7, 7, 8, 6, 7 y 8 son las valoraciones obtenidas en seis habitaciones de un hotel respecto al servicio. Su media es 7.16 y su desviación típica

$$\sigma = \sqrt{(7 - 7{:}16)^2 + (7 - 7{:}16)^2 + (8 - 7{:}16)^2 + \cdots)/5} = 0{:}75$$

Si las medidas hubieran sido 10, 10, 10, 2, 2, y 5, es decir, muy heterogéneas, la media hubiera sido 6.5 pero la desviación típica es de 4. Esto indica que existe mucha variabilidad entre las observaciones, y habría que preguntarse por qué.

En el caso de distribuciones bimodales la desviación típica o cualquier otra medida de dispersión suele ser grande ya que las observaciones están en dos grupos relativamente separados.

8 Coe…ciente de Variación

El coe…ciente de variación es el cociente entre la desviación típica y la media e indica precisamente si hay mucha o poca variabilidad para el nivel de la muestra.

$$cv = \frac{\text{desviación}}{\text{media}}$$

9 Diagrama de Caja (Box-Plot)

El diagrama de caja o Box-Plot es uno de los grá…cos más completos y útiles para resumir información: Detecta muy bien la existencia de valores atípicos y permite comparar varias muestras de forma muy e…ciente.

El diagrama de Caja representa la mediana y los cuartiles de la muestra en una caja. Los cuartiles son los valores que dejan por debajo el 25% (cuartil inferior) y el 75% (cuartil superior) de la muestra.

La construcción del diagrama de caja es simple:

1. Se calcula el valor del cuartil inferior y superior $Q_i; Q_S$
2. Se calcula el Valor de la Mediana: Med
3. Se Calcula el Rango intercuartílico: $RI = Q_s - Q_i$

4. Se dibuja la caja entre Q_i y Q_S. Se hace una raya vertical en la mediana.

5. Se calculan los puntos de corte para datos atípicos:

Se consideran datos atípicos los menores del primer cuartil o mayores que el tercero que están a una distancia superior a una vez y media del RI de su cuartil.

Es decir serán puntos atípicos por ser muy bajos los que sean menores que

$$Q_i \text{ ¡ } 1{:}5RI$$

Y serán atípicos por ser demasiado altos

$$Q_s + 1{:}5RI$$

- Si no hay datos atípicos el diagrama de caja se representa mediante la caja y dos líneas a cada lado de ella que llegan al máximo y al mínimo de la muestra respectivamente
- Si existen atípicos se pintan las líneas hasta el límite de los atípicos: Q_i ¡ 1:5RI y Q_s + 1:5RI: Y se marcan los puntos atípicos mediante cruces.

Ejemplo:

Se tiene la siguiente relación de pesos de alumnos de la Universidad:

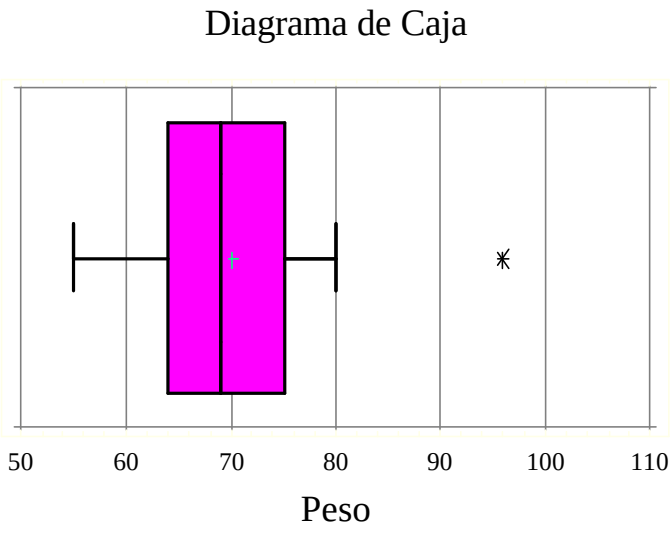
55 59 61 62 64 64 67 68 68 68 69 70 72 73 74 75 75 76 78 80 96.

Hay 21 observaciones.

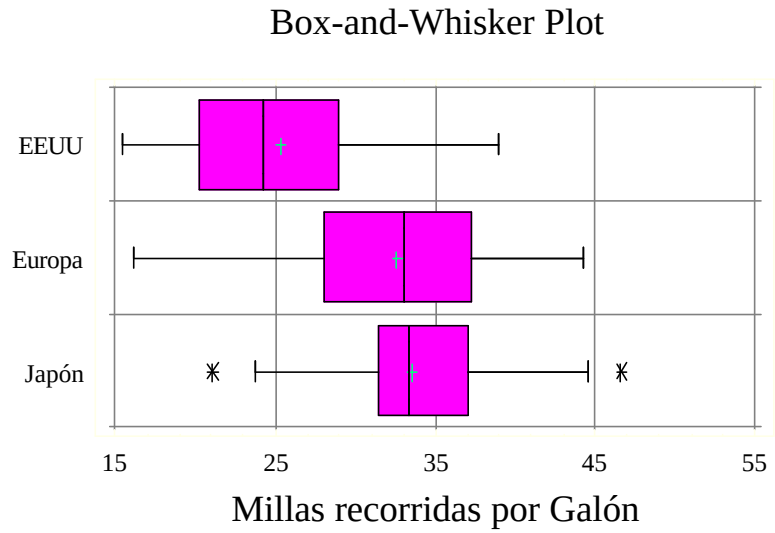
- La mediana será la observación 11: Med = 69:
- El cuartil inferior será la media de las observaciones 5 y 6, 64 y 64, Q_i = 64
- El Cuartil superior será la media de las observaciones 16 y 17, 75 y 75Q_s = 75:

- El Rango Intercuartílico será: $RI = 75 - 64 = 11$
- Corte inferior de atípicos: $Q_i - 1{:}5RI = 64 - 1{:}5x11 = 64 - 11 = 47{:}5$
- Corte superior de atípicos: $Q_s + 1{:}5RI = 75 + 1{:}5x11 = 75 + 11 = 86$

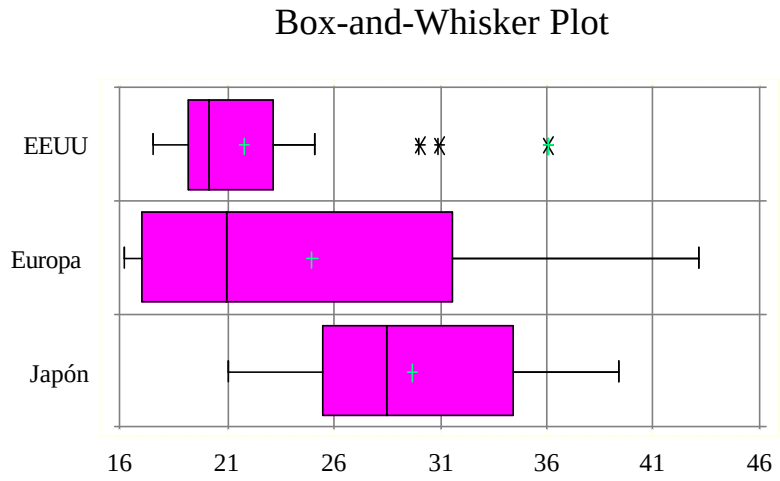
Como puede verse, no existen atípicos inferiores y si superiores. Así la raya inferior llegará hasta el valor mínimo y la superior hasta el corte de atípicos. Pintaremos 96 con una cruz:

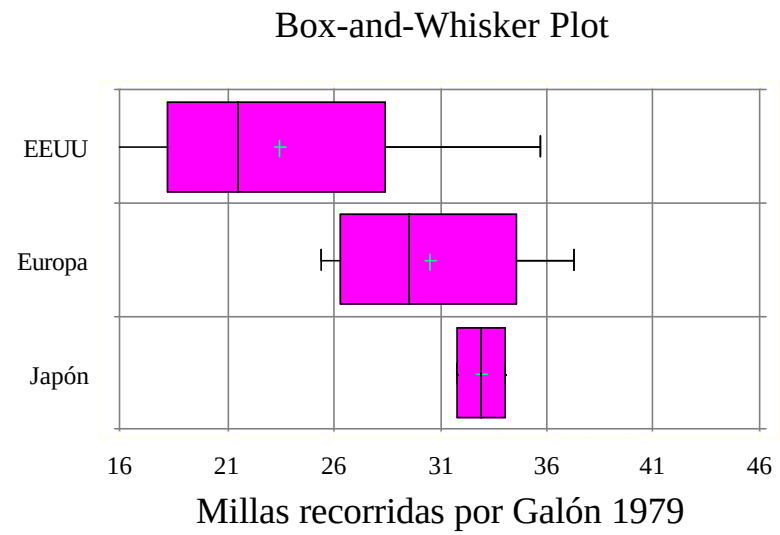


El siguiente Box-Plot presenta los consumos de automóviles vendidos en EE.UU. en los años 1978 y sucesivos según su origen.

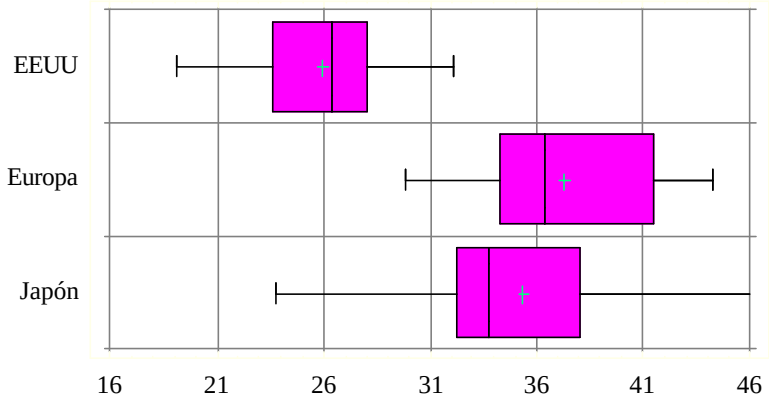


Como puede verse rápidamente los coches Norteamericanos no se ajustaron al aumento de precio de los combustibles. Vamos a realizar estos Box-Plots año a año, desde 1979 a 1982 para estudiar con más detalle su evolución


Box-and-Whisker Plot
EEUU
Europa
Japón
16
21
26
31
36
41
46
Millas recorridas por Galón 1978

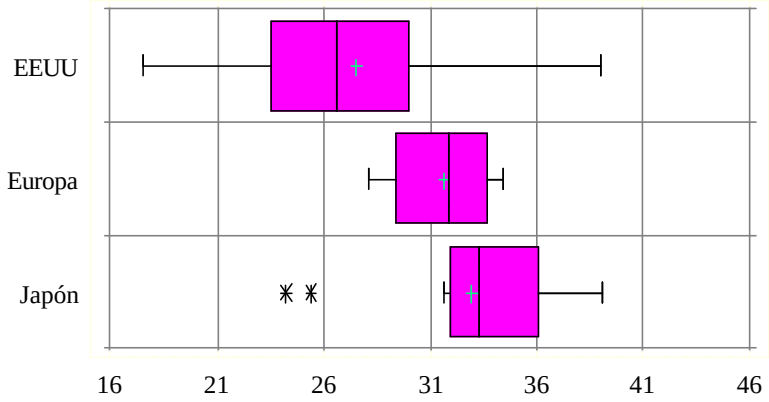

Box-and-Whisker Plot
EEUU
Europa
Japón
16
21
26
31
36
41
46
Millas recorridas por Galón 1979

Box-and-Whisker Plot

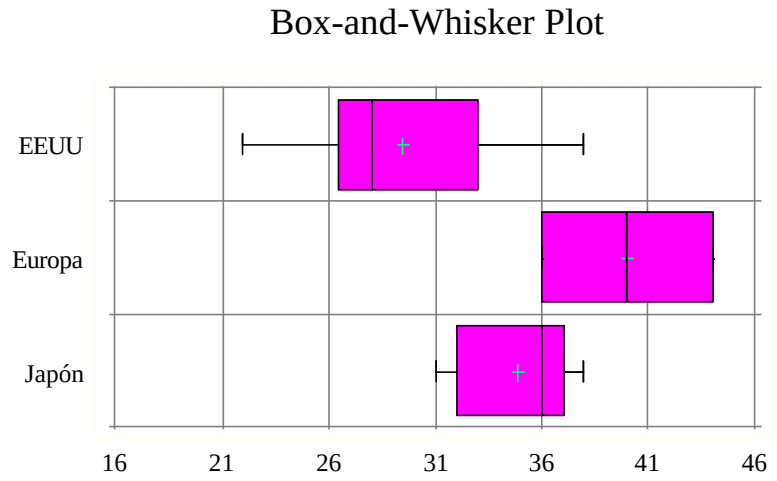


Millas recorridas por Galón 1980

Box-and-Whisker Plot



Millas recorridas por Galón 1981



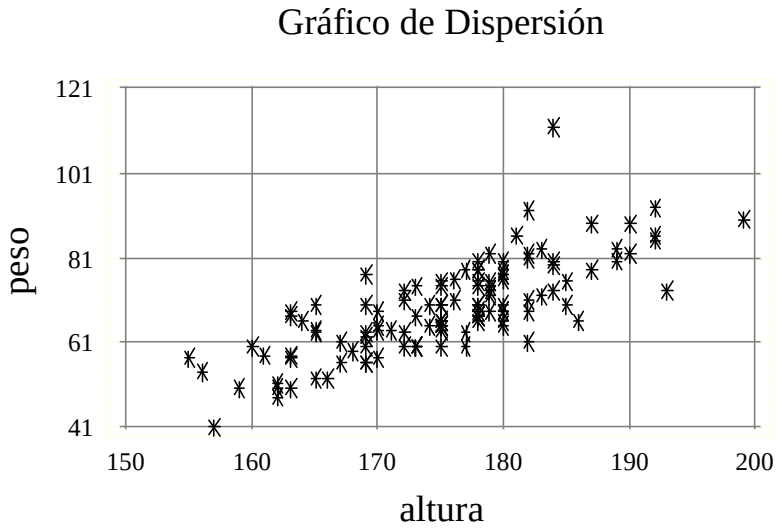
Ejercicio:

Estudiar el proceso de adaptación de los automóviles al encarecimiento del precio de la gasolina.

10 Relación entre dos Variables.

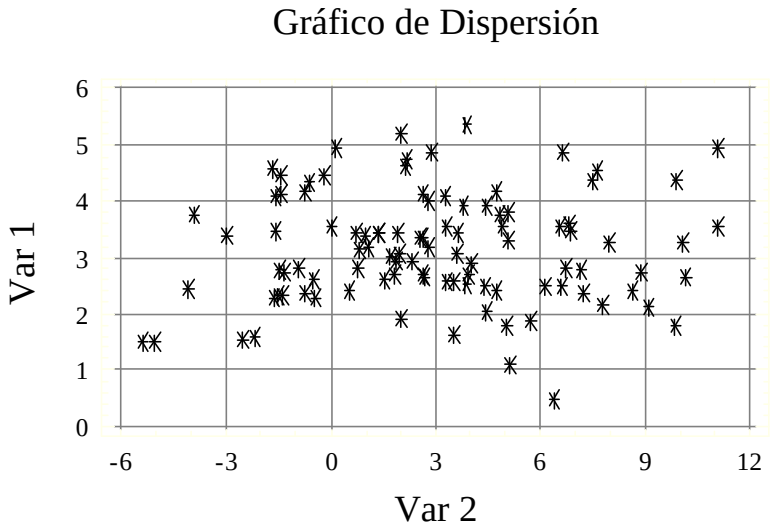
Habitualmente en la mayoría de los problemas que se estudian, no sólo se analiza una sola variable, sino que se estudian varias variables a la vez. En este caso es muy útil realizar Diagramas de Dispersión (Scatter plots) que van a proporcionar información sobre si existe relación entre dos variables.

El grá...co siguiente muestra la relación entre el peso y la altura de 104 estudiantes de la Universidad Politécnica de Madrid



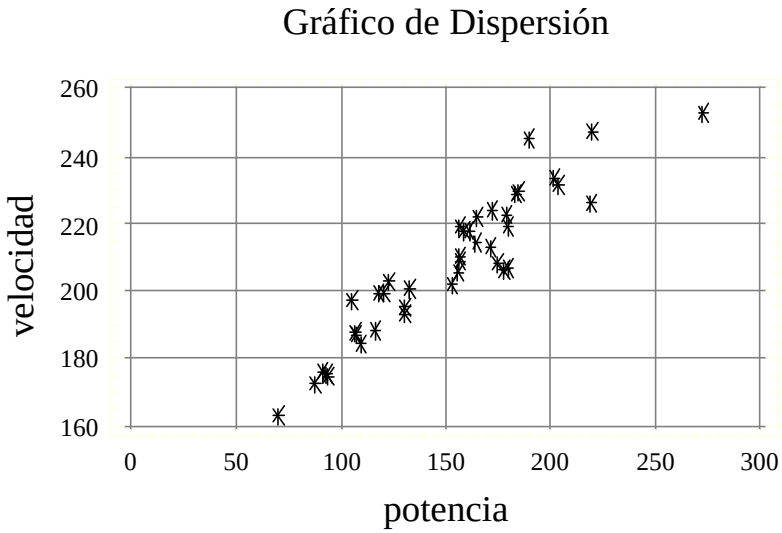
Como puede observarse existe una relación entre ambas variables, que es lineal creciente.

Si entre dos variables no existiera relación alguna, el grá…co de dispersión tendría el aspecto siguiente:



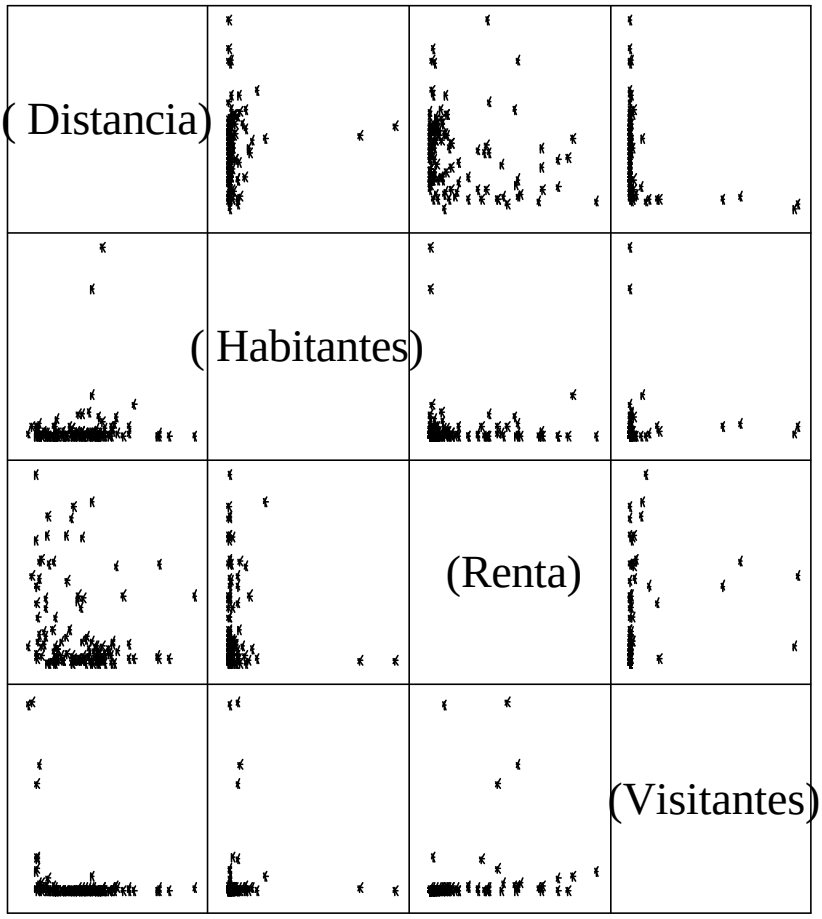
Es decir para cualquier rango de valores de la variable 2, la variable 1 tiene unos valores semejantes. Esto indica que Var 1 no depende de Var 2.

El siguiente grá...co de dispersión muestra la relación entre velocidad punta y potencia de una serie de automóviles. ¿Existe relación? ¿Es razonable el aspecto del grá...co?



10.1 Transformaciones.

Existen relaciones entre dos variables que no son lineales como hemos visto en los grá…cos anteriores. Un ejemplo de relaciones no lineales se presenta en los siguientes datos. Los grá…cos contenidos en la siguiente matriz representan el número de visitantes anuales que vienen a España procedentes de una serie de países. Se recoge para cada país el número de visitantes, Número de habitantes, distancia a España en Km y Renta per Cápita de esos países. Puede verse en los diagramas de dispersión que no existe una relación lineal entre las variables. Pero además no están claras las relaciones que pudiera haber entre las variables.



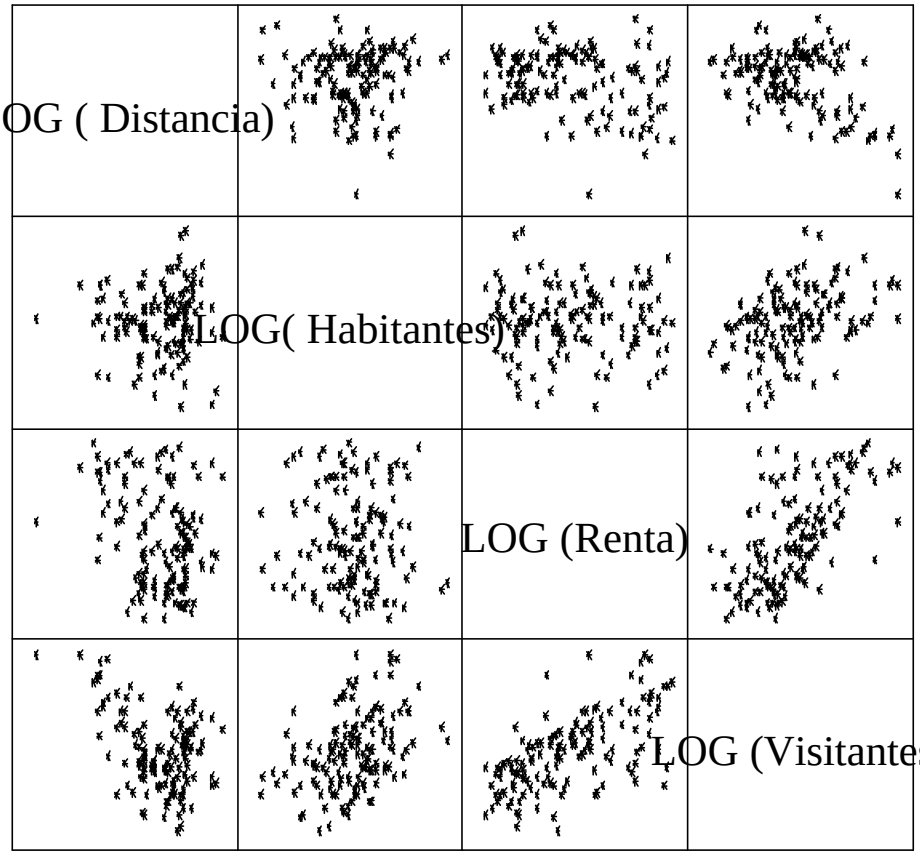
Sin embargo si transformamos las variables a logaritmos, tal como se hace en la siguiente …gura, puede observarse que hay relaciones lineales entre algunas variables (Log Renta y Log Visitantes) y entre otras variables no hay relación (Log Distancia y Log Habitantes)

Este resultado es interesante por varios motivos. En primer lugar simpli…ca notablemente la comprensión del fenómeno, ya que, por ejemplo, se ve claramente que si aumenta la renta del país de origen, también lo hace el

número de visitantes que recalan en España. Esto normal teniendo en cuenta que España es uno de los primeros destinos turísticos del mundo.

No hay mucha relación entre la distancia y los habitantes, lo cual parece bastante razonable.

Pero, además, la transformación logarítmica tiene una propiedad fundamental: Representa relaciones entre las tasas de crecimiento de las variables. Así, el que exista una relación lineal entre $\log x$ y $\log y$, indica que existe una relación lineal entre la tasa de crecimiento de x, es decir $\Delta x=x$ y la tasa de crecimiento de y, $\Delta y=y$.



10.2 Correlación.

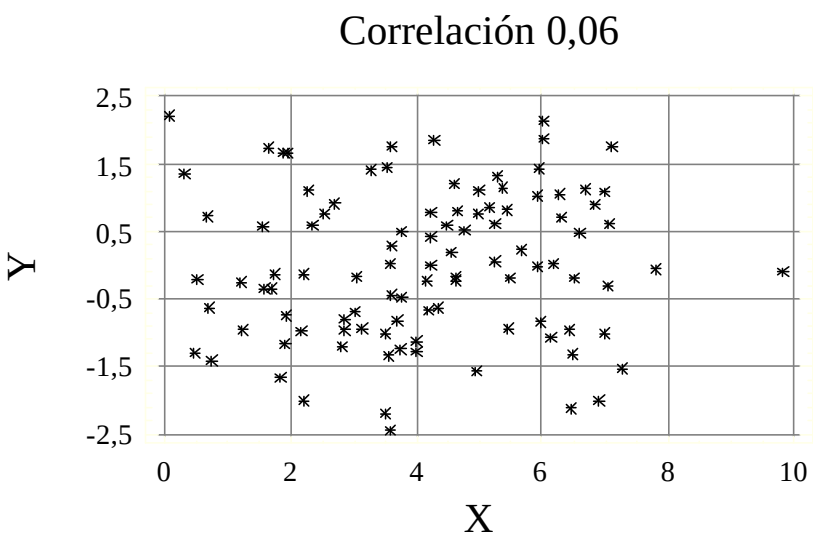
Una medida analítica de la existencia de relación entre dos variables es el coe…ciente de correlación. Se de…ne como

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

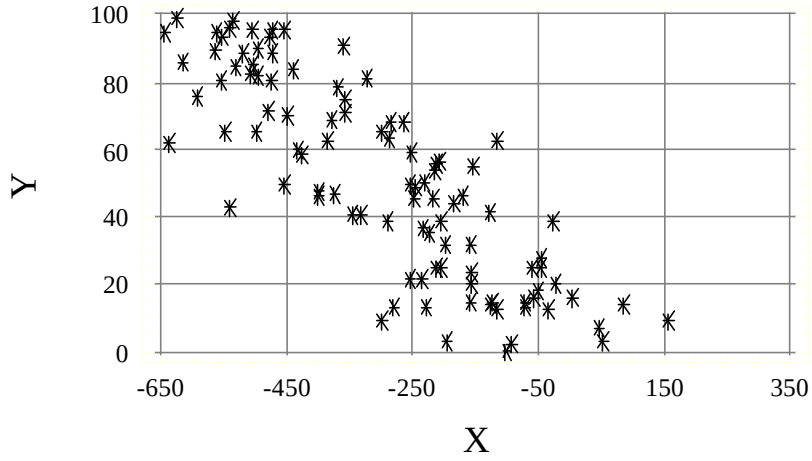
donde s_x y s_y representan las desviaciones típicas de x e y:

El coe…ciente de correlación toma valores entre ¡1 y +1. Cuando $r_{xy} = 0$ quiere decir que no existe relación entre x e y: Si $r_{xy} = 1$ existe una relación lineal positiva perfecta entre x e y. Finalmente si $r_{xy} = ¡1$, existe una relación lineal perfecta pero de pendiente negativa entre ambas variables.

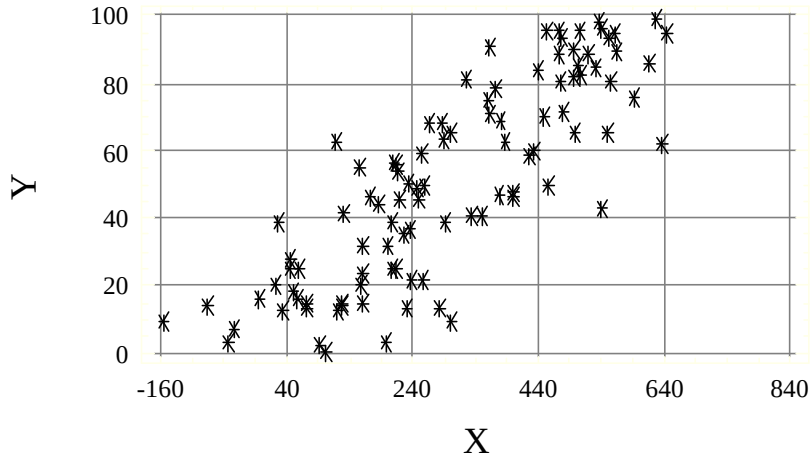
Los siguientes grá…cos ilustran estas ideas.



Correlación -0.83



Correlación +0.83



Es impotante destacar que la correlación es una buena medida para datos lineales pero si los datos son no lineales puede dar lugar a equívocos. El

ejemplo siguiente muestra unos datos con una estructura muy clara y sin que exista apenas correlación entre ellos. El coe…ciente de correlación es de 0.05, muy bajo. Es muy necesario por tanto estudiar el coe…ciente de

correlación entre las observaciones. Pero es imprescindible completarlo con un buen análisis grá…co