ESTADÍSTICA

INTRODUCCIÓN

ESTADÍSTICA

TEMA 0. Introducción

¿Qué es la Estadística?

Estadística es la ciencia de:

- Recolectar
- Describir
- Organizar
- Interpretar

} **Datos**

con el fin de transformar dichos datos en información y conseguir una toma de decisiones más eficiente.

ESTADÍSTICA

¿Quienes usan la estadística?

- Investigadores: científicos, ingenieros, ...
- Médicos
- Administradores.
- Organismos oficiales.
- Diarios y revistas.
- Marketing.
- Políticos.
- Deportes.
- etc.

ESTADÍSTICA

La Estadística en la formación de un Ingeniero

- *"En general, la aplicación de técnicas estadística puede considerarse como uno de los 20 desarrollos científicos más significativos del siglo XX, por su impacto sobre nuestra forma de vida y sobre nuestra forma de conocernos y conocer el mundo que nos rodea". (Science)*
- *"No hay conocimiento que pueda contribuir tanto a mejorar la calidad, productividad y competitividad de una empresa como el de los métodos estadísticos". (W.E. Deming)*
- *"Las herramientas estadísticas básicas deben ser conocidas y utilizadas por todo el mundo en una empresa, desde la alta gerencia a los operarios en las líneas". (Ishikawa)*

ESTADÍSTICA

La Estadística en la formación de un Ingeniero

- *"Los conocimientos estadísticos constituyen un aporte esencial para el planteamiento y la resolución de muchos problemas, pero la estadística es mucho más efectiva cuando se combinan con el apropiado conocimiento del tema al que se aplican, en definitiva, la estadística es una herramienta muy útil, pero no es un sustituto de la destreza natural del investigador". (Box, Hunter & Hunter)*
- *Como ejemplo, la multinacional Ford recoge entre los 14 principios operativos que inspiran toda la actividad de la compañía los dos siguientes:*

  *"**7**-Proporcionar al personal directivo un amplio conocimiento y sentido de los métodos estadísticos, dado que estos constituyen herramientas poderosas para determinar las medidas a adoptar para una mejora continua".*

  *"**8**-Proporcionar como mínimo formación básica en Estadística a todos los empleados".*

ESTADÍSTICA

La Estadística en la formación de un Ingeniero

- La responsabilidad básica de un ingeniero es la de liderar la mejora continua de la calidad y de la productividad en todos los procesos que dependan de él. Para ello es necesario cambiarlos y estos cambios son el fruto del análisis de datos. Cuestiones que se plantean: **¿cómo generar datos que contengan información relevante?** y **¿cómo extraer dicha información de los datos?**.
- Entre las áreas sobre las cuales puede tener más impacto la Estadística en los próximos años destacamos las siguientes:
  - Mejora de la calidad y de la productividad.
  - Recogida y uso de información para la agricultura, la industria y la Administración.
  - Integración de la Estadística en la planificación empresarial
  - Desarrollo de nuevos productos y todos el proceso de la innovación.

ESTADÍSTICA

Partes de la Estadística

- *Estadística Descriptiva:* Métodos de organizar, resumir y presentar los datos de manera informativa.
- *Probabilidad:* Estudio de los fenómenos aleatorios.
- *Inferencia Estadística:* Métodos usados para emitir conclusiones acerca de una *población*, basándose en los datos de una *muestra.*

  - *Población* es el conjunto de individuos objeto de estudio.

  - *Muestra* es un subconjunto de la población de interés.

ESTADÍSTICA

Algunos Ejemplos

- ¿Cuál es el número de llamadas telefónicas recibidas en una centralita durante un día? No existe un número fijo que pueda ser conocido a priori, s ino un conjunto de posibles valores, cada uno de ellos con un cierto grado de certeza.
- ¿Cuál es el tamaño de un paquete de información que se transmite a través de HTTP? No existe un número fijo, sino que éste es desconocido a priori.
- ¿Cuál es la posición de un objeto detectado mediante GPS? Dicho sistema transmite una estimación de dicha posición, pero existen márgenes de error que determinan una región del plano donde el objeto se encuentra con alta probabilidad.
- ¿Qué ruido se adhiere a una señal que se envía desde un emisor a un receptor? Dependiendo de las características del canal, dicho ruido será más o menos relevante. Su presencia deberá ser diferenciada de la señal primitiva, sin que se conozca ésta, teniendo en cuenta que se trata de un ruido aleatorio.
- ¿Cuál fue el programa de televisión más visto la pasada noche? Los índices de audiencia se obtienen mediante estimaciones a partir de muestras representativas.

ESTADÍSTICA

TEMA 1. ESTADÍSTICA DESCRIPTIVA

ESTADÍSTICA

TEMA 1. Estadística Descriptiva

OBJETIVO:

- Resumir la información contenida en un conjunto de datos, usando para ello **métodos gráficos** y **medidas numéricas** que informan de lo más relevante.

- Un dato puede consistir en un solo número {58}, en un par de números {(1.66, 58)}, una terna {(1.66, 58, M)}, etc.

ESTADÍSTICA

1.1. Primeros pasos

Cuando disponemos de un conjunto de datos, debemos identificar:

1. La **característica** que representan dichos datos (**variable**).
2. La **población** de la que proceden los datos (conjunto total de individuos de interés).
3. La **naturaleza** de los datos:

   3.1. Variables cualitativas o atributos: Expresan una cualidad y no un valor numérico. Ejemplos: Sexo, Nacionalidad, Marcas de coche, Grado de Satisfacción con la Universidad, etc..

   3.2. Variables cuantitativas: Toma valores numéricos

   a) *Cuantitativas Discretas* sólo pueden asumir ciertos valores y normalmente hay huecos entre ellos. Son conteos normalmente. Ejemplos: nº de asignaturas aprobadas, cantidad de hijos.

   b) *Cuantitativas Continuas* puede asumir cualquier valor dentro de un intervalo. Normalmente representan magnitudes como longitud, superficie, volumen, peso, tiempo, dinero.

ESTADÍSTICA

Formas de presentar y resumir la información de un conjunto de datos:

A) Tabla de frecuencias

- A.1) Datos no agrupados
- A.2) Datos agrupados

B) Descripción gráfica

- B.1) Gráficos para v. cualitativas o cuantitativas discretas
- B.2) Gráficos para v. cuantitativas continuas
- B.3) Diagramas acumulados
- B.4) Gráfico temporal

C) Descripción numérica

- C.1) Medidas de localización o centralización
- C.2) Medidas de dispersión o variabilidad
- C.3) Medidas de forma

ESTADÍSTICA

A) Tabla de Frecuencias

Intentan resumir la información recogida en la muestra, de forma que no se pierda nada de información (o poca).

- Frecuencias absolutas: Contabilizan el número de individuos de cada modalidad o **clase.**
- Frecuencias relativas (porcentajes): Es el cociente entre la frecuencia absoluta y el número total de datos. Contabilizan el porcentaje de individuos de cada modalidad.
- Frecuencias acumuladas: Contabilizan el número de individuos que toman un valor menor o igual que el dado en una modalidad. Sólo tienen sentido para variables cuantitativas (numéricas)
- **Ejemplos** de tablas de frecuencias para datos cualitativos y para datos cuantitativos discretos (transparencia 1)

ESTADÍSTICA

EJEMPLO

- ¿Cuántos individuos tienen menos de 2 hijos?
  - **frec. indiv. sin hijos +frec. indiv. con 1 hijo = 419 + 255 = 674**
- ¿Qué porcentaje de individuos tiene 6 hijos o menos?
  97,3%
- ¿Qué cantidad de hijos es tal que al menos el 50% de la muestra tiene una cantidad inferior o igual?
  2 hijos

Número de hijos

	Frec.	Porcent. (válido)	Porcent. acum.	
0	419	27,8	27,8	
1	255	16,9	44,7	
2	375	24,9	69,5	>50%
3	215	14,2	83,8	
4	127	8,4	92,2	
5	54	3,6	95,8	
6	24	1,6	97,3	
7	23	1,5	98,9	
Ocho+	17	1,1	100,0	
Total	1509	100,0		

La Tabla 2.1 presenta un ejemplo de una distribución de frecuencias para una variable cualitativa: se indican las clases o atributos y sus frecuencias observadas. Cuando los atributos no corresponden a una escala ordinal (por ejemplo alto, medio, bajo) conviene ordenarlos por su frecuencia de aparición.

TABLA 2.1. *Distribución de defectos en libros en una imprenta*

Clases	*Frecuencia*	*Frecuencia relativa*
Corte de las hojas	60	0,43
Mala impresión	40	0,29
Tinta irregular	20	0,14
Encuadernación	12	0,09
Portada	6	0,04
Lomo	2	0,01
TOTAL	140	1

La Tabla 2.2 presenta esta misma idea para una variable discreta. Esta representación es útil cuando el número de valores posibles es pequeño. En otro caso, conviene agrupar los datos, como indicamos a continuación.

TABLA 2.2. *Distribución de frecuencias de la variable: número de llamadas recibidas en una centralita en períodos de un minuto*

X	*(f)* *frecuencia*	*(fr)* *frecuencia relativa*
0	40	0,44
1	26	0,29
2	14	0,16
3	6	0,07
4	3	0,03
5	0	0,00
6	1	0,01
TOTAL	90	1

ESTADÍSTICA

Datos agrupados:

- Para datos cuantitativos continuos, los datos se suelen agrupar en **clases**, que son intervalos que no se solapan y cuya unión cubre todo el rango de los datos.
- Suelen elegirse de la misma longitud, de modo que basta con seleccionar el número de clases a tomar.
- La elección del número de clases puede influir en la posterior interpretación de los datos.
- Una regla empírica, sugiere que el número de clases sea aproximadamente $\sqrt{n}$ donde n = nº total de datos.
- **Ejemplos** de tablas de frecuencias para datos cuantitativos agrupados en clases (transparencia 2).

ESTADÍSTICA

Distribución del peso (en Kg) de una muestra de 500 alumnos varones de una Universidad

peso	f_i	h_i	%	F_i	H_i	% acumulado
Menos de 45	1	0.002	0.20	1	0.002	0.2
[45 - 50)	3	0.006	0.60	4	0.008	0.8
[50 - 55)	12	0.024	2.40	16	0.032	3.2
[55 - 60)	75	0.150	15.00	91	0.182	18.2
[60 - 65)	103	0.206	20.60	194	0.388	38.8
[65 - 70)	155	0.310	31.00	349	0.698	69.8
[70 - 75)	101	0.202	20.20	450	0.900	90.0
[75 - 80)	29	0.058	5.80	479	0.958	95.8
[80 - 85)	11	0.022	2.20	490	0.980	98.0
[85 - 90)	8	0.016	1.60	498	0.996	99.6
90 o más	2	0.004	0.40	500	1.000	100.0
Total	500	1.000	100			

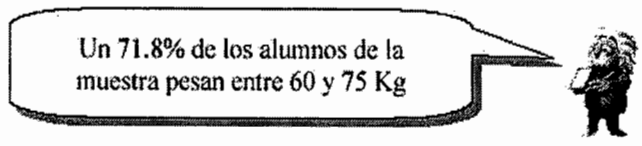


Tabla 1-1 Resistencia a la tensión de 80 muestras de aleación de aluminio-litio

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Tallo	Hoja	Frecuencia
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Figura 1-3 Diagrama de tallo y hoja para los datos de resistencia a la tensión de la tabla 1-1.

Tabla 1-2 Distribución de frecuencias para los datos de resistencia a la tensión, de la tabla 1-1

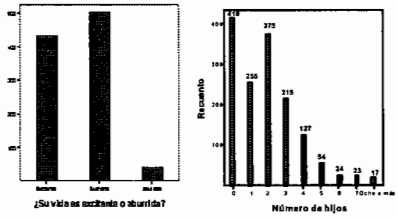
Intervalo de clase (psi)	Conteo	Frecuencia	Frecuencia relativa	Frecuencia relativa acumulativa
$70 \leq x < 90$	||	2	0.0250	0.0250
$90 \leq x < 110$	|||	3	0.0375	0.0625
$110 \leq x < 130$	𝍸 |	6	0.0750	0.1375
$130 \leq x < 150$	𝍸 𝍸 ||||	14	0.1750	0.3125
$150 \leq x < 170$	𝍸 𝍸 𝍸 𝍸 ||	22	0.2750	0.5875
$170 \leq x < 190$	𝍸 𝍸 𝍸 ||	17	0.2125	0.8000
$190 \leq x < 210$	𝍸 𝍸	10	0.1250	0.9250
$210 \leq x < 230$	||||	4	0.0500	0.9750
$230 \leq x < 250$	||	2	0.0250	1.0000

ESTADÍSTICA

B) Descripción Gráfica

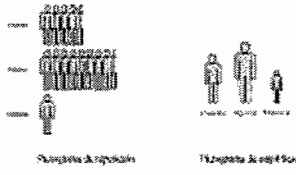
B.1) Gráficos para v. cualitativas o cuantitativas discretas

- Diagramas de barras
  - Alturas proporcionales a las frecuencias (abs. o rel.)



- Diagramas de sectores (tartas, polares)
  - El área de cada sector es proporcional a su frecuencia (abs. o rel.)

- Pictogramas
  - El área de cada modalidad debe ser proporcional a la frecuencia.

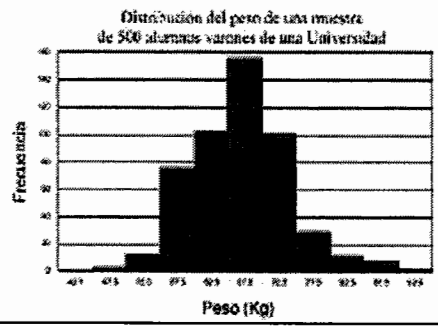


ESTADÍSTICA

B.2) Gráficos para v. cuantitativas continuas

- **Diagrama de puntos**
  - Para conjuntos con menos de 25 datos (transp. 3)
- **Diagrama de tallo-hojas**
  - Para conjuntos de datos de tamaño moderado (transp. 4)
- **Histograma**

  - Para conjuntos con gran número de datos. Es la representación gráfica de la tabla de frecuencias para datos agrupados en clases. El área que hay bajo el histograma entre dos puntos cualesquiera indica la cantidad (porcentaje o frecuencia) de individuos en dicho intervalo.



1-2.1 Diagrama de puntos y diagrama tallo y hoja

Montgomery (1991a) describe un experimento en el que un ingeniero agrega un polímero de látex a un mortero de cemento portland, para determinar los efectos del polímero sobre la resistencia a la tensión (en kgf/cm^2). Los datos obtenidos de este experimento son 16.85, 16.40, 17.21, 16.35, 16.52, 17.04, 16.96, 17.15, 16.59 y 16.57. En la figura 1-1 aparecen

Figura 1-1 Diagrama de puntos para la resistencia a la tensión de un mortero de cemento portland modificado.

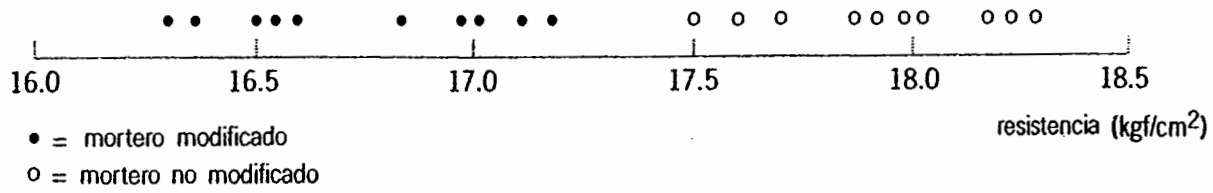


Figura 1-2 Diagrama de puntos para los datos de resistencia a la tensión.

El diagrama de puntos es una gráfica muy útil para visualizar un conjunto pequeño de datos; por ejemplo, de unas 20 observaciones. La gráfica permite ver con rapidez y facilidad la **ubicación** o **tendencia central** de los datos, así como su **dispersión** o **variabilidad**. Por ejemplo, nótese que la parte media de los datos está muy próxima a 16.8, y que los valores de resistencia a la tensión caen dentro del intervalo definido por los valores 16.3 y 17.2 kgf/cm^2.

A menudo, los diagramas de puntos son útiles al comparar dos o más conjuntos de datos. Por ejemplo, los siguientes son diez datos de resistencia a la tensión de un mortero portland sin modificar: 17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22, 17.90, 17.96 y 18.15. El diagrama de puntos de la figura 1-2 muestra los dos conjuntos de mediciones de resistencia a la tensión, donde los puntos sólidos corresponden al mortero modificado, y los círculos, al mortero no modificado. Nótese que el diagrama de puntos revela de inmediato que el mortero modificado parece tener una menor resistencia a la tensión, pero que la variabilidad inherente *a* ambos grupos de mediciones es casi la misma.

La figura 1-5 presenta la gráfica de tallo y hoja de los datos de resistencia a la tensión de la tabla 1-1, producida por el paquete Statgraphics. El software utiliza tallos un poco diferentes de los de la figura 1-2. (El usuario no tiene control sobre este aspecto.) El programa también forma dos categorías adicionales de observaciones: "LO", que contiene los valores extremos más pequeños, 76 y 87, y "HI", con los valores extremos más grandes, 237 y 245. Nótese también que la computadora ordena las hojas de cada tallo de menor a mayor. Esta forma de la gráfica se conoce como **diagrama de tallo y hoja ordenado en pantalla**. Esto no se hace cuando la gráfica se construye de manera manual, ya que es una labor que consume mucho tiempo. La computadora añade una columna en la parte izquierda de los tallos; esta columna proporciona un conteo de las observaciones que están en cada tallo y por encima de él en la mitad superior del diagrama, así como un conteo de las observaciones que están en cada tallo y por debajo de él en la mitad inferior del diagrama. En el tallo de la parte media (16), la columna indica el número de observaciones que corresponden a dicho tallo.

Stem-and-leaf display for strength: unit = 1 1:2 represents 12

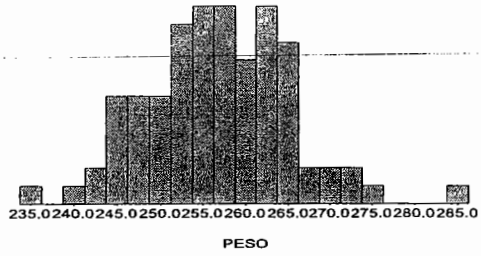
```
                LO : 76, 87

  3              9 : 7
  5             10 : 1 5
  8             11 : 0 5 8
 11             12 : 0 1 3
 17             13 : 1 3 3 4 5 5
 25             14 : 1 2 3 5 6 8 9 9
 37             15 : 0 0 1 3 4 4 6 7 8 8 8 8
(10)            16 : 0 0 0 3 3 5 7 7 8 9
 33             17 : 0 1 1 2 4 4 5 6 6 8
 23             18 : 0 0 1 1 3 4 6
 16             19 : 0 3 4 6 9 9
 10             20 : 0 1 7 8
  6             21 : 8
  5             22 : 1 8 9

                HI : 237, 245
```

Figura 1-5 Diagrama de tallo y hoja generado por Statgraphics.

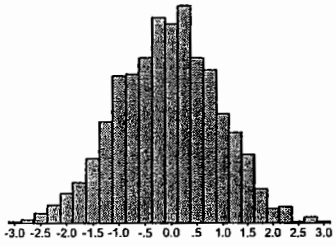
Histograma

Se representan mediante cajas las frecuencias de los resultados agrupados en intervalos.

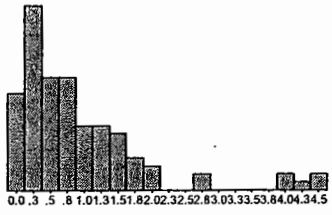
Se utiliza para: **Variables cuantitativas continuas**



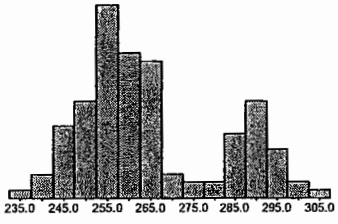
Describe la forma en que se DISTRIBUYEN los datos



normal



asimétrico



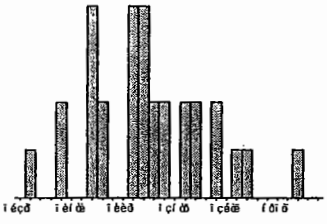
bimodal

Las frecuencias se pueden representar en:

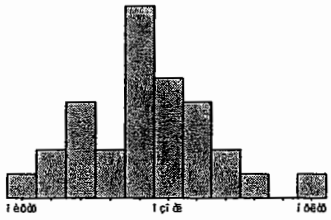
- *valores absolutos en la altura*
- *valores relativos en la altura*
- ***valores relativos en el área***
- *valores acumulados ...*

Consejos:

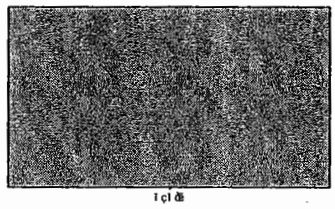
1. **Usar intervalos de la misma longitud**
2. **Los intervalos no pueden solaparse**
3. **Cada observación sólo puede pertenecer a un intervalo**
4. **Todos los datos deben pertenecer a algún intervalo**
5. **La forma del histograma depende de la amplitud del intervalo que se elija.** *Regla: amplitud correspondiente a $\sqrt{n}$ intervalos cubriendo todo el rango de valores*



30 clases



automático



1 clase

Los histogramas pueden proporcionar mucha información respecto a la estructura de los datos: simetría, asimetría, variabilidad, unimodalidad..

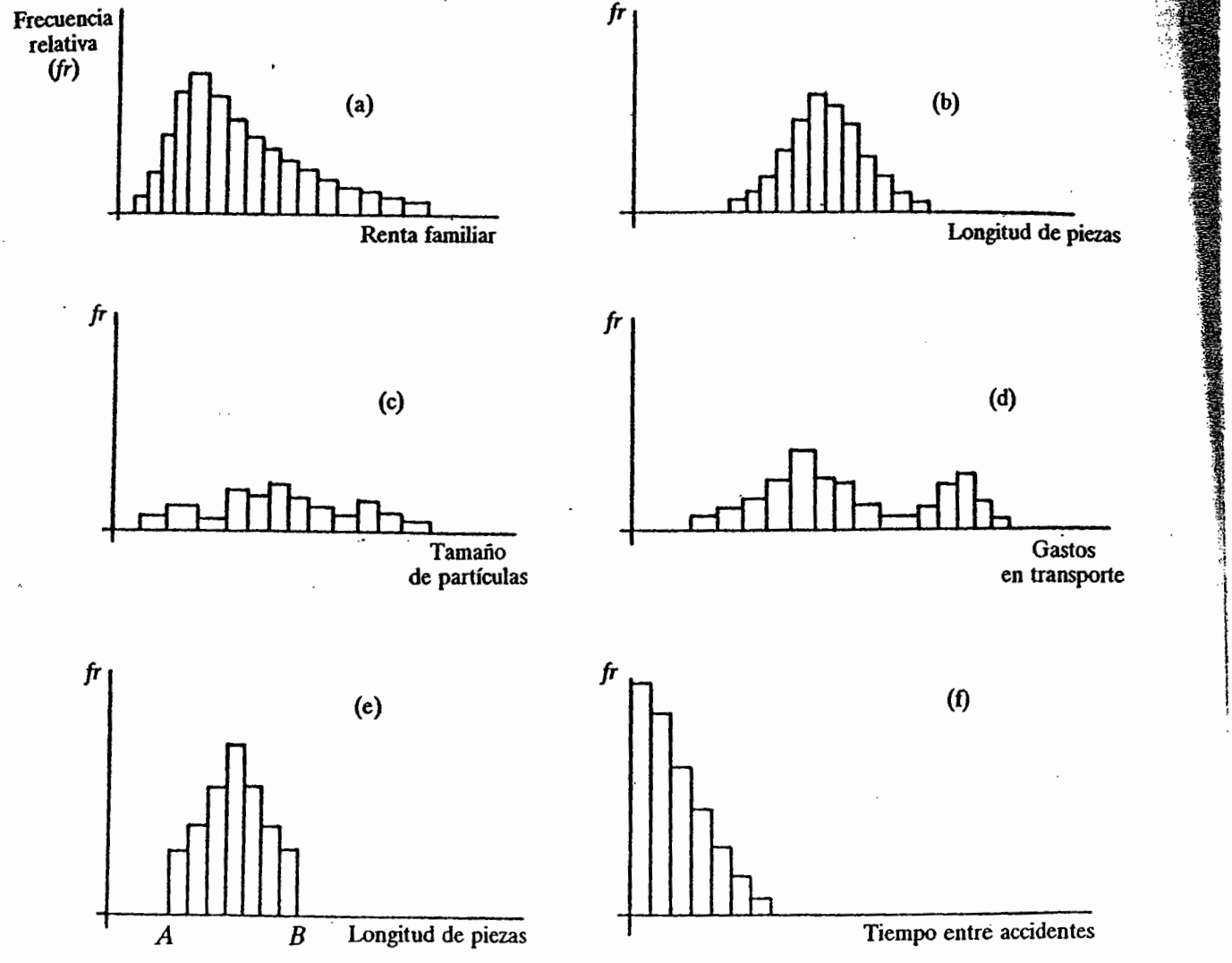
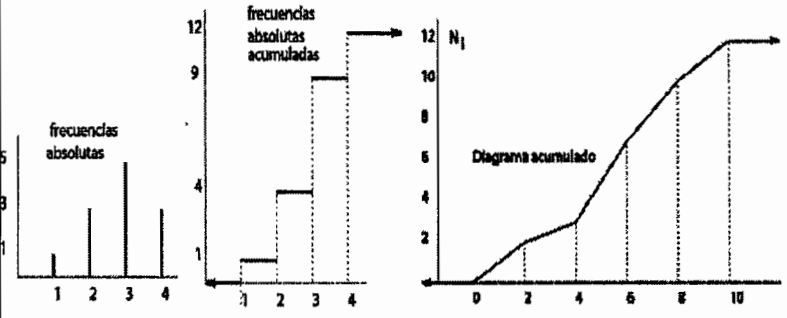


FIGURA 2.4. *Algunos histogramas típicos.*

ESTADÍSTICA

B.3) Diagramas acumulados

Algunos de los diagramas anteriores tiene su correspondiente diagrama acumulado. Se realizan a partir de las frecuencias acumuladas. Indican, para cada valor de la variable, la cantidad (frecuencia) de individuos que poseen un valor inferior o igual al mismo.

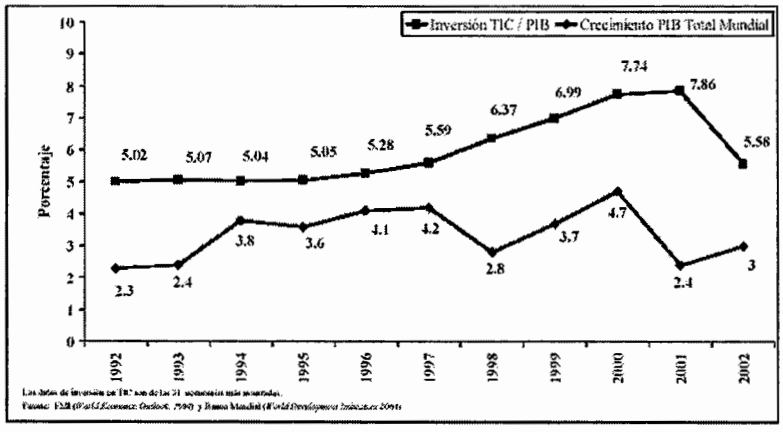


Frec. Absolutas | **Frec. Abs Acumuladas** | **Polígono de Frec. Abs Acumuladas**

ESTADÍSTICA

B.4) Gráfico temporal

Gráfico 1: Evolución del Sector TIC e IMPACTO en PIB Mundial 1992-2002*



- El crecimiento TICs va de la mano con crecimiento PIB.

ESTADÍSTICA

B.4) Gráfico temporal

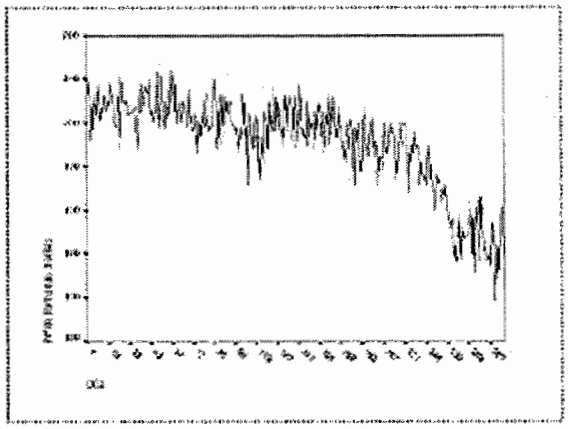


Figura 7.2: Número de llamadas recibidas en un día. Secuencia asociada a 276 días.

ESTADÍSTICA

C) Descripción Numérica

Objetivo: Resumir la información más relevante de la muestra o población en unos pocos números (parámetros).

C.1) Medidas de Centralización o Localización

- Indican valores con respecto a los que los datos parecen agruparse.
  - Media, mediana y moda

C.2) Medidas de Posición

- Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
  - Cuantiles, percentiles, cuartiles, deciles,...

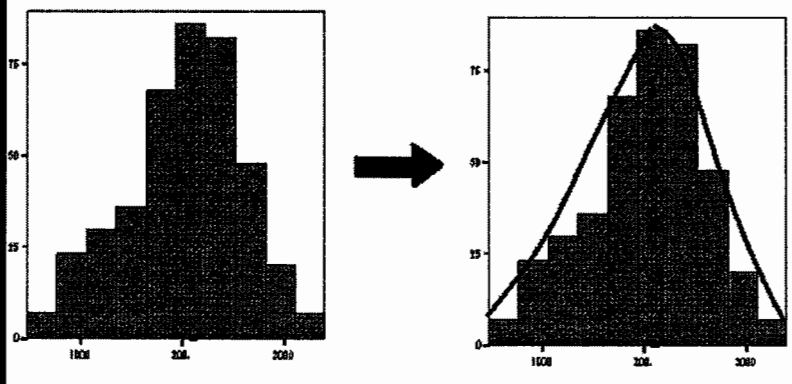
C.3) Medidas de Dispersión o Variabilidad

- Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
  - Rango, varianza, desviación típica, rango intercuartílico, coeficiente de variación

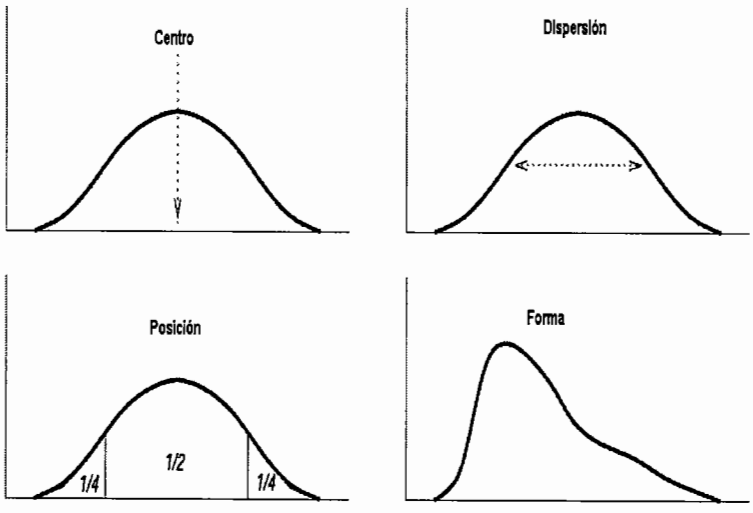
C.4) Medidad de Forma

- Indican la forma en que se distribuyen los datos
  - Coeficientes de asimetría y de apuntamiento o curtosis

ESTADÍSTICA



ESTADÍSTICA



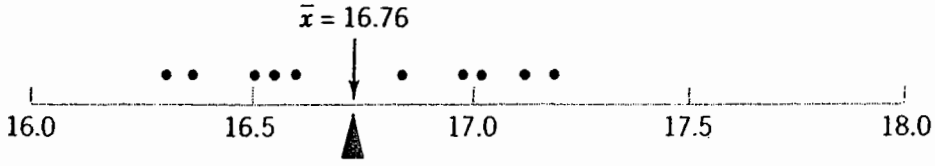


Figura 1-11 La media muestral como punto de equilibrio de un sistema de pesos.

Para los datos de resistencia de la aleación de aluminio-litio de la tabla 1-1, la suma de las 80 observaciones es

$$\sum_{i=1}^{80} x_i = 13\,013$$

de modo que la media muestral es

$$\bar{x} = \frac{\sum_{i=1}^{80} x_i}{80} = \frac{13\,013}{80} = 162.7$$

Si se examina el diagrama de tallo y hoja de la figura 1-3 o el histograma de la figura 1-6, se observa que la media muestral de 162.7 psi es un valor "típico" de la resistencia a la tensión, ya que éste se presenta casi en la parte media de los datos, donde se concentran las observaciones. Sin embargo, esta impresión puede ser errónea. Supóngase que el histograma tiene la apariencia de la figura 1-12. La media de estos datos sigue siendo una medida de

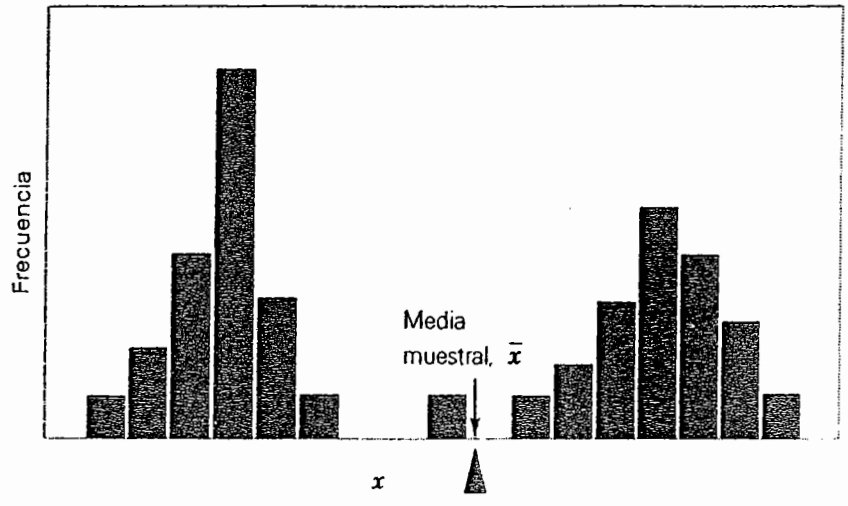


Figura 1-12 Histograma

EJEMPLO 1-6

A continuación se presentan 20 observaciones en orden del tiempo de falla, en horas, de un material aislante eléctrico (adaptadas del trabajo de Nelson, *Applied Life Data Analysis*, 1982):

204	228	252	300	324	444	624	720	816	912
1176	1296	1392	1488	1512	2520	2856	3192	3528	3710

Nótese que la mediana es

$$\tilde{x} = q_2 = \frac{912 + 1176}{2} = 1044$$

lo que corresponde a un valor que está entre la décima y la undécima observaciones. El primer cuartil debe tener al menos el 25% de los datos o por lo menos cinco observaciones en su valor o por debajo de él, y al menos el 75% de los datos o al menos 15 observaciones en su valor o por encima de él. Tanto la quinta como la sexta observación satisfacen esta definición, de modo que se define a q_1 como la media de estas observaciones:

$$q_1 = \frac{324 + 444}{2} = 384$$

De manera similar, el tercer cuartil debe tener al menos el 75% de los datos, o por lo menos 15 observaciones en su valor o por debajo de él, y al menos el 25% de los datos, o por lo menos cinco observaciones, en su valor o por encima de él. Las observaciones 15 y 16 satisfacen esta definición, así que se toma a q_3 como la media de estos valores:

$$q_3 = \frac{1512 + 2520}{2} = 2016$$

1-4 MEDIDAS DE VARIABILIDAD

La localización o tendencia central no necesariamente proporciona información suficiente para describir datos de manera adecuada. Por ejemplo, considérense los datos de resistencia a la tensión (en psi) de dos muestras de aleación de aluminio-litio:

Muestra 1: 130, 150, 145, 158, 165, 140

Muestra 2: 90, 128, 205, 140, 165, 160

La media de ambas muestras es 148 psi. Sin embargo, con respecto al diagrama de puntos de la figura 1-14, se observa que la dispersión o variabilidad de la muestra 2 es mucho mayor que la de la muestra 1.

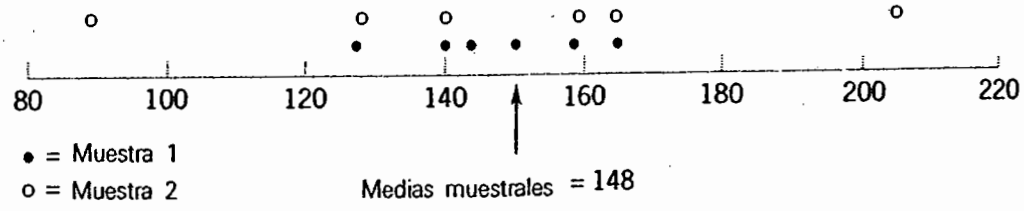


Figura 1-14 Datos de resistencia a la tensión.

EJEMPLO 1-10

Con un micrómetro, se realizan mediciones del diámetro de un balero, que tienen una media de 4.03 mm y una desviación estándar de 0.012 mm; con otro micrómetro se toman mediciones de la longitud de un tornillo, que tienen una media de 1.76 pulgadas y una desviación estándar de 0.0075 pulgadas. Los coeficientes de variación son

$$cv_{\text{balero}} = \frac{0.012}{4.03} = 0.003$$

y

$$cv_{\text{tornillo}} = \frac{0.0075}{1.76} = 0.004$$

respectivamente. En consecuencia, las mediciones hechas con el primer micrómetro exhiben una variabilidad relativamente menor que las efectuadas con el otro micrómetro.

Dpto. Mat. Aplicada y Estadística

Diagrama de caja-bigotes

El diagrama de caja-bigotes es un resumen gráfico que permite visualizar, para un conjunto de datos, la tendencia central, la dispersión y la presencia posible de datos atípicos. Para realizarlo se necesita calcular la mediana, el primer cuartil, y el tercer cuartil de los datos:

La mediana

La mediana es el punto que deja el 50% de los datos a su izquierda y el otro 50% a su derecha. Es una medida de centralización más adecuada que la media en el caso en que la distribución de los datos es asimétrica (lo que se ve en el histograma) o si hay datos atípicos. Si la distribución es simétrica, la media y la mediana coinciden.

Para calcular la mediana de un conjunto de n datos, $x_1, x_2, \ldots, x_n$. Empiezo por ordenar los datos por orden creciente. La mediana es el dato ordenado n° $(n+1)/2$.

Ejemplo: 125, 129, 134, 185, 200. La mediana es el dato ordenado n° 3, y es igual a 134.

11, 15, 20, 23: la mediana es el dato ordenado n° 2.5, que tomamos por convención igual al punto medio entre el dato n° 2 y el dato n°3. En este caso, la mediana es igual a 17.5.

Los cuartiles.

El primer cuartil Q_1 es el punto que deja el 25% de los datos a su izquierda y el otro 75% a su derecha, mientras que el tercer cuartil Q_3 es el punto que deja el 75% de los datos a su izquierda y el otro 25% a su derecha. Por lo tanto el par (Q_1, Q_3) nos proporciona información sobre la dispersión presente en los datos: cuanto más alejados estén los cuartiles, más dispersos están los datos. Por ello, calculamos el rango intercuartílico RIC como la diferencia entre Q_3 y Q_1.

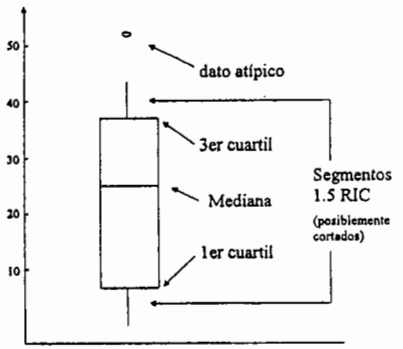
Para calcular los cuartiles, empezamos por calcular la mediana Me de los datos. El primer cuartil es la mediana del grupo de datos que queda a la izquierda de Me (Me excluida), mientras que el tercer cuartil se calcula como la mediana del grupo que queda a su derecha (Me excluida).

Datos atípicos.

Un dato atípico es un dato que se aleja del patrón global de nuestro conjunto. Para detectarlos, se utiliza el RIC : se considera atípico un dato situado fuera del intervalo

$$[Q_1 - 1.5 \times RIC;\ Q_3 + 1.5 \times RIC]$$

El diagrama de caja-bigotes presenta de manera gráfica estas informaciones:

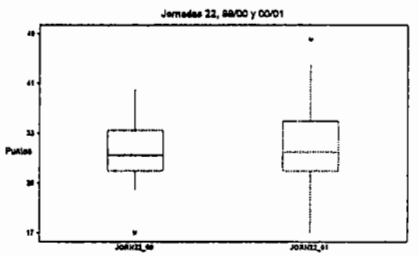


Los segmentos 1.5 RIC (llamados bigotes) se recortan hasta : el dato del conjunto inmediatamente superior a $Q_1 - 1.5 \times RIC$ para el bigote inferior, y el dato inmediatamente inferior a $Q_3 + 1.5 \times RIC$, para el bigote superior.

La mayor utilidad de los diagramas caja-bigotes es para comparar dos o más conjuntos de datos.

Ejemplo

La puntuación de los equipos de la liga española después de la vigésimasegunda jornada en la liga 99/00 y en la liga 00/01 se pueden comparar con un diagrama caja-bigotes:



Comentarios: Hay un dato atípico para cada temporada: en la temporada 99/00, corresponde al Sevilla que se descolgaba del resto de los equipos, mientras que en la temporada 00/01, destaca el R. Madrid que se aleja del patrón general de la liga. Por otra parte el centro de los datos es el mismo. La dispersión es sensiblemente la misma, (los equipos están algo más apelotonados en la liga 99/00). En las dos temporadas, hay más dispersión entre los equipos situados en el segundo cuarto de la tabla, que entre los equipos situados en el tercer cuarto de la tabla. (es decir, Q_3 está más alejado de Me que Q_1)

Clasificación

		TOTAL						CASA						FUERA			
		P	PL	F	C	G	E	P	F	C	PL	G	E	P	F	C	PL
1	R.Sociedad	22	44	38	27	6	4	0	19	12	22	6	4	2	19	15	22
2	R.Madrid	22	42	46	23	7	3	0	29	10	24	4	6	2	17	13	18
3	Valencia	22	42	36	17	8	2	2	22	6	26	4	4	2	14	11	16
4	Deportivo	22	39	31	25	5	3	2	15	10	18	6	3	3	16	15	21
5	Celta	22	36	28	20	6	3	3	17	9	21	5	0	5	11	11	15
6	Betis	22	36	35	30	7	2	3	19	12	23	3	4	3	16	18	13
7	Atlético	22	32	35	28	6	4	2	20	12	22	2	4	4	15	16	10
8	Athletic	22	29	34	37	6	2	4	18	13	20	2	3	5	16	24	9
9	Alavés	22	28	27	33	5	4	2	14	8	19	2	3	6	13	25	9
10	Villareal	22	28	22	24	5	2	3	12	10	17	2	5	5	10	14	11
11	Barcelona	22	27	31	30	4	4	2	20	14	16	3	2	7	11	16	11
12	Málaga	22	27	28	28	5	4	2	19	12	19	1	5	5	9	16	8
13	Osasuna	22	27	24	27	6	1	5	13	11	19	1	5	4	11	16	8
14	Valladolid	22	27	21	25	5	2	4	13	12	17	3	1	7	8	13	10
15	Mallorca	22	27	25	37	3	2	5	10	20	11	5	1	6	15	17	16
16	Sevilla	22	26	17	18	3	5	2	10	7	14	3	3	6	7	11	12
17	Racing	22	26	26	29	5	2	4	16	10	17	3	0	8	10	19	9
18	Español	22	22	24	33	5	2	5	18	16	17	1	2	7	6	17	5
19	Rayo	22	22	22	32	3	3	5	14	15	12	3	1	7	8	17	10
20	Recreativo	22	15	18	45	2	4	5	12	19	10	1	2	8	6	26	5

CHAMPIONS UEFA DESCIENDEN

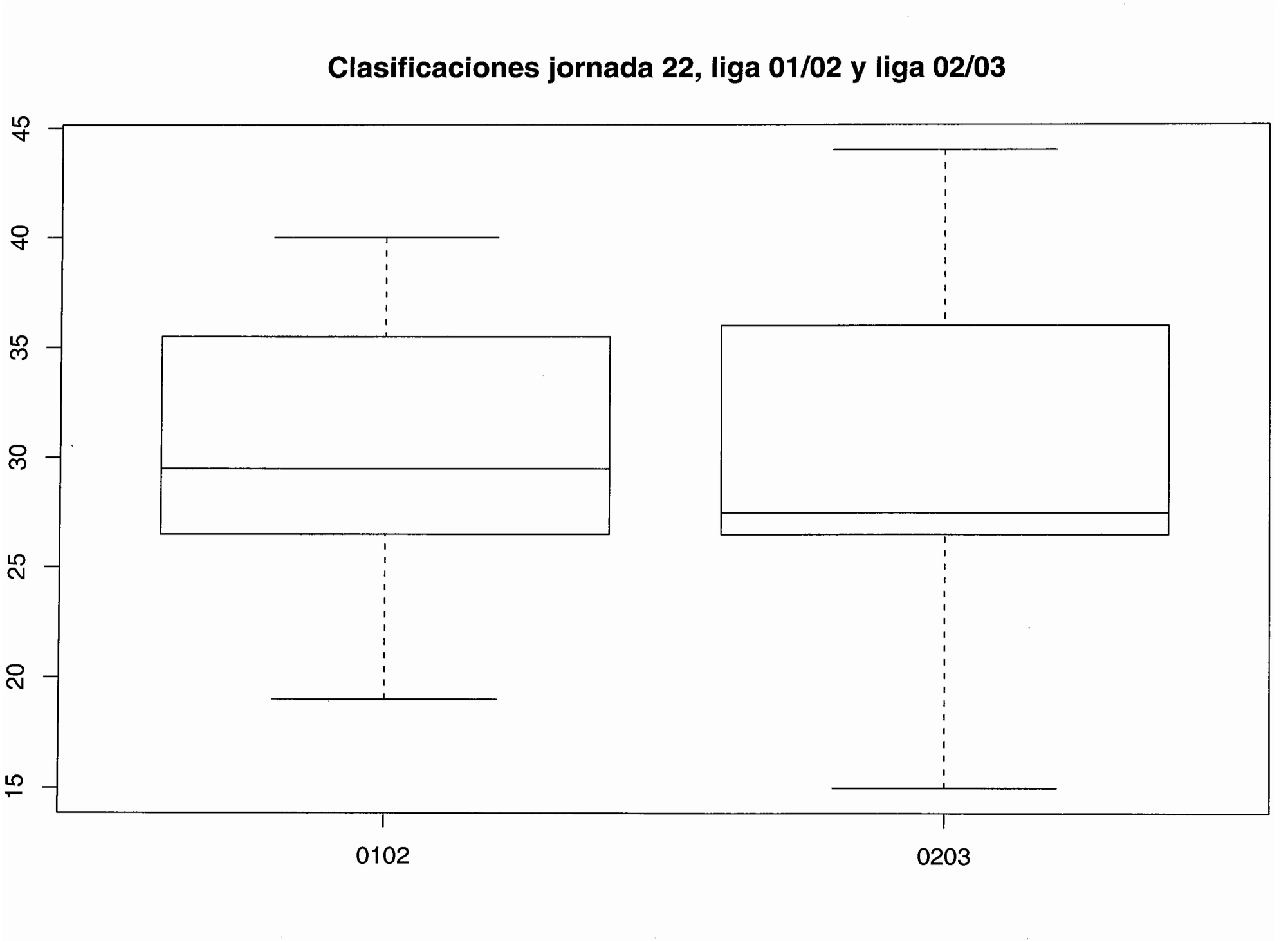

Clasificaciones jornada 22, liga 01/02 y liga 02/03
45
40
35
30
25
20
15
0102
0203

TABLA 2.5. *Relación entre color de los ojos de padres e hijos en una muestra de 1.000 personas*

Hijo	*Padre*	
	Claro	*Oscuro*
Claro	0,25	0,08
Oscuro	0,12	0,55

TABLA 2.6. *Relación entre averías mensuales de una máquina y temperatura media de funcionamiento*

Averías	*Temperatura*			
	40°	*50°*	*60°*	
2	0,20	0,15	0,10	0,45
3	0,12	0,07	0,05	0,24
4	0,04	0,10	0,02	0,16
5	—	0,05	0,10	0,15
	0,36	0,37	0,27	1

En el interior de cada casilla (x_i, y_j) aparece la frecuencia relativa $fr(x_i, y_j)$ correspondiente a los dos valores que definen la casilla. Por tanto:

$$\sum_i \sum_j fr(x_i, y_j) = 1$$

Esta misma idea puede extenderse para cualquier número de variables, aunque la representación gráfica no sea posible para más de tres.

TABLA 2.6. *Relación entre averías mensuales de una máquina y temperatura media de funcionamiento*

Averías	*Temperatura*			
	40°	*50°*	*60°*	
2	0,20	0,15	0,10	0,45
3	0,12	0,07	0,05	0,24
4	0,04	0,10	0,02	0,16
5	—	0,05	0,10	0,15
	0,36	0,37	0,27	1

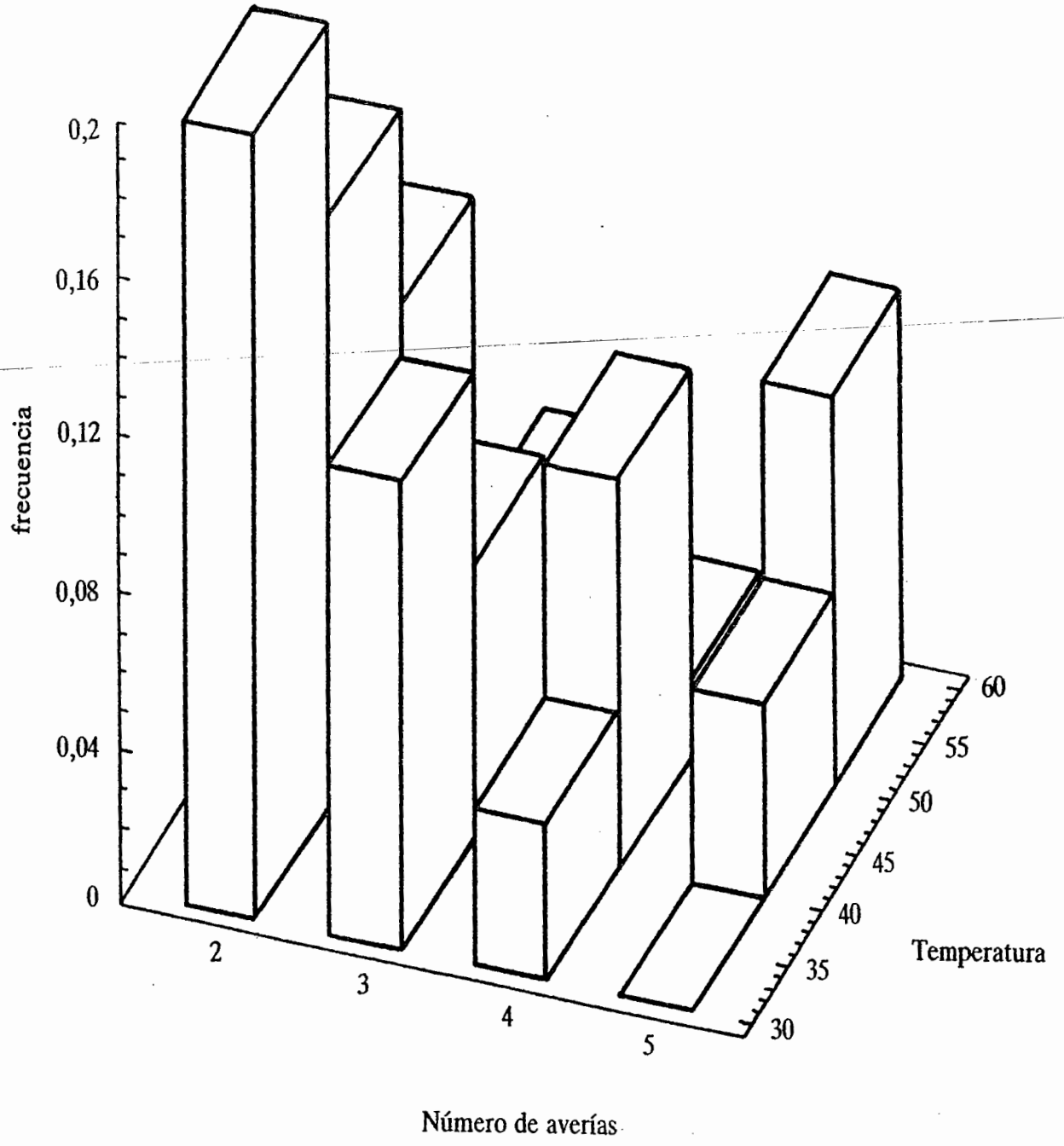


FIGURA 2.19. *Histograma tridimensional de los datos de la tabla 2.6.*

4.2. Representaciones gráficas

La representación gráfica más útil de dos variables continuas sin agrupar es el diagrama de dispersión, que se obtiene representado cada observación bidimensional $(x_i\, y_i)$ como un punto en el plano cartesiano. Este diagrama es especialmente útil para indicar si existe o no relación entre las varibles. La figura 2.18 presenta algunos ejemplos.

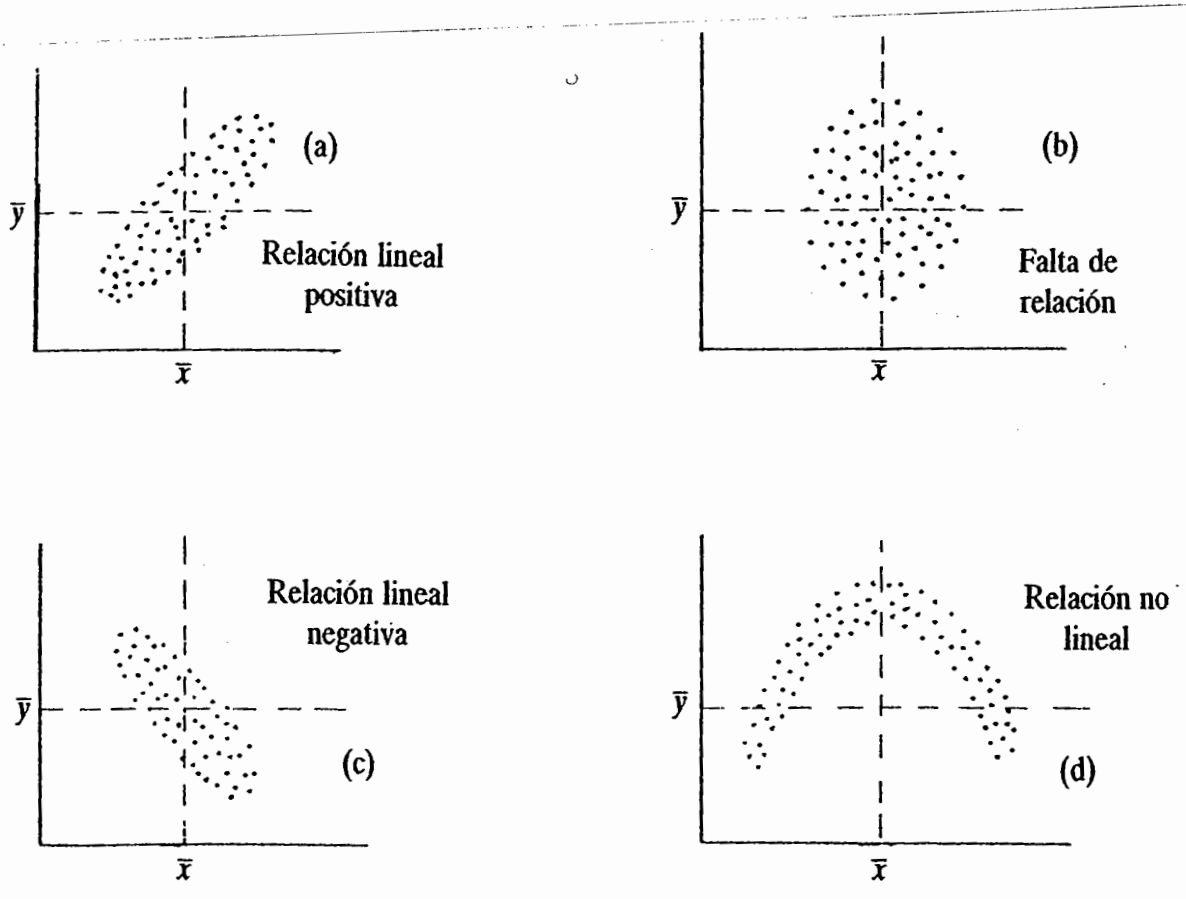
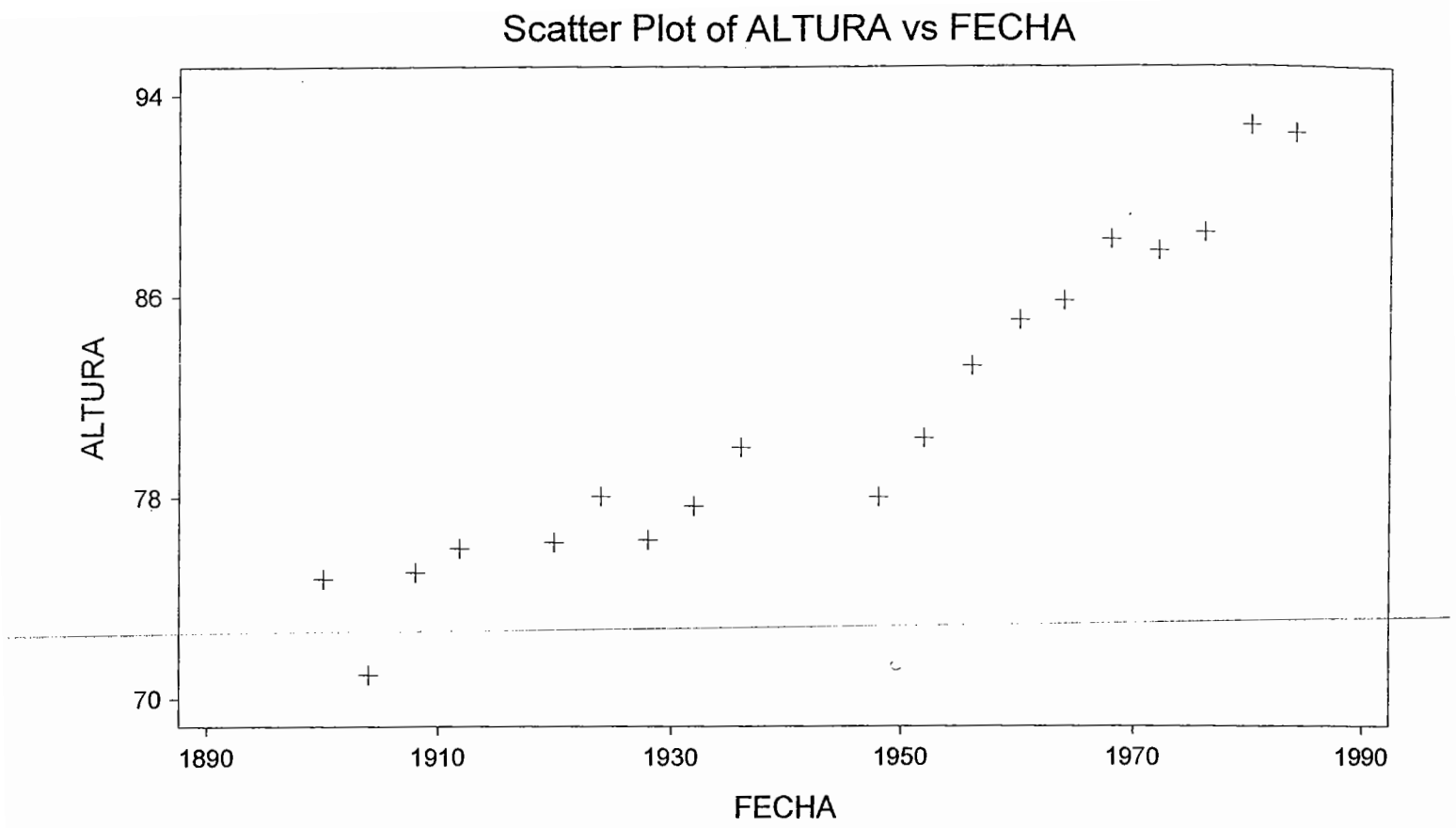
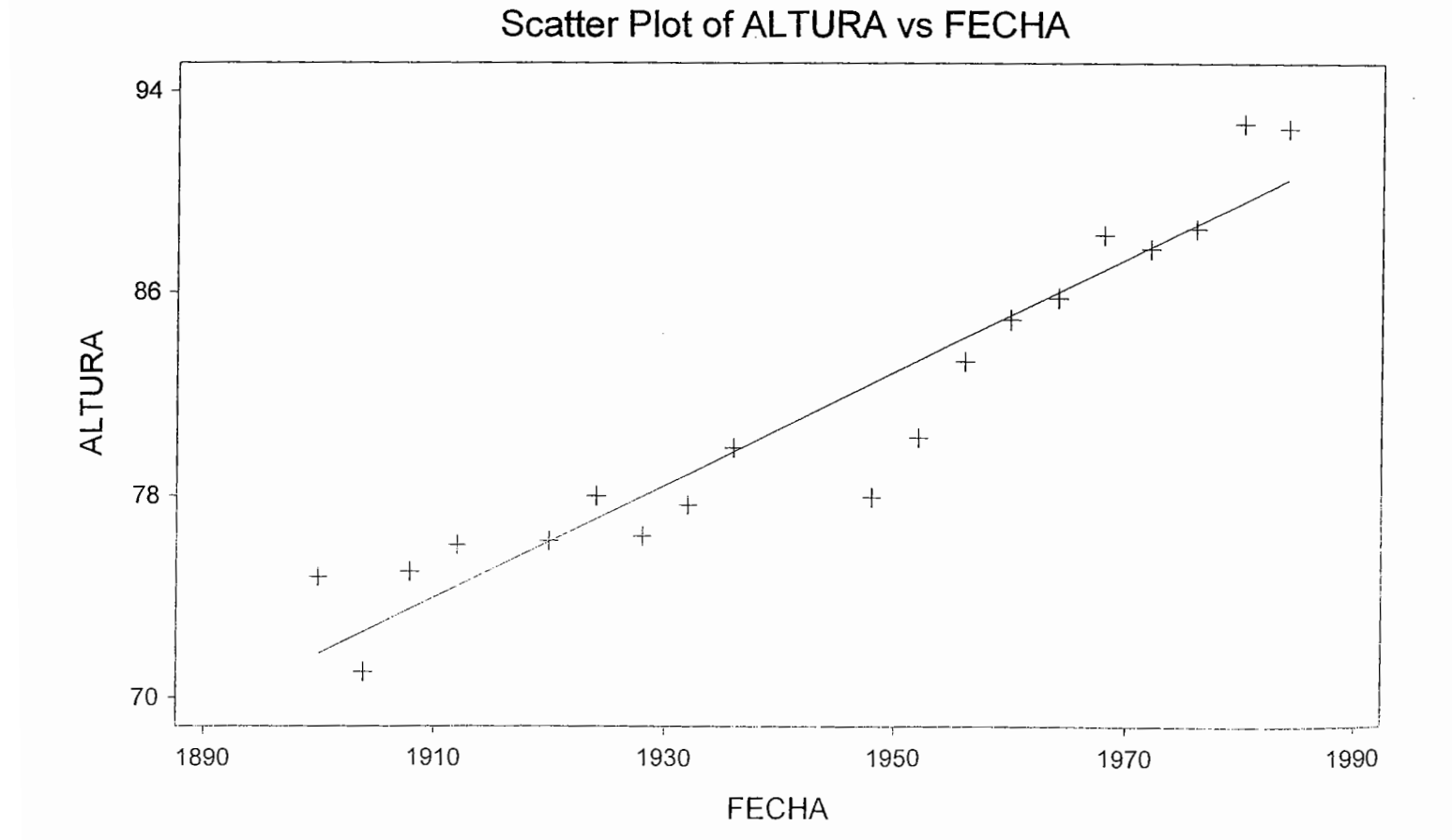


FIGURA 2.18. *Distintos tipos de relación entre los variables.*


Scatter Plot of ALTURA vs FECHA
94
86
78
70
ALTURA
1890
1910
1930
1950
1970
1990
FECHA


Scatter Plot of ALTURA vs FECHA
94
86
78
70
ALTURA
1890
1910
1930
1950
1970
1990
FECHA

2.4.2 Características de la regresión mínimo-cuadrática

La regresión mínimo-cuadrática tiene en cuenta las distancias de los puntos a la recta sólo en la dirección de y. Por tanto, en una regresión las variables x e y juegan papeles distintos.

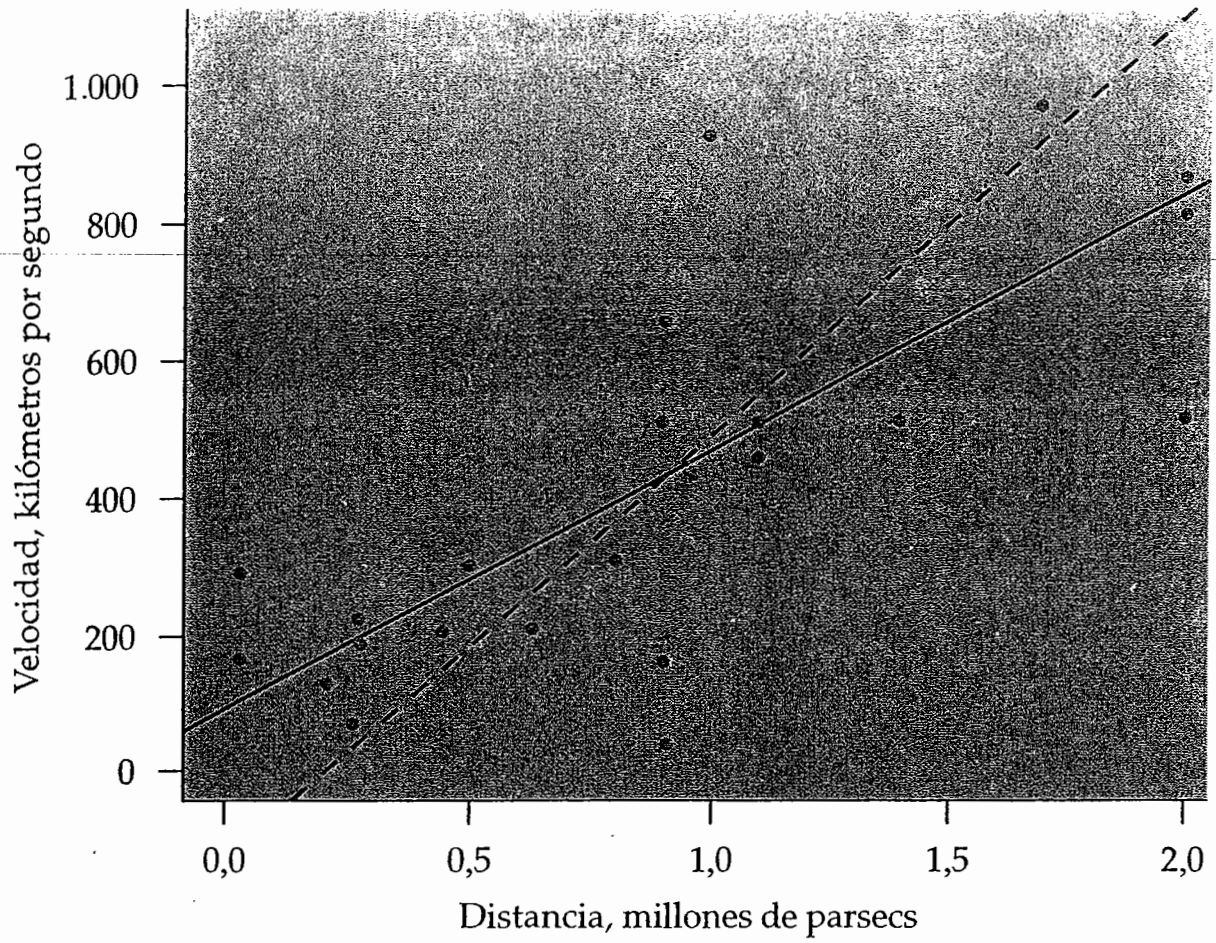


Figura 2.11. El diagrama de dispersión de los datos de Hubble sobre la distancia a la Tierra de 24 galaxias y la velocidad con la que éstas se alejan de nosotros. Las dos rectas representadas son las dos rectas de regresión mínimo-cuadráticas: la de la velocidad en relación a la distancia (línea continua) y la de la distancia en relación a la velocidad (línea discontinua).

EJEMPLO 2.11

La figura 2.11 es un diagrama de dispersión de los datos que sirvieron de base para descubrir que el Universo se está expandiendo. Son las distancias a la Tierra de 24