Estadística descriptiva

José Gabriel Palomo Sánchez
gabriel.palomo@upm.es

E.U.A.T.
U.P.M.

Julio de 2011

Índice I

1. Introducción
   1. Generalidades
   2. Tipos de datos
   3. Objetivos de la estadística descriptiva
   4. Frecuencias
2. Tablas y gráficos
   1. Tablas
   2. Gráficos y transformaciones
3. Medidas numéricas
   1. Medidas de centralización
   2. Medidas de dispersión
   3. Desigualdad de Chebychef
   4. Comparación de dispersiones
   5. Otras medidas de dispersión
   6. Percentiles
   7. El diagrama de cajas. Puntos atípicos

Índice II

3. Medidas numéricas. (Continuación)
   8. El coeficiente de asimetría
   9. El coeficiente de curtosis
4. Variables bidimensionales
   1. Variables marginales
   2. Variables condicionadas

Generalidades I

La estadística aplicada tiene como objetivo, en muchas ocasiones, dar respuesta a preguntas concretas sobre el comportamiento de conjuntos muy amplios, o inaccesibles, de individuos.

Generalidades II

Algunos ejemplos de estas preguntas podrían ser los siguientes:

- ¿Qué proporción de ciudadanos votaría a un determinado partido político, si hubiese hoy elecciones?
- ¿Qué porcentaje de españoles gasta más del 50 % del presupuesto familiar en la adquisición de su vivienda?
- ¿Cumple una clase de cemento las especificaciones de una norma ISO?

Generalidades III

En general, para estudiar preguntas como las expuestas anteriormente, se selecciona una parte del conjunto de individuos que se quiere investigar, y se toman datos coherentes con el contenido del problema. El análisis de estos datos ofrece, generalmente, alguna respuesta a la pregunta planteada.

El conjunto de todos los individuos objeto del estudio se denomina **población**. Y el conjunto de individuos seleccionados para la obtención de datos se denomina **muestra**.

Generalidades IV

Así pues, el trabajo fundamental que se realiza en estadística aplicada requiere analizar una colección de datos extraídos de un conjunto de individuos.

Al conjunto de datos se le asigna también el nombre de **muestra**. Del mismo modo es usual en la literatura denominar **variable estadística** a un conjunto de datos.

Tipología de los conjuntos de datos I

El tipo de datos, así como el problema que originó su recogida, condiciona la clase de análisis estadístico que conviene realizar. Los conjuntos de datos, de manera general, se clasifican como:

- **Datos cualitativos:** cada dato es una cualidad, como por ejemplo un color, un estado civil, una posición,...
- **Datos numéricos:** cada dato es un número.

Tipología de los conjuntos de datos II

A su vez los datos numéricos se clasifican como:

- **Datos discretos:** sólo pueden tomar valores en un conjunto asimilable a un subconjunto de los números enteros. Por ejemplo:
  - Número de hijos de una persona, número de veces que alguien va al cine al cabo de un año, ...
- **Datos continuos:** pueden tomar cualquier valor en un rango. Por ejemplo:
  - Resistencia de un material, duración de un aparato, ...

Objetivos de la Estadística descriptiva

Las técnicas de la estadística descriptiva y del análisis exploratorio de datos tienen como objetivo ordenar los datos, en base a obtener el máximo de información, y a orientar la investigación. Para ello se usan herramientas tales como:

1. Tablas
2. Gráficos: Diagramas de barras, histogramas, diagramas de cajas,...
3. Medidas numéricas:
   - De centralización: Media, mediana, moda ...
   - De dispersión: Rango, varianza, desviación típica, ...
   - Otros índices: Percentiles, asimetría, curtosis, ...

Frecuencias

La realización de tablas y gráficos de un conjunto de datos requiere de algunas definiciones previas.

Definiciones

- La **frecuencia absoluta** de un dato, f_a, es el número de veces que dicho dato se repite en el conjunto de la muestra.
- La **frecuencia relativa** de un dato, f_r, es el número de veces que dicho dato se repite en el conjunto de la muestra, comparado con el número total de datos, n,

$$f_r = \frac{f_a}{n}$$

Frecuencias acumuladas

Sea un conjunto de datos ordenado: $x_1, x_2, \ldots x_n$. Con frecuencias absolutas y relativas respectivas:

$$f_{a1}, f_{a2}, \ldots, f_{an}. \quad \text{Y} \quad f_{r1}, f_{r2}, \ldots, f_{rn}.$$

Definiciones

- Las **frecuencia absolutas acumuladas y relativas** de un dato x_i se definen, respectivamente, como:

$$F_{ac}(x_i) = \sum_{j=1}^{i} f_{aj}. \quad \text{Y} \quad Fr_{ac}(x_i) = \sum_{j=1}^{i} f_{rj}.$$

Tablas

Una tabla presenta las frecuencias de los datos, agrupados en intervalos o **clases**, cuyo punto medio es la marca de clase.

La siguiente tabla, por ejemplo, resume las longitudes de los pétalos de cincuenta iris de la variedad versicolor:

Clase	Lim. Inf.	Lim. Sup.	Marca	F_a	F_r	F_{ac}	Fr_{ac}
1	$-\infty$	2'8		0	0	0	0
2	2'8	3'15	2'97	1	0'02	1	0'02
3	3'15	3'51	3'33	4	0'08	5	0'10
4	3'51	3'87	3'69	3	0'06	8	0'16
5	3'87	4'22	4'05	15	0'3	23	0'46
6	4'22	4'58	4'40	13	0'26	36	0'72
7	4'58	4'94	4'76	12	0'24	48	0'96
8	4'94	5'30	5'12	2	0'04	50	1
9	5'3	$+\infty$		0	0	50	1

Tablas. Observaciones

- Cuando el volumen de datos es importante, las tablas pueden resultar confusas.
- Un gráfico es generalmente más intuitivo, aunque contenga la misma información que la tabla.
- Las tablas y los gráficos contienen menos información que el conjunto de datos.

Gráficos I

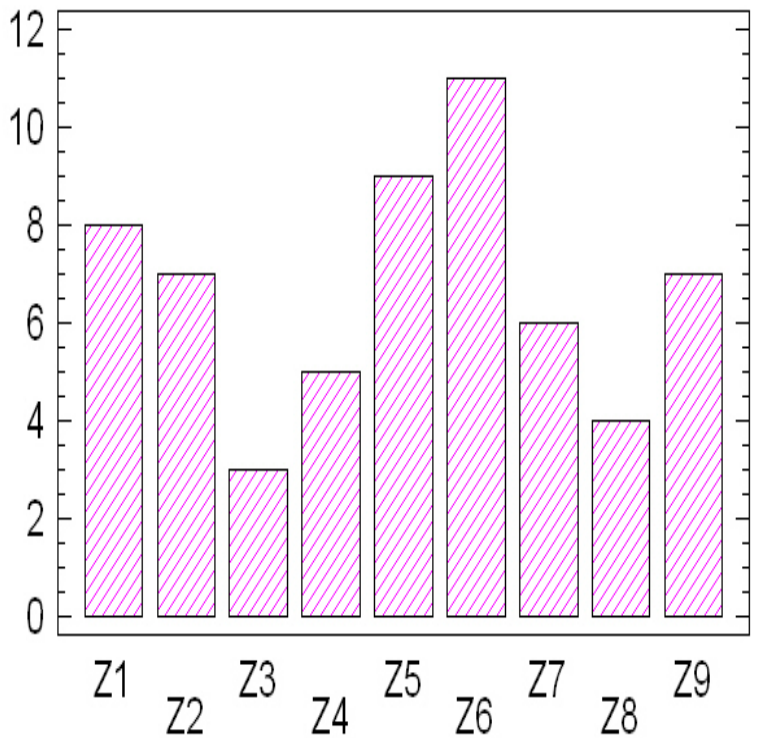
La clase de gráfico que se emplea para representar un conjunto de datos depende de la tipología de los mismos.

- Cuando se analizan variables cualitativas o discretas son útiles:
  1. Los diagramas de barras.
  2. Los diagramas de sectores.
  3. Los diagramas de Pareto.

Gráficos II

En un **diagrama de barras** se representan directamente las frecuencias, absolutas o relativas, de todos los datos.

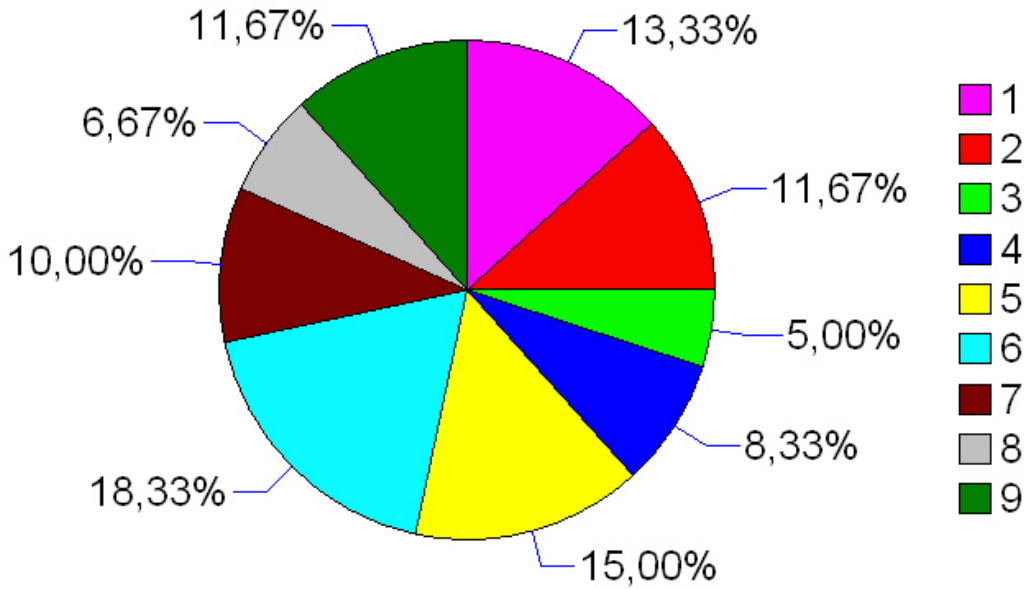
Por ejemplo, en el gráfico de la figura se representa el número de proyectiles caídos en nueve zonas diferentes de Londres, durante un bombardeo de esta ciudad en la segunda guerra mundial.



Gráficos III

En un **diagrama de sectores** se representan las frecuencias, absolutas o relativas, de todos los datos mediante la superficie de sectores circulares.

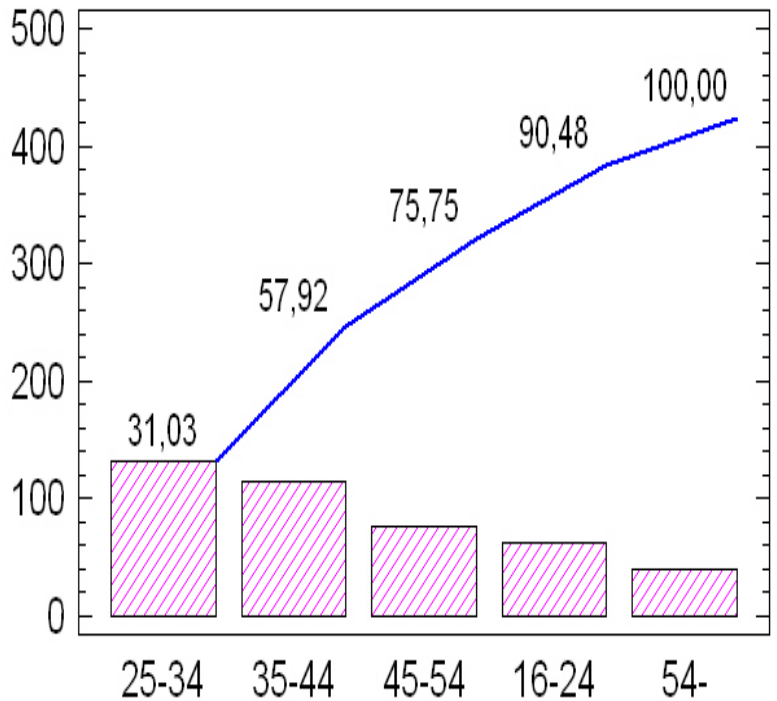
Por ejemplo, este gráfico representa los datos del ejemplo anterior.



Gráficos IV

Un **diagrama de Pareto** es un diagrama de barras, en el que los datos aparecen ordenados por el valor de sus frecuencias. En ocasiones se representan también en el gráfico las frecuencias acumuladas de los datos.

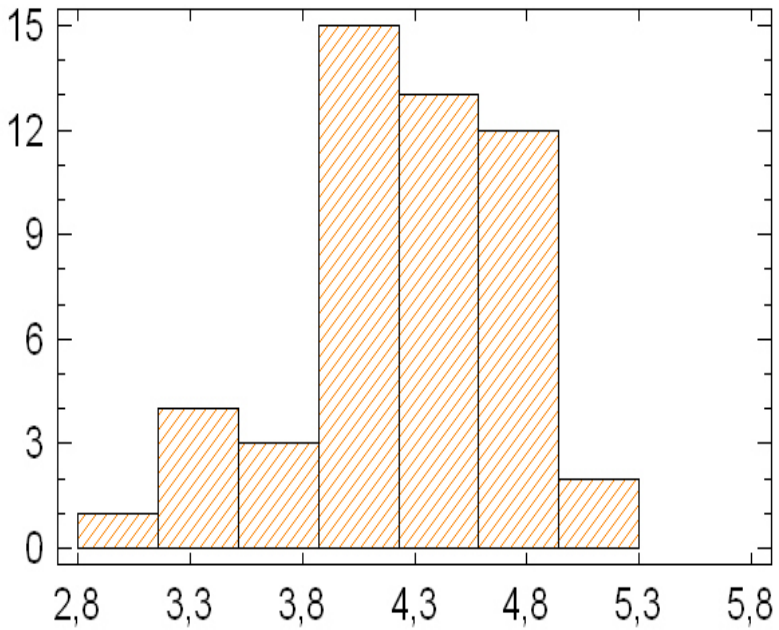
Ejemplo: la siguiente figura describe el número de accidentes, en miles y clasificados por edades, ocurridos en España durante el año 2005. (INE, Noviembre 2007).



Gráficos V

- Cuando la variable es **numérica** de tipo continuo, el **histograma** es el gráfico más empleado. Para construirlo se divide el conjunto de datos en clases, y se representan verticalmente las frecuencias, absolutas o relativas, de las distintas clases.

El siguiente histograma representa los datos de los iris versicolor empleados anteriormente.



Gráficos VI

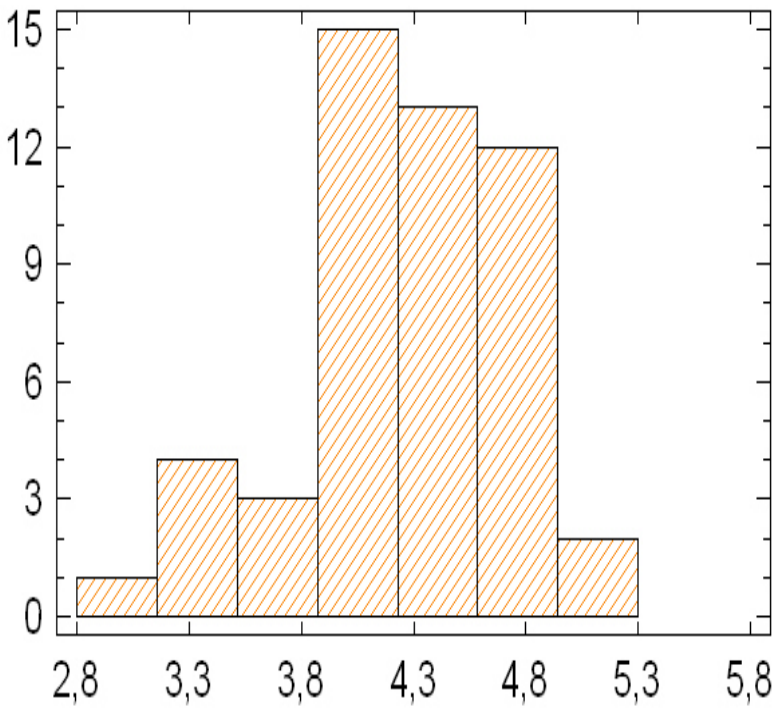
Observaciones

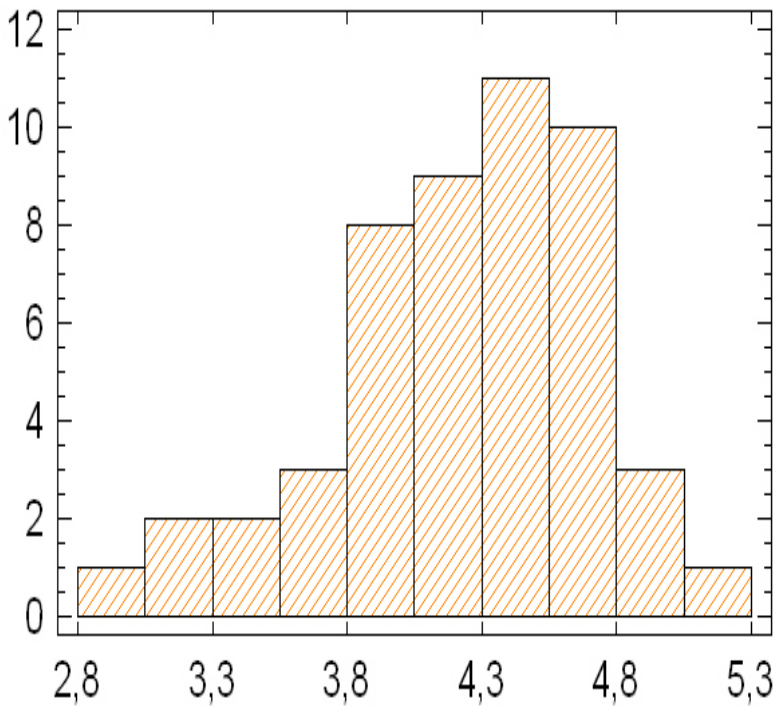
1. La apariencia del histograma puede cambiar si se modifica el número de clases, no existiendo una regla óptima para la elección de este número. En general, se tantea con un número de clases comprendido entre $\sqrt{n}$ y $2\sqrt{n}$.

2. La apariencia del histograma no depende de la elección de la frecuencia absoluta o de la relativa para el eje de ordenadas.

Gráficos VII

Los siguientes histogramas, con siete y diez clases, respectivamente, representan los datos de la variable pétalos versicolor.





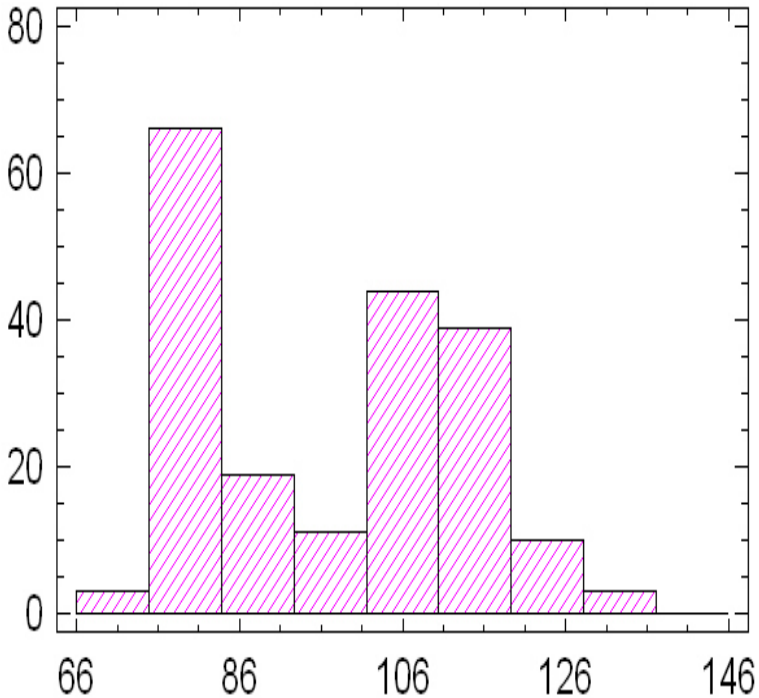
Gráficos VIII

- En un histograma conviene observar, al menos:
  1. Las zonas de concentración de los datos, una o varias.
  2. La variabilidad de los datos.
  3. La simetría.
  4. La existencia de cortes.
  5. Los posibles puntos atípicos.

Gráficos IX

El histograma adjunto representa las longitudes de los élitros de una determinada clase de coelópteros capturados en la isla de Tabarca.

Obsérvese la presencia de dos zonas de acumulación de datos, histograma bimodal, que sugiere heterogeneidad en los mismos.



Gráficos X. Transformaciones.

En ocasiones es más informativo estudiar un conjunto de datos transformados que los propios datos. **Una transformación** de un conjunto de datos consiste en sustituir cada uno de ellos por el resultado de aplicarle una función monótona.

Algunos ejemplos de transformaciones corrientes son: la logarítmica, la inversa, la raíz cuadrada o el cuadrado.

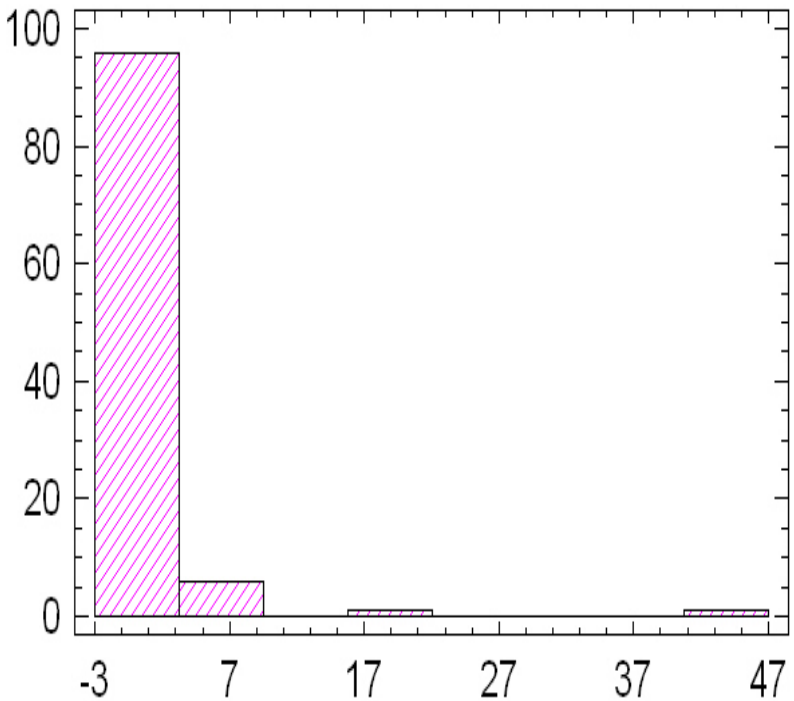
Gráficos XI. Transformaciones.

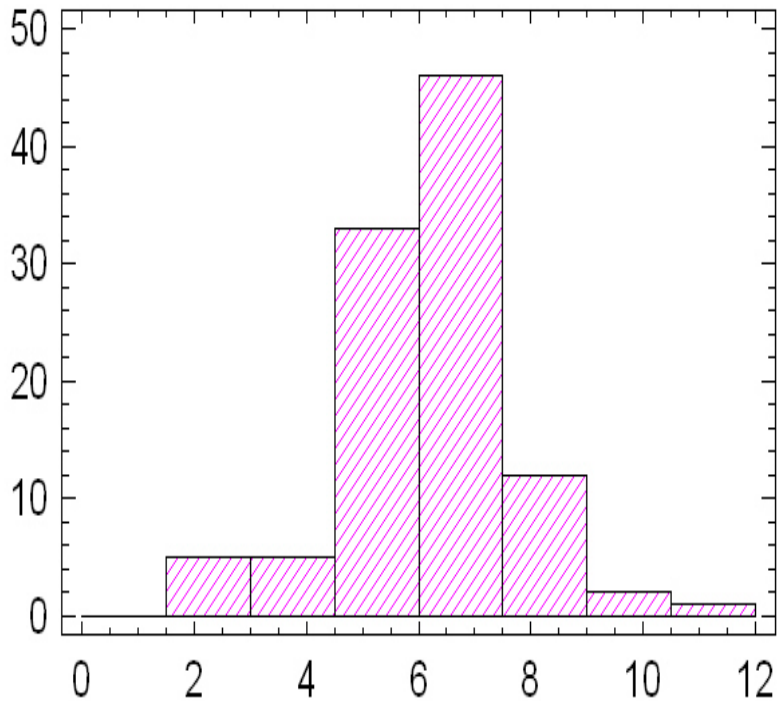
Observaciones

Lo importante en cualquier transformación, T, es que la proporción de datos que se encuentra en cualquier intervalo (a, b) es la misma que la que se encuentra en el intervalo $(T(a), T(b))$.

Gráficos XII

Los siguientes histogramas representan, respectivamente, el número de mujeres en los distintos municipios de Sevilla consignadas en el censo de Floridablanca, y la misma variable en logaritmos.





Gráficos XIII

Observación

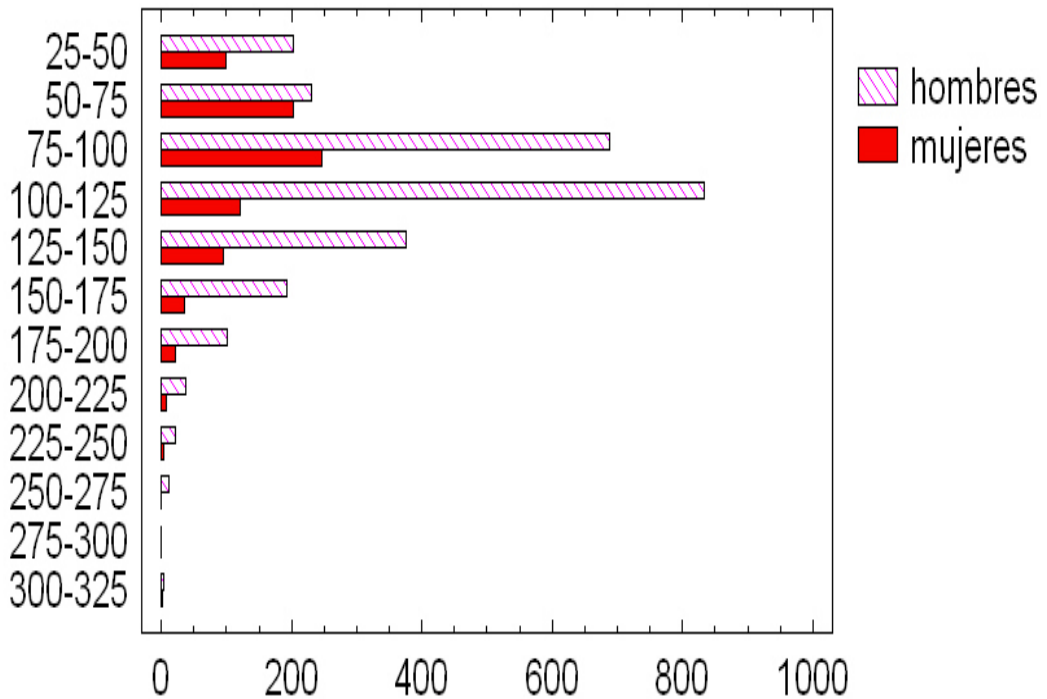
El aspecto de los gráficos y, consecuentemente, el resultado de un análisis puede depender de algunas elecciones que haga el investigador. Considérese el siguiente ejemplo:

- Se afirma, en general, que los salarios de las mujeres es inferior al de los hombres. Para comprobar esta conjetura se han recogido los salarios de 833 mujeres y 2694 hombres, que trabajan en una gran empresa.

Gráficos XIV

En el diagrama de barras múltiple de la figura se representan estos datos, indicando en el eje de ordenadas el rango salarial de los distintos individuos.

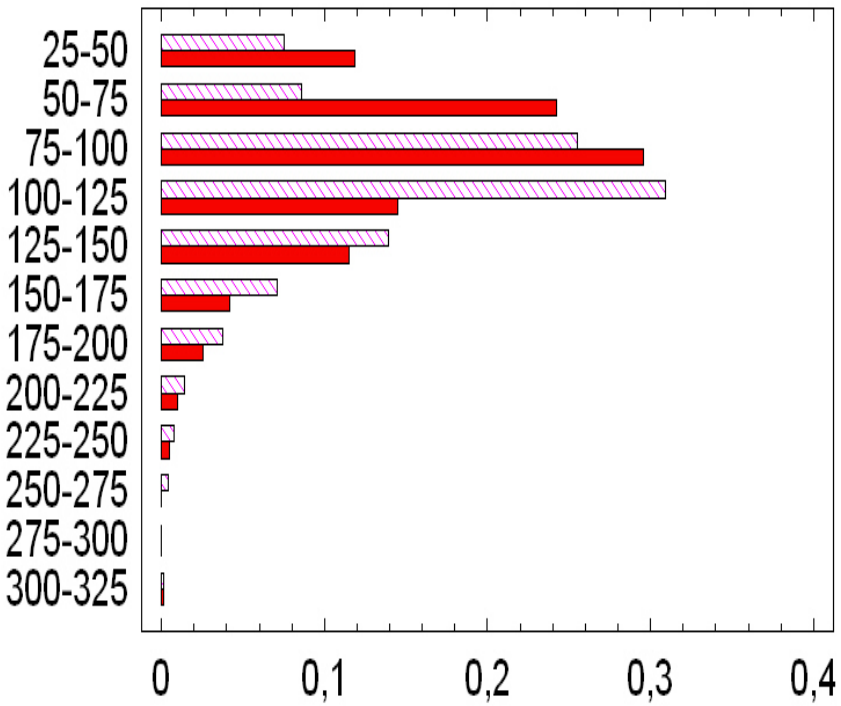
¿Considera admisible la veracidad de la conjetura?



Gráficos XV

Este nuevo diagrama representa los datos relativizados por el número de hombres y de mujeres de la muestra.

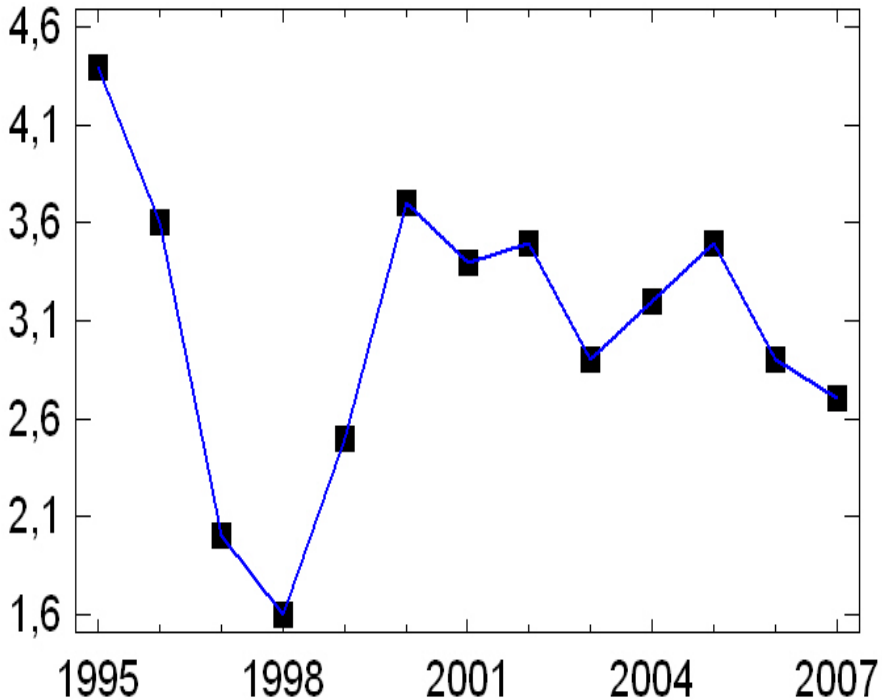
¿Se modifica el juicio emitido anteriormente?



Gráficos XVI

En general, diferentes situaciones pueden requerir la construcción de gráficos distintos a los descritos anteriormente.

Sirva como ejemplo este diagrama temporal que representa la evolución del IPC en España, en periodos de Septiembre a Septiembre, desde 1995 hasta 2007. (INE, Noviembre 2007).



Medidas numéricas

Las medidas numéricas de un conjunto de datos numéricos son números calculados a partir de los propios datos, con objeto de que informen sobre alguna característica del propio conjunto. Se subdividen en:

- **Medidas de centralización** que informan acerca del valor en torno al cual se situa la muestra.
- **Medidas de dispersión** que informan sobre la separación de los individuos de la muestra respecto de alguna medida de centralización.
- **Otras medidas** que aportan información sobre otros aspectos, tales como simetría de los datos, apuntamiento, etc.

Medidas de centralización I

- La principal medida de centralización de un conjunto de datos numéricos $\{x_1, x_2, \ldots, x_n\}$ es la media, que se define por:

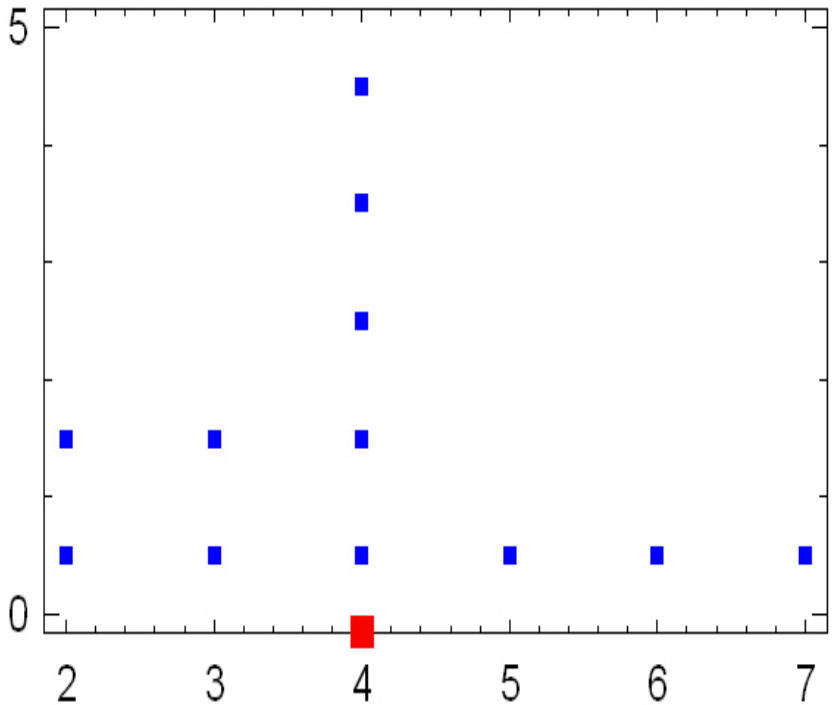
$$\bar{x} = \frac{\sum x_i}{n}$$

Incoveniente de la media

La media es sensible a la presencia de valores atípicos.
El siguiente ejemplo tiene por objeto poner en evidencia la sensibilidad de la media frente a la presencia de puntos atípicos.

Medidas de centralización II. Ejemplo

El diagrama de puntos de la figura muestra la media del conjunto de datos formado por los valores:
$\{2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 6, 7\}$
cuyo valor es $\bar{x} = 4$.

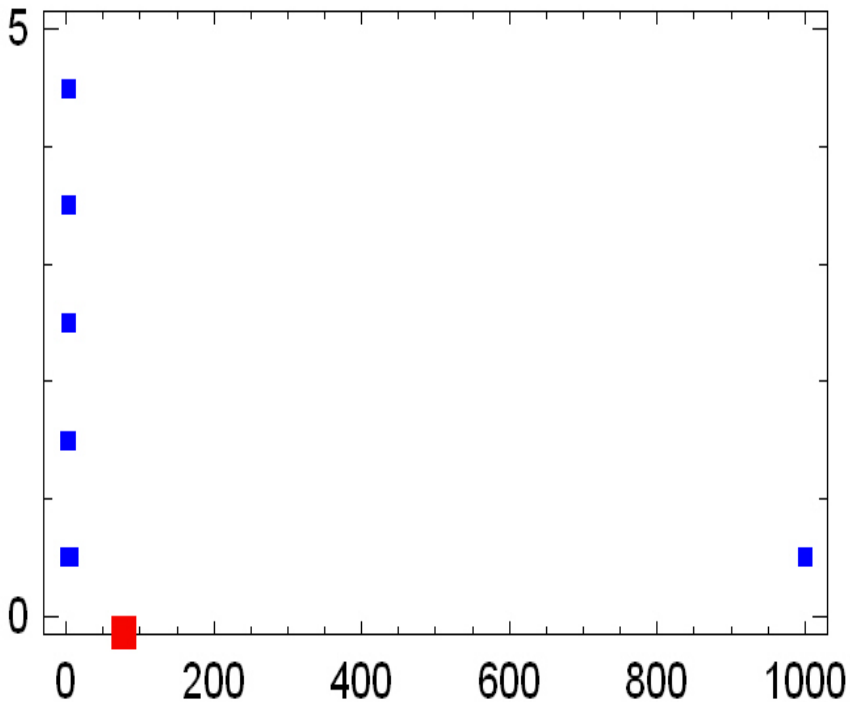


Medidas de centralización III. Ejemplo

En este nuevo diagrama se observa el desplazamiento de la media si al anterior conjunto de datos se le añade un nuevo valor igual a 1000.

El valor de la nueva media es es $\bar{x} = 80'61$.

Obsérvese cómo el dato atípico ha desplazado la media hacia la derecha.



Medidas de centralización IV

- Una medida de centralización más «robusta» frente a la presencia de datos atípicos es la mediana.
- La mediana es un valor que divide el conjunto **ordenado** de los datos en dos grupos con el mismo número de elementos. Así, si los datos ordenados son $\{x_1, \ldots, x_n\}$ y n es impar:

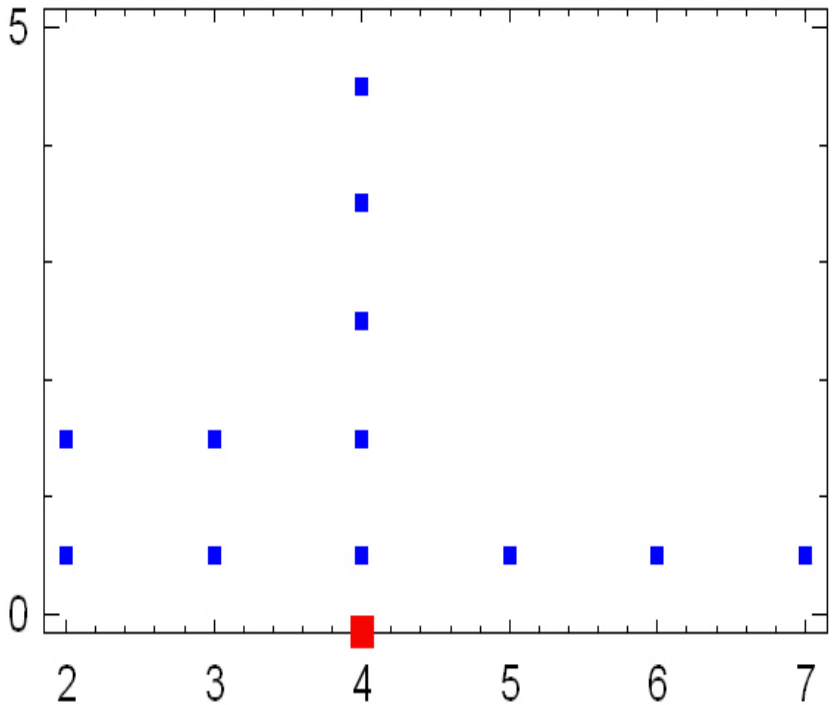
$$Me = x_{\left(\frac{n+1}{2}\right)}$$

Mientras que si n es par:

$$Me = \frac{x_{\frac{n}{2}} + x_{(\frac{n}{2}+1)}}{2}$$

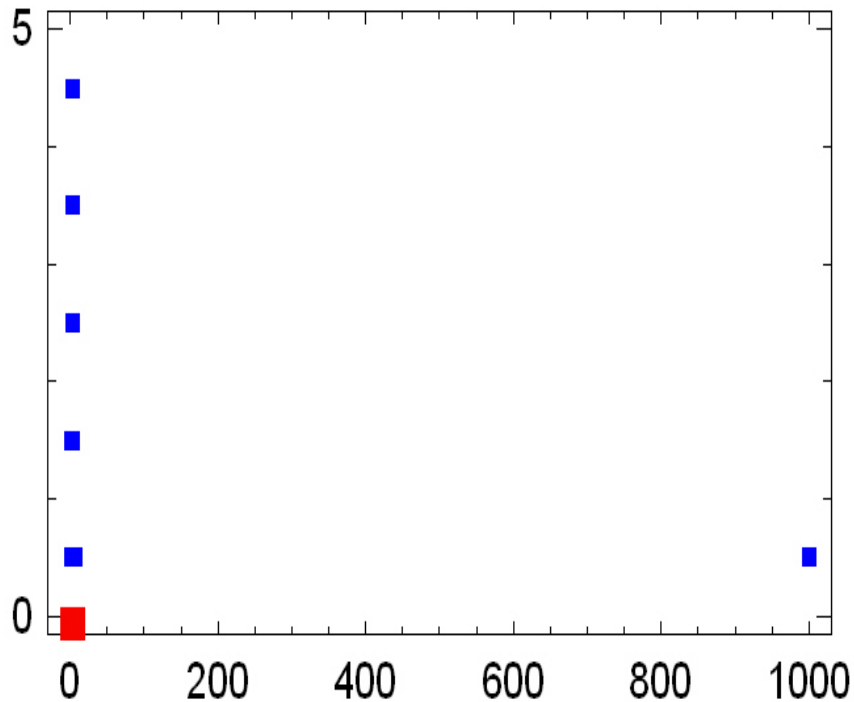
Medidas de centralización V

El siguiente diagrama de puntos muestra la mediana del conjunto de datos formado por los valores:
$\{2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 6, 7\}$
cuyo valor es $Me = 4$.



Medidas de centralización VI

En este nuevo diagrama se observa cómo la mediana permanece inalterable si al anterior conjunto de datos se le añade un nuevo valor igual a 1000.



Medidas de centralización VII

- Otras medidas de centralización son:

  1. **La moda, o clase modal**, que representa el valor, o clase, de mayor frecuencia.
  2. **La media recortada** al α por ciento, que es la media de los datos resultantes de eliminar el $\alpha\,\%$ de los datos extremos por la derecha y por la izquierda.

Medidas de centralización VIII

Observaciones

- Valores próximos de las distintas medidas de centralización es un síntoma de simetría en los datos.
- Cuando no hay valores atípicos la mejor medida de centralización es la media, porque es la que más información emplea en su cálculo.

Medidas de dispersión I

- Una medida de la dispersión de un conjunto de datos, $\{x_1, \ldots, x_n\}$, respecto de la media, es la varianza:

$$s_X^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Obsévese que la varianza es un promedio de los cuadrados de las distancias de todos los datos a la media.

Medidas de dispersión II

Observaciones

- La razón por la que no se emplea como medida de dispersión, lo que sería más natural, el promedio de las diferencias de los datos a la media es que, en cualquier conjunto de datos $\{x_1, \ldots, x_n\}$, se verifica que:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

- La varianza tiene las unidades de los datos al cuadrado.

Medidas de dispersión III

- Para construir una medida de dispersión con las unidades de los datos se define la desviación típica en la forma:

$$s_X = +\sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

La desviación típica tiene las unidades de los datos, y genera una «unidad» de distancia entre los mismos a través de la desigualdad de Chebychef.

La desigualdad de Chebychef I

- **La desigualdad de Chebychef** establece que si un conjunto de datos tiene media $\bar{x}$ y desviación típica s, para todo $k \neq 1$, en el intervalo:

$$(\bar{x} - ks, \bar{x} + ks),$$

se encuentra, **al menos**, el $(1 - \frac{1}{k^2}) \times 100\,\%$ de los datos.

- Como consecuencia, para cualquier conjunto de datos en los intervalos $(\bar{x} - 2s, \bar{x} + 2s)$ y $(\bar{x} - 3s, \bar{x} + 3s)$ se encuentran, como mínimo, el $75\,\%$ ó el $88'88\,\%$ de los datos, respectivamente.

La desigualdad de Chebychef II. Ejemplo

- En la empresa **A** el salario medio anual de los empleados es 35000 euros y la desviación típica 5000 euros.
- En la empresa **B** el salario medio anual de los empleados es 35000 euros y la desviación típica 1000 euros.

¿En cuál de las dos empresas preferiría trabajar?

La desigualdad de Chebychef III. Ejemplo

Para contestar a esta pregunta considere que, según la desigualdad de Chebychef:

- En la empresa **A**, por lo menos el 88′88 % de los empleados tiene un salario comprendido en el intervalo:

$$(20000, 50000) \text{ Euros.}$$

- En la empresa **B**, por lo menos el 88′88 % de los empleados tiene un salario comprendido en el intervalo:

$$(32000, 38000) \text{ Euros.}$$

Comparación de dispersiones I

- Tanto la varianza como la desviación típica carecen de escala.
- Cuando se desea comparar variabilidades entre dos conjuntos de datos conviene tener en cuenta la magnitud de los mismos, no siendo razonable comparar variabilidades de conjuntos de datos muy heterogéneos.

Comparación de dispersiones II

Una medida de dispersión, adimensional, que permite comparar dispersiones es el **coeficiente de variación:**

$$CV_x = \frac{s_X}{|\bar{x}|}$$

Observaciones

- El coeficiente de variación mide cuántas veces contiene la desviación típica de un conjunto de datos a su media.
- También se emplea como medida de dispersión el inverso del coeficiente de variación, denominado **coeficiente señal ruido:**

$$CSR_x = \frac{|\bar{x}|}{s_X}$$

Otras medidas de dispersión

Observaciones

1. Frecuentemente se emplean como medidas de dispersión la varianza muestral corregida y la desviación típica muestral corregida, cuyas expresiones son, respectivamente:

$$\hat{s}_X^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad \text{y} \quad \hat{s}_X = +\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

2. Muchos paquetes estadísticos y calculadoras manuales ofrecen estos parámetros directamente. El cálculo de cualquiera de ellos, conocido otro cualquiera, no ofrece dificultad.

3. Cuando el número de datos es muy grande, los valores de los parámetros y de los parámetros corregidos son muy próximos.

Percentiles I

- Los percentiles son medidas numéricas que aportan información tanto sobre la concentración de los datos, como sobre su dispersión.
- Los percentiles son medidas de orden. (Como la mediana.)
- A través de los percentiles se genera un criterio para identificar puntos atípicos.

Percentiles II

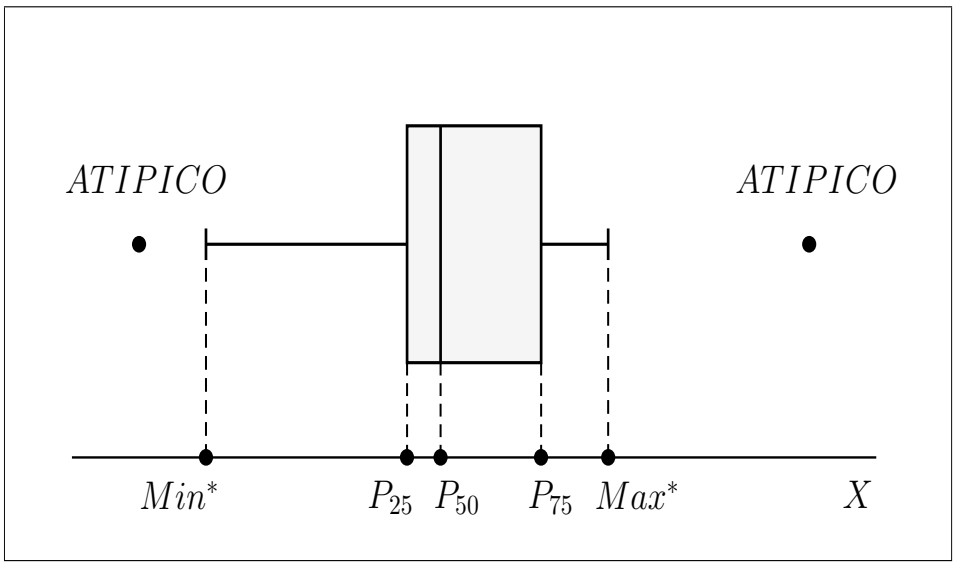
Definición

Se denomina **percentil** α de un conjunto ordenado de datos, al menor dato que es mayor o igual que el $\alpha\,\%$ de todos ellos, se representa por P_α.

- La mediana de un conjunto de datos es el percentil 50.
- Los percentiles 25, 50, y 75 conforman los **cuartiles**, y junto con el mínimo y el máximo dividen a los datos en cuatro grupos que contienen, cada uno de ellos, el 25 % de los mismos.

El diagrama de cajas I. Puntos atípicos

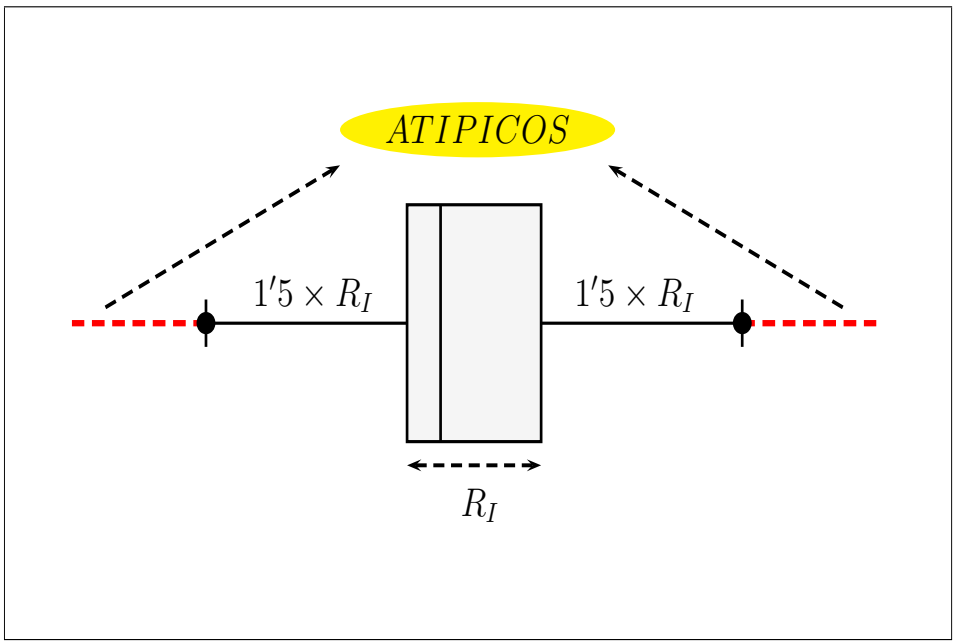
Un diagrama de caja y bigotes representa los valores de los cuartiles, del máximo y el mínimo de los datos no atípicos (Min^* y Max^*), así como los valores atípicos, segun el criterio del **rango intercuartílico**, como se muestra en la figura.



El diagrama de cajas II. Puntos atípicos

Se denomina **rango intercuartílico** a la diferencia: $R_I = P_{75} - P_{25}$.

Esta figura muestra cómo el criterio del rango intercuartílico considera atípicos aquellos valores que sea alejan del P_{25}, o del P_{75}, más de $1'5 \times R_I$, por la izquierda o por la derecha, respectivamente.



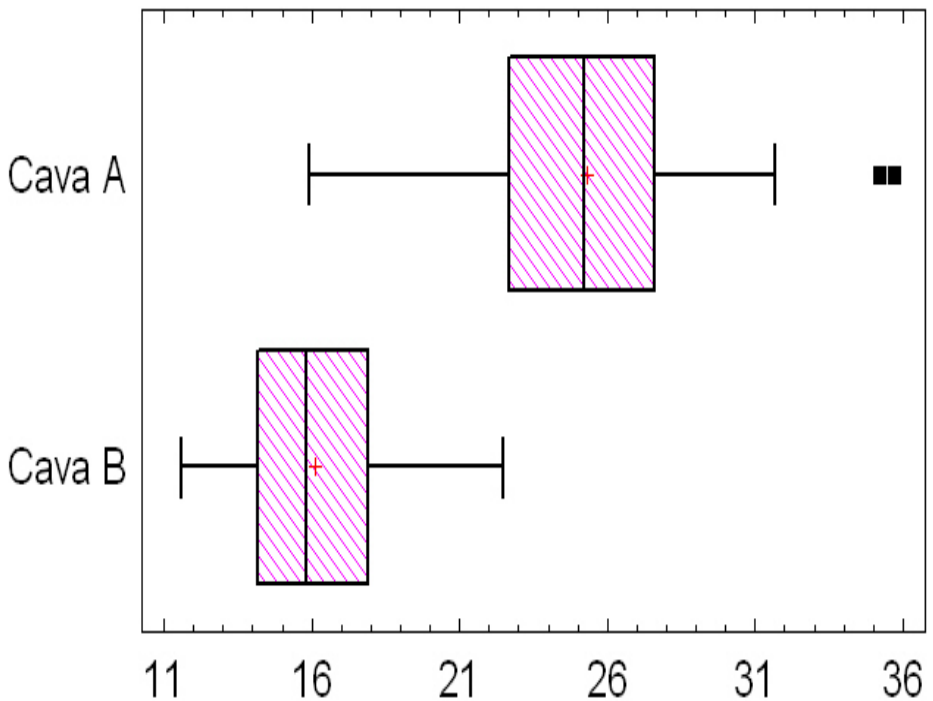
El diagrama de cajas III

- Los diagramas de caja informan sobre:
  1. La simetría de los datos.
  2. La concentración de los datos.
  3. La dispersión.
  4. La presencia de puntos atípicos.

El diagrama de cajas IV

Una aplicación del diagrama de cajas es la comparación de variables.

Por ejemplo, en el diagrama de cajas múltiple adjunto se compara la altura, en milímetros, de las capas de gas producidas al escanciar 37 muestras de dos clases diferentes de cava.



El coeficiente de asimetría

- El coeficiente de asimetría mide la simetría de los datos con respecto de la media. Se define por medio de la expresión:

$$As = \frac{\sum(x_i - \bar{x})^3}{ns_X^3}$$

1. Este coeficiente no tiene unidades.
2. Cuando la simetría es perfecta respecto de la media, $As = 0$.

El coeficiente de curtosis I

- El coeficiente de curtosis mide el apuntamiento de los datos. Se define por medio de la expresión:

$$K = \frac{\sum (x_i - \bar{x})^4}{n s_X^4}$$

1. Este coeficiente no tiene unidades.
2. Cuando el apuntamiento es: $K = 3$, la distribución de los datos se denomina mesocúrtica.
3. Cuando el apuntamiento es: $K > 3$, la distribución de los datos se denomina leptocúrtica.
4. Cuando el apuntamiento es: $K < 3$, la distribución de los datos se denomina platicúrtica.

El coeficiente de curtosis II

Observaciones

Distintos autores y programas definen el coeficiente de curtosis como el valor $K - 3$, lo que explica que puedan obtenerse coeficientes de apuntamiento negativos.

Variables bidimensionales I

En ocasiones se desea estudiar conjuntamente el comportamiento de dos variables, **variable bidimensional**. Por ejemplo la estatura, E, y el peso, P, de un conjunto de individuos. En ese caso, los datos disponibles forman un conjunto de pares.

	E	P
Individuo 1	E_1	P_1
Individuo 2	E_2	P_2
$\vdots$	$\vdots$	$\vdots$
Individuo n	E_n	P_n

Variables bidimensionales II

Para describir conjuntamente una variable bidimensional se emplea una tabla bidimensional, que en la casilla (i, j) dispone la frecuencia absoluta, o relativa, de los individuos que en la primera variable toman el valor i y en la segunda el valor j.

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}
	$\vdots$				$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}
	$\vdots$				$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}

Variables bidimensionales III

En el caso en que las variables X e Y sean continuas, éstas se suelen agrupar en clases.
Por ejemplo en la siguiente tabla se resumen las frecuencias absolutas de la variable bidimensional estatura peso, (E, P), en metros y en kilos de un conjunto de 100 individuos:

	$P \leq 50$	$50 < P \leq 70$	$70 < P \leq 90$	$90 < P$
$E \leq 1'5$	2	1	1	0
$1'5 < E \leq 1'65$	2	7	25	6
$1'65 < E \leq 1'8$	0	6	15	5
$1'8 < E \leq 1'95$	1	4	12	4
$E > 1'95$	0	1	2	6

Variables bidimensionales IV. Variables marginales

- La presentación de una tabla bidimensional permite obtener tablas de las variables unidimensionales correspondientes, sumando las filas o columnas según convenga.

- Estas variables se suelen denominar marginales, porque habitualmente se presentan en los márgenes de la tabla bidimensional.

Variables bidimensionales V. Variables marginales

Por ejemplo, en el margen derecho de esta tabla se representa la distribución de frecuencias de la variable X.

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k	
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}	$\sum_{s=1}^{k} f_{1s}$
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}	$\sum_{s=1}^{k} f_{2s}$
	$\vdots$				$\vdots$	$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}	$\sum_{s=1}^{k} f_{is}$
	$\vdots$				$\vdots$	$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}	$\sum_{s=1}^{k} f_{rs}$

Variables bidimensionales VI. Variables marginales

Similarmente para la variable Y, en el margen inferior,

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}
	$\vdots$		$\vdots$		$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}
	$\vdots$		$\vdots$		$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}
	$\sum_{s=1}^{r} f_{s1}$	$\cdots$	$\sum_{s=1}^{r} f_{sj}$	$\cdots$	$\sum_{s=1}^{r} f_{sk}$

Variables bidimensionales VII. Variables condicionadas

Si se observan los distintos valores de la variable X para un valor fijo de la variable Y, Y_j, se obtiene la distribución de X **condicionada** a $Y = Y_j$. Esta variable es unidimensional. En la columna marcada aparecen las frecuencias **absolutas** de la misma.

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}
	$\vdots$				$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}
	$\vdots$				$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}

Variables bidimensionales VIII. Variables condicionadas

Similarmente, la distribución de frecuencias **absolutas** de la variable Y, condicionada por el valor i de la variable X vendría dada por la fila:

	Y_1	$\cdots$	Y_j	$\cdots$	Y_k
X_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1k}
X_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2k}
	$\vdots$		$\vdots$		$\vdots$
X_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ik}
	$\vdots$		$\vdots$		$\vdots$
X_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rk}

Variables bidimensionales IX

Observaciones

- El cálculo de las frecuencias relativas de la variable X condicionada por el valor Y_j de la variable Y requiere dividir las frecuencias absolutas por la suma $\sum_{i=1}^{r} f_{ij}$, mientras que para calcular las frecuencias relativas de la variable Y condicionada por el valor X_i de la variable X se han de dividir las frecuencias absolutas por la suma $\sum_{j=1}^{k} f_{ij}$.

- El interés habitual, cuando se estudian variables bidimensionales, consiste en analizar la posible relación de dependencia entre las variables unidimensionales. Éste será el objeto del capítulo de la asignatura dedicado a la Regresión lineal.