Estadística Empresarial I

Tema 1

Concepto de Estadística

¿Qué es la Estadística?

Concepto de Estadística:

La **Estadística** forma parte de los métodos cuantitativos que utiliza la Ciencia Económica para describir, analizar, predecir y modelizar la realidad. El término **estadística** tiene su raíz en la palabra *estadista*, que a su vez proviene del término latín *status.*

Diccionario de la Lengua Española → Censo o recuento de población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación del Estado, provincia, pueblo o clase.

Estadística → Colección de datos numéricos ordenados.

Ejemplos: Estadísticas de empleo, del censo de un municipio, de un acontecimiento deportivo, ...

Sin embargo, la **Estadística**, además, incluye:

- -Diseño de experimentos
- -Reducción y procesamiento de los datos
- -Toma de decisiones

Para comprender mejor la Estadística, hablaremos de la existencia de dos tipos de *fenómenos*.

FENÓMENOS

CAUSALES O DETERMINISTAS

Son aquellos en los que se puede saber el resultado final siempre que se realice en las mismas condiciones.

Ejemplo: Medir la altura de una mesa.

ALEATORIOS O ESTADÍSTICOS

Son aquellos en los que no se puede prever el resultado final al repetirlos en análogas condiciones.

Ejemplo: Lanzar una moneda, IPC.

CON REGULARIDAD ESTADÍSTICA

Se pueden repetir tantas veces como se quiera en iguales condiciones. Si bien no se puede predecir el resultado final, las frecuencias relativas de cada posible resultado se estabilizan alrededor de un valor determinado. Esa **regularidad estadística** o **ley del azar** es la base de la **Tª de Probabilidades.**

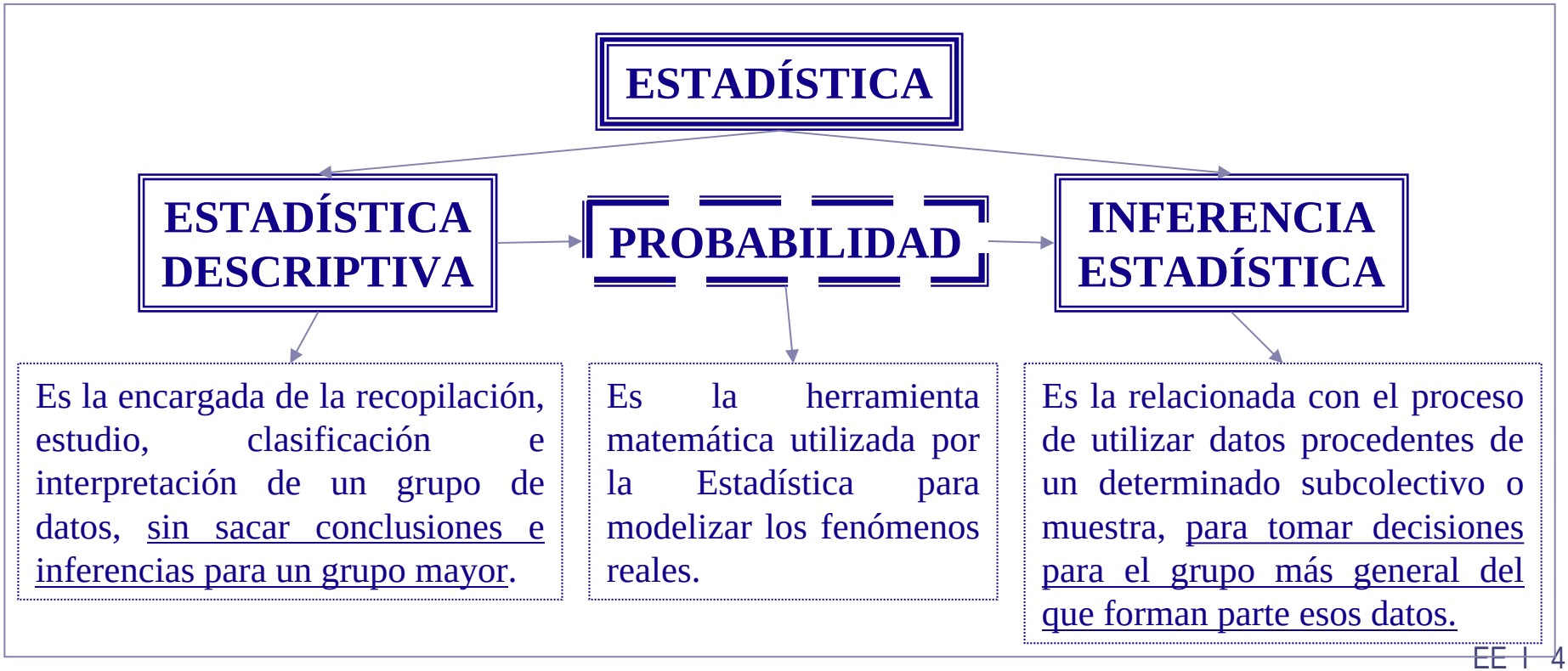
SIN REGULARIDAD ESTADÍSTICA

En ellos intervienen, además del **azar**, estrategias o posiciones humanas, surgiendo así el **concepto subjetivo de probabilidad**, que se realiza en términos de grados de creencia, opiniones, ..., dentro de lo que se conoce por **Estadística Bayesiana**.

ESTADÍSTICA

Ciencia que estudia los fenómenos (aleatorios o estadísticos), prescindiendo de los casos aislados y considerando las regularidades y propiedades del conjunto, infiriendo en su caso sobre la totalidad del fenómeno o población, a partir de los resultados que aporta una subpoblación o muestra, con un grado de certeza o fiabilidad medida probabilísticamente.

Se puede dividir la **Estadística** en dos grandes ramas, unidas por la **Teoría de la Probabilidad**.



<u>Ejemplo:</u> Se quiere llevar a cabo un estudio sobre la edad de los estudiantes universitarios canarios. Para ello, se obtiene una muestra representativa de manera que se obtiene una edad media de 22 años. ¿Podría asegurarse que la edad media de todos los estudiantes canarios está en torno a ese valor?

- En resumen:
  - Si se quiere resumir la distribución de los caracteres observados, usaremos la **Estadística Descriptiva**.
  - Si, por el contrario, se espera generalizar las características obtenidas a la población, estaremos ante la **Estadística Inferencial**.

- Hoy en día, el desarrollo de la informática ha permitido poner a disposición de los estadísticos nuevos y potentes instrumentos de observación y análisis de la realidad multidimensional, englobadas en lo que se conoce como **Técnicas de Análisis Multivariante**. Dichas técnicas permiten analizar, verificar, probar y poner a prueba ciertas hipótesis, renovando y generalizando los métodos de la **Estadística Descriptiva**, utilizando numerosos resultados de la **Inferencia Estadística.**

Evolución histórica de los contenidos de la Estadística:

1ª ETAPA
Civilizaciones anteriores al s. XVI

Censos de población y de bienes del Estado. Eran meros recuentos ya que no se extraían conclusiones.

2ª ETAPA
Del s. XVII a fines del s. XIX

Mejoras en el conocimiento cuantitativo de las cosas del Estado, en las facetas de recogida de la información, descripción y análisis de la misma, extrayendo conclusiones y realizando predicciones.

A partir de los juegos de azar, se incorpora el **Cálculo de Probabilidades** como instrumento para el estudio de fenómenos económicos y sociales.

3ª ETAPA
De fines del s. XIX a primeros del s. XX

Surge la **Estadística Inferencial**, gracias a la fusión de las dos vertientes existentes hasta ese momento: la **Estadística Descriptiva** y la **Teoría de la Probabilidad**.

4ª ETAPA
s. XX

Formalización rigurosa de la modelización matemática y el desarrollo teórico de la **Tª de la Probabilidad** y la **Inferencia Estadística**.

Introducción de la informática en el análisis estadístico, Tª de los Procesos Estocásticos, Tª de la Decisión, Análisis Multivariante, aplicaciones de la Estadística en la Economía (Econometría, Control de Calidad, Simulación y Análisis Conjunto).

Aplicaciones de los métodos estadísticos a la economía y la empresa:

- Describir la realidad socioeconómica (producción, costes, mercado, ...), obteniendo de los mismos sus principales características.
- Utilización del **Muestreo** y la **Inferencia Estadística** para inferir características de una muestra a la población que representan. Es útil para:
  - Realización de auditorías, control interno y verificación de la empresa, la estimación sobre el total o el importe medio de una cuenta, contrastar el valor probable de la misma.
  - El control de calidad, ya sea en los procesos de producción, el diseño de nuevos productos, o la calidad de los servicios públicos o privados.
  - El análisis financiero, en la simulación de proyectos de inversión.
- Mediante las técnicas de predicción, cualquier organización puede realizar predicciones de las actividades futuras y elegir las acciones a tomar a partir de ellas.
- Las técnicas multivariantes son de gran utilidad en el campo comercial y de mercados, donde será necesario investigar el consumo de un producto en una determinada zona, realizar sondeos sobre la aceptación de un producto, etc.
- Las técnicas de decisión clásicas (estimación y contraste de hipótesis), así como las técnicas de decisión bayesianas y deterministas, se utilizan en la toma de decisiones para la administración de empresas, en el sector producción, etc.

Estadística Empresarial I

Tema 2

Series Estadísticas.

Tabulación y Representación

Introducción

El estudio estadístico de cualquier fenómeno conlleva una serie de etapas:

- **Definición de los objetivos del estudio**, lo cual permitirá al investigador decidir sobre cuáles son los datos y la documentación estadística que necesita.
- **Elaboración de los datos**. Para ello, necesita realizar una serie de observaciones sobre las cuales poder analizar e interpretar los datos obtenidos. Se requiere que esos datos puedan ser:
  - Ordenados mediante una tabulación adecuada.
  - Presentados en base a representaciones gráficas.
- **Utilización de los datos para su análisis, interpretación y, si es posible, predicción**, para los que podrán ser caracterizados con medidas que resumen la cantidad de información observada, con las que interpretar posteriormente los datos.

Conceptos Previos

Toda investigación estadística empieza esencialmente por observar y anotar las características del fenómeno que se quiere estudiar. Por ello, partiremos de una serie de conceptos:

Unidad estadística: Es el dato individual, objeto de la observación, cualquiera que sea su naturaleza. Puede ser un ser vivo, un objeto o un hecho, y debe ser definido sin ambigüedad.

Población: Se entiende por *población estadística* un conjunto de *unidades estadísticas* sobre las que se verifica un determinado criterio, de manera que tengan alguna característica en común.

- Las **poblaciones** pueden estar formadas por *unidades estadísticas* variables o invariables a lo largo del tiempo.

Ejemplos: Grupo de alumnos de 1º, municipios de Tenerife.

- Según su tamaño, las **poblaciones** pueden ser finitas (que poseen un número finito de *unidades estadísticas*) o infinitas (poseen un número infinito).

Ejemplos: Profesores de estadística de empresariales de la ULL, bolígrafos fabricados.

Muestra estadística: Se trata de un subconjunto de la *población* elegido de una forma representativa.

Caracteres: Son las distintas características o cualidades que poseen las *unidades estadísticas* de una determinada *población*, y que se pueden estudiar desde el punto de vista estadístico.

CARACTERES

- **CUALITATIVOS O ATRIBUTOS**: No pueden describirse numéricamente, sino con letras. No son susceptibles de medidas y son observables sólo cualitativamente.
  Ejemplos: Profesión, sexo, nacionalidad
  Atributos: A, B, C,... Modalidades: a_1, a_2,...
- **CUANTITATIVOS O VARIABLES ESTADÍSTICAS**: Son descritos numéricamente, por lo que son medibles y cuantificables.
  Variables: X, Y, Z,... Valores: x_1, x_2,... *Ejemplos:* Altura, edad, peso, nº de hijos

VARIABLES ESTADÍSTICAS

- **DISCRETAS:** Sólo pueden tomar valores numéricos aislados.
- **CONTINUAS:** Pueden tomar cualquier valor dentro de un intervalo.

NOTA: En realidad, la distinción entre variable discreta y variable continua es en muchos casos arbitraria, ya que todas las medidas pueden convertirse en discretas. Además, en el caso de muchas variables, estamos limitados por los instrumentos de medida.

Ordenación y Tabulación

Al estudiar los datos de una **población** o **muestra**, lo más frecuente es que se obtenga un gran volumen de información. Una vez ordenados los valores de forma creciente o decreciente, se lleva a cabo una reducción de las observaciones llamada **tabulación**, obteniendo así una **tabla estadística**.

La **tabla estadística** debe reunir la máxima información posible del objeto de estudio, lo cual requiere:

- Un título que precise su contenido.
- Una indicación sobre las unidades utilizadas.
- Una especificación clara de los subtítulos de cada columna.
- Notas aclaratorias al pie de la tabla sobre la fuente de los datos o sobre algún término ambiguo.

Conceptos:

FRECUENCIA TOTAL (N): Es el número total de datos o unidades estadísticas consideradas.

FRECUENCIA ABSOLUTA (n_i): Es el número de veces que se repite cada una de las modalidades de un atributo, o cada uno de los valores de una variable.

FRECUENCIA RELATIVA (f_i):

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

Refleja la proporción, en tantos por uno, de los individuos de cada modalidad o valor

FRECUENCIA ABSOLUTA ACUMULADA (N_i): Es la suma acumulada de las frecuencias absolutas una vez ordenados los valores (o modalidades) de la variable (o atributo) de forma creciente.

$$N_i = \sum_{j=1}^{i} n_j$$

FRECUENCIA RELATIVA ACUMULADA (F_i): Es el cociente entre la frecuencia absoluta acumulada y la frecuencia total.

$$F_i = \frac{N_i}{N}$$

***INTERVALOS* [L_i, L_{i+1}):** Cuando el número de valores de la variable es muy elevado, se pueden reducir agrupándolos en intervalos. Por convenio, se consideran los intervalos solapados y semiabiertos por la derecha. Al agrupar en intervalos se pierde información.

AMPLITUD DE UN INTERVALO (a_i): Es la diferencia existente entre el límite inferior y superior del intervalo.

MARCA DE CLASE (X_i): Es el punto medio de cada intervalo. Se utiliza como representante del intervalo a la hora de hacer cálculos.

$[L_i, L_{i+1})$

$$a_i = L_{i+1} - L_i$$

→ Puede ser **constante** para todos los intervalos o variar de uno a otro (**variable**).

$$X_i = \frac{L_i + L_{i+1}}{2}$$

DENSIDAD DE FRECUENCIA (d_i): Es el cociente entre la frecuencia absoluta y la amplitud del intervalo.

$$d_i = \frac{n_i}{a_i}$$

→ Este concepto sólo se utiliza en el caso de variables cuyos valores están agrupados en intervalos de amplitud variable.

Ejemplo 1: Distribución de frecuencias del número de hijos de 150 familias en Canarias.

Nº de hijos	ni	fi	Ni	Fi
0	45	0,3	45	0,3
1	60	0,4	105	0,7
2	21	0,14	126	0,84
3	15	0,1	141	0,94
4	6	0,04	147	0,98
5	3	0,02	150	1
	150			

Ejemplo 2: Distribución de frecuencias de las estaturas de un grupo de 150 personas.

Intervalos	ni	fi	Ni	Fi	Xi	ai	di
[140,160)	9	0,06	9	0,06	150	20	0,45
[160,170)	75	0,5	84	0,56	165	10	7,5
[170,175)	45	0,3	129	0,86	172,5	5	9
[175,180)	15	0,1	144	0,96	177,5	5	3
[180,200)	6	0,04	150	1	190	20	0,3
	150						

Clasificación de las series estadísticas: Las **series estadísticas** pueden clasificarse según diversos criterios:

SERIES ESTADÍSTICAS (según dependencia del tiempo)	**TEMPORALES:** Las unidades estadísticas dependen del intervalo de tiempo tomado como unidad. Se considera como una tabla estadística con dos variables, siendo una de ellas el tiempo. **ATEMPORALES:** Las unidades estadísticas se recogen en un momento determinado, sin que interese su evolución en el tiempo.
SERIES ESTADÍSTICAS (según nº de caracteres estudiado)	**SIMPLES:** Cuando en ellas se estudia un solo carácter. También se denominan distribuciones de frecuencias unidimensionales. **MÚLTIPLES:** Se estudian varios caracteres simultáneamente. Se conocen como distribuciones de frecuencias n-dimensionales.
SERIES ESTADÍSTICAS (según su constitución)	**DE VARIABLES:** Están constituidas por variables discretas o continuas. Se tabula cuántas veces se repite cada valor o n-upla de valores. Según los tipos de frecuencias, pueden ser: distribuciones de frecuencias unitarias, distribuciones de frecuencias no agrupadas en intervalos o distribuciones de frecuencias agrupadas. **DE ATRIBUTOS:** Se tabula cuántas veces se repite cada modalidad o combinación de modalidades. **MIXTAS:** Dentro de la tabla aparecen las veces que se repiten las combinaciones de valores y modalidades.

Representaciones gráficas

Constituyen un conjunto de herramientas que permiten representar las observaciones estadísticas mediante magnitudes o figuras geométricas. El objetivo de la **representación gráfica** es proporcionar una imagen de los datos numéricos que complemente a la **tabla estadística**.

Ventajas:

- Permiten realizar una labor de síntesis buscando las regularidades y periodos.
- Constituyen un método de control ya que descubren las variaciones anormales debidas a alguna razón o a un error.
- Se pueden descubrir errores de imprenta o de cálculo.
- En un único gráfico se pueden representar varias tablas estadísticas, lo que permitirá el estudio y comparación de fenómenos relacionados entre sí o contrapuestos.

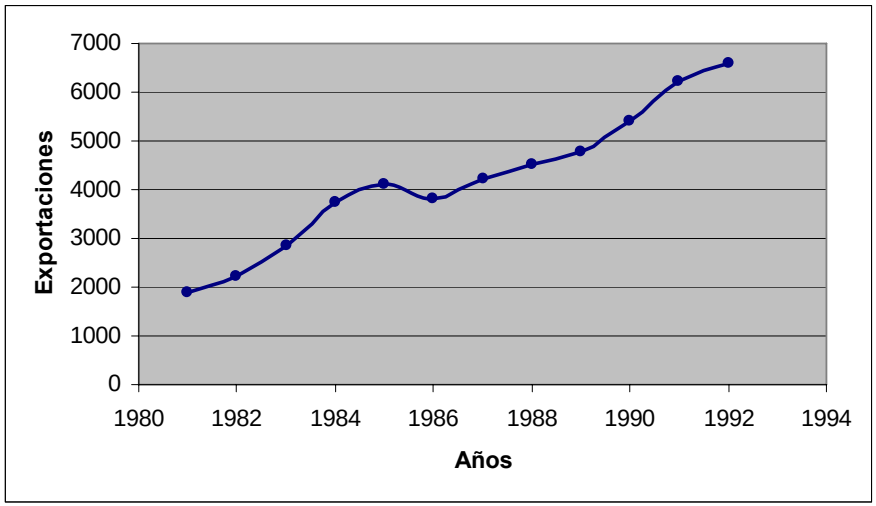
Inconvenientes:

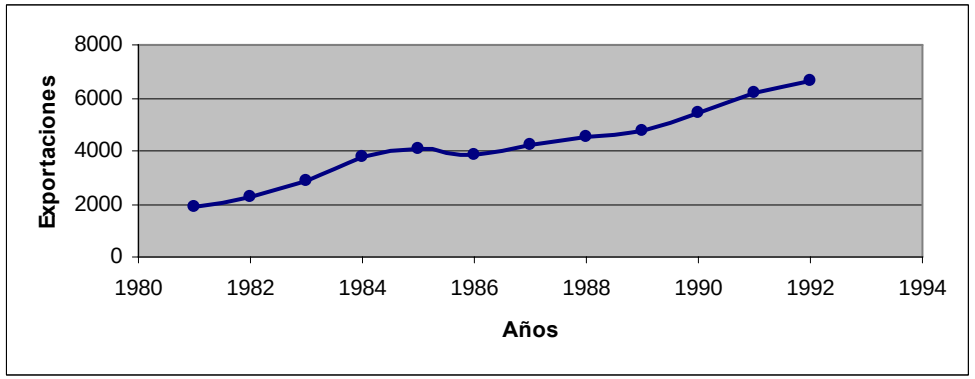
- No sustituyen a la tabla estadística, sino que la completan.
- Deben rotularse con un título adecuado, en el que estén perfectamente delimitados los hechos observados en el espacio y en el tiempo.
- La lectura de un gráfico es menos precisa que la de una tabla estadística, ya que se basa en impresiones visuales de longitud, áreas o diversas tonalidades cromáticas.
- Las unidades de las escalas de los gráficos pueden ampliarse o reducirse, exagerando hechos insignificantes o atenuando los importantes.

<u>Ejemplo:</u> Las exportaciones en miles de millones de ptas en España entre el año 1981 y 1992 fueron las que se presentan a continuación.

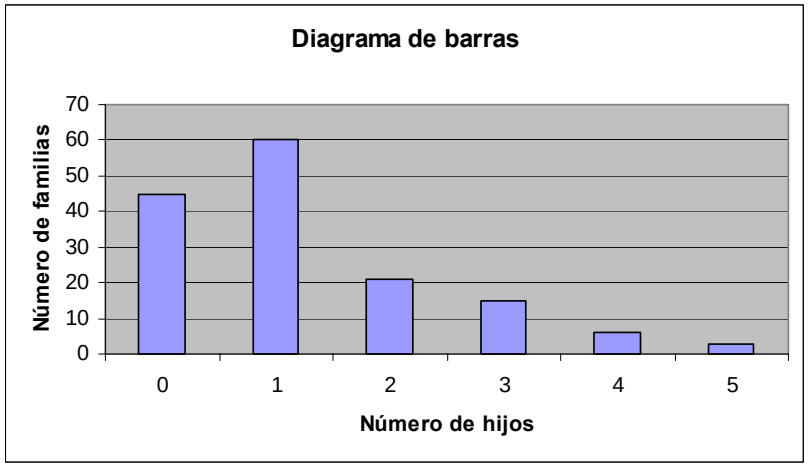
AÑOS	EXPORTACIONES MM PESETAS
1981	1890
1982	2234
1983	2847
1984	3744
1985	4109
1986	3816
1987	4212
1988	4507
1989	4972
1990	5421
1991	6226
1992	6606

Analizando los datos adjuntos, se obtiene que las exportaciones en España se incrementaron en un 249’52 %, entre 1981 y 1992. ¿Por qué en el segundo gráfico no se aprecia que aumenten tanto?





Distribuciones de frecuencias unidimensionales: variables no agrupadas en intervalos:



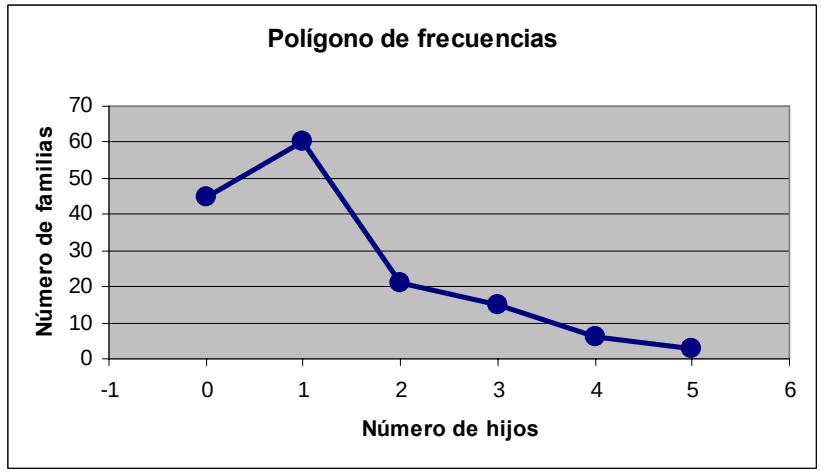


DIAGRAMA DE BARRAS: Para cada valor x_i de la variable, se levanta una barra de altura n_i o f_i.

POLÍGONO DE FRECUENCIAS: Se obtiene uniendo los extremos superiores de cada barra del diagrama de barras.

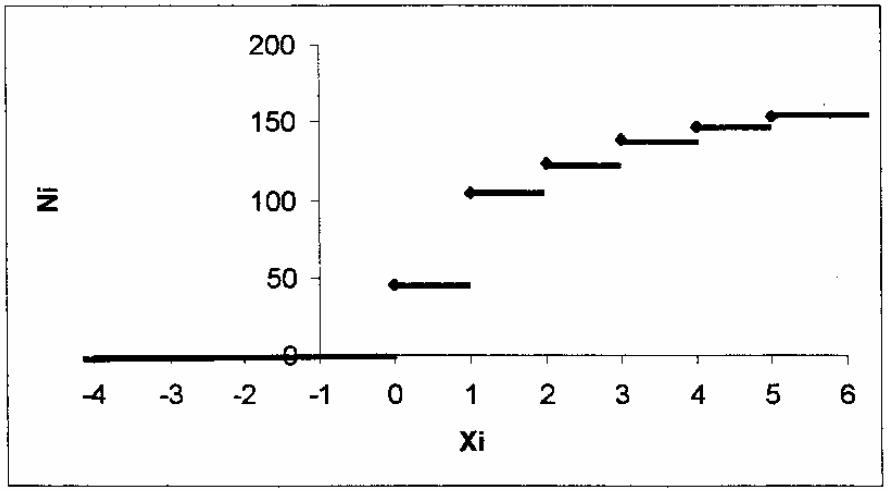
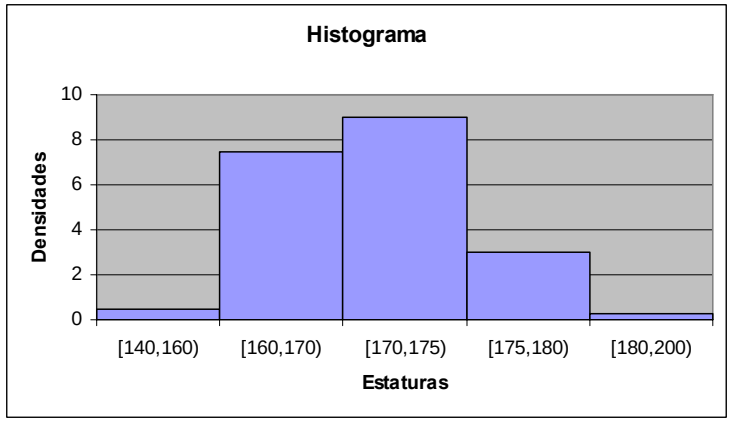
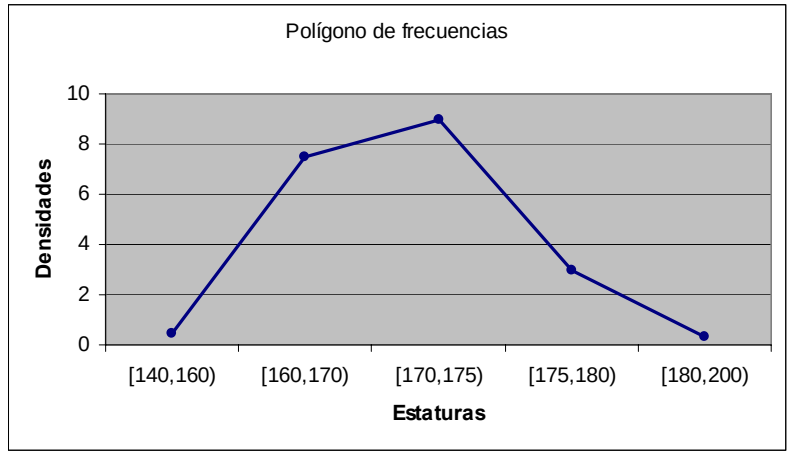


DIAGRAMA ACUMULATIVO: Se representan los valores de la variable frente a las N_i o F_i. El gráfico se confecciona mediante escalones entre un valor de la variable y el siguiente.

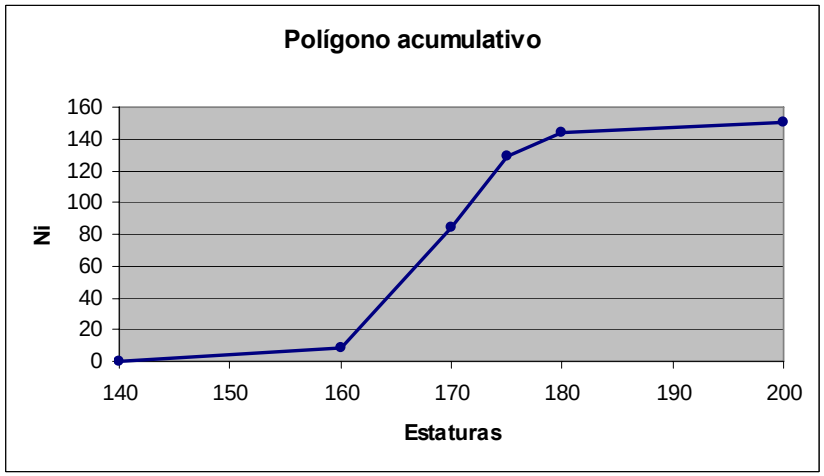
Distribuciones de frecuencias unidimensionales: variable agrupada en intervalos.





HISTOGRAMA: Para cada intervalo, se levanta una barra de altura $\mathbf{n_i}$ o $\mathbf{f_i}$ si los intervalos son de amplitud constante. Si la amplitud es variable, se usa $\mathbf{d_i}$.

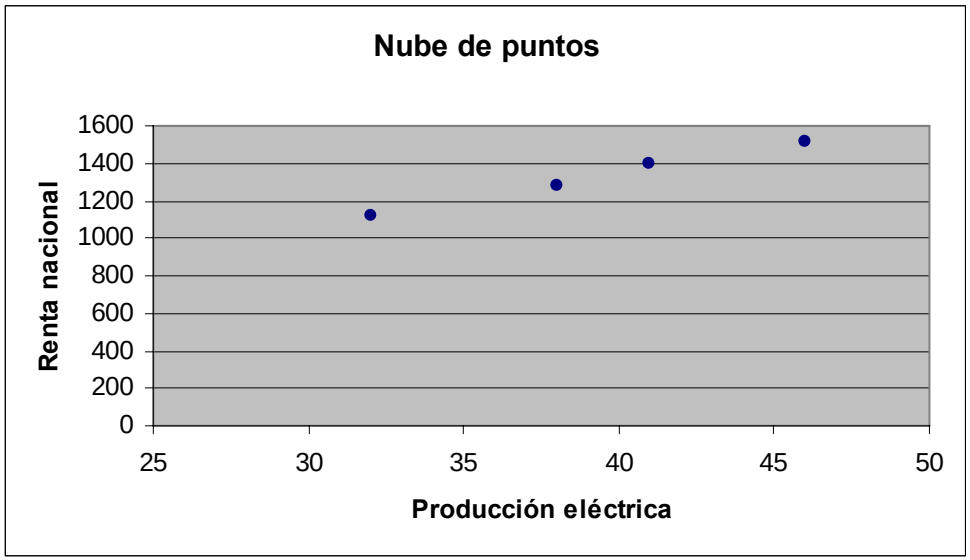
POLÍGONO DE FRECUENCIAS: Se construye sobre el histograma uniendo los puntos medios superiores a cada barra.

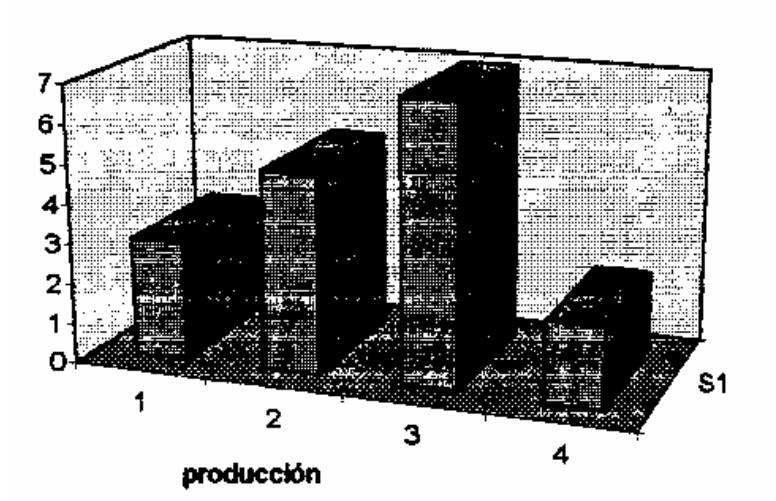


POLÍGONO ACUMULATIVO: Se construye trazando, sobre cada intervalo, líneas hasta la altura $\mathbf{N_i}$ o $\mathbf{F_i}$ de cada uno.

Distribuciones de frecuencias bidimensionales:

Producción energía eléctrica. Miles millones Kw/h	32	38	41	46
Renta Nacional. Miles millones pts año	1118	1275	1400	1513

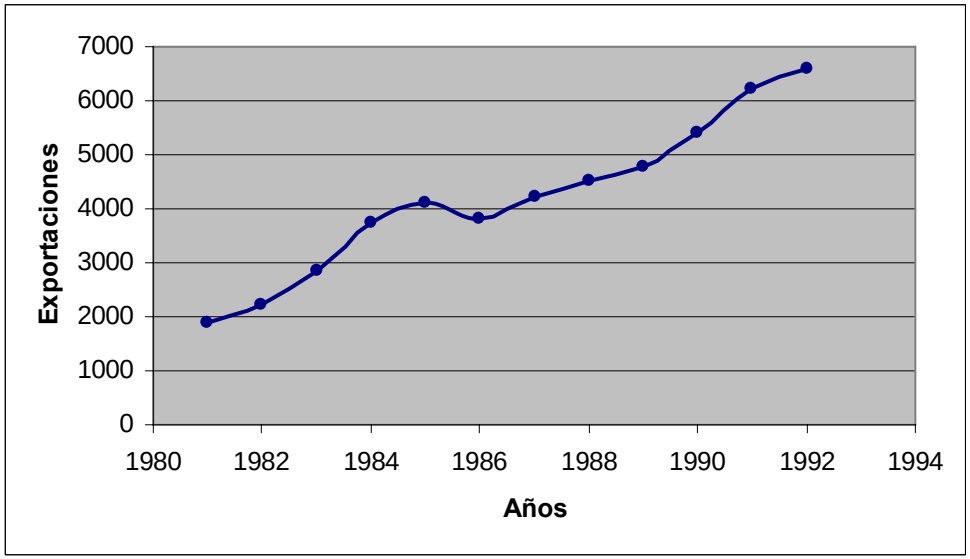


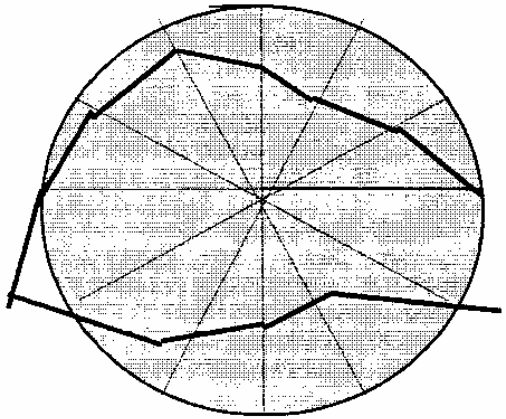


NUBE DE PUNTOS O DIAGRAMA DE DISPERSIÓN: Se representa mediante un punto cada uno de los pares de valores de las variables.

DEONDOGRAMA: Se realiza en un espacio tridimensional, de forma que en dos de los ejes se representan los valores de la variable bidimensional, y en el tercer eje, los $\mathbf{n_{ij}}$ o $\mathbf{f_{ij}}$ (si los datos no están agrupados en intervalos) o $\mathbf{d_{ij}}$ (si están agrupados). Asociado a cada par de valores se levanta un paralelepípedo.

Representaciones gráficas de series temporales:



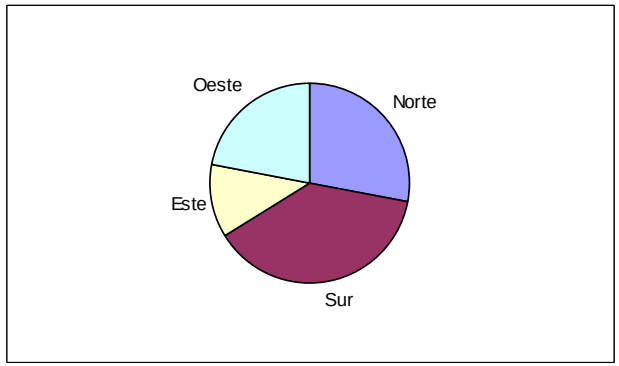


COORDENADAS POLARES: Se utilizan para fenómenos que presentan movimientos periódicos de 1 año.

COORDENADAS CARTESIANAS: Se representan los periodos de tiempo frente a los valores de la variable a estudiar.

Otras representaciones:

PIRÁMIDES DE EDADES: Son histogramas de frecuencias, pero con los ejes cambiados. Se usan mucho para estudiar la distribución de los habitantes según su edad y sexo.



DIAGRAMAS DE SECTORES: Se trata de un círculo dividido en tantos sectores como modalidades del atributo.

$$N^{o}\ de\ grados = \frac{n_i}{N}.360^{o}$$

Estadística Empresarial I

Tema 3

Distribuciones de frecuencias unidimensionales

Introducción

La **tabla estadística** obtenida mediante la clasificación de los datos nos ofrece toda la información disponible y su estructura fundamental.

Sin embargo, en muchas ocasiones resulta complicado interpretar toda esa extensa información, por lo que se intentará resumir mediante una serie de medidas obtenidas a partir de las distribuciones de frecuencias.

TIPOS DE MEDIDAS

- **Medidas de posición**: Sintetizan la información obtenida reduciéndola a un solo valor.
- **Medidas de dispersión**: Determinan si las medidas de posición son representativas o no del conjunto de datos.
- **Medidas de forma**: Establecen una distinción de las distribuciones según la forma de su representación gráfica.
- **Medidas de concentración**: Hacen referencia al mayor o menor grado de equidad en el reparto total de los valores de la variable.

Medidas de posición

Para tener un valor que represente un fenómeno, en lugar de manejar todos los datos, la distribución de frecuencias se puede caracterizar mediante las **medidas de posición**, alrededor de las cuales, se encuentran distribuidos los valores de la variable.

Las **medidas de posición** incluyen a las **medidas de tendencia central** o **promedios** (media aritmética, geométrica, armónica, mediana y moda) y a las **medidas no centrales** (cuantiles).

Con respecto a las **medidas de tendencia central**, éstas deben reunir las siguientes características:

- La característica del valor central debe ser definida objetivamente, a partir de los datos de la distribución de frecuencias.
- Debe basarse en todas las observaciones de la serie, para que represente a la distribución.
- No debe tener un carácter matemático muy abstracto, debe ser concreta y sencilla.
- Debe ser fácil de calcular.
- Ha de adaptarse con facilidad a cálculos algebraicos posteriores.

Media aritmética: Es la suma de todas las observaciones dividida entre el tamaño de la población o muestra.

$$\bar{x} = \frac{x_1.n_1 + x_2.n_2 + \ldots + x_k.n_k}{N} = \sum_{i=1}^{k} \frac{x_i.n_i}{N}$$

Nota: Para distribuciones de frecuencias agrupadas en intervalos, se utilizarán las marcas de clase X_i en lugar de los valores de la variable.

Ejemplo:

x_i	n_i	$x_i.n_i$
2	3	6
3	4	12
5	2	10
6	1	6
	10	**34**

$$\bar{x} = 3'4$$

PROPIEDADES:

1) La suma de las desviaciones de los valores respecto a su media es cero.

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

2) Si sumamos o restamos a todos los valores una constante **k**, la media aumentará o se reducirá en esa constante. Luego, la media aritmética queda afectada por los cambios de origen.

3) Multiplicando o dividiendo los valores de **X** por una constante **k**, la media quedará multiplicada o dividida por dicha constante. Por tanto, también le afectan los cambios de escala.

Ejercicio: Sea X una variable de media $\bar{x}$ y sea $Z = \frac{X - a}{b}$ (a y b constantes). Demostrar que: $\bar{x} = a + b.z$

Ventajas	Inconvenientes
- Es fácil de calcular. - Intervienen todos los valores de la variable	- Es bastante sensible a valores extremos, lo cual puede distorsionar su valor y su representatividad.

Ejercicio: Sean las calificaciones (entre 0 y 50) obtenidas para 5 alumnos las siguientes: 0.4, 0.8, 1.0, 1.4, 50. *Obtener la media aritmética y estudiar la representatividad de la misma.*

$$\bar{x} = 10.72$$

Media geométrica: Es la raíz N-ésima del producto de los valores de la variable elevados a sus respectivas frecuencias absolutas. Es de utilidad en problemas relativos a números índices.

$$G = \sqrt[N]{x_1^{n_1}.x_2^{n_2}...x_k^{n_k}} = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}} \rightarrow \log G = \frac{\sum_{i=1}^{k} n_i.\log x_i}{N}$$

Es la media aritmética de los logaritmos de los valores de la variable

Ventajas	Inconvenientes
- Es menos sensible que la media aritmética a valores extremos. - Intervienen todos los valores de la variable	- Su significado estadístico es menos intuitivo que el de la media aritmética. - Su cómputo es más difícil que el de la media aritmética. - Si un valor de la variable es 0, la media geométrica no será representativa.

NOTA: ¿Qué ocurrirá si alguno de los valores de la variable es negativo? ¿Se podrá determinar?

Ejemplo:

x_i	n_i	$\log x_i$	$n_i . \log x_i$
2	3	0.30103	0.90309
3	4	0.47712	1.90848
5	2	0.69897	1.39794
6	1	0.77815	0.77815
	10		**4.98766**

$$\log G = 0.498766 \rightarrow G = anti\log 0.498766 = 3.1533$$

Media armónica: Es la inversa de la media aritmética de los inversos de los valores de la variable. Su aplicación resulta adecuada cuando se promedian velocidades y tasas de tiempo.

$$H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_k}{x_k}}$$

Ventajas	Inconvenientes
- Intervienen todos los valores de la variable. - En algunos casos, es más representativa que la media aritmética.	- Influencia de los valores pequeños de la variable, destacando su no determinación cuando alguno de los valores de la variable es igual a 0.

Ejemplo: Un coche recorre 60 Km a 50 Km/h y 40 Km a 70 Km/h. Obtener la velocidad media.

Usando la media aritmética: $\frac{50+70}{2} = 60\ Km/h$

$v = \frac{s}{t} \rightarrow t_1 = \frac{60}{50}\ horas \quad t_2 = \frac{40}{70}\ horas \quad Tiempo\ total: t = t_1 + t_2 = \frac{60}{50} + \frac{40}{70}$

Velocidad media

$$v = \frac{s}{t} = \frac{100}{\frac{60}{50} + \frac{40}{70}} = 56.4\ Km/h$$

Usando la media armónica:

$$H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2}} = \frac{100}{\frac{60}{50} + \frac{40}{70}} = 56.4\ Km/h$$

RELACIÓN ENTRE LOS TRES PROMEDIOS: $H \leq G \leq \bar{x}$

Moda: Es el valor de la variable que más veces se repite, luego será el que tenga una mayor frecuencia absoluta asociada (o mayor densidad de frecuencia) en la distribución de frecuencias.

DISTRIBUCIONES NO AGRUPADAS EN INTERVALOS:

$$Mo = x_j \ / \ n_j = \max_i n_i$$

Ejemplos: Determinar la moda de cada una de las distribuciones de frecuencias siguientes:

x_i	n_i
1	3
2	4
8	8
15	10
21	1
	26

x_i	n_i
0	2
2	9
3	9
8	8
9	6
	34

DISTRIBUCIÓN AGRUPADA EN INTERVALOS: En este caso, primero se determinará el **intervalo modal**, que será aquel que tenga asociado una mayor frecuencia absoluta (si la amplitud es constante) o densidad de frecuencia (si la amplitud es variable).

INTERVALO MODAL
$[L_j, L_{j+1})$

Distribución de frecuencias agrupada en intervalos de amplitud constante

$$[L_j, L_{j+1}) / n_j = \max_{i=1,\ldots,k} n_i$$

Distribución de frecuencias agrupada en intervalos de amplitud variable

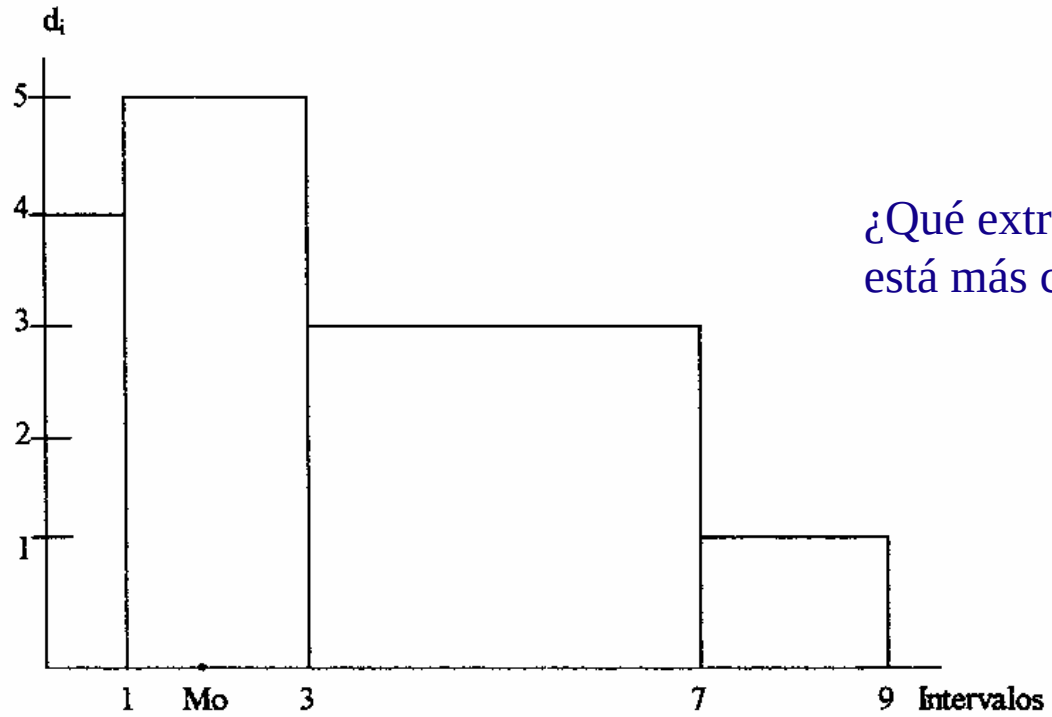
$$[L_j, L_{j+1}) / d_j = \max_{i=1,\ldots,k} d_i$$

Una vez determinado el **intervalo modal**, habrá que darle a la **moda** un valor puntual dentro de ese intervalo. Para ello, usaremos dos métodos basados en el principio de que la moda estará más cerca del de aquel intervalo contiguo que posea una frecuencia absoluta o densidad de frecuencia mayor, según sean los intervalos de amplitud constante o variable.

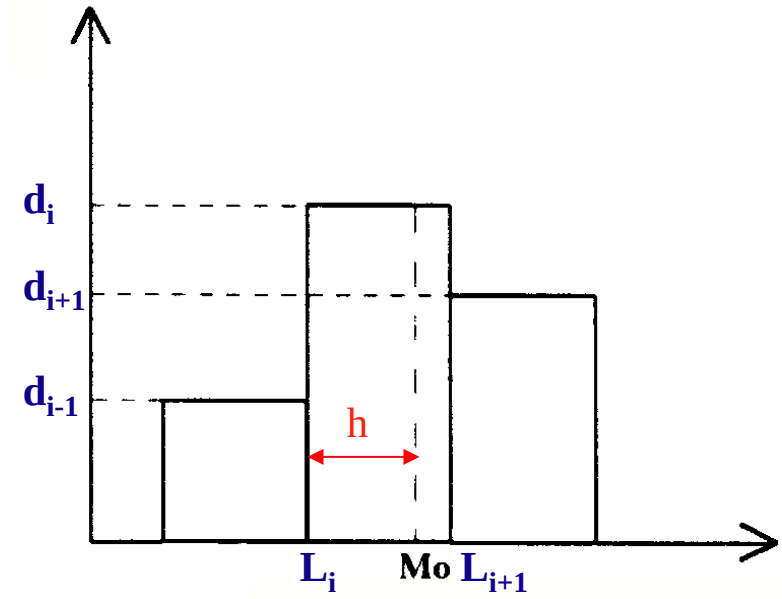
<u>Ejemplo:</u> Determinar el intervalo modal para la siguiente distribución de frecuencias.

Distribución de edades de 28 niños.

Intervalo	n_i	d_i
0-1	4	4
1-3	10	5
3-7	12	3
7-9	2	1
	28	



¿Qué extremo del intervalo modal está más cercano a la moda?



Los métodos utilizados para obtener la **moda** son los siguientes:

(a) Método de las frecuencias: Las distancias de la **moda** a los intervalos contiguos son inversamente proporcionales a las frecuencias (o densidades de frecuencias) contiguas.

$$\frac{h}{a_i - h} = \frac{d_{i+1}}{d_{i-1}} \Rightarrow h = \frac{d_{i+1}}{d_{i-1} + d_{i+1}} a_i \rightarrow \boxed{Mo = L_i + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} a_i}$$

(b) Método de la diferencia de frecuencias: Las distancias de la **moda** a los intervalos contiguos son directamente proporcionales a las diferencias contiguas de frecuencias (o densidades de frecuencia).

$$\frac{h}{a_i - h} = \frac{h_{i-1}}{h_{i+1}} \Rightarrow h = \frac{h_{i-1}}{h_{i-1} + h_{i+1}} a_i \longrightarrow \boxed{Mo = L_i + \frac{h_{i-1}}{h_{i-1} + h_{i+1}} a_i}$$

$$con\ \ h_{i-1} = d_i - d_{i-1}\ \ y\ \ h_{i+1} = d_i - d_{i+1}$$

NOTAS:

- El valor de la **moda** no coincide por ambos métodos, ya que son ambos métodos aproximados.
- Si la amplitud de los intervalos es constante, las densidades de frecuencia se sustituyen por las frecuencias absolutas.

Ejemplo: Para el ejemplo de la distribución de edades se obtiene el siguiente valor de la moda en cada caso..

$$Mo = 1 + \frac{3}{4+3} 2 = 1{,}857$$

$$Mo = 1 + \frac{5-4}{(5-4)+(5-3)} 2 = 1 + \frac{1}{1+2} 2 = 1{,}666$$

Mediana: Es aquel valor tal que, una vez ordenados los valores de la variable en orden creciente, deja a su izquierda y a su derecha igual número de frecuencias.

DISTRIBUCIONES NO AGRUPADAS EN INTERVALOS:

N impar → La **mediana** será el dato que ocupa la posición **(N+1)/2**

N par → La **mediana** será la media aritmética de los datos que ocupan las posiciones **N / 2** y **N / 2 + 1**.

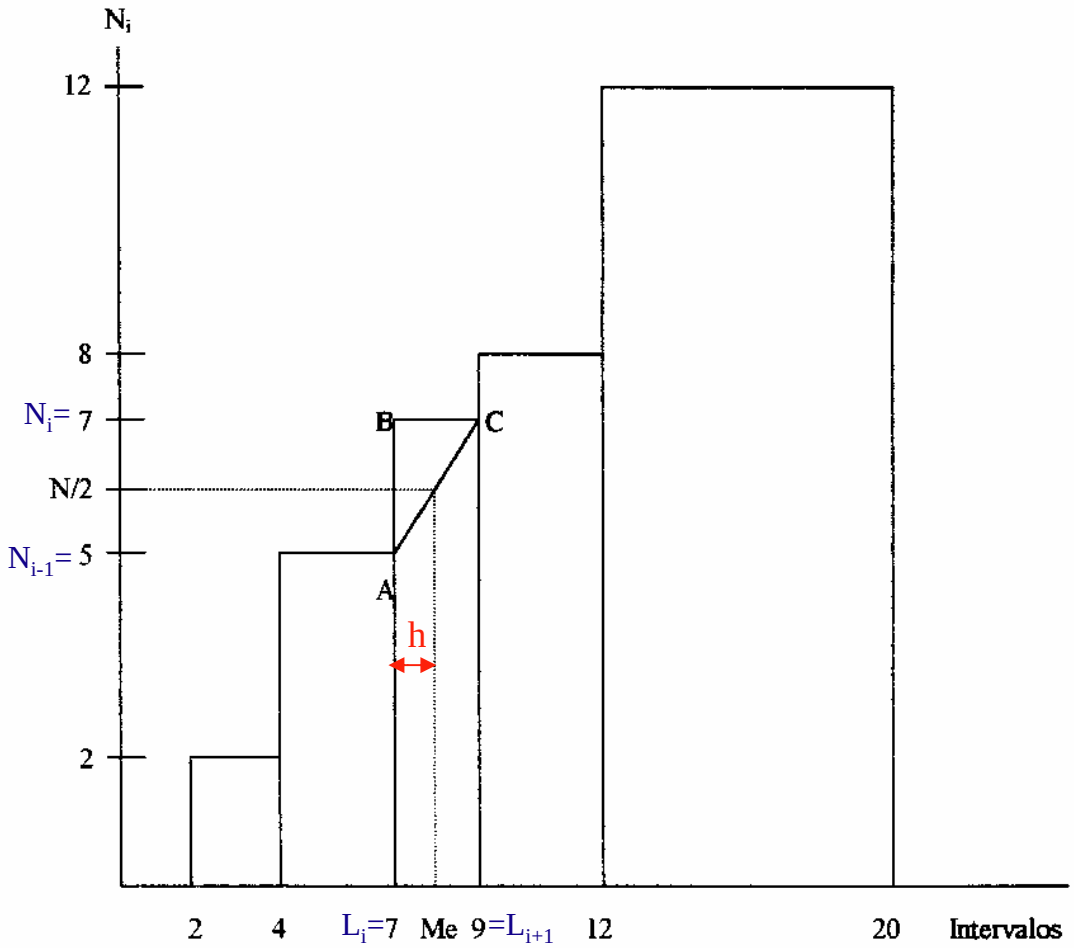
Ejemplos: Obtener la mediana en cada distribución de frecuencias.

x_i	n_i	N_i
1	25	25
2	10	35
3	15	50
4	2	52
5	3	55
	55	

x_i	n_i	N_i
2	2	2
3	3	5
4	1	6
5	5	11
6	1	12
	12	

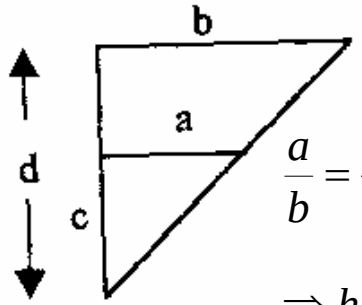
DISTRIBUCIONES AGRUPADAS EN INTERVALOS:

Usando el **polígono acumulativo de frecuencias**, determinaremos el **intervalo mediano**, buscando el valor en el eje de las abscisas al que le corresponde una valor de **N / 2** en el polígono acumulativo.



Distribución de las edades de 12 jóvenes.

$[L_i, L_{i+1})$	n_i	N_i
[2,4)	2	2
[4,7)	3	5
[7,9)	2	7
[9,12)	1	8
[12,20)	4	12
	12	



Teorema de Tales sobre ABC

$$\frac{a}{b} = \frac{c}{d} \Rightarrow \frac{h}{a_i} = \frac{N/2 - N_{i-1}}{n_i}$$

$$\Rightarrow h = \frac{N/2 - N_{i-1}}{n_i} a_i$$

Por tanto, para obtener la **mediana**, usaremos la expresión:

$$Me = L_i + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

<u>Ejemplo:</u> Obtener la mediana asociada a la distribución de frecuencias del ejemplo anterior.

$$Me = 7 + \frac{6-5}{2} 2 = 8$$

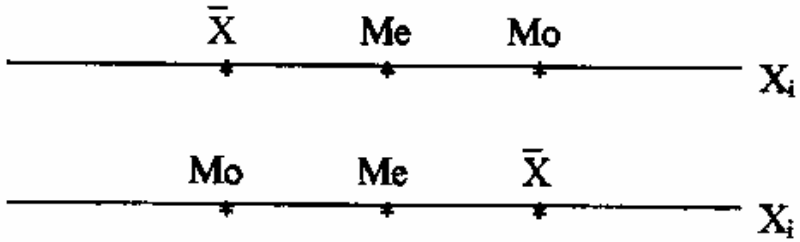
Ventajas	Inconvenientes
- Facilidad de cálculo. - No es sensible a valores extremos, ya que no los tiene en cuenta.	- En su determinación no intervienen todos los valores de la variable, por lo que no utiliza toda la información disponible.

*<u>NOTA:</u> Las ventajas e inconvenientes coinciden también para el caso de la **moda**.*

RELACIONES ENTRE LAS MEDIDAS DE TENDENCIA CENTRAL:

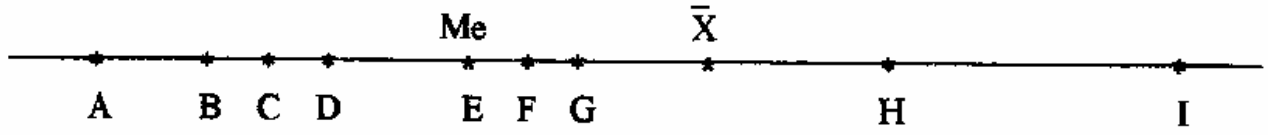
- La media aritmética da mucha importancia a los valores extremos de la distribución, mientras que la media geométrica y la armónica destacan la influencia de los valores pequeños y reducen la de los grandes.

- En las distribuciones unimodales la **mediana** siempre está comprendida entre la **media aritmética** y la **moda**, pudiendo llega a coincidir con alguna o con ambas.



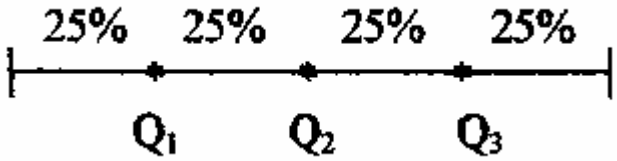
- La conveniencia de una u otra medida dependerá del tipo de variable analizada y de los fines de la investigación. Así, en el caso de los atributos, sólo tendrá sentido el cálculo de la **moda**, que será la modalidad más frecuente.

Ejemplo: Supongamos una distribución sobre los Km en los que están situados los barrios de un municipio. ¿Dónde localizarías el ayuntamiento y el hospital?



Cuantiles: Son los valores de la distribución que la dividen en partes iguales. Dentro de ellos tenemos los **cuartiles**, **deciles** y **percentiles**.

CUARTILES: Son 3 valores de la distribución que la dividen en 4 partes, de modo que cada una engloba el 25 % de los datos.



DECILES: Son 9 valores de la distribución que la dividen en 10 partes, de modo que cada una engloba el 10 % de los datos.

PERCENTILES: Son 99 valores de la distribución que la dividen en 100 partes, de modo que cada una engloba el 1 % de los datos.

En el caso de distribuciones no agrupadas en intervalos, para obtener $\mathbf{Q_k}$, $\mathbf{D_k}$ y $\mathbf{P_k}$, se procederá de manera similar al caso de la **mediana**, pero ahora con **k.N/4**, **k.N/10** y **k.N/100,** respectivamente. Para las distribuciones agrupadas, se usarán las expresiones:

$$Q_k = L_i + \frac{\frac{kN}{4} - N_{i-1}}{n_i} a_i \qquad D_k = L_i + \frac{\frac{kN}{10} - N_{i-1}}{n_i} a_i \qquad P_k = L_i + \frac{\frac{kN}{100} - N_{i-1}}{n_i} a_i$$

Ejemplo: Obtener, para la distribución de edades anterior, $\mathbf{Q_1}$, $\mathbf{D_6}$ y $\mathbf{P_{73}}$.

Medidas de dispersión

Las **medidas de posición** permitían sintetizar la información proporcionada por la distribución de frecuencias, sin embargo conviene estudiar el grado de representatividad que poseen como síntesis de toda la información. Medir la representatividad de estas medidas equivale a cuantificar la separación de los valores de la distribución respecto a esa medida *(dispersión* o *variabilidad).* De esta forma se introducen las **medidas de dispersión**, con el fin de mostrar el grado de representatividad de las *medidas de posición*.

Ejemplo: Supongamos dos situaciones distintas en las que la edad media del fallecimiento en carretera es de 40 años. ¿En cuál de los dos casos será la media aritmética más representativa?

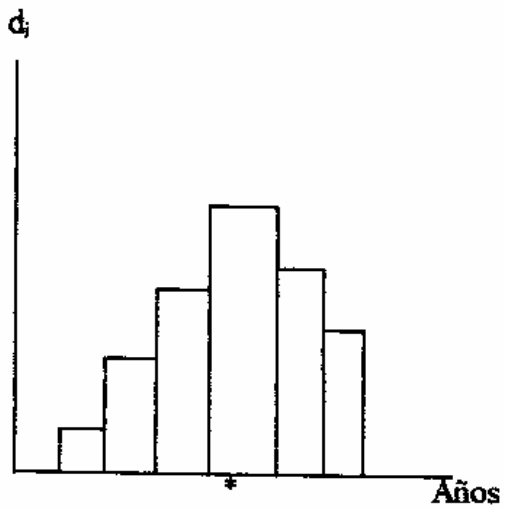


FIGURA A

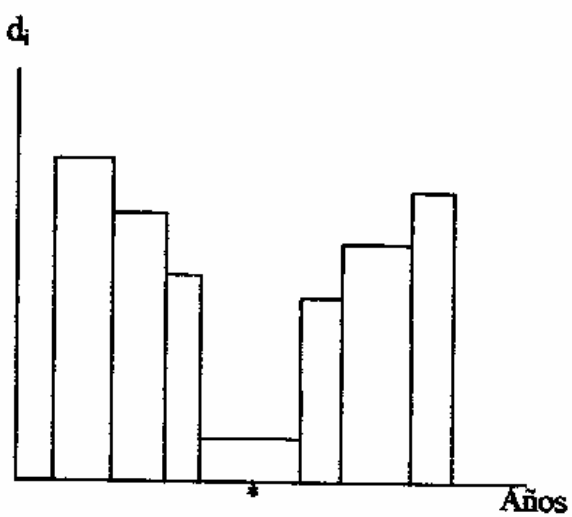


FIGURA B

MEDIDAS DE DISPERSIÓN

ABSOLUTAS: Son aquellas que vienen expresadas en unas determinadas unidades.

RELATIVAS: Son aquellas que carecen de unidades (son *adimensionales*).

Medidas de dispersión absolutas:

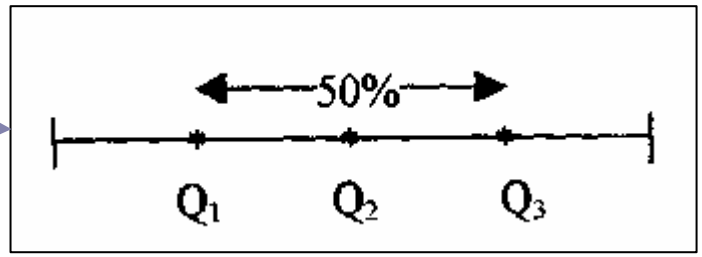
- Existen algunas medidas que hacen referencia a la *dispersión* de la distribución, pero que no indican nada sobre la representatividad de las medidas de posición.

Rango o recorrido

$$R = x_k - x_1$$

Recorrido intercuartílico

$$RI = Q_3 - Q_1$$

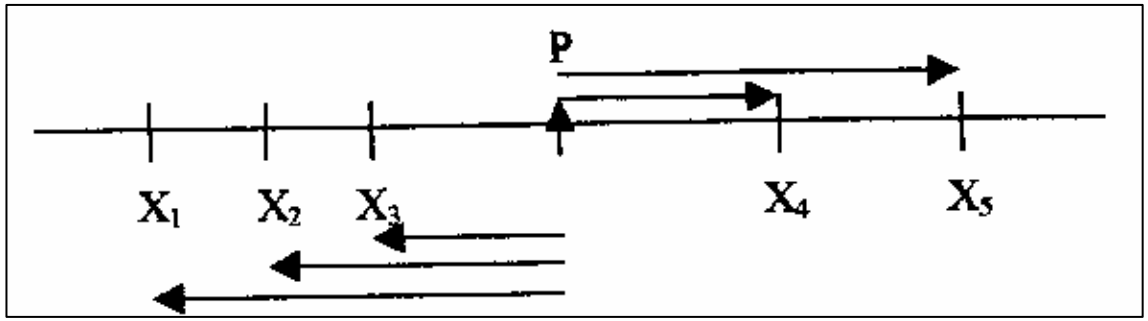


Ejemplo: Obtener el rango y el recorrido intercuartílico de la siguiente distribución de frecuencias:

$R = 5 - 2 = 3 \quad RI = 4 - 2 = 2$

x_i	n_i
2	3
3	4
4	2
5	1

- Para medir la representatividad de una medida de tendencia central **P** parece lógico emplear las distancias de todas la observaciones respecto de ella.



$(x_i - P).n_i \longrightarrow \sum_{i=1}^{k} \frac{(x_i - P).n_i}{N}$ (Media de las desviaciones respecto a **P**)

Sin embargo, algunas desviaciones **(x_i-P)** serán positivas y otras negativas, con lo que se compensarán, obteniéndose una dispersión inferior a la real. Para evitar esto, se consideran *desviaciones absolutas y cuadráticas*.

$$D = \sum_{i=1}^{k} \frac{|x_i - P|.n_i}{N}$$

$$D^2 = \sum_{i=1}^{k} \frac{(x_i - P)^2.n_i}{N}$$

Sustituyendo **P** por medidas de posición concretas, se obtendrán varias **medidas de dispersión.**

Desviación media respecto a la media

$$D_{\bar{x}} = \sum_{i=1}^{k} \frac{|x_i - \bar{x}|\, n_i}{N}$$

Desviación media respecto a la mediana

$$D_{Me} = \sum_{i=1}^{k} \frac{|x_i - Me|\, n_i}{N}$$

Desviación media respecto a la moda

$$D_{Mo} = \sum_{i=1}^{k} \frac{|x_i - Mo|\, n_i}{N}$$

NOTA: Estas tres medidas de dispersión vienen expresadas en las mismas unidades de los valores de la variable.

Ejemplo: Determinar las medidas de dispersión anteriores para la siguiente distribución de frecuencias:

x_i	n_i	$x_i n_i$	$\lvert x_i - \bar{x} \rvert . n_i$	$\lvert x_i - Me \rvert . n_i$	$\lvert x_i - Mo \rvert . n_i$
2	3	6	3.3	3	3
3	4	12	0.4	0	0
4	2	8	1.8	2	2
5	1	5	1.9	2	2
	10	**31**	**7.4**	**7**	**7**

$\bar{x} = 3.1 \quad Me = 3 \quad Mo = 3$ → $D_{\bar{x}} = 0.74 \quad D_{Me} = 0.7 \quad D_{Mo} = 0.7$

Varianza

$$S^2 = \sum_{i=1}^{k} \frac{(x_i - \bar{x})^2 n_i}{N}$$

$$S^2 = \sum_{i=1}^{k} \frac{x_i^2 n_i}{N} - \bar{x}^2$$

Ejemplo: Obtener la varianza de la siguiente distribución de frecuencias:

x_i	n_i	$x_i n_i$	$x_i^2 \cdot n_i$
2	3	6	12
3	4	12	36
4	2	8	32
5	1	5	25
	10	**31**	**105**

$$S_x^2 = \frac{105}{10} - 3.1^2 = 0.89\ unidades^2$$

Ventajas	Inconvenientes
- Es una buena medida de dispersión cuando se ha utilizado la media como medida de posición.	- Viene expresada en una unidad distinta a la de la variable, concretamente, en las unidades de la variable al cuadrado.

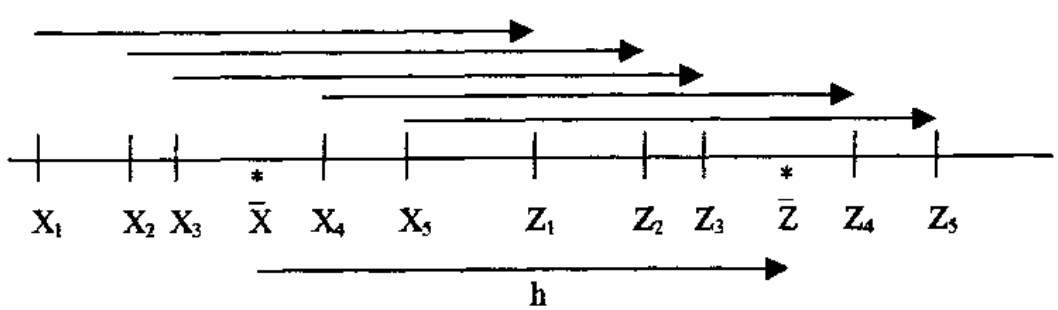
PROPIEDADES DE LA VARIANZA:

(1) La **varianza** nunca puede ser negativa, es decir, $\mathbf{0 \leq S^2 < +\infty}$.

(2) A mayor **varianza**, mayor **dispersión** de los valores en torno a la media.

(3) Si a todos los valores de la variable le sumamos una constante **h**, la **varianza** permanece inalterada.

Sea **X** una v.a., y definimos $\mathbf{Z = X + h}$. Entonces $\mathbf{S_Z^2 = S_X^2}$.



Intuitivamente, las desviaciones en torno a la media se mantienen.

(4) Si multiplicamos todos los valores de la variable por una constante **h**, la **varianza** se multiplicará por el cuadrado de dicha constante.

Sea **X** una v.a., y definimos $\mathbf{Z = X \cdot h}$. Entonces $\mathbf{S_Z^2 = h^2 \cdot S_X^2}$.

PROBLEMA:

Sea una variable X y sea $Z = \frac{X - P}{a}$ (a y P constantes). Demostrar que $S_X^2 = a^2 \cdot S_Z^2$

$$S_Z^2 = \sum_{i=1}^{k} \frac{(Z_i - \bar{Z})^2 n_i}{N} = \sum_{i=1}^{k} \frac{\left(\frac{X_i - P}{a} - \frac{\bar{X} - P}{a}\right)^2 n_i}{N} = \frac{1}{a^2} S_X^2$$

$$S_X^2 = a^2 S_Z^2$$

Desviación típica o estándar

$$S_X = +\sqrt{S_X^2}$$

Al considerar la raíz cuadrada de la varianza, se obtiene una medida que viene expresada en las mismas unidades que los valores de la variable.

Ejemplo: Calcular la desviación típica de la distribución de frecuencias del ejemplo anterior.

$$S_X = \sqrt{0.89} = 0.943$$

Medidas de dispersión relativas:

Estas medidas se caracterizan por su *adimensionalidad* (ausencia de unidades), lo que permite comparar la representatividad de las medidas de posición en dos distribuciones de frecuencias, aún cuando vengan expresadas en diferentes unidades de medida.

Ejemplo: El dinero que gasta diariamente en máquinas tragaperras un rico ludópata tiene por media 40.000 ptas y por desviación típica 5.000 ptas, mientras que la distribución del dinero gastado por otro vicioso más moderado tiene por media 800 ptas y por desviación típica 500 ptas. ¿Cuál presentará una mayor dispersión?

Las **medidas de dispersión relativas** son el cociente entre una medida de dispersión absoluta y su correspondiente medida de posición.

Coeficiente de variación de Pearson

$$CVP = \frac{S}{|\bar{x}|} \qquad CVP = \frac{S}{|\bar{x}|}100$$

Este coeficiente mide el número de veces en tantos por uno o en porcentaje, según se exprese, que la desviación típica S_X, contiene a la media aritmética. Por tanto, cuanto mayor sea CVP, más dispersos estarán los datos y por tanto menos representativa será la media aritmética.

Ejemplo: *Para comparar la dispersión en el ejemplo anterior, utilizaremos el CVP.*

$$CVP_X = \frac{5000}{40000} = 0'125 \quad CVP_Y = \frac{500}{800} = 0'625$$

Luego, la dispersión del dinero gastado por el vicioso moderado es mayor que la del ludópata rico. Así, el ludópata rico es más constante en su gasto, estando sus gastos diarios más próximos al gasto medio.

Coeficiente de variación respecto a la media

$$CVM(\bar{x}) = \frac{D_{\bar{x}}}{|\bar{x}|}$$

$$CVM(\bar{x}) = \frac{D_{\bar{x}}}{|\bar{x}|}.100$$

Coeficiente de variación respecto a la mediana

$$CVM(Me) = \frac{D_{Me}}{|Me|}$$

$$CVM(Me) = \frac{D_{Me}}{|Me|}.100$$

Coeficiente de variación respecto a la moda

$$CVM(Mo) = \frac{D_{Mo}}{|Mo|}$$

$$CVM(Mo) = \frac{D_{Mo}}{|Mo|}.100$$

Este índice mide el número de veces (o porcentaje) que la desviación media respecto a cada medida de posición **P** contiene a dicha medida **P**. Cuanto mayor sea **CVM**, menos representativa será la medida de posición **P**.

Momentos

Los **momentos** son valores que caracterizan a una distribución, de manera que dos distribuciones son iguales si todos sus **momentos** lo son.

Momento de orden r respecto a P

$$M_r(P) = \sum_{i=1}^{k} \frac{(x_i - P)^r n_i}{N}$$

P = 0

Momentos respecto al origen

$$a_r = \sum_{i=1}^{k} \frac{x_i^r n_i}{N}$$

Casos particulares:

$$a_0 = 1 \quad a_1 = \sum_{i=1}^{k} \frac{x_i n_i}{N} = \bar{x} \quad a_2 = \sum_{i=1}^{k} \frac{x_i^2 n_i}{N}$$

P = $\bar{x}$

Momentos centrales o respecto a la media

$$m_r = \sum_{i=1}^{k} \frac{(x_i - \bar{x})^r n_i}{N}$$

Casos particulares:

$$m_0 = 1 \quad m_1 = 0 \quad m_2 = S^2 \quad m_3 = \sum_{i=1}^{k} \frac{(x_i - \bar{x})^3 n_i}{N}$$

Relaciones entre los momentos:

- $$S^2 = \sum_{i=1}^{k} \frac{(X_i - \overline{X})^2 n_i}{N} = \sum_{i=1}^{k} \frac{X_i^2 n_i}{N} - \overline{X}^2 \quad \Rightarrow m_2 = a_2 - a_1^2$$

- $$m_r = \frac{\sum_{i=1}^{k} (X_i - \overline{X})^r n_i}{N} = \frac{\sum_{i=1}^{k} \sum_{h=0}^{r} (-1)^h \binom{r}{h} X_i^{r-h} \overline{X}^h n_i}{N} = \frac{\sum_{h=0}^{r} (-1)^h \binom{r}{h} \overline{X}^h \sum_{i=1}^{k} X_i^{r-h} n_i}{N} =$$

$$= \sum_{h=0}^{r} (-1)^h \binom{r}{h} a_1^h a_{r-h}$$

- $$m_3 = \sum_{i=1}^{k} \frac{(X_i - \overline{X})^3 n_i}{N} = \frac{\sum_{i=1}^{k} X_i^3 n_i - 3\sum_{i=1}^{k} X_i^2 \overline{X} n_i + 3\sum_{i=1}^{k} X_i \overline{X}^2 - \sum_{i=1}^{k} \overline{X}^3 n_i}{N} =$$

$$= \sum_{i=1}^{k} \frac{X_i^3 n_i}{N} - 3\overline{X} \sum_{i=1}^{k} \frac{X_i^2 n_i}{N} + 3\overline{X}^2 \sum_{i=1}^{k} \frac{X_i}{N} - N\frac{\overline{X}^3}{N} =$$

$$= \sum_{i=1}^{k} \frac{X_i^3 n_i}{N} - 3\overline{X} \sum_{i=1}^{k} \frac{X_i^2 n_i}{N} + 2\overline{X}^3 = a_3 - 3a_1 a_2 + 2a_1^3$$

Medidas de forma

Las **medidas de forma** establecen una tipología de las distribuciones según la forma de su representación gráfica. Se van a clasificar en: **medidas de asimetría** y **medidas de curtosis o apuntamiento**.

MEDIDAS DE ASIMETRÍA: Su finalidad es elaborar un indicador que permita establecer el **grado de asimetría** de los valores de la variable en la distribución sin necesidad de llevar a cabo su representación gráfica.

Se dice que la distribución de frecuencias es **simétrica** si existen pares de valores equidistantes a la **media aritmética** y los valores de cada par tienen las mismas frecuencias. Entonces, si la distribución es *unimodal*, se verificará que:

$$\bar{x} = Me = Mo$$

Distribución simétrica unimodal

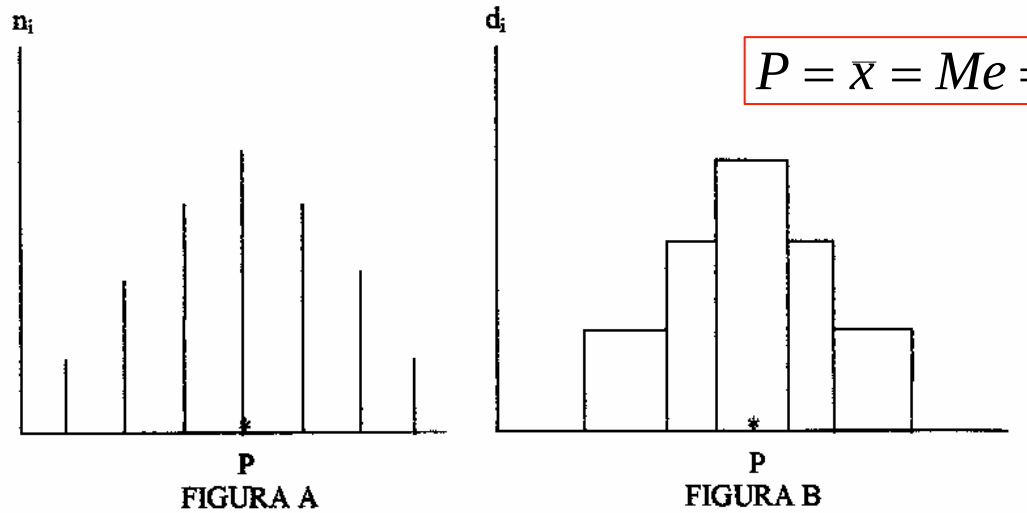


FIGURA A FIGURA B

Distribución simétrica bimodal: Puede ser **campaniforme** o **en forma de U**.

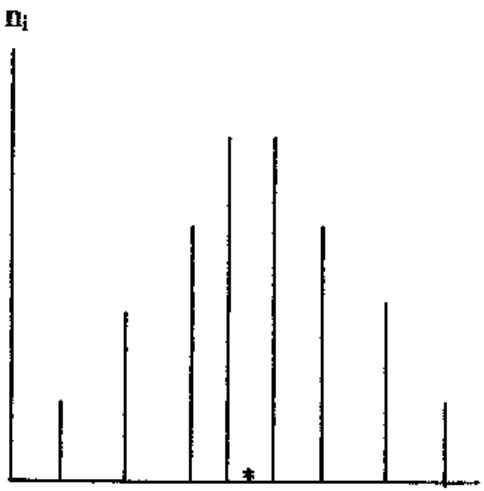


FIGURA C

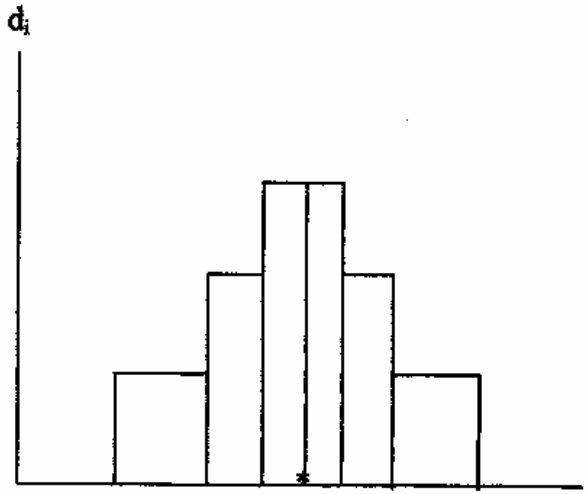


FIGURA D

CAMPANIFORME

$P = \bar{x} = Me$

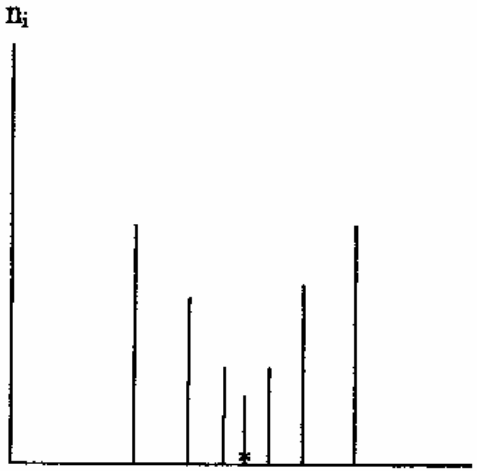


FIGURA E

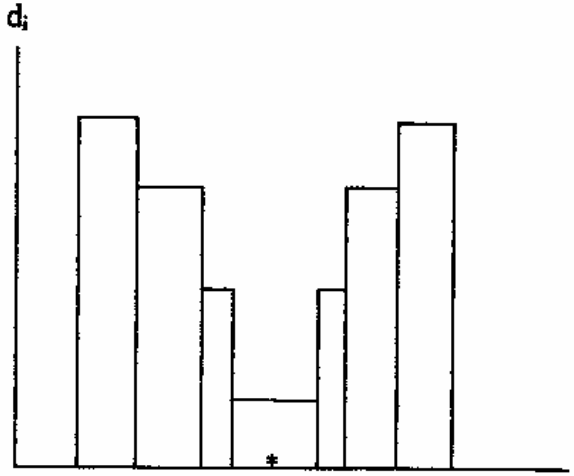


FIGURA F

EN FORMA DE U

$P = \bar{x} = Me$

Distribución asimétrica : Puede serlo **a la derecha** o **a la izquierda**.

Una distribución es **asimétrica a la derecha o positiva** si la distribución se orienta más hacia la derecha que a la izquierda de la media aritmética (los datos están más dispersos a la derecha de la media).

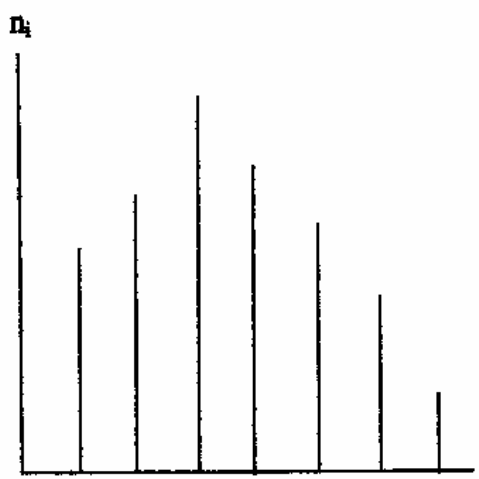


FIGURA G

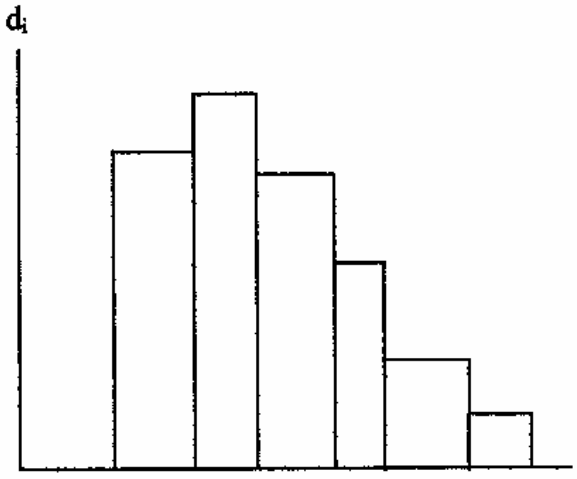


FIGURA H

Una distribución es **asimétrica a la izquierda o negativa** si la distribución se orienta más hacia la izquierda que a la derecha de la media aritmética (los datos están más dispersos a la izquierda de la media).

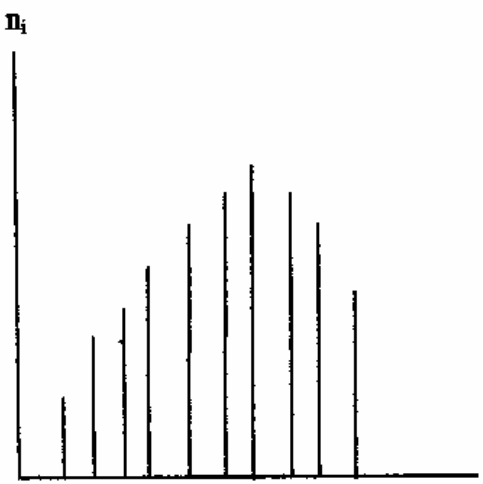


FIGURA I

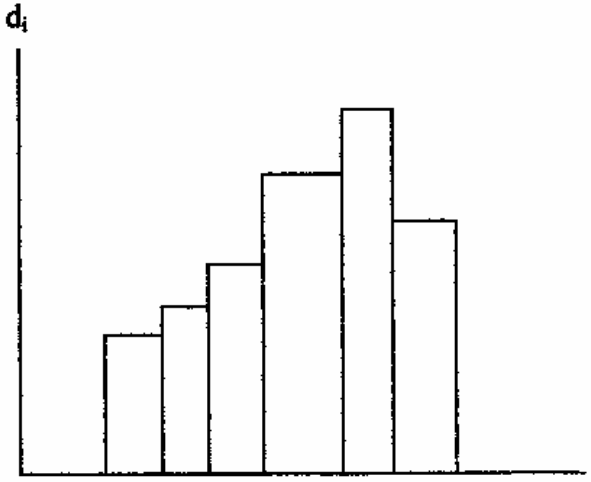
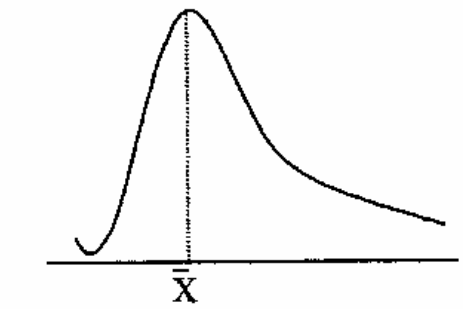
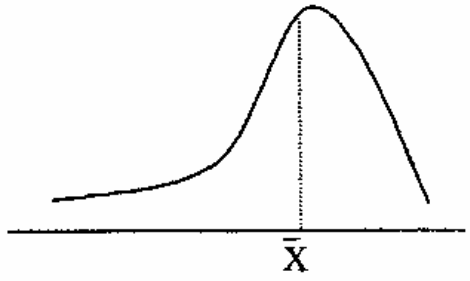


FIGURA J

- Si la distribución es **asimétrica a la derecha**, de las dos ramas de la curva que separa la media, la de la derecha es más larga que la de la izquierda. Si es **asimétrica a la izquierda**, ocurrirá lo contrario.



Distribución asimétrica a la derecha



Distribución asimétrica a la izquierda

- Para medir el **grado de asimetría** de una distribución o compararlo con el de otra, podemos utilizar el **coeficiente de asimetría de Pearson** y **el de Fisher**.

COEFICIENTE DE ASIMETRÍA DE PEARSON

$$A_p = \frac{\bar{x} - Mo}{S}$$

$A_p < 0 \Rightarrow$ *Asimétrica a la izquierda*

$A_p = 0 \Rightarrow$ *Simétrica*

$A_p > 0 \Rightarrow$ *Asimétrica a la derecha*

COEFICIENTE DE ASIMETRÍA DE FISHER

$$g_1 = \frac{m_3}{S^3} = \frac{a_3 - 3a_1a_2 + 2a_1^3}{S^3}$$

$g_1 < 0 \Rightarrow$ *Asimétrica a la izquierda*

$g_1 = 0 \Rightarrow$ *Simétrica*

$g_1 > 0 \Rightarrow$ *Asimétrica a la derecha*

Coeficiente de asimetría de Pearson	Coeficiente de asimetría de Fisher
VENTAJAS	***VENTAJAS***
- Facilidad de cálculo	- Es más preciso que el de Pearson, pudiendo aplicarse en cualquier caso.
INCONVENIENTES	***INCONVENIENTES***
- Sólo se puede utilizar si la distribución es unimodal y campaniforme. - Al basarse sólo en la distancia entre la media y la moda, no es muy precisa.	- Su cálculo no es tan inmediato como el de Pearson.

Ejemplo: Indicar el grado de asimetría que presenta la siguiente distribución de frecuencias:

$$A_p = \frac{\bar{x} - Mo}{S} = \frac{3-3}{1'211} = 0$$

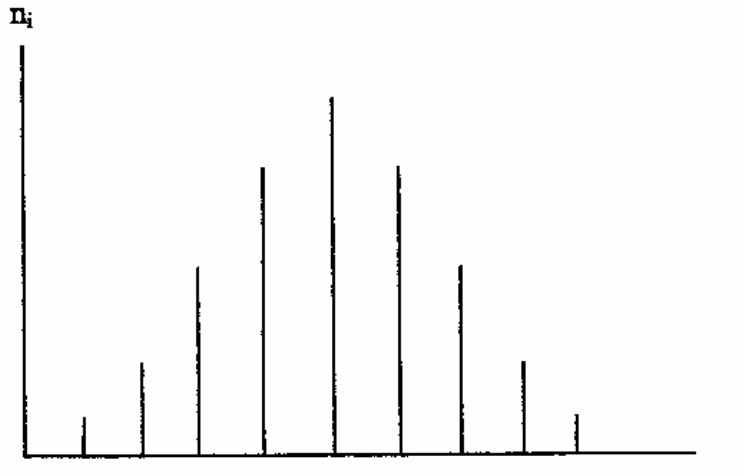
$$g_1 = \frac{a_3 - 3a_1a_2 + 2a_1^3}{S^3} = \frac{\frac{603}{15} - 3\frac{45}{15}\frac{157}{15} + 2\left(\frac{45}{15}\right)^3}{1'211^3} = 0$$

x_i	n_i	$x_i.n_i$	$x_i^2 \cdot n_i$	$x_i^3 \cdot n_i$	$x_i^4 \cdot n_i$
1	2	2	2	2	2
2	3	6	12	24	48
3	5	15	45	135	405
4	3	12	48	192	768
5	2	10	50	250	1250
	15	**45**	**157**	**603**	**2473**

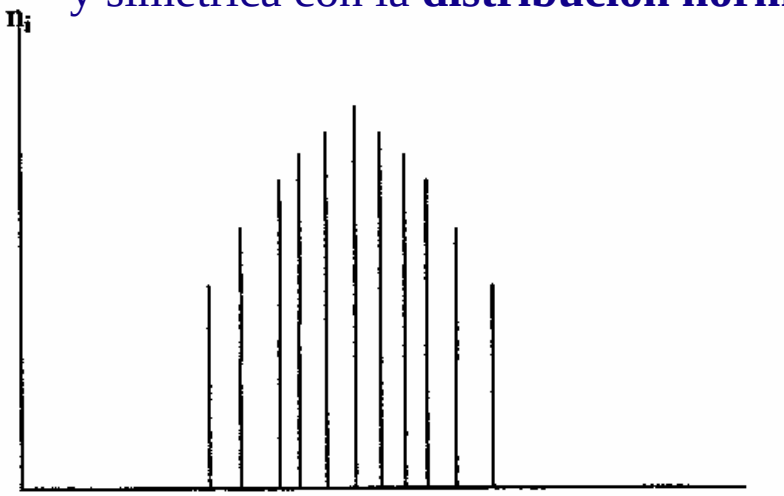
MEDIDAS DE CURTOSIS O APUNTAMIENTO:

Existe un tipo de distribución campaniforme y simétrica, de manera que la mayoría de los valores están cerca de la media, y a medida que nos alejamos de ésta, las frecuencias disminuyen. Es la **distribución normal**.

Las **medidas de apuntamiento** comparan cualquier distribución de forma campaniforme y simétrica con la **distribución normal**.



Distribución Normal
(mesocúrtica)



Distribución Leptocúrtica
(más apuntada que la normal)

Distribución Platicúrtica
(menos apuntada que la normal)

- Para medir el **apuntamiento** y comparar éste con el de otra distribución se utiliza el **coeficiente de apuntamiento de Fisher**.

COEFICIENTE DE APUNTAMIENTO DE FISHER

$$g_2 = \frac{m_4}{S^4} = \frac{a_4 - 4a_1a_3 + 6a_1^2a_2 - 3a_1^4}{S^4}$$

$$g_2 < 3 \Rightarrow Platicúrtica$$
$$g_2 = 3 \Rightarrow Mesocúrtica$$
$$g_2 > 3 \Rightarrow Leptocúrtica$$

- El **coeficiente de apuntamiento de Fisher** también nos permite determinar, sin necesidad de la representación gráfica, si la distribución es campaniforme o en forma de U. La frontera entre ambos tipos de distribuciones es la **distribución uniforme**, para la que $\mathbf{g_2 = 1'8}$. Así:

$$g_2 < 1'8 \Rightarrow En\ forma\ de\ U$$
$$g_2 = 1'8 \Rightarrow Uniforme$$
$$g_2 > 1'8 \Rightarrow Campaniforme$$

Ejemplo: Para el ejemplo anterior se obtiene un valor $\mathbf{g_2 = 2'17}$*, ¿qué podrías comentar acerca de su apuntamiento y forma?*

Medidas de concentración

Las **medidas de concentración** reflejan el mayor o menor grado de igualdad o equidad en el reparto total de los valores de la variable.

Ejemplo: En una distribución estadística de rentas, desde el punto de vista de la equidad económica, ni la media ni la varianza son significativas. Lo que verdaderamente interesa es la mayor o menor igualdad en su reparto entre los componentes de la población.

Sean **h** individuos cuyos salarios son $\mathbf{x_1, x_2, ..., x_h}$.

$$P = \sum_{i=1}^{h} x_i = "Dinero\ \ total\ \ repartido\ \ entre\ \ los\ \ h\ \ individuos".$$

Las situaciones que se pueden presentar están entre dos situaciones extremas:

Concentración máxima o menor equidad en el reparto

$$x_i = \begin{cases} 0 & para\ i = 1, 2, ..., h-1 \\ P & para\ i = h \end{cases}$$

Concentración mínima o mayor equidad en el reparto

$$x_i = \frac{P}{h} \quad i = 1, 2, ..., h.$$

- Para medir la **concentración** se utilizan dos tipos de medidas: una de tipo gráfico (**curva de Lorenz**) y otra en forma de coeficiente (**índice de Gini**).

CURVA DE LORENZ:

Sea la distribución de frecuencias **(x_i,n_i), i=1,...,k**, cuyos valores están ordenados de menor a mayor, $x_1 < x_2 < ... < x_n$, donde **X** representa los "niveles de salarios percibidos por N individuos". Se definen los pares **(p_i,q_i), i=1,...,k**, como:

$$p_i = F_i = \frac{N_i}{N}100 \qquad p_i = \text{" porcentaje que representan los } N_i \text{ primeros individuos"}$$

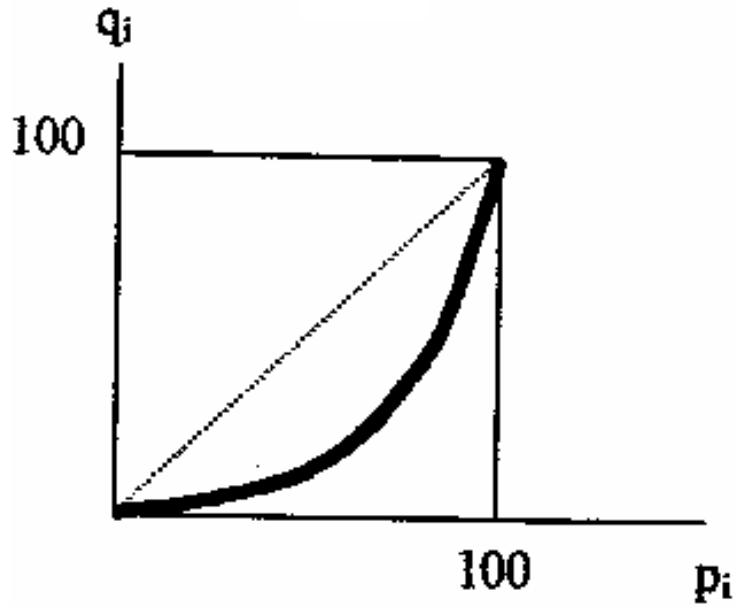
$$q_i = \frac{u_i}{u_n}100, \text{ donde } u_i = \sum_{j=1}^{i} x_j n_j \qquad q_i = \text{" porcentaje que representa el salario } u_i \text{ sobre el total de salarios } u_k\text{"}$$

$$0 \le p_i, q_i \le 100$$

El par **(p_i,q_i)** informa del porcentaje de individuos, **p_i,** que percibe un porcentaje de salarios, **q_i**, del salario total.

x_i	n_i	$x_i.n_i$	N_i	u_i	$p_i=(N_i/N).100$	$q_i=(u_i/u_k).100$
x_1	n_1	$x_1.n_1$	N_1	u_1	p_1	q_1
x_2	n_2	$x_2.n_2$	N_2	u_2	p_2	q_2
:	:	:	:	:	:	:
:	:	:	:	:	:	:
x_k	n_k	$x_k.n_k$	N_k	u_k	$p_k=100$	$q_k=100$
	N	$\mathbf{u_k}$				

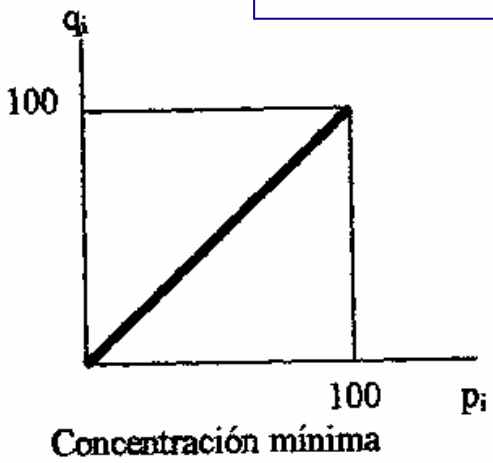
Esta distribución de rentas se puede materializar gráficamente mediante la **curva de concentración** o **curva de Lorenz**. Para obtenerla se dibuja un cuadrado cuyos lados están divididos en una escala de 0 a 100. En el eje de abscisas se representa $\mathbf{p_i}$ y en el de ordenadas $\mathbf{q_i}$. A continuación, representamos los puntos $\mathbf{(p_i, q_i)}$, que al unirlos darán lugar a la **curva de Lorenz**.



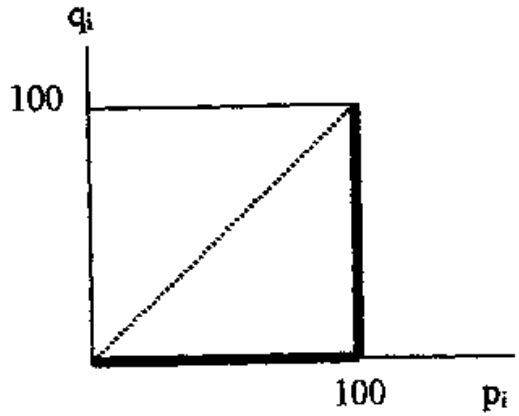
PROPIEDADES:

- Si los valores de la variable están ordenados de menor a mayor, se verifica que $p_i \geq q_i$.
- La **curva de Lorenz** se situará entre los dos casos extremos que se consideran.

CASOS EXTREMOS



Concentración mínima

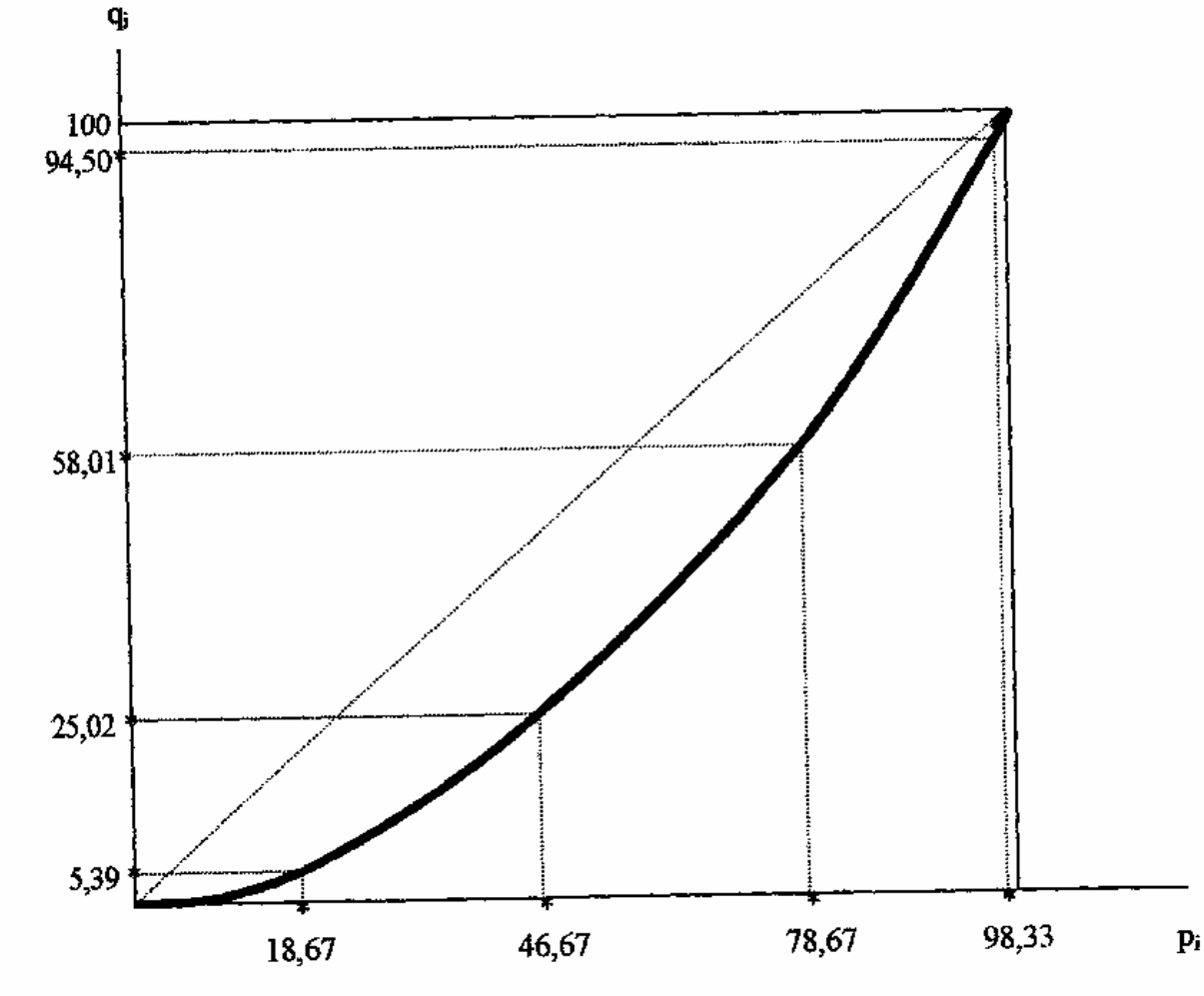


Concentración máxima

<u>Ejemplo:</u> Distribución de los sueldos percibidos por los 300 trabajadores de una empresa.

Sueldos (miles de ptas)	X_i	n_i	$x_i.n_i$	N_i	u_i	p_i	q_i
0 – 70	35	56	1960	56	1960	18’67	5’39
70 – 100	85	84	7140	140	9100	46’67	25’02
100 – 150	125	96	12000	236	21100	78’67	58’01
150 – 300	225	59	13275	295	34375	98’33	94’50
300 – 500	400	5	2000	300	36375	100	100
		300	**36375**				

*Para el ejemplo anterior, la **curva de Lorenz** obtenida será:*



ÍNDICE DE GINI:

Con el **índice de Gini** se pretende obtener un indicador que exprese el grado de concentración manifestado, desde el punto de vista gráfico, con la **curva de Lorenz**.

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i}, 0 \le I_G \le 1$$

Cuanto más próximo esté el **índice de Gini** a 0, menor concentración existirá, por lo que habrá una mayor equidad en el reparto de salarios.

CASOS EXTREMOS

Concentración mínima:

$p_i = q_i$, i = 1, ..., k.

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{0}{\sum_{i=1}^{k-1} p_i} = 0$$

Concentración máxima:

$q_i = 0$, i = 1, ..., k-1.

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{\sum_{i=1}^{k-1} p_i}{\sum_{i=1}^{k-1} p_i} = 1$$

Ejemplo: Obtener el índice de Gini para la distribución de frecuencias anterior.

	p_i	q_i	$p_i - q_i$
	18’67	5’39	13’28
	46’67	25’02	21’65
	78’67	58’01	20’66
k-1 →	98’33	94’50	3’83
	242’32		**59’42**

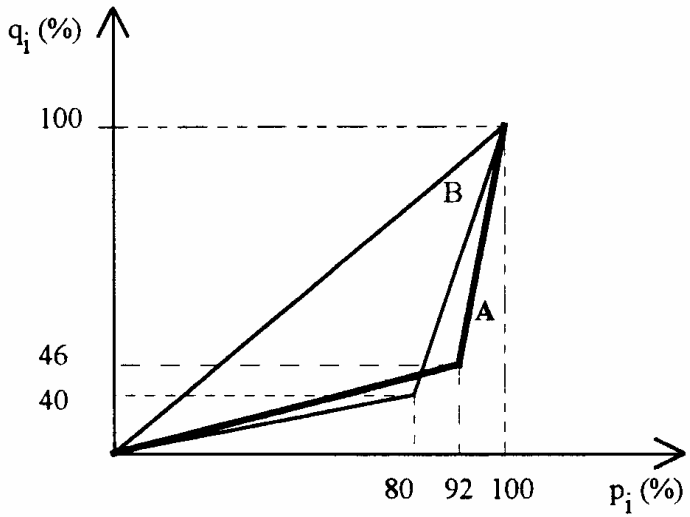
$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{59'42}{242'32} = 0'24$$

Por tanto, podemos concluir que la distribución está poco concentrada, estando los salarios bastante bien repartidos.

Si bien el **índice de Gini** tiene la ventaja de resumir en una sola cifra las complejas informaciones expresadas en la **curva de Lorenz,** puede darse el caso de que dos distribuciones de frecuencias diferentes presenten el mismo valor del **índice de Gini**, aún siendo la estructura del reparto de los valores de cada variable diferentes.

Ejemplo: Las distribuciones de frecuencias A y B generan las curvas de Lorenz siguientes, que muestran una estructura de reparto distinta. Sin embargo, puede comprobarse que: .

$$I_G^A = I_G^B$$



Estadística Empresarial I

Tema 4

Distribuciones de frecuencias q-dimensionales

Introducción

En el tema anterior estudiamos las características más importantes que presentaba una variable **X** considerada de forma aislada. Sin embargo, para una población o muestra determinada, se pueden estudiar simultáneamente dos o más caracteres diferentes.

Ejemplo: Sobre un grupo de empresas podemos observar sus ingresos (X) y sus gastos (Y), o bien, su número de trabajadores (X), los salarios que perciben (Y) y las horas de trabajo que realizan (Z). Sobre un grupo de personas estudiamos su altura (X) y su peso (Y).

El objetivo de este análisis simultáneo de 2 o más caracteres es estudiar las posibles relaciones entre ellos para detectar algún tipo de *dependencia* o variación conjunta (*covariación*).

En este tema vamos a estudiar cuestiones generales como son la **tabulación**, **representación gráfica**, **distribuciones marginales y condicionadas**, así como los **momentos**, tanto para el caso bidimensional como para el q-dimensional.

Distribuciones bidimensionales: Tabulación

Una **distribución bidimensional** está formada por el conjunto de pares de valores de dos caracteres (x_i,y_j), dispuestos mediante una tabla de doble entrada llamada **tabla de correlación**.

X\Y	y_1	y_2	. .	y_j	. .	y_k	$n_{i.}$
x_1	n_{11}	n_{12}	. .	n_{1j}	. .	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	. .	n_{2j}	. .	n_{2k}	$n_{2.}$
:	:	:	. .	:	. .	:	:
x_i	n_{i1}	n_{i1}	. .	n_{ij}	. .	n_{ik}	$n_{i.}$
:	:	:	. .	:	. .	:	:
x_h	n_{h1}	n_{h2}	. .	n_{hj}	. .	n_{hk}	$n_{h.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	. .	$n_{.j}$	. .	$n_{.k}$	N

Frecuencia absoluta conjunta n_{ij}:
Nº de veces que se presenta el par (x_i,y_j).

Frecuencia relativa conjunta: $f_{ij} = \frac{n_{ij}}{N}$

Frecuencia absoluta marginal:

$$n_{i.} = \sum_{j=1}^{k} n_{ij} \qquad n_{.j} = \sum_{i=1}^{h} n_{ij}$$

Frecuencia relativa marginal

$$f_{i.} = \frac{n_{i.}}{N} \qquad f_{.j} = \frac{n_{.j}}{N}$$

$$\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = N \qquad \sum_{i=1}^{h}\sum_{j=1}^{k} f_{ij} = 1 \qquad \sum_{i=1}^{h} n_{i.} = N \qquad \sum_{j=1}^{k} n_{.j} = N$$

<u>Ejemplo:</u> En una determinada oposición se quiere estudiar la relación entre la edad de los 15 aspirantes (X) y la calificación que han obtenido (Y). A partir de las observaciones obtener la tabla de correlación.

X	23	25	26	26	25	21	28	23	28	22	26	26	22	26	26
Y	3	3	4	3	8	4	5	5	5	4	3	3	6	3	4

Cuando la distribución tenga <u>pocas observaciones</u>, aunque la **tabla de correlación** siga siendo válida, resulta más cómodo tabular tabular los datos en columnas de la siguiente forma:

x_i	y_j	n_i
x_1	y_1	n_1
x_2	y_2	n_2
:	:	:
x_k	y_k	n_k
		N

<u>Ejemplo:</u> A continuación se muestran las edades (X) y número de hijos (Y) de un grupo de mujeres.

X	28	29	29	29	30	32
Y	2	1	1	3	4	1

Tabular los datos de manera adecuada.

Distribuciones marginales y condicionadas

DISTRIBUCIONES MARGINALES

Partiendo de una **distribución bidimensional**, nos puede interesar estudiar aisladamente cada una de las variables sin hacer referencia alguna a los valores de la otra. De esta manera, obtenemos dos **distribuciones marginales**, una **respecto de X** y otra **respecto de Y**.

DISTRIBUCIONES CONDICIONADAS

Partiendo de una **distribución bidimensional**, podemos determinar otro tipo de distribuciones unidimensionales, fijando una determinada condición. Así, obtendremos la **distribución de X condicionada a que Y = y_j**, así como la de **Y condicionada a que X = x_i.**

X	
x_i	$n_{i.}$
x_1	$n_{1.}$
x_2	$n_{2.}$
:	:
x_h	$n_{h.}$
	N

Y	
y_j	$n_{.j}$
y_1	$n_{.1}$
y_2	$n_{.2}$
:	:
y_k	$n_{.k}$
	N

X	
x_i / Y=y_j	$n_{i/j}$
x_1	n_{1j}
x_2	n_{2j}
:	:
x_h	n_{hj}
	$n_{.j}$

Y	
y_j / X=x_i	$n_{j/i}$
y_1	n_{i1}
y_2	n_{i2}
:	:
y_k	n_{ik}
	$n_{i.}$

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{\frac{n_{ij}}{N}}{\frac{n_{.j}}{N}} = \frac{f_{ij}}{f_{.j}}$$

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{\frac{n_{ij}}{N}}{\frac{n_{i.}}{N}} = \frac{f_{ij}}{f_{i.}}$$

Ejemplo: *Para el ejemplo anterior de la oposición, obtener:*

X (edad)	23	25	26	26	25	21	28	23	28	22	26	26	22	26	26
Y (nota)	3	3	4	3	8	4	5	5	5	4	3	3	6	3	4

(a) Distribuciones marginales respecto de X y de Y.

(b) Distribución de las edades de los aspirantes que obtuvieron un 4 de puntuación.

(c) Distribución de las puntuaciones para los aspirantes de 22 años.

(d) ¿Son X e Y independientes?

INDEPENDENCIA ESTADÍSTICA:

$$\text{X e Y son independientes} \Leftrightarrow f_{ij} = \frac{n_{ij}}{N} = \frac{n_{i.}}{N}\frac{n_{.j}}{N} = f_{i.}f_{.j}, \forall i, j$$

Si **X** e **Y** son **independientes estadísticamente**, entonces $\mathbf{f_{i/j} = f_{i.}}$ y $\mathbf{f_{j/i} = f_{.j}}$

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{\frac{n_{ij}}{N}}{\frac{n_{.j}}{N}} \underset{indep}{=} \frac{\frac{n_{i.}}{N}\frac{n_{.j}}{N}}{\frac{n_{.j}}{N}} = \frac{n_{i.}}{N} = f_{i.} \qquad f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{\frac{n_{ij}}{N}}{\frac{n_{i.}}{N}} \underset{indep}{=} \frac{\frac{n_{i.}}{N}\frac{n_{.j}}{N}}{\frac{n_{i.}}{N}} = \frac{n_{.j}}{N} = f_{.j}$$

Distribuciones Q-dimensionales

Habitualmente, en los problemas reales intervienen más de dos características, por lo que se hace necesario el estudio de las **distribuciones Q-dimensionales**.

Dada una variable Q-dimensional **$(X_1, X_2, ..., X_Q)$**, el conjunto de observaciones de esta variable acompañadas de sus correspondientes frecuencias absolutas conjuntas, constituye la **distribución conjunta Q-dimensional**, que se tabula de la siguiente forma:

X_1	X_2	...	X_Q	$n_{(X_1,X_2,...,X_Q)}$
x_{11}	x_{21}	...	x_{Q1}	n_1
x_{12}	x_{22}	...	x_{Q2}	n_2
...	...	...	...	...
x_{1i}	x_{2i}	...	x_{Qi}	n_i
...	...	...	...	...
x_{1h}	x_{2h}	...	x_{Qh}	n_h
				N

Q = 3

X	Y	Z	$n_{(X,Y,Z)}$
x_1	y_1	z_1	n_1
x_2	y_2	z_2	n_2
...	...	...	...
x_i	y_i	z_i	n_i
...	...	...	...
x_h	y_h	z_h	n_h
			N

DISTRIBUCIONES MARGINALES Y CONDICIONADAS DE (X,Y,Z)

Distribuciones marginales

Unidimensionales: **Marginales respecto de X, de Y y de Z.** Se obtienen considerando individualmente cada variable, prescindiendo de los valores de las otras dos.

Bidimensionales: **Marginales respecto de (X,Y), de (X,Z) y de (Y,Z)**. Se obtienen prescindiendo de los valores de una de las tres componentes y considerando la distribución conjunta de las otras dos.

Análogamente, las **distribuciones condicionadas** podrán ser unidimensionales y bidimensionales.

Ejemplo: Para la siguiente distribución de frecuencias tridimensional, obtener:

(a) Distribución marginal respecto de X.

(b) Distribución respecto de (X,Y).

(c) Distribución de Z condicionada a que X=2 e Y=3.

(d) Distribución de (X,Y) condicionada a que Z=1.

X	Y	Z	$n_{(X,Y,Z)}$
1	2	3	2
2	3	1	1
3	1	2	3
2	3	4	2
4	1	1	1
3	4	2	4
1	4	3	2

Momentos bidimensionales. Independencia.

Momentos de órdenes r y s respecto a los parámetros P y Q

$$M_{rs}(P,Q) = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x_i - P)^r (y_j - Q)^s n_{ij}}{N}$$

$P = 0, Q = 0$ $\qquad$ $P = \bar{x}, Q = \bar{y}$

Momentos de órdenes r y s respecto al origen

$$a_{rs} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{x_i^r\, y_j^s\, n_{ij}}{N}$$

Momentos de órdenes r y s respecto a la media (o centrales)

$$m_{rs} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x_i - \bar{x})^r (y_j - \bar{y})^s n_{ij}}{N}$$

Casos particulares:

$a_{10} = \bar{x} \quad a_{01} = \bar{y}$

$$a_{20} = \sum_{i=1}^{h} \frac{x_i^2\, n_{i.}}{N} \qquad a_{02} = \sum_{j=1}^{k} \frac{y_j^2\, n_{.j}}{N}$$

Casos particulares:

$m_{10} = 0 \quad m_{01} = 0$

$$m_{20} = \sum_{i=1}^{h} \frac{(x_i - \bar{x})^2\, n_{i.}}{N} = S_X^2 \qquad m_{02} = \sum_{j=1}^{k} \frac{(y_j - \bar{y})^2\, n_{.j}}{N} = S_Y^2$$

COVARIANZA

$$m_{11} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x_i - \bar{x})(y_j - \bar{y})\, n_{ij}}{N} = S_{XY}$$

Se trata de una medida que hace referencia a la **dependencia lineal** existente entre ambas variables. Si la **covarianza** es positiva, las dos variables varían en el mismo sentido, y si es negativa, lo harán en sentido opuesto.

Relaciones entre los momentos centrales y los momentos respecto al origen

$$m_{20} = a_{20} - a_{10}^2 \qquad m_{02} = a_{02} - a_{01}^2 \qquad m_{11} = a_{11} - a_{10}.a_{01}$$

Ejercicio: Sea una distribución bidimensional (X,Y), y otra (Z,W) construida a partir de la anterior de manera que: $Z = \frac{X - P}{a}$ y $W = \frac{Y - Q}{b}$

Comprobar que: $S_{XY} = a.b.S_{ZW}$

INDEPENDENCIA Y COVARIACIÓN:

Si **X** e **Y** son independientes $\Rightarrow$ $\mathbf{S_{XY} = 0}$

Comprobar que $a_{11} = a_{10}.a_{01}$

Nota: El recíproco, en general, no es cierto.

Momentos Q-dimensionales. Matriz de covarianzas.

Momentos de órdenes $r_1, r_2, ..., r_Q$ respecto a los parámetros $P_1, P_2, ..., P_Q$

$$M_{r_1 r_2 ... r_Q}(P_1, P_2, ..., P_Q) = \sum_{i=1}^{k} \frac{(x_{1i} - P_1)^{r_1} (x_{2i} - P_2)^{r_2} ... (x_{Qi} - P_Q)^{r_Q} n_i}{N}$$

Momentos de órdenes $r_1, r_2, ..., r_Q$ respecto al origen

$$a_{r_1 r_2 ... r_Q} = \sum_{i=1}^{k} \frac{x_{1i}^{r_1} x_{2i}^{r_2} ... x_{Qi}^{r_Q} n_i}{N}$$

Momentos de órdenes $r_1, r_2, ..., r_Q$ respecto a la media

$$m_{r_1 r_2 ... r_Q} = \sum_{i=1}^{k} \frac{(x_{1i} - \bar{x}_1)^{r_1} (x_{2i} - \bar{x}_2)^{r_2} ... (x_{Qi} - \bar{x}_Q)^{r_Q} n_i}{N}$$

Casos particulares:

$a_{100...0} = \bar{x}_1 \quad a_{010...0} = \bar{x}_2 \cdots a_{00...01} = \bar{x}_Q$

$m_{200...0} = S_{11} \quad m_{020...0} = S_{22} \cdots m_{00...02} = S_{QQ}$

$m_{1100...0} = S_{12} \quad m_{10100...0} = S_{13} \cdots m_{00...011} = S_{Q-1Q}$

MATRIZ DE COVARIANZAS

$$S = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1Q} \\ S_{21} & S_{22} & \cdots & S_{2Q} \\ \vdots & \vdots & \vdots & \vdots \\ S_{Q1} & S_{Q2} & \cdots & S_{QQ} \end{pmatrix}$$

Estadística Empresarial I

Tema 5

Regresión y correlación bidimensional y múltiple

Introducción

A partir de una distribución de frecuencias bidimensional **(X,Y)** podemos determinar el grado de **dependencia estadística** que existe entre las distribuciones marginales **X** e **Y**, y analizar la relación existente entre ellas. Esto se llevará a cabo en dos procedimientos:

- Explicar los valores que toma una de las variables (**variable dependiente**) en función de los valores de la otra (**variable independiente**). De esto se encargará la **regresión**.
- Medir el grado de dependencia existente entre las variables, para lo que se estudiará la **correlación**.

Las técnicas estadísticas de **regresión** y **correlación** deben aplicarse sobre variables entre las que se sepa que existe algún tipo de influencia, ya que podría ocurrir que la dependencia estadística fuera debida *al azar* o bien fuera *indirecta* (existe una tercera variable que influye sobre ambas).

Ejemplos: Número de nacimientos y número de aprobados en EE I; el gasto en vacaciones y el gasto en electrodomésticos pueden moverse en la misma dirección debido a la renta.

La **regresión de Y sobre X** consistirá en encontrar una función que explique el comportamiento de la variable **Y** a partir de los valores que toma la variable **X**. De análoga forma, la **regresión de X sobre Y** explicará el comportamiento de **X** a partir de los valores de **Y**.

Y = f (X)

Variable dependiente — Variable independiente

Para encontrar estas funciones se suelen aplicar distintos **métodos de ajuste**. Por tanto, el ajuste consistirá en encontrar la ecuación de la curva que más se aproxime a las observaciones.

- Elegir el tipo de función que mejor se adapte a los datos representados en la nube de puntos.

¿Qué tipo de ajuste plantearías en cada caso?

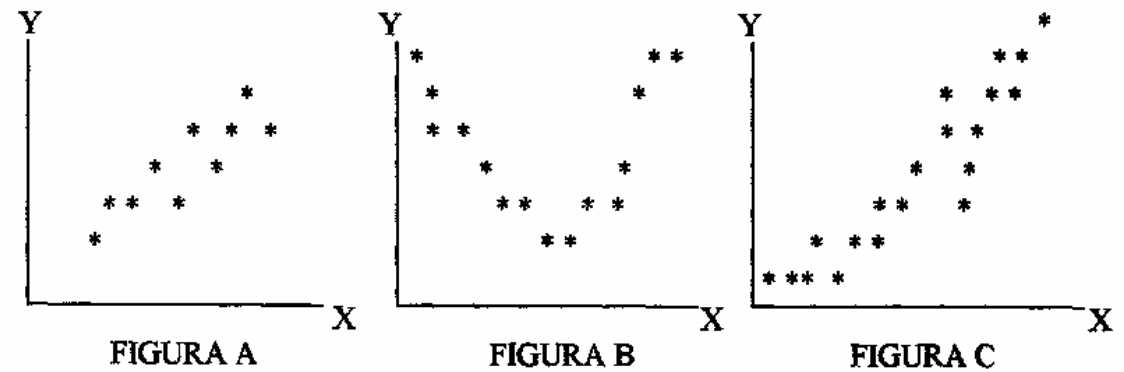


FIGURA A FIGURA B FIGURA C

- Calcular los parámetros que caracterizan la función ajustada, mediante el **método de los mínimos cuadrados** (es el más representativo).

Ajuste mínimo-cuadrático

Sean **N** observaciones (x_i,y_i), i = 1, ..., N, con frecuencia unitaria (podemos suponerlo sin pérdida de generalidad), de manera que al representar su correspondiente diagrama de dispersión o nube de puntos, decidimos ajustarle una función que depende de **R** parámetros.

$$\mathbf{Y = f (X, a_1, a_2, ..., a_R)}$$

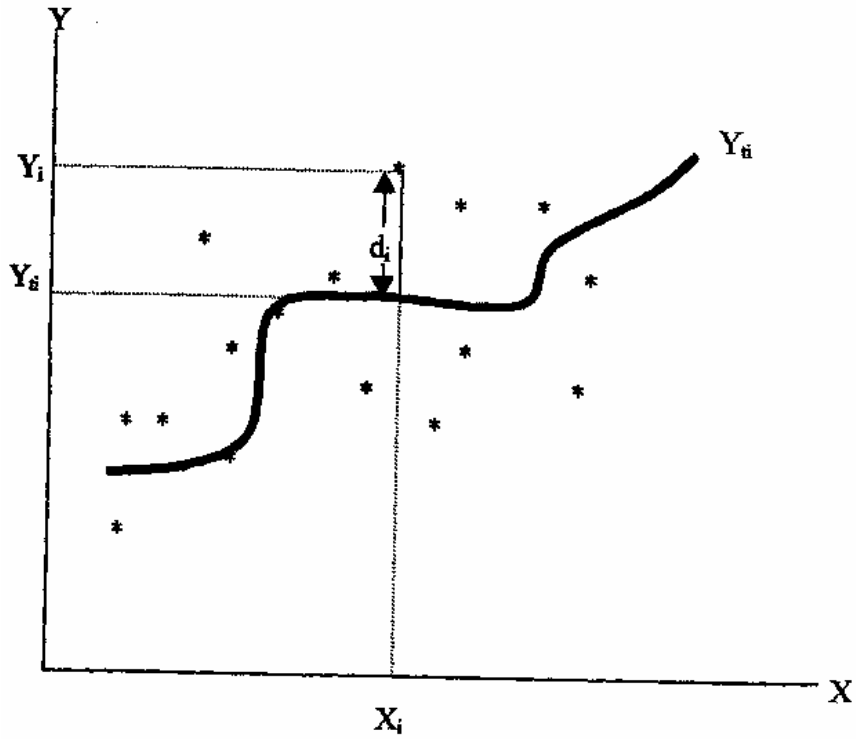
Dado $\mathbf{x_i}$
- $\mathbf{y_i}$: valor observado
- $\mathbf{y_{ti}}$: valor teórico o ajustado

$$\mathbf{y_{ti} = f (x_i, a_1, a_2, ..., a_R)}$$

RESIDUO: $d_i = y_i - y_{ti} = y_i - f (x_i, a_1, a_2, ..., a_R)$

La función de ajuste o curva de regresión de **Y** sobre **X** será aquella que minimice:

$$H = \sum_{i=1}^{N} d_i^2$$



$$\min H = \min \sum_{i=1}^{N} d_i^2 = \min \sum_{i=1}^{N} (y_i - f(x_i, a_1, ..., a_R))^2$$

$$\frac{\partial H}{\partial a_1} = 0 \quad \frac{\partial H}{\partial a_2} = 0 \quad \cdots \quad \frac{\partial H}{\partial a_R} = 0$$

AJUSTE LINEAL:

$y_{ti} = f(x_i, a, b) = a + b\, x_i$

$$\begin{cases} \dfrac{\partial H}{\partial a} = 0 \Rightarrow \displaystyle\sum_{i=1}^{N} y_i = N a + b \sum_{i=1}^{N} x_i \\ \dfrac{\partial H}{\partial b} = 0 \Rightarrow \displaystyle\sum_{i=1}^{N} x_i\, y_i = a \sum_{i=1}^{N} x_i + b \sum_{i=1}^{N} x_i^2 \end{cases}$$

AJUSTE PARABÓLICO:

$y_{ti} = f(x_i, a, b, c) = a + b\, x + c\, x_i^2$

$$\begin{cases} \dfrac{\partial H}{\partial a} = 0 \Rightarrow \displaystyle\sum_{i=1}^{N} y_i = N a + b \sum_{i=1}^{N} x_i + c \sum_{i=1}^{N} x_i^2 \\ \dfrac{\partial H}{\partial b} = 0 \Rightarrow \displaystyle\sum_{i=1}^{N} x_i\, y_i = a \sum_{i=1}^{N} x_i + b \sum_{i=1}^{N} x_i^2 + c \sum_{i=1}^{N} x_i^3 \\ \dfrac{\partial H}{\partial c} = 0 \Rightarrow \displaystyle\sum_{i=1}^{N} x_i^2\, y_i = a \sum_{i=1}^{N} x_i^2 + b \sum_{i=1}^{N} x_i^3 + c \sum_{i=1}^{N} x_i^4 \end{cases}$$

AJUSTE EXPONENCIAL:

$y_{ti} = f(x_i, a, b) = a\, b^{x_i}$

$\ln y_{ti} = \ln a + x_i \ln b$

$$\begin{cases} \sum_{i=1}^{N} \ln y_i = N \ln a + \ln b \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i \ln y_i = \ln a \sum_{i=1}^{N} x_i + \ln b \sum_{i=1}^{N} x_i^2 \end{cases}$$

AJUSTE POTENCIAL:

$y_{ti} = f(x_i, a, b) = a\, x_i^{b}$

$\ln y_{ti} = \ln a + b \ln x_i$

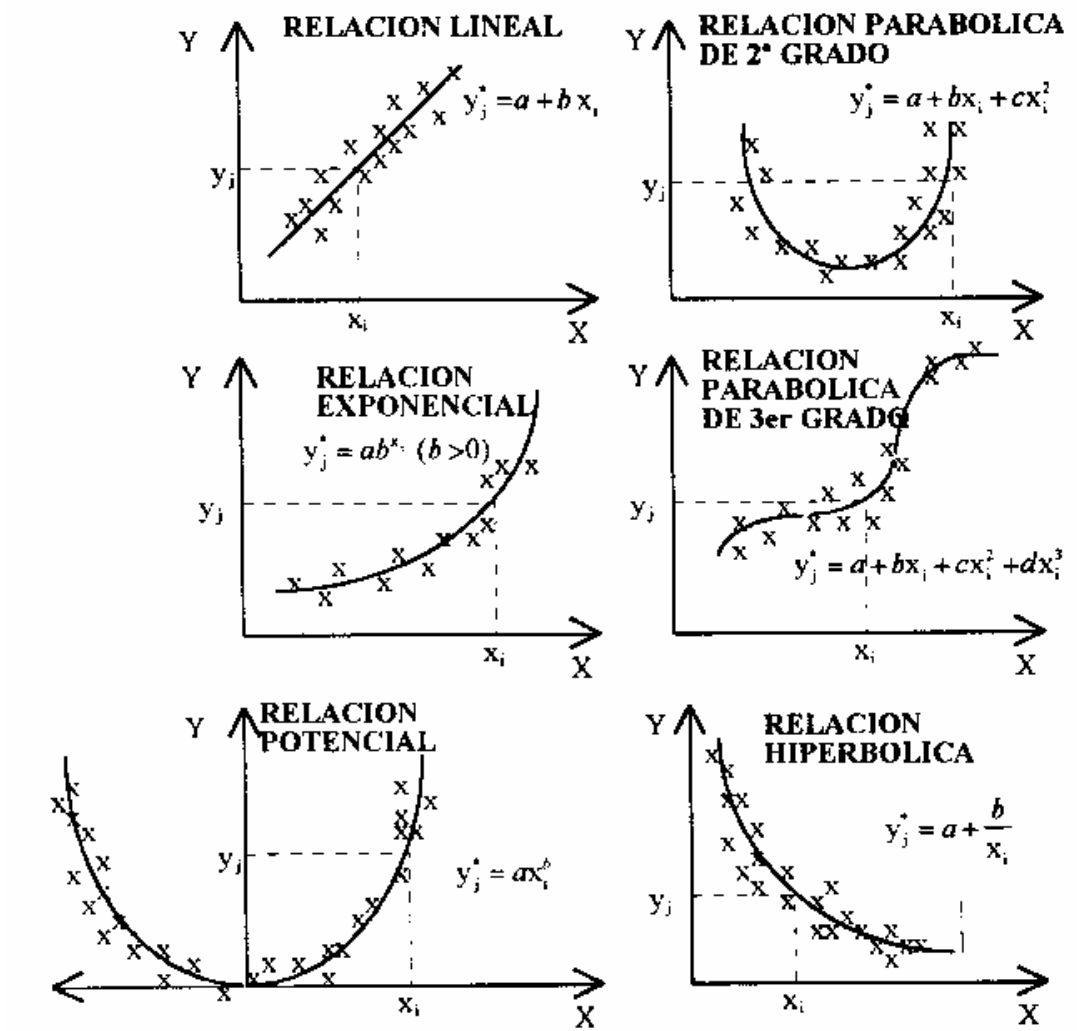
$$\begin{cases} \sum_{i=1}^{N} \ln y_i = N \ln a + b \sum_{i=1}^{N} \ln x_i \\ \sum_{i=1}^{N} \ln x_i \ln y_i = \ln a \sum_{i=1}^{N} \ln x_i + b \sum_{i=1}^{N} (\ln x_i)^2 \end{cases}$$

AJUSTE HIPERBÓLICO:

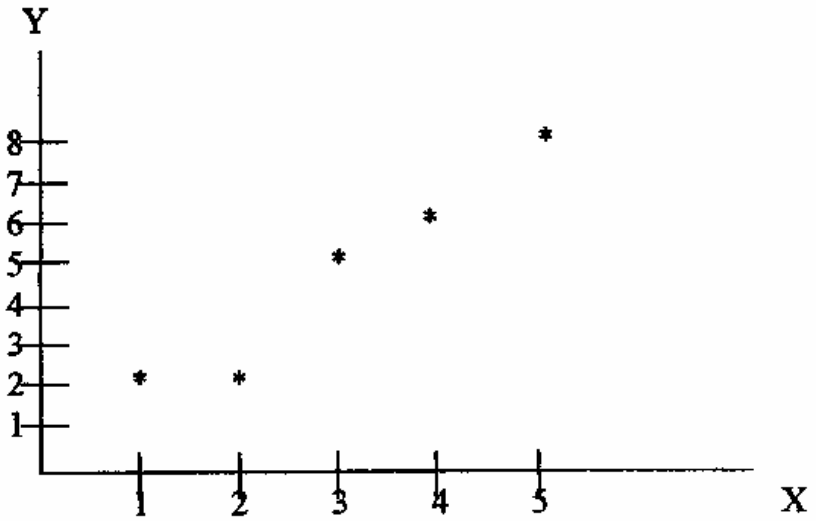
$y_{ti} = f(x_i, a, b) = a + b \dfrac{1}{x_i}$

$$\begin{cases} \sum_{i=1}^{N} y_i = N a + b \sum_{i=1}^{N} \dfrac{1}{x_i} \\ \sum_{i=1}^{N} \dfrac{1}{x_i}\, y_i = a \sum_{i=1}^{N} \dfrac{1}{x_i} + b \sum_{i=1}^{N} \dfrac{1}{x_i^2} \end{cases}$$

TIPOS DE AJUSTES



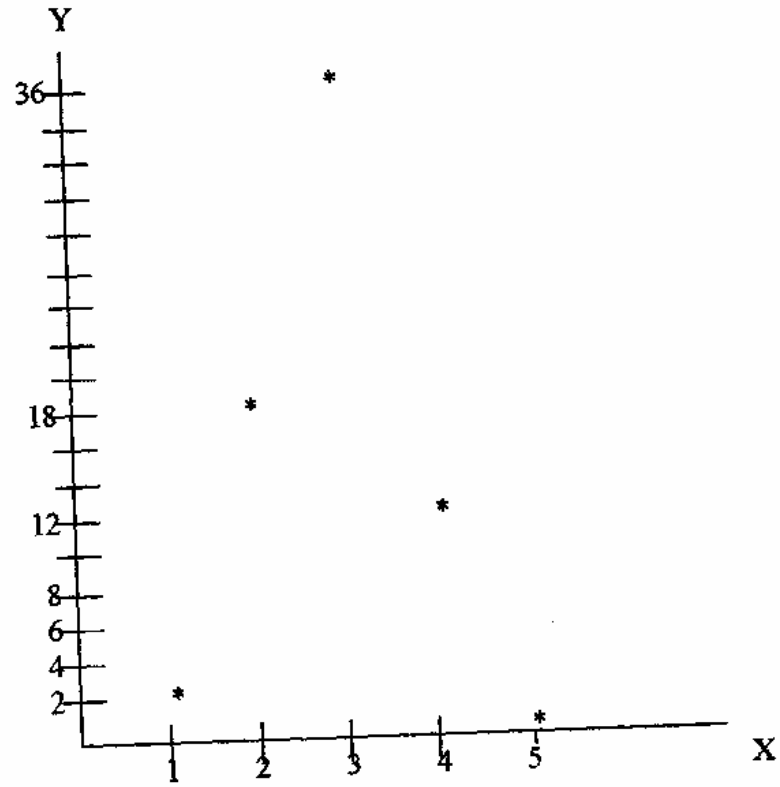
<u>Ejemplo 1:</u> Ajustar a las siguientes observaciones la función que mejor explique Y a partir de X. Los datos son: (1,2), (2,2), (3,5), (4,6), (5,8).



Y = - 0’2 + 1’6 X

<u>Ejercicio:</u> Ajustar a las siguientes observaciones la función que mejor explique Y a partir de X. Los datos son: (1,0’5), (1’5,1’7), (2,4), (2’5,8), (3,13’5).

<u>Ejemplo 2:</u> Ajustar a las siguientes observaciones la función que mejor explique Y a partir de X. Los datos son: (1,2), (2,18), (3,36), (4,12), (5,1).



Y = - 33 + 41’2 X – 7 X²

Regresión Lineal. Coeficientes de regresión.

A partir de las ecuaciones del **ajuste lineal por mínimos cuadrados** se van a obtener las **rectas de regresión de Y sobre X** y **de X sobre Y**:

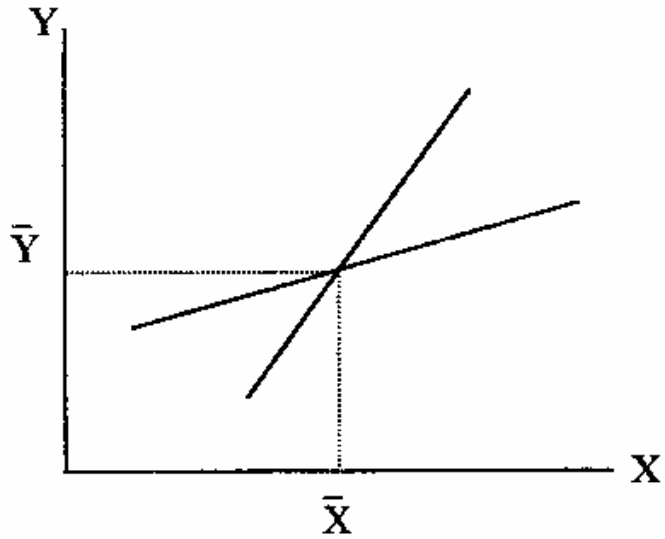
$$\left.\begin{aligned}\sum_{i=1}^{N} y_i &= N\,a + b\sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i\,y_i &= a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2\end{aligned}\right\} \Rightarrow \left\{\begin{aligned} b &= \frac{S_{XY}}{S_X^2} \\ a &= \bar{y} - \frac{S_{XY}}{S_X^2}\,\bar{x}\end{aligned}\right.$$

Recta de regresión de Y sobre X

$$y - \bar{y} = \frac{S_{XY}}{S_X^2}(x - \bar{x})$$

Recta de regresión de X sobre Y

$$x - \bar{x} = \frac{S_{XY}}{S_Y^2}(y - \bar{y})$$



Condición suficiente de minimización:

$$\frac{\partial^2 H}{\partial a^2} > 0 \quad y \quad \overline{H} = \begin{vmatrix} \frac{\partial^2 H}{\partial a^2} & \frac{\partial^2 H}{\partial a \partial b} \\ \frac{\partial^2 H}{\partial b \partial a} & \frac{\partial^2 H}{\partial b^2} \end{vmatrix} > 0$$

Ajuste lineal

$$\frac{\partial^2 H}{\partial a^2} = 2.N > 0$$

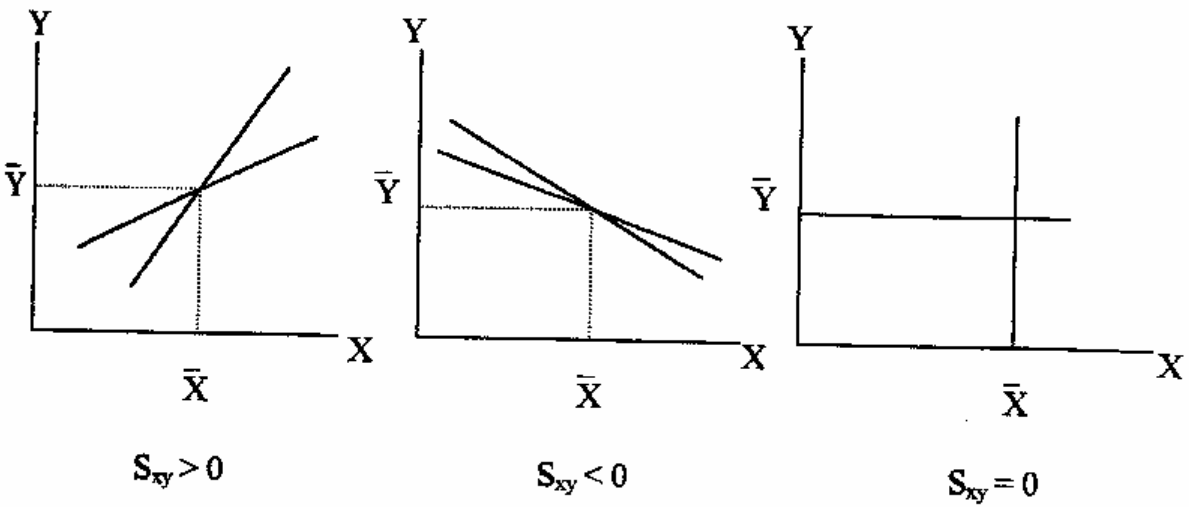
$$\overline{H} = 2.N.S_X^2 > 0$$

COEFICIENTES DE REGRESIÓN: Indican la pendiente de la recta de regresión correspondiente.

Recta de regresión de Y sobre X → $b = \frac{S_{XY}}{S_X^2}$

Recta de regresión de X sobre Y → $b' = \frac{S_{XY}}{S_Y^2}$

Ejercicio: Indicar cómo se comportan las pendientes de la rectas de regresión según el signo de la covarianza.



Coeficiente de determinación y de correlación lineal

Una vez que se ha realizado un ajuste para tratar de explicar una variable **Y** en función de otra variable **X**, necesitaremos obtener un indicador de la **bondad del ajuste** planteado.

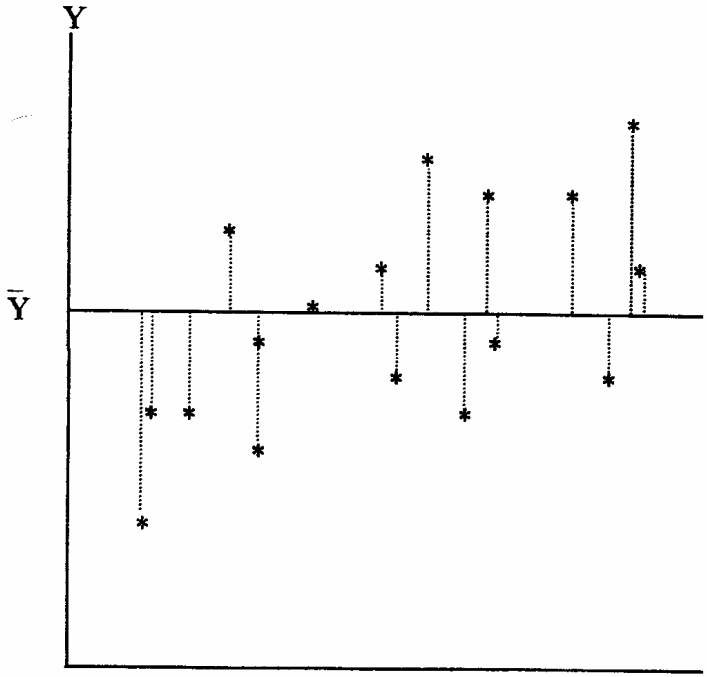


FIGURA A

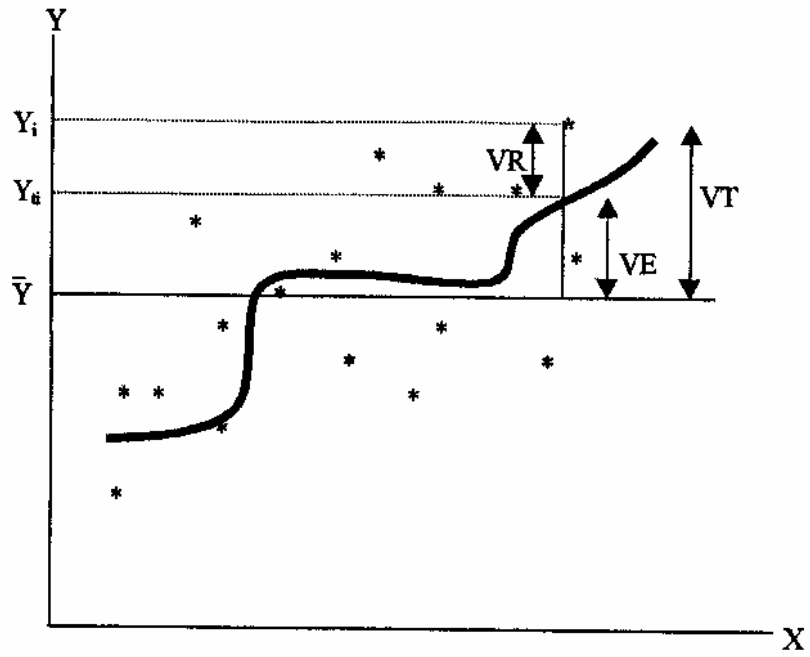


FIGURA B

Varianza total (VT)

$$VT = S_Y^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{N}$$

Varianza residual (VR)

$$VR = S_{ry}^2 = \frac{\sum_{i=1}^{N} d_i^2}{N} = \frac{\sum_{i=1}^{N} (y_i - y_{ti})^2}{N}$$

Varianza explicada (VE)

$$VT = VE + VR$$

Para medir la **bondad del ajuste** planteado podría considerarse la proporción de la varianza total que queda explicada por la regresión.

Coeficiente de determinación → $$R^2 = \frac{VE}{VT} = \frac{VT - VR}{VT} = 1 - \frac{VR}{VT} = 1 - \frac{S^2_{ry}}{S^2_Y}$$

Así, cuanto mayor sea el valor de **R^2**, mejor será el ajuste realizado, ya que la varianza residual sería pequeña.

$0 \leq R^2 \leq 1$

- **$R^2 = 0 \Leftrightarrow S^2_{ry} = S^2_Y$ (Ajuste pésimo)**
- **$R^2 = 1 \Leftrightarrow S^2_{ry} = 0$ (Ajuste perfecto)**

Para el caso de la **regresión lineal**, la varianza residual tomará el valor siguiente: $$S^2_{ry} = S^2_Y - \frac{S^2_{XY}}{S^2_X}$$

Coeficiente de determinación lineal → $$r^2 = 1 - \frac{S^2_{ry}}{S^2_Y} = \frac{S^2_Y - S^2_{ry}}{S^2_Y} = \frac{S^2_Y - \left(S^2_Y - \frac{S^2_{XY}}{S^2_X}\right)}{S^2_Y} = \frac{S^2_{XY}}{S^2_X S^2_Y}$$

NOTA: En el caso lineal, el coeficiente de determinación R^2 coincide para ambas rectas de regresión.

Coeficiente de correlación lineal simple

$$r = \sqrt{r^2} \qquad r = \frac{S_{XY}}{S_X\, S_Y}$$

Este coeficiente está directamente relacionado con los **coeficientes de regresión lineal**, **b** y **b'**, ya que:

$$r^2 = b.b' \quad b = r.\frac{S_Y}{S_X} \quad b' = r.\frac{S_X}{S_Y}$$

Usando estas relaciones, las **rectas de regresión** pueden expresarse de la siguiente forma:

Recta de regresión de Y sobre X

$$y - \overline{y} = r\,\frac{S_Y}{S_X}(x - \overline{x})$$

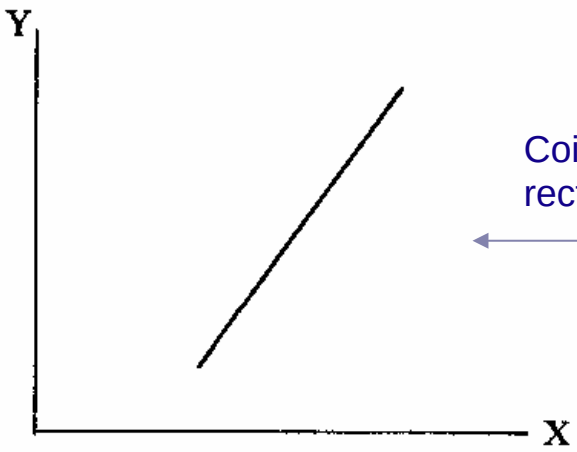
Recta de regresión de X sobre Y

$$x - \overline{x} = r\,\frac{S_X}{S_Y}(y - \overline{y})$$

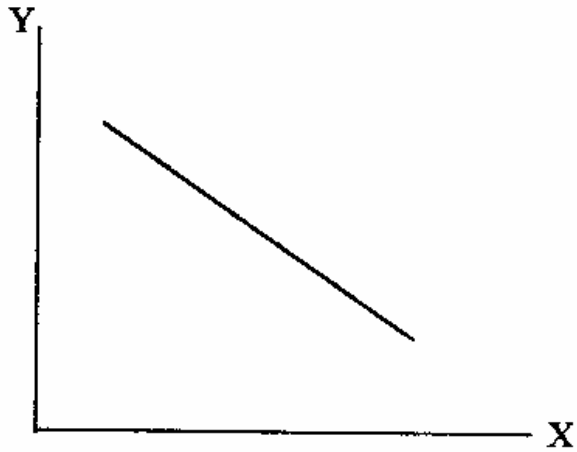
$-1 \le r \le 1$

El signo de **r** vendrá dado por el de la **covarianza S_{XY}**, por lo que, si **X** e **Y** varían en el mismo sentido, **r** será positivo, y si lo hacen en sentido opuesto, **r** será negativo.

CASOS POSIBLES:

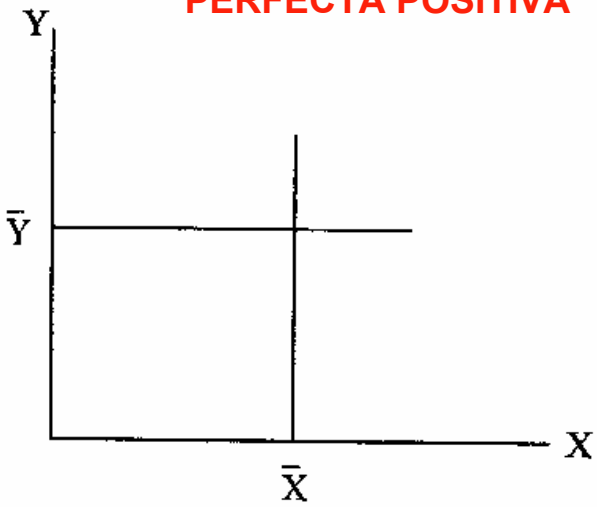


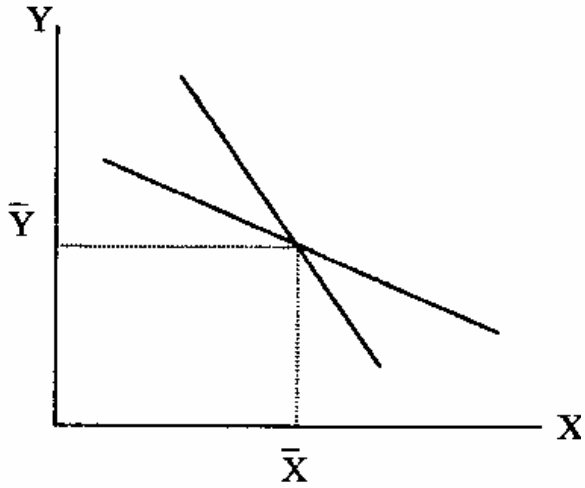
Coinciden las dos rectas de regresión

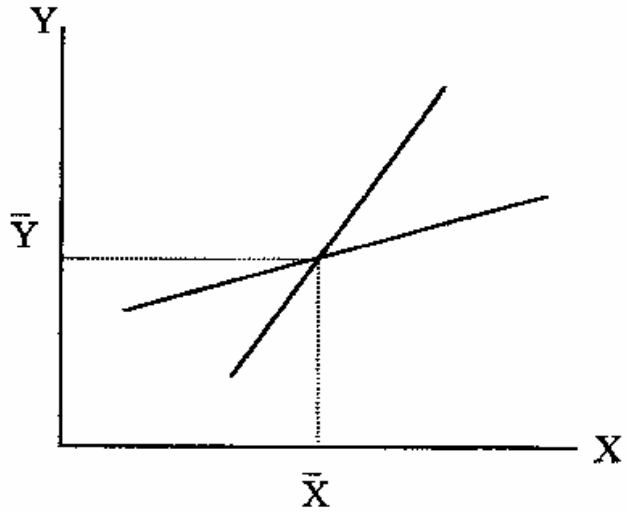


r = 1 → CORRELACIÓN LINEAL PERFECTA POSITIVA

r = -1 → CORRELACIÓN LINEAL PERFECTA NEGATIVA







r = 1 → CORRELACIÓN LINEAL NULA (No existe relación lineal)

-1 < r < 0 → CORRELACIÓN LINEAL POSITIVA

0 < r < 1 → CORRELACIÓN LINEAL NEGATIVA

Cuanto más se aleje **r** de 0, mejor será el ajuste lineal planteado entre ambas variables. El signo de **r** sólo nos indicará el sentido de la variación entre **X** e **Y**.

En resumen:

- A efectos de interpretar la bondad del ajuste lineal entre dos variables, se suele utilizar con más frecuencia el **coeficiente de determinación lineal r^2** en lugar del **coeficiente de correlación lineal r**.
- Si queremos obtener el sentido de variación de ambas variables, sí que debemos recurrir a **r** (o bien a S_{XY}).

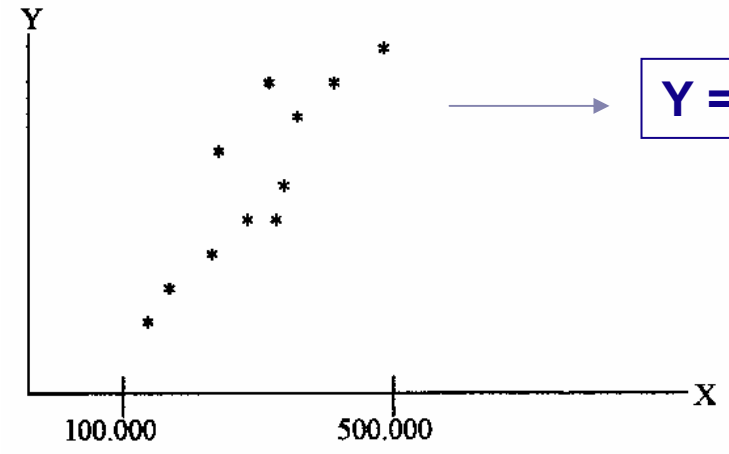
- Cuando se plantea un ajuste no lineal entre dos variables, debemos obtener el **coeficiente de determinación general R^2** para poder analizar la bondad de dicho ajuste.
- En este caso, no tiene mucho sentido hablar de **coeficiente de correlación general R**, ya que el signo carece de interpretación.

Predicción

La aplicación más interesante de la técnica de regresión es la de predecir valores de la variable dependiente para determinados valores de la variable independiente, que no aparezcan en la distribución de frecuencias.

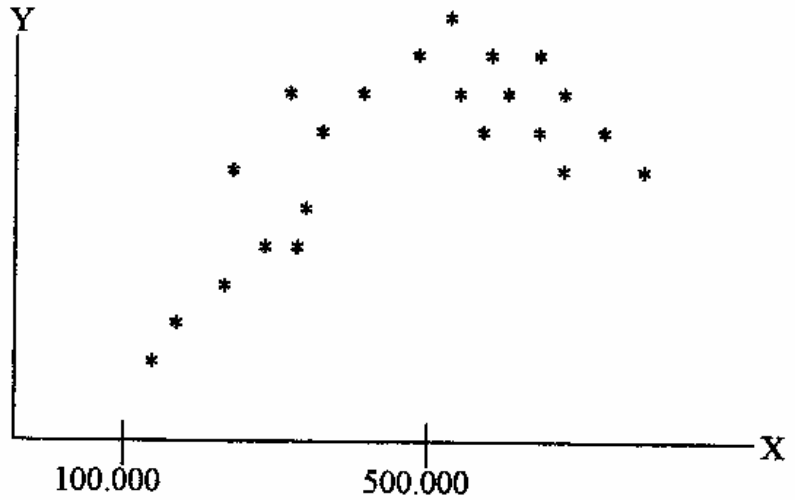
Cuando la predicción se realiza para valores de la variable independiente que pertenecen al intervalo de variación de los datos observados, de denomina **interpolación**. Si la predicción se hace para valores de la variable independiente situados fuera de dicho intervalo, recibe el nombre de **extrapolación**.

<u>Ejemplo:</u> Supongamos que hemos obtenido una recta de regresión que nos explica el gasto mensual por individuo en bebidas alcohólicas (Y) en función del sueldo mensual (X). Predecir el gasto en bebidas alcohólicas para un individuo que gana mensualmente 300000 ptas, y para otro que gana 700000 ptas.



Y = - 20000 + 0’22 . X

A continuación, comparar los valores obtenidos con los obtenidos a partir del gráfico siguiente, al que se le han añadido más observaciones.



Algunas consideraciones que hay que tener en cuenta a la hora de realizar predicciones son:

- La fiabilidad de la predicción será mayor cuanto mejor sea el ajuste, es decir, cuanto mayor sea el **R^2**.

- La fiabilidad de la predicción disminuye a medida que nos alejamos de los datos de partida.

- Al ir más allá de los datos originales, la predicción debe contemplarse desde una perspectiva inferencial para abordarla correctamente, quedando encuadrado fuera del marco de la Estadística Descriptiva.

Estadística Empresarial I

Tema 7

Números Índices

Introducción

Los **números índices** tratan de establecer una comparación de una serie de observaciones de una variable estadística (normalmente económica) respecto a una situación inicial fijada arbitrariamente.

Ejemplos: Para la variable precio de un artículo determinado:

-¿Cuánto se ha incrementado el precio con respecto al año 1995?

- Según el nivel de vida de cada año, ¿cuándo es más caro, ahora o en 1995?

Habrá que tener en cuenta dos aspectos:

- Fijación arbitraria del periodo inicial al que se referirán las comparaciones, lo más adecuada posible a los objetivos perseguidos.
- Comparación de magnitudes simples y complejas, lo que supone en muchos casos la agregación de magnitudes.

En definitiva, un **número índice** es una medida estadística abstracta que muestra los cambios de una variable en un **periodo** actual respecto a un **periodo base** o **de referencia**, temporal o espacial. La magnitud o variable que se estudia suele ser el precio **p**, la cantidad **q** o el valor **v=p.q**.

Índices Simples

Los **índices simples** son aquellos que hacen referencia a una magnitud medible. Dada una magnitud **X** y su evolución temporal (espacial):

T	0	1	2	...	t
X	x_0	x_1	x_2	...	x_t

Índice simple de la magnitud **X** en el periodo actual **t** respecto al periodo base **0**.

$$I_0^t(X) = \frac{x_t}{x_0} 100$$

El **índice simple** recoge el porcentaje de incremento o disminución de la magnitud de un solo bien o servicio. Según el tipo de magnitud con la que se trabaje, se obtiene:

Índice de precios

$$I_t^0(P) = \frac{p_t}{p_0} 100$$

Índice cuántico

$$I_t^0(Q) = \frac{q_t}{q_0} 100$$

Índice de valor

$$I_t^0(V) = \frac{v_t}{v_0} 100 = \frac{p_t q_t}{p_0 q_0} 100 = I_0^t(P).I_0^t(Q)$$

NOTA: Los números índices pueden expresarse en tanto por ciento, pero a la hora de trabajar con ellos se hace en tantos por uno.

PROPIEDADES DE LOS ÍNDICES SIMPLES:

- **Existencia:** Todo número índice simple debe existir y tomar un valor finito no nulo.
- **Identidad:** $I_0^0(X) = I_t^t(X) = 1$
- **Inversión:** $I_t^0(X) = \frac{1}{I_0^t(X)}$
- **Circularidad:** $I_t^{t'}(X).I_{t'}^{t''}(X).I_{t''}^{t}(X) = 1$
- **Cambio de base:** Podemos obtener los índices respecto a otro periodo base o de referencia **t'**:

$$I_{t'}^{t}(X) = \frac{I_0^t(X)}{I_0^{t'}(X)}$$

- **Índice de producto de magnitudes:** $I_0^t(X.Y) = I_0^t(X)\, I_0^t(Y)$
- **Índice de cociente de magnitudes:** $I_0^t\left(\frac{X}{Y}\right) = \frac{I_o^t(X)}{I_o^t(Y)}$
- **Proporcionalidad:** Si $x_t' = (1+k)\, x_t$, entonces $I'^t_0(X) = (1+k)\, I_0^t(X)$
- **Homogeneidad:** A un número índice no le afectan las unidades de medida.

Ejemplo: A continuación se muestran los precios en miles de ptas de un determinado artículo en varios años diferentes:

T	1998	1999	2000	2001
X	107	110	116	119

(a) Indicar cúanto ha variado el precio de dicho artículo para cada año con respecto al año 1998.

(b) Calcular los índices de precios para cada año considerando como periodo base el año 1999.

Para cada uno de los artículos que integran un determinado sector, se puede calcular un **número índice simple** que indique la evolución de su precio, cantidad o valor; pero puede ser interesante obtener un número índice único que represente de manera conjunta a todos los artículos, a partir de los números índices simples calculados. A esos número índices que representan a un conjunto de magnitudes se les llama **números índices complejos**.

Índices Complejos

Los **índices complejos** son los que hacen referencia a una magnitud compleja. Se van a obtener a partir de un conjunto de **índices simples**, resumiéndolos de manera que refleje el comportamiento global de la magnitud.

Sea la magnitud **X** referida a **N** artículos:

Artículo/Periodo	0	t	Índices simples
1	x_{10}	x_{1t}	$I_1 = x_{1t} / x_{10}$
2	x_{20}	x_{2t}	$I_2 = x_{2t} / x_{20}$
:	**:**	**:**	**:**
N	x_{N0}	x_{Nt}	$I_N = x_{Nt} / x_{N0}$

Los **índices complejos** pueden ser **no ponderados** o **ponderados**. La **ponderación** recoge la importancia relativa de cada magnitud simple dentro del conjunto de todas ellas.

ÍNDICES COMPLEJOS NO PONDERADOS:

Índice media aritmética

$$\bar{I} = \frac{I_1 + \ldots + I_N}{N} = \sum_{i=1}^{N} \frac{I_i}{N}$$

Índice media geométrica

$$I_G = \sqrt[N]{I_1 \ldots I_N} = \sqrt[N]{\prod_{i=1}^{N} I_i}$$

Índice media armónica

$$I_H = \frac{N}{\frac{1}{I_1} + \ldots + \frac{1}{I_N}} = \frac{1}{\sum_{i=1}^{N} \frac{1}{I_i}}$$

Índice media agregativa

$$I_A = \frac{x_{1t} + \ldots + x_{Nt}}{x_{10} + \ldots + x_{N0}} = \frac{\sum_{i=1}^{N} x_{it}}{\sum_{i=1}^{N} x_{i0}}$$

- No es un índice obtenido a partir de los números índices simples, tiene sentido en aquellos casos en que alguno de los índices simples no está definido (da un valor 0 o ∞).
- Sólo puede emplearse si las magnitudes vienen expresadas en la mismas unidades.

ÍNDICES COMPLEJOS PONDERADOS: Tienen en cuenta la importancia relativa de cada magnitud simple dentro del conjunto de ellas.

Índice media aritmética ponderado

$$\bar{I}^* = \frac{I_1 w_1 + \ldots + I_N w_N}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i}$$

Índice media geométrica ponderado

$$I_G^* = \sqrt[\sum_{i=1}^{N} w_i]{I_1^{w_1} \ldots I_N^{w_N}} = \sqrt[\sum_{i=1}^{N} w_i]{\prod_{i=1}^{N} I_i^{w_i}}$$

Índice media armónica ponderado

$$I_H^* = \frac{\sum_{i=1}^{N} w_i}{\frac{1}{I_1} w_1 + \ldots + \frac{1}{I_N} w_N} = \frac{\sum_{i=1}^{N} w_i}{\sum_{i=1}^{N} \frac{1}{I_i} w_i}$$

Índice media agregativa ponderado

$$I_A^* = \frac{x_{1t}w_1 + \ldots + x_{Nt}w_N}{x_{10}w_1 + \ldots + x_{N0}w_N} = \frac{\sum_{i=1}^{N} x_{it}w_i}{\sum_{i=1}^{N} x_{i0}w_i}$$

Si la magnitud **X** considerada es el **precio**, se han considerado cuatro sistemas de ponderación:

(1) $\mathbf{w_i = p_{i0}\, q_{i0}}$ → valor de la cantidad del bien i-ésimo en el periodo base, a precios de dicho periodo.
(2) $\mathbf{w_i = p_{it}\, q_{it}}$ → valor de la cantidad del bien i-ésimo en el periodo actual, a precios de dicho periodo.
(3) $\mathbf{w_i = p_{i0}\, q_{it}}$ → valor de la cantidad del bien i-ésimo en el periodo actual, a precios del periodo base.
(4) $\mathbf{w_i = p_{it}\, q_{i0}}$ → valor de la cantidad del bien i-ésimo en el periodo base, a precios del periodo actual.

NOTAS:

- Los sistemas **(1)** y **(2)** corresponden a situaciones reales, mientras que **(3)** y **(4)** no.
- Los sistemas **(2)** y **(3)** tienen el gran inconveniente de que necesitan conocer las cantidades consumidas en el periodo actual, lo cual no es simple posible.
- Los sistemas más utilizados son el **(1)** y **(3)**.

Si la magnitud **X** es la **cantidad**, se utilizan los mismos sistemas de ponderación, cambiando precios (**p**) por cantidades (**q**).

Índice de precios de Laspeyres: Se trata de un índice media aritmética ponderado obtenido usando el sistema de ponderación **(1)**: $\mathbf{w_i = p_{i0}\, q_{i0}}$

$$L_p = \bar{I}(P) = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} p_{i0} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} = \frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}}$$

Determina el incremento de valor que experimenta un conjunto de artículos o bienes entre los periodos **0** y **t**, suponiendo que las cantidades consumidas son las mismas para ambos periodos e iguales a $\mathbf{q_{i0}}$.

Índice de cuántico de Laspeyres: Se trata de un índice media aritmética ponderado obtenido usando el sistema de ponderación **(1)**: $\mathbf{w_i = q_{i0}\, p_{i0}}$

$$L_q = \bar{I}(Q) = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}} q_{i0} p_{i0}}{\sum_{i=1}^{N} q_{i0} p_{i0}} = \frac{\sum_{i=1}^{N} q_{it} p_{i0}}{\sum_{i=1}^{N} q_{i0} p_{i0}}$$

Determina el incremento de valor que experimenta un conjunto de artículos o bienes entre los periodos **0** y **t**, suponiendo que los precios son las mismos para ambos periodos e iguales al del periodo **0**, $\mathbf{p_{i0}}$.

Índice de precios de Paasche: Se trata de un índice media aritmética ponderado obtenido usando el sistema de ponderación **(3)**: $\mathbf{w_i = p_{i0}\, q_{it}}$

$$P_p = \bar{I}(P) = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} p_{i0} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{it}} = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{it}}$$

Determina el incremento de valor que experimenta un conjunto de artículos o bienes entre los periodos **0** y **t**, suponiendo que las cantidades consumidas son las mismas para ambos periodos e iguales a $\mathbf{q_{it}}$.

Índice de cuántico de Paasche: Se trata de un índice media aritmética ponderado obtenido usando el sistema de ponderación **(3)**: $\mathbf{w_i = q_{i0}\, p_{it}}$

$$P_q = \bar{I}(Q) = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}} q_{i0} p_{it}}{\sum_{i=1}^{N} q_{i0} p_{it}} = \frac{\sum_{i=1}^{N} q_{it} p_{it}}{\sum_{i=1}^{N} q_{i0} p_{it}}$$

Determina el incremento de valor que experimenta un conjunto de artículos o bienes entre los periodos **0** y **t**, suponiendo que los precios son las mismos para ambos periodos e iguales al del periodo **t**, $\mathbf{p_{it}}$.

DEFLACTACIÓN:

A partir del **índice de precios de Paasche P_p** se puede estimar el valor de los bienes y servicios del periodo actual en unidades monetarias del periodo base.

$$\sum_{i=1}^{N} p_{i0}\, q_{it} = \frac{\sum_{i=1}^{N} p_{it}\, q_{it}}{P_{p_0}^{\ t}}$$

A esta propiedad se le conoce como **deflactación**, y permite corregir el efecto de la pérdida del valor del dinero y hacer comparaciones en una unidad común.

- Cuando se valoran los bienes y servicios a precios de un mismo periodo, hablaremos de **precios constantes** o **reales**.
- Cuando se valoran los bienes y servicios a precios de cada periodo, hablaremos de **precios constantes** o **reales**.

En la práctica, se presenta el problema de que el **Índice de Paasche** no se suele obtener, ya que necesita las cantidades consumidas en el periodo actual (q_{it}). Por ello, se suele utilizar como **deflactor** el **Índice de Laspeyres** o el **Índice de Precios al Consumo (IPC).**

DEFLACTACIÓN → $$\text{Valor real o constante} = \frac{\text{Valor monetario o corriente}}{\text{Deflactor}}$$

Ejemplo: El precio del kilogramo de plátanos entre los años 1995 y 1998 y el IPC de cada año (con respecto al año 1995) aparecen en la tabla adjunta. ¿En qué año estuvo más barato y más caro el Kg de plátanos?

t	Precio (kg)	IPC
1995	50	100
1996	55	110
1997	60	116
1998	62	125

Ejemplo: Conocidos los precios y cantidades de 2 artículos de consumo correspondientes a tres años, determinar los índices de precios y de cantidades de Laspeyres y Paasche con base 1998.

Años	Artículo A		Artículo B	
	Precio	Cantidad	Precio	Cantidad
1998	2	10	5	12
1999	3	15	6	10
2000	4	20	7	6

Índice de Precios de Consumo

El **IPC** es un índice de precios que se obtiene en España por parte del INE, a nivel nacional, por comunidades autónomas y por provincias, con una periodicidad mensual, recogiendo el incremento de valor de un grupo representativo de los productos y servicios consumidos por todas las familias del país, que forman la **cesta de la compra**.

Hasta el año 1997, se utilizaba un **índice de Laspeyres** con periodo base fijo. El principal problema que presenta es que la estructura de ponderaciones pierde vigencia con el paso del tiempo.

A partir del segundo trimestre de 1997 se implantó la **Encuesta Continua de Presupuestos Familiares (ECPF)**, que permite disponer de información sobre el gasto de las familias de forma más detallada y con una periodicidad menor que antes. Este nuevo sistema es más **dinámico**, al permitir:

- Actualizar las **ponderaciones** en periodos cortos de tiempo.
- Incluir nuevos productos cuando su consumo comience a ser significativo, así como eliminar los que sean poco significativos.

De esta forma, se crea un sistema de actualización continua de la estructura de consumo, basado en un flujo de información entre el **IPC** y el **ECPF**. Esta actualización se materializa en:

- Una revisión anual de las ponderaciones.
- Un completo cambio de base cada 5 años: composición de la cesta de la compra, revisión profunda de las ponderaciones y de la definición del IPC.

Para obtener el **IPC** base 2001, se utilizará una **cesta de la compra** que clasifica los productos y servicios en 12 grupos:

1. Alimentos y bebidas no alcohólicas.
2. Bebidas alcohólicas y tabaco.
3. Vestido y calzado.
4. Vivienda.
5. Menaje.
6. Medicina.
7. Transporte.
8. Comunicaciones.
9. Ocio y cultura.
10. Enseñanza.
11. Hoteles, café y restaurantes.
12. Otros.

Para calcular el nuevo **IPC** se utilizará un **índice de Laspeyres encadenado**, que consiste en referir los precios del periodo corriente a los precios del año anterior, actualizándose las ponderaciones con información de la **ECPF**.

$$IPC_{t-1}^{t} = \frac{\sum_{i=1}^{N} I_i \, w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{it-1}} p_{it-1} q_{it-1}}{\sum_{i=1}^{N} p_{it-1} q_{it-1}}$$

$$I_{01}^{04} = I_{03}^{04} . C_{01}^{03}$$

Estos índices se enlazan a través de un **coeficiente de enlace C**.

Estadística Empresarial I

Tema 8

Series Temporales

Introducción

Hasta ahora, se han estudiado las observaciones de una determinada variable estadística, organizadas mediante una distribución de frecuencias, sin tener en cuenta el instante en el tiempo en que fueron tomadas. Sin embargo, en muchos problemas económicos, interesa disponer de datos registrados en intervalos de tiempo sucesivos, que constituyen una **serie temporal**.

Un hecho que distingue las observaciones ordenadas en el tiempo del resto es que las diferentes observaciones que forman una **serie temporal** no son independientes una de otras.

Ejemplo: El número de automóviles fabricados en enero de 1989 no es independiente de los que se fabricaron en diciembre de 1988.

Por tanto, las variables que se estudian en las ciencias sociales y económicas están sujetas a cambios a lo largo del tiempo.

Análisis de Series Temporales

Una **serie temporal** es una sucesión de observaciones numéricas referidas a un fenómeno, mediante una variable o conjunto de variables, dispuestas en orden cronológico de ocurrencia.

Así, la **serie temporal** describe la variación de los valores de la variable en el tiempo, como resultado del **comportamiento sistemático** o **aleatorio** de dicha variable. Si una serie muestra alguna tendencia en su variación durante un periodo de tiempo prolongado del pasado, parece lógico suponer que tales regularidades seguirán existiendo en el futuro, y podrán establecerse así **predicciones** sobre valores futuros.

Las observaciones pueden obtenerse:

- En un momento dado. *Ejemplos: nº de coches en la cola de una gasolinera, precio de un producto, etc.*
- Como suma de cantidades asociadas a un periodo. *Ejemplos: producción anual de energía, nº de nacimientos al mes, etc.*
- Como promedio de un periodo. *Ejemplos: Media mensual de trabajadores afiliados a la S.S., tasa trimestral de actividad, etc.*

Las publicaciones de datos estadísticos contienen en su mayor parte **series temporales** que vienen expresadas en cifras absolutas y en cifras relativas.

PRODUCCIÓN NACIONAL DE AUTOMÓVILES						
Años	1984	1985	1986	1987	1988	1989
Turismos	119510	154954	249405	273524	310556	368991

ÍNDICES DE PRODUCCIÓN MENSUAL DE TV FABRICADOS						
Años (media mensual)	1984	1985	1986	1987	1988	1989
Índices	100	264,2	221	171,8	296,2	379,8

Aunque los datos de las **series temporales** requieren una menor organización preliminar que los datos asociados a una **distribución de frecuencias**, conviene tomar ciertas precauciones:

- Las fechas a las que se aplican las cifras deberán entenderse claramente y estar definidas de forma precisa.
- Los datos correspondientes a los distintos periodos considerados deben ser comparables entre sí, y obtenidos en las mismas condiciones y unidades.

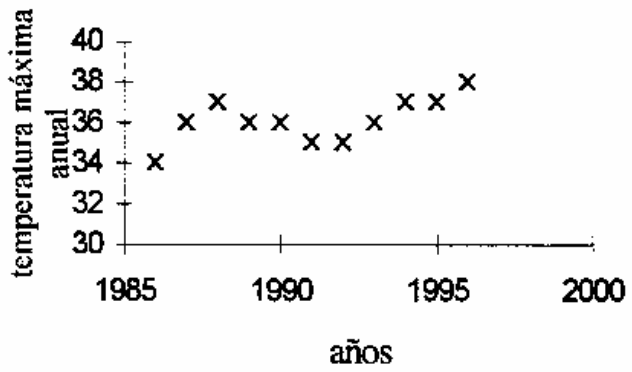
Cuando se dispone de datos correspondientes a una **serie temporal**, conviene comenzar su análisis mediante una **representación gráfica**, siendo la más utilizada el gráfico en coordenadas cartesianas, considerando en el eje de las abscisas la variable tiempo y en el de ordenadas la variable estudiada.

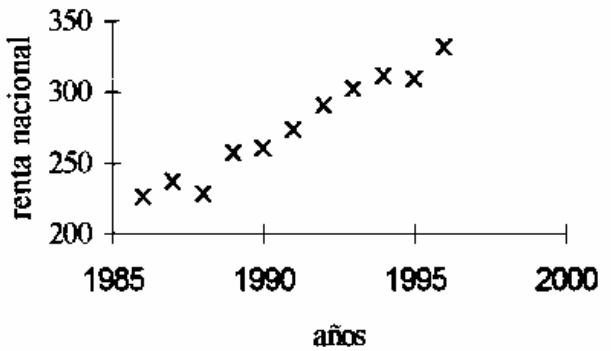
AÑOS	TEMPERATURA MÁXIMA ANUAL DE ESPAÑA*
1986	34
1987	36
1988	37
1989	36
1990	36
1991	35
1992	35
1993	36
1994	37
1995	37
1996	38

*En grados centígrados

AÑOS	RENTA NACIONAL ESPAÑOLA*
1986	226
1987	237
1988	228
1989	257
1990	260
1991	273
1992	290
1993	302
1994	311
1995	309
1996	331

*En miles de millones de pesetas

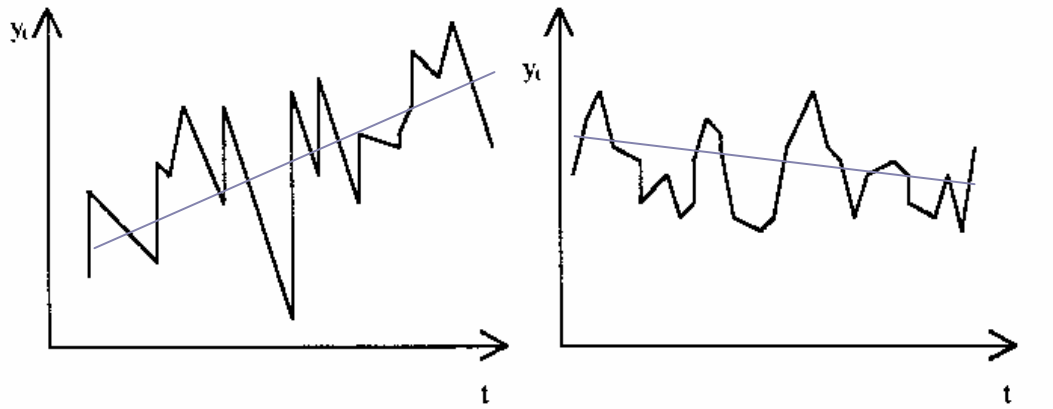




Naturaleza de las Series Temporales

Una **serie temporal** está formada por varias **componentes**, que son las encargadas de explicar los cambios observados en la variable a lo largo del tiempo. La descomposición más común es la que distingue las componentes **tendencial**, **estacional**, **cíclica** y **aleatoria**, propuesta por el **enfoque clásico** de las **series temporales**.

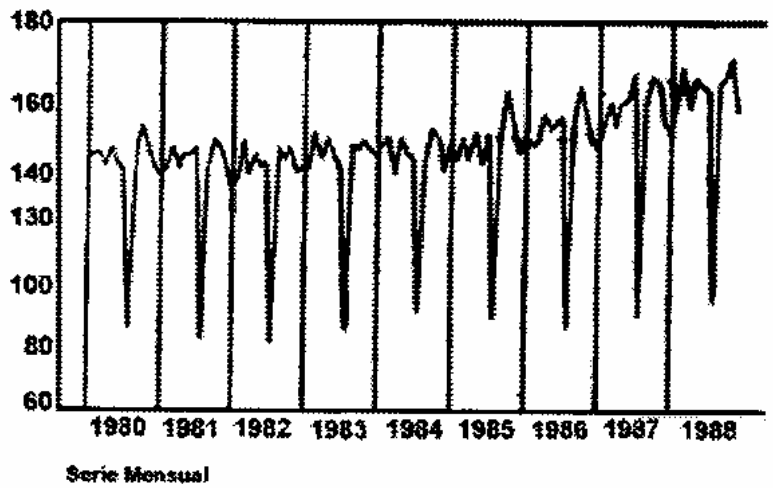
1. Tendencia regular o secular: Es el comportamiento a largo plazo que presenta la serie, ignorando las fluctuaciones a corto y medio plazo. Esta *componente tendencial* puede presentar pautas de crecimiento, decrecimiento o estabilidad.



2. Variaciones estacionales: Son las oscilaciones a corto plazo que se reproducen de forma periódica más o menos regular con periodo constante igual o inferior al año, debidas principalmente a las influencias de las estaciones del año, causas climatológicas, costumbres, etc.

Ejemplos: Las temperaturas medias mensuales tienen cada año un máximo en verano y un mínimo en invierno, por lo que presentan periodicidad anual; el volumen de compras diarias que se realiza en un supermercado presenta máximos y mínimos a principios y a finales de mes, respectivamente, luego presenta periodicidad mensual,...

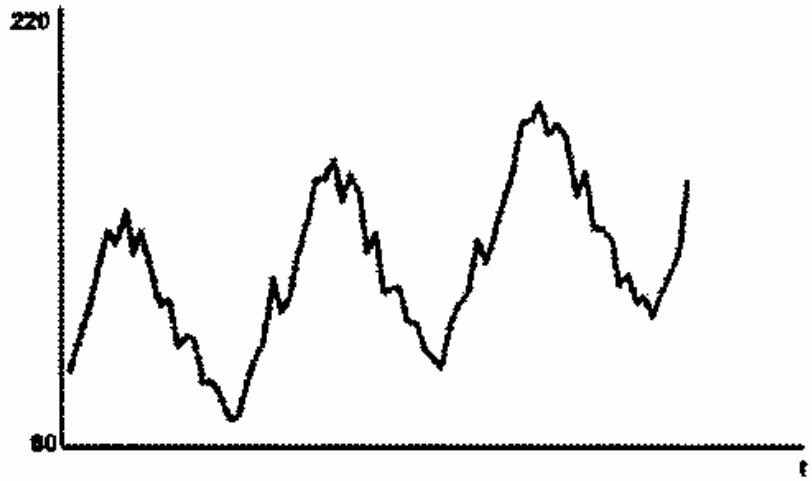
En la gráfica siguiente se representa la serie del IPI (Índice de Producción Industrial) y se observa la caída del mes de agosto, comportamiento que se repite de forma regular y periódica.

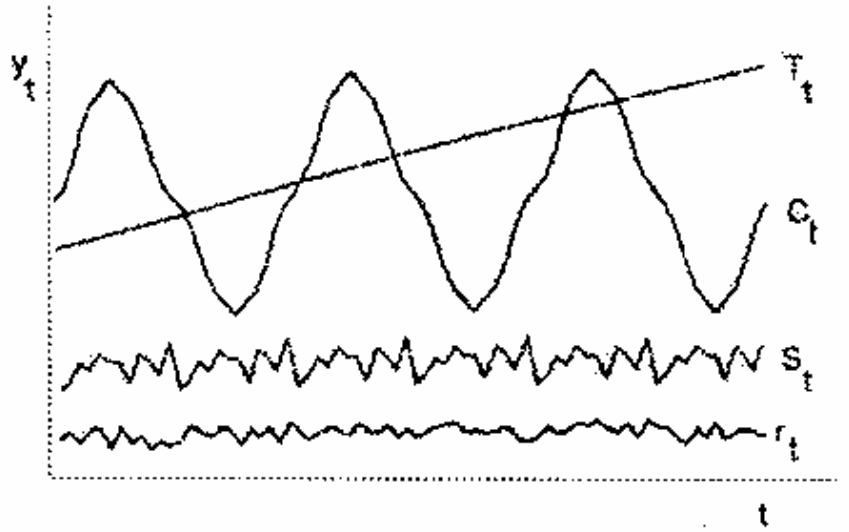


3. Movimientos cíclicos: Son movimientos a medio plazo que se reproducen de manera periódica, pero no tan regular como los de la *componente estacional*. Con un periodo no constante y más amplio que los periodos estacionales, los ciclos observados en series económicas están asociados principalmente a la alternancia de etapas de prosperidad y depresión de la actividad económica.

4. Variaciones irregulares o aleatorias: Son comportamientos que no muestran carácter periódico ni regular y que se deben a fenómenos catastróficos o fortuitos que afectan de manera casual a la variable, como pueden ser inundaciones, terremotos, incendios, accidentes, huelgas,...

Dada una **serie temporal**, el objetivo será descomponerla en cada una de las cuatro **componentes** consideradas.





Generalmente, las **componentes** de la **serie temporal** se pueden combinar mediante tres esquemas o modelos:

MODELO ADITIVO:
$$Y_i = T_i + E_i + C_i + I_i$$

MODELO MULTIPLICATIVO I:
$$Y_i = T_i . E_i . C_i . I_i$$

MODELO MULTIPLICATIVO II:
$$Y_i = T_i . E_i . C_i + I_i$$

Un supuesto fundamental en el análisis clásico de las series temporales es la independencia de las variaciones residuales respecto a las demás componentes. Este supuesto se verifica en el **modelo aditivo** y en el **multiplicativo II**.

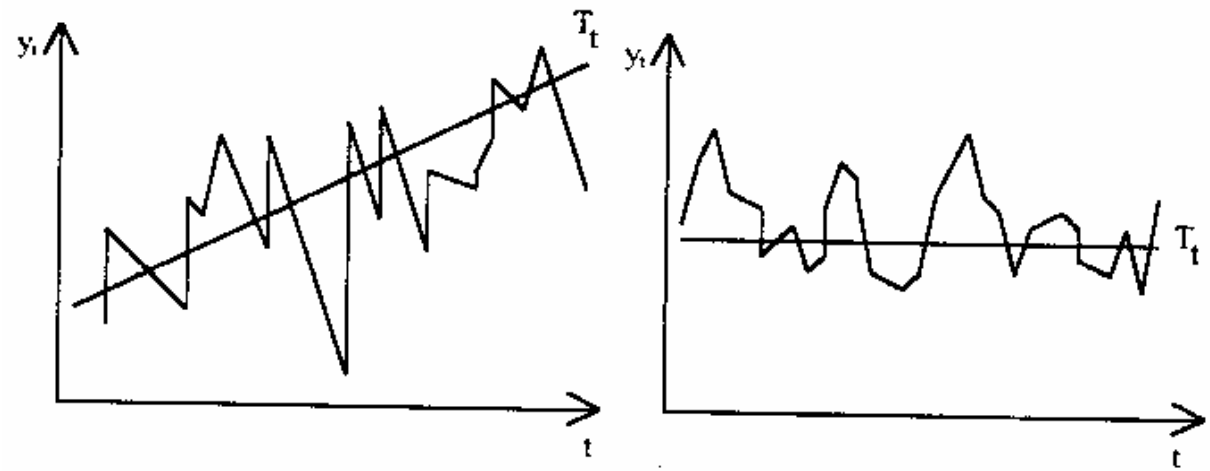
De los dos citados, se utiliza más en la práctica el modelo **multiplicativo**, ya que las variaciones relativas o porcentuales representan mejor las situaciones que las variaciones absolutas. En él, sólo la **componente tendencial** viene expresada en términos absolutos, mientras que las demás componentes vienen expresadas en forma de números índice.

Análisis de la Tendencia Secular

Los procedimientos estadísticos que se utilizan para estimar la **componente tendencial** (responsable del comportamiento a largo plazo de la serie) se dividen en **analíticos** y **no analíticos**.

MÉTODOS NO ANALÍTICOS:

1. Ajuste gráfico: Consiste en trazar una línea, ajustada a las observaciones, que refleje el comportamiento a largo plazo de la serie, ignorando fluctuaciones a corto y medio plazo.



2. Ajuste por medias móviles: Consiste en hallar las medias de cada grupo de **R** observaciones consecutivas, siendo **R** generalmente el número de observaciones anuales de las que se dispone.

MEDIAS MÓVILES DE ORDEN **R**

R IMPAR

$$\bar{y}_{\frac{R+1}{2}} = \frac{y_1 + y_2 + \ldots + y_p}{R}$$

$$\bar{y}_{\frac{R+3}{2}} = \frac{y_2 + y_3 + \ldots + y_{p+1}}{R}$$

$$\bar{y}_{\frac{R+5}{2}} = \frac{y_3 + y_4 + \ldots + y_{p+2}}{R}$$

$$\vdots$$

R PAR

$$\bar{\bar{y}}_{\frac{R+2}{2}} = \frac{\bar{y}_{\frac{R+1}{2}} + \bar{y}_{\frac{R+3}{2}}}{2}$$

$$\bar{\bar{y}}_{\frac{R+4}{2}} = \frac{\bar{y}_{\frac{R+3}{2}} + \bar{y}_{\frac{R+5}{2}}}{2}$$

$$\bar{\bar{y}}_{\frac{R+6}{2}} = \frac{\bar{y}_{\frac{R+5}{2}} + \bar{y}_{\frac{R+7}{2}}}{2}$$

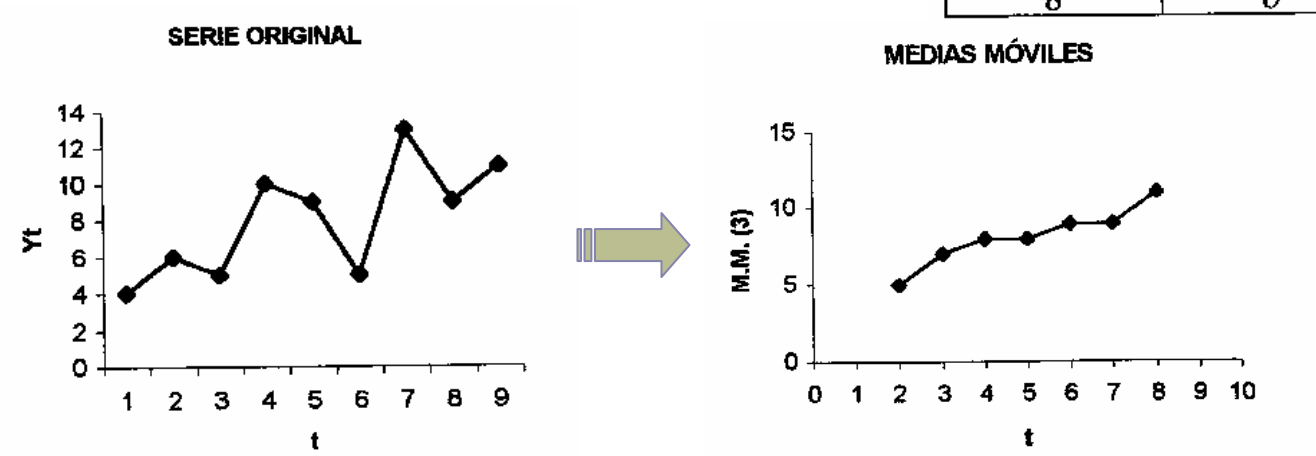
$$\vdots$$

Si **R** es par, las **medias móviles** quedan descentradas, por lo que habrá que calcular de nuevo las **medias móviles de orden 2**.

Ejemplos: Consideremos las dos series siguientes, una cuatrimestral y otra trimestral.

AÑOS	t	Y_t	MM (3)
1998	1	4	-
	2	6	5
	3	5	7
1999	4	10	8
	5	9	8
	6	5	9
2000	7	13	9
	8	9	11
	9	11	-

t	Y_t	MM (4)	MM(2)
1	3		-
2	2		-
		3	
3	5		3,5
		4	
4	2		4,5
		5	
5	7		4,5
		4	
6	6		4,5
		5	
7	1		
8	6		



Con este método se suaviza la serie, ya que se consiguen eliminar las **oscilaciones estacionales**. Sin embargo, cuanto mayor sea el orden **R**, más observaciones se pierden.

MÉTODOS ANALÍTICOS:

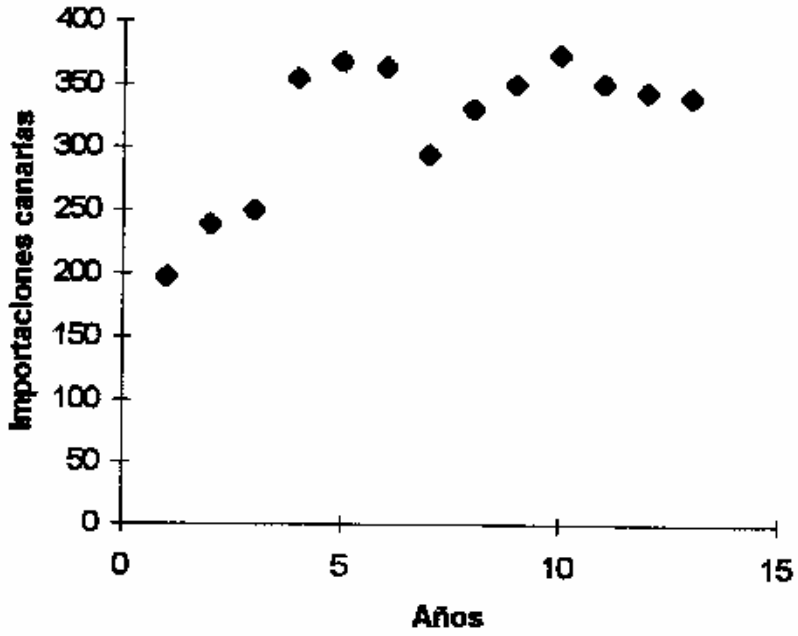
En muchos casos, la **tendencia** puede representarse mejor mediante una función matemática **Y = f(t)** que mediante la poligonal de las **medias móviles**.

Para obtener dicha función, primero habrá que representar gráficamente la **serie temporal**, y decidir qué tipo de ajuste es el más adecuado para la regresión de **Y** sobre **t**. A continuación, se obtienen los coeficientes de la curva de regresión a través del **método de los mínimos cuadrados**, pudiéndose medir la **bondad del ajuste** planteado mediante el **coeficiente de determinación R^2**.

Es interesante señalar que si la **serie temporal** presenta un cambio brusco en su **tendencia**, es aconsejable ajustar diferentes funciones a cada conjunto de datos que presenten una **tendencia** homogénea.

Ejemplo: La siguiente serie refleja la evolución de las importaciones del extranjero en Canarias en miles de millones de ptas entre 1980 y 1992.

AÑOS	IMPORTACIONES DEL EXTRANJERO EN CANARIAS (miles millones pts)
1980	196,93
1981	238,76
1982	249,98
1983	354,55
1984	367,86
1985	363,64
1986	293,89
1987	331,21
1988	350,28
1989	374,52
1990	351,47
1991	344,96
1992	340,95



AÑOS	t	Y_i	$Y_i * t$	t^2
1980	-6	196,93	-1181,58	36
1981	-5	238,76	-1193,80	25
1982	-4	249,98	-999,92	16
1983	-3	354,55	-1063,65	9
1984	-2	367,86	-735,72	4
1985	-1	363,64	-363,64	1
1986	0	293,89	0	0
1987	1	331,21	331,21	1
1988	2	350,28	700,56	4
1989	3	374,52	1123,56	9
1990	4	351,47	1405,88	16
1991	5	344,96	1724,80	25
1992	6	340,95	2045,70	36
	0	4159	1793,40	182

Ajuste de una recta: **$Y = a + b \cdot t$**

$$\sum_{i=1}^{N} Y_i = N.a + b\sum_{i=1}^{N} t_i$$

$$\sum_{i=1}^{N} t_i Y_i = a\sum_{i=1}^{N} t_i + b\sum_{i=1}^{N} t_i^2$$

$Y = 319'92 + 9'85\ t$

Con el fin de simplificar los cálculos, se considera como "año 0" el año 1986, ya que es el año central entre los trece. Si hubiera un nº par de años, se escoge uno de los dos años centrales como "año 0".

Variaciones Estacionales

Las **variaciones estacionales** son oscilaciones periódicas de periodo fijo igual o inferior al año, debidas principalmente a las influencias de las estaciones del año, causas climatológicas, costumbres, etc.

En su estudio se presentan dos problemas fundamentales:

- **¿Cómo medir las variaciones estacionales?** Existen muchas formas de medir las variaciones estacionales, aunque todas tienen como objetivo básico la obtención de un índice que pueda utilizarse para ajustar los datos originales a las variaciones estacionales. Dichos índices permiten interpretar el comportamiento de la variable estudiada en los periodos considerados respecto a la media del año, comparando de forma relativa.

- **¿Cómo eliminar la influencia de las variaciones estacionales en el análisis de la tendencia?** Se realiza mediante el proceso de la **desestacionalización**, que consiste en dividir cada valor de la serie original entre el índice de variación estacional correspondiente.

Cálculo de los índices de variación estacional:

Uno de los métodos más utilizados para la obtención de los **índices de variación estacional** es el **método de la medias móviles**. Se trata de obtener una medida generalizada y en términos relativos del comportamiento de la serie en cada uno de los periodos considerados. El método consiste en:

(1) Obtener las **medias móviles**, utilizando tantos valores **R** como periodos considerados dentro del año. Si el número de periodos **R** es par, las medias móviles obtenidas no estarán centradas, por lo que habrá que centrarlas utilizando la semisuma de cada par de las anteriores.

(2) A partir de las **medias móviles centradas**, se obtienen las **razones de las medias móviles**, que relacionan los valores reales de la variable con las **medias móviles centradas**.

$$\text{Razón medias móviles} = \frac{\text{Valor original}}{\text{Media móvil centrada}}$$

(3) Ordenando las razones de las medias móviles por periodos, se obtendrán los **índices generales de variación estacional**, calculando la media asociada a cada periodo.

La media de los **índices generales de variación estacional (IGVE)** en el año deberá ser igual a 100 (o 1 en tantos por uno), por lo que la suma de todos ellos será igual a **R.100** (**R**, en tantos por uno)**,** siendo **R** el número de periodos considerados en el año. Si por razones de redondeo, dicha suma no alcanzara el valor citado, se podrían conseguir los **IGVE** mediante simples reglas de tres.

Si las **RMM** no tienen un comportamiento similar en igual periodo en la mayoría de los años considerados, no tiene sentido obtener los **IGVE**, ya que significaría que su influencia dentro de la tendencia de la serie no es grande. En tal caso, nos conformaríamos con las **RMM**, que actuarían como índices estacionales definitivos.

Ejemplo: Se cuenta con una serie temporal de los precios medios por trimestre de un determinado producto, entre los años 1995 y 1998.

AÑO	PERÍODO	t	Y_t
1995	I	1	19,5
	II	2	15,2
	III	3	17,7
	IV	4	20,6
1996	I	5	19,5
	II	6	15,4
	III	7	18,5
	IV	8	21,5
1997	I	9	20,6
	II	10	16,5
	III	11	19,3
	IV	12	22,7
1998	I	13	21,1
	II	14	17,7
	III	15	21,5
	IV	16	24,4

AÑO	t	Y_t	M.M.	M.M.C.	R.M.M.
1995	1	19,5			
	2	15,2			
	3	17,7	18,250	18,250	96,99
	4	20,6	18,250	18,275	112,72
1996	5	19,5	18,300	18,400	105,98
	6	15,4	18,500	18,613	82,74
	7	18,5	18,725	18,863	98,08
	8	21,5	19,000	19,138	112,34
1997	9	20,6	19,275	19,375	106,32
	10	16,5	19,475	19,625	84,08
	11	19,3	19,775	19,838	97,29
	12	22,7	19,900	20,050	113,22
1998	13	21,1	20,200	20,475	103,05
	14	17,7	20,750	20,963	84,43
	15	21,5	21,175		
	16	24,4			

	I	II	III	IV
1995			96,99	112,72
1996	105,98	82,74	98,08	112,34
1997	106,32	84,08	97,29	113,22
1998	103,05	84,44		
I.G.V.E.	105,12	83,75	97,45	112,75

- Los **IGVE** miden el nivel porcentual del *componente estacional* con respecto al nivel medio o *tendencia*.
- La media de los **IGVE** debe ser igual a 100 (o a 1), indicando que en un periodo anual las fluctuaciones *estacionales* se deben compensar; al no poder existir fluctuaciones *estacionales* superiores al año.
- La magnitud en estudio sufre un incremento de un 5'12 % respecto al valor de la *tendencia*, debido a la *estacionalidad* observada en el 1er trimestre. Se produce, también, una disminución del 16'25 % respecto al valor *tendencia*l, debido a la *estacionalidad* del 2º. Se obtiene una disminución del 2'55 % respecto al valor *tendencial*, causada por la *estacionalidad* del 3º; y, por último, se produjo un incremento del 12'75 % respecto a la *tendencia*, debido a la *estacionalidad* del 4º trimestre.

Desestacionalización:

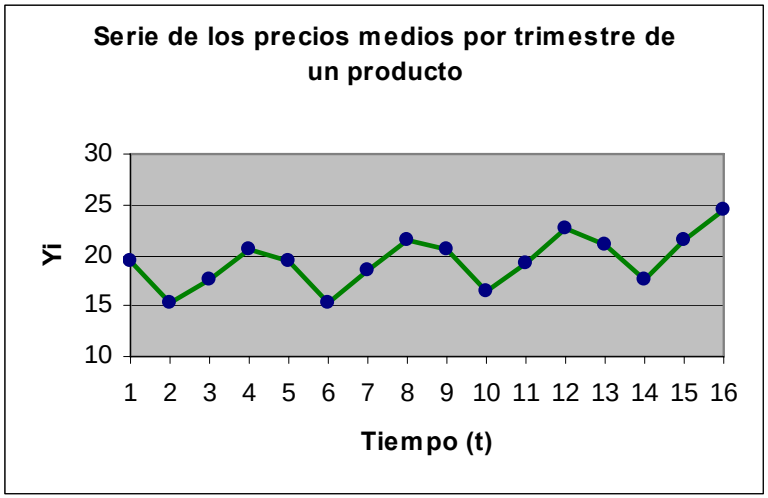
El proceso de **desestacionalización** consiste en suprimir la influencia de las **variaciones estacionales** en una **serie temporal**. Para ello, se divide cada valor de la serie original entre el correspondiente **IGVE** expresado en tantos por uno.

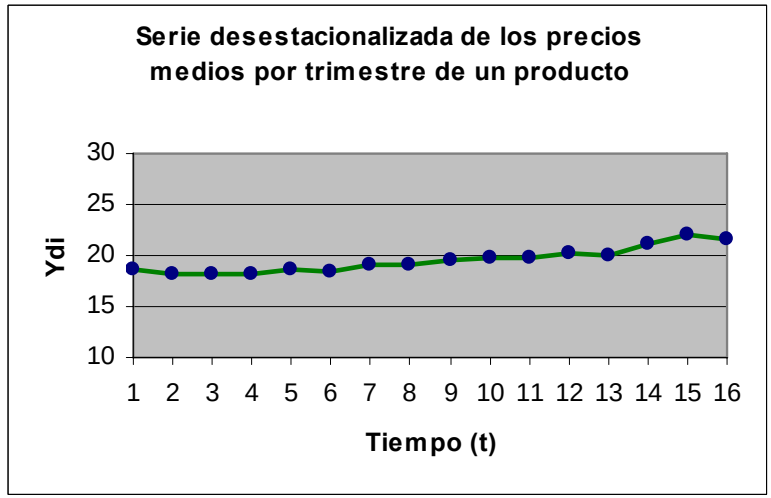
$$Y_{di} = \frac{Y_i}{I.G.V.E.} = \frac{Y_i}{E_i}$$

Una vez **desestacionalizada** la serie temporal, se debe obtener la **tendencia**, ya que en la serie resultante $\mathbf{Y_{d\,i}}$ se ha eliminado la influencia de las variaciones estacionales. Para ello, se planteará un ajuste $\mathbf{Y_d = f(t)}$, obtenido aplicando el método de los mínimos cuadrados.

A continuación, habrá que determinar los valores de la **tendencia**, que serán los valores desestacionalizados teóricos obtenidos a partir del modelo de regresión considerados.

$$T_i = f(t_i)$$





AÑO	t	Y_i	E (I.G.V.E.)	Y_d
1995	1	19,5	1,0512	18,55
	2	15,2	0,8375	18,15
	3	17,7	0,9745	18,16
	4	20,6	1,1275	18,27
1996	5	19,5	1,0512	18,55
	6	15,4	0,8375	18,39
	7	18,5	0,9745	18,98
	8	21,5	1,1275	19,07
1997	9	20,6	1,0512	19,60
	10	16,5	0,8375	19,70
	11	19,3	0,9745	19,80
	12	22,7	1,1275	20,13
1998	13	21,1	1,0512	20,07
	14	17,7	0,8375	21,13
	15	21,5	0,9745	22,06
	16	24,4	1,1275	21,64

t	t'	Y_d	t'^2	$Y_d * t'$
1	-7	18,55	49	-129,85
2	-6	18,15	36	-108,90
3	-5	18,16	25	-90,80
4	-4	18,27	16	-73,08
5	-3	18,55	9	-55,65
6	-2	18,39	4	-36,78
7	-1	18,98	1	-18,98
8	0	19,07	0	0
9	1	19,60	1	19,60
10	2	19,70	4	39,40
11	3	19,80	9	59,40
12	4	20,13	16	80,52
13	5	20,07	25	100,35
14	6	21,13	36	126,78
15	7	22,06	49	154,42
16	8	21,64	64	173,12
	8	312,27	344	239,55

$$Y_d = 19'39 + 0'24\,t' \longrightarrow r^2 = 0'88$$

Los valores de la **tendencia** se obtendrán a partir del ajuste lineal:

$$T_i = 19'39 + 0'24\, t'_i \;,\; i = 1, ..., 16$$

t'	Y_d	T (Y_d teórica)	E (I.G.V.E.)
-7	18,55	17,72	1,0512
-6	18,15	17,96	0,8375
-5	18,16	18,20	0,9745
-4	18,27	18,44	1,1275
-3	18,55	18,68	1,0512
-2	18,39	18,92	0,8375
-1	18,98	19,16	0,9745
0	19,07	19,40	1,1275
1	19,60	19,64	1,0512
2	19,70	19,88	0,8375
3	19,80	20,12	0,9745
4	20,13	20,36	1,1275
5	20,07	20,60	1,0512
6	21,13	20,84	0,8375
7	22,06	21,08	0,9745
8	21,64	21,32	1,1275

Supongamos que se quiere predecir cuál va a ser el comportamiento de los precios del producto en los cuatro trimestres de 1999.

AÑO 1999	t'	T	E	Previsiones de $Y = T \bullet E$
I	9	21,56	1,0512	22,66
II	10	21,80	0,8375	18,26
III	11	22,04	0,9745	21,48
IV	12	22,28	1,1275	25,12

Movimientos Cíclicos

Son movimientos a medio plazo que se reproducen de manera periódica, con un periodo no constante y más amplio que los periodos estacionales.

Puesto que la **componente cíclica** no siempre presenta un carácter tan sistemático como en el caso de las componentes tendencial y estacional, no existen muchos métodos que permitan su obtención. Un método que puede ser válido es el siguiente:

- A partir de un **esquema multiplicativo I**, se despeja **C.I**.

$$Y_i = T_i \,.\, E_i \,.\, C_i \,.\, I_i \Rightarrow C_i \,.\, I_i = \frac{Y_i}{T_i \,.\, E_i} = \frac{Y_{di}}{T_i}$$

- Los índices de los movimientos cíclicos se obtendrán a través de las **medias móviles de orden 3** de los valores $\mathbf{C_i.I_i}$.

Los valores de la **componente cíclica** del ejemplo se muestran a continuación:

Y_i	Y_t	T	$C*I$	C
19,5	18,55	17,72	104,67	
15,2	18,15	17,96	101,06	101,83
17,7	18,16	18,20	99,77	99,97
20,6	18,27	18,44	99,09	99,38
19,5	18,55	18,68	99,29	98,51
15,4	18,39	18,92	97,16	98,51
18,5	18,98	19,16	99,09	98,19
21,5	19,07	19,40	98,31	99,05
20,6	19,60	19,64	99,76	99,06
16,5	19,70	19,88	99,10	99,09
19,3	19,80	20,12	98,42	98,80
22,7	20,13	20,36	98,87	98,25
21,1	20,07	20,60	97,46	99,25
17,7	21,13	20,84	101,43	101,19
21,5	22,06	21,08	104,67	102,53
24,4	21,64	21,32	101,50	

Movimientos Irregulares

Son comportamientos que no muestran carácter periódico ni regular y que se deben a fenómenos catastróficos o fortuitos que afectan de manera casual a la variable.

Para la estimación de la **componente irregular** se dividirán los valores de **C.I** entre los índices obtenidos para la **componente cíclica**:

$$I_i = \frac{C_i \,.\, I_i}{C_i}$$

Cuanto más cerca esté cada valor $\mathbf{I_i}$ a 100 (o a 1, en tantos por uno), menor será el residuo asociado a esa observación.

Los valores obtenidos para la **componente irregular** son los siguientes:

Y_i	$C*I$	C	I
19,5	104,67		
15,2	101,06	101,83	99,24
17,7	99,77	99,97	99,80
20,6	99,09	99,38	99,71
19,5	99,29	98,51	100,79
15,4	97,16	98,51	98,63
18,5	99,09	98,19	100,92
21,5	98,31	99,05	99,25
20,6	99,76	99,06	100,71
16,5	99,10	99,09	100,01
19,3	98,42	98,80	99,62
22,7	98,87	98,25	100,63
21,1	97,46	99,25	98,20
17,7	101,43	101,19	100,24
21,5	104,67	102,53	102,09
24,4	101,50		

COMPONENTES OBTENIDAS PARA LA SERIE TEMPORAL DEL EJEMPLO

AÑOS	TR.	t'	Y_t	M.M.	M.M.C.	R.M.M.	E	Y_d	t'^2	Y_d*t'	T	C*I	C	I
1995	I	-7	19,5				1,0512	18,55	49	-129,85	17,72	104,67		
	II	-6	15,2	18,250			0,8375	18,15	36	-108,90	17,96	101,06	101,83	99,24
	III	-5	17,7	18,250	18,250	96,99	0,9745	18,16	25	-90,80	18,20	99,77	99,97	99,80
	IV	-4	20,6	18,300	18,275	112,72	1,1275	18,27	16	-73,08	18,44	99,09	99,38	99,71
1996	I	-3	19,5	18,500	18,400	105,98	1,0512	18,55	9	-55,65	18,68	99,29	98,51	100,79
	II	-2	15,4	18,725	18,613	82,74	0,8375	18,39	4	-36,78	18,92	97,16	98,51	98,63
	III	-1	18,5	19,000	18,863	98,08	0,9745	18,98	1	-18,98	19,16	99,09	98,19	100,92
	IV	0	21,5	19,275	19,138	112,34	1,1275	19,07	0	0	19,40	98,31	99,05	99,25
1997	I	1	20,6	19,475	19,375	106,32	1,0512	19,60	1	19,60	19,64	99,76	99,06	100,71
	II	2	16,5	19,775	19,625	84,08	0,8375	19,70	4	39,40	19,88	99,10	99,09	100,01
	III	3	19,3	19,900	19,838	97,29	0,9745	19,80	9	59,40	20,12	98,42	98,80	99,62
	IV	4	22,7	20,200	20,050	113,22	1,1275	20,13	16	80,52	20,36	98,87	98,25	100,63
1998	I	5	21,1	20,750	20,475	103,05	1,0512	20,07	25	100,35	20,60	97,46	99,25	98,20
	II	6	17,7	21,175	20,963	84,43	0,8375	21,13	36	126,78	20,84	101,43	101,19	100,24
	III	7	21,5				0,9745	22,06	49	154,42	21,08	104,67	102,53	102,09
	IV	8	24,4				1,1275	21,64	64	173,12	21,32	101,50		

Estadística Empresarial I

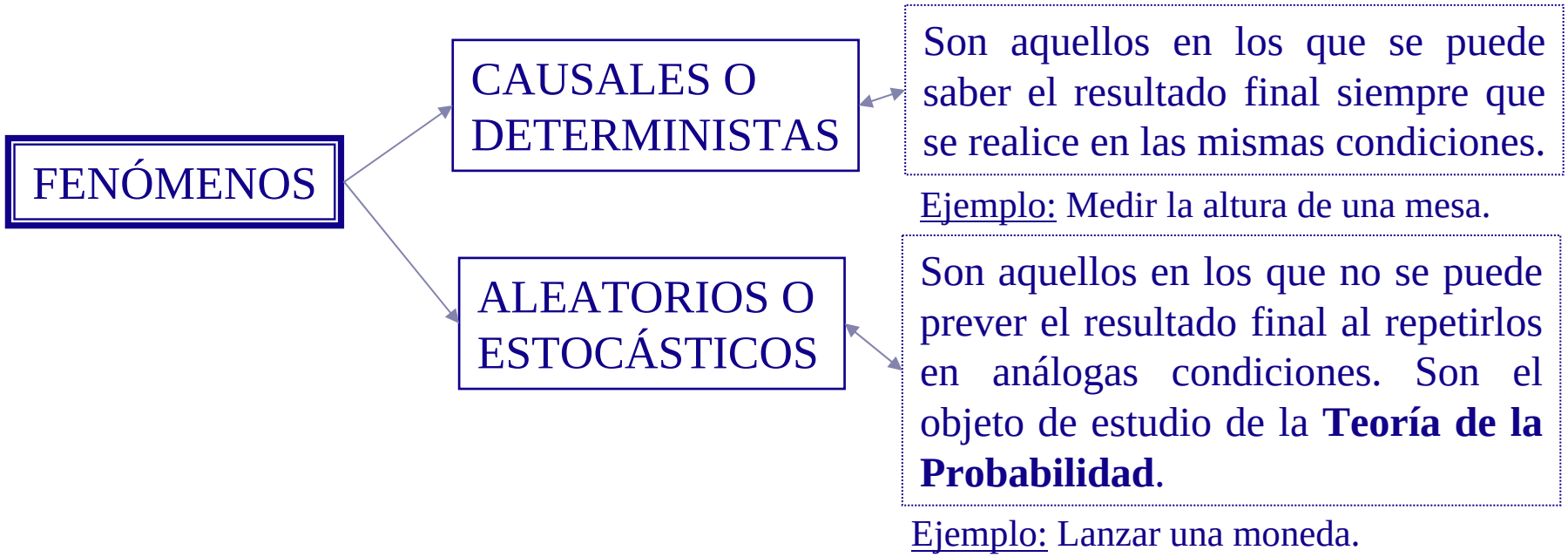
Tema 9

Teoría de la probabilidad

Introducción

Se entiende por **fenómeno** o **experimento** cualquier situación u operación en la que se puede presentar un conjunto de posibles resultados.

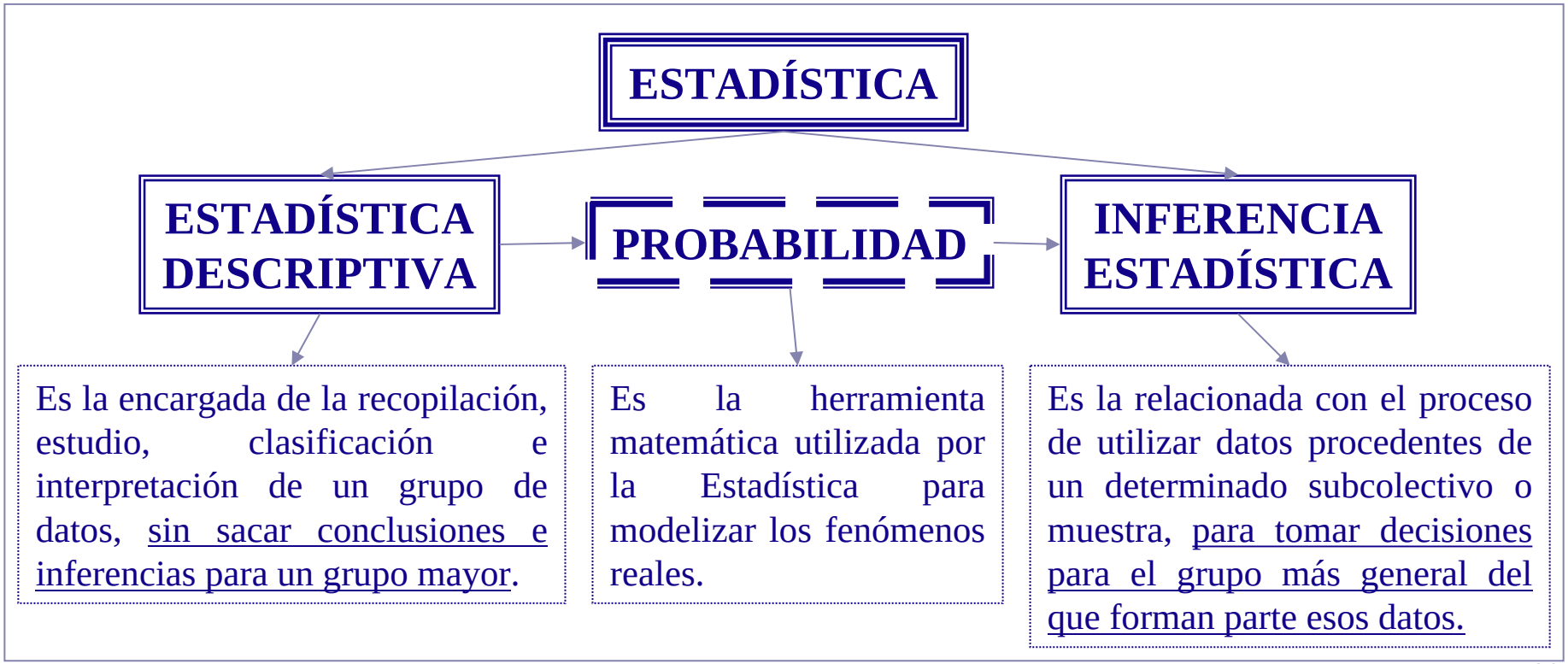
La **Estadística** estudia dos tipos de **fenómenos** o **experimentos**:



En el campo de la economía y de la empresa, los **fenómenos o experimentos aleatorios** son los más comunes, y sus principales características son las siguientes:

- Se conocen previamente los posibles resultados del experimento.
- Es imposible predecir el resultado del experimento antes de realizarlo.
- En sucesivas realizaciones del experimento en las mismas condiciones iniciales, se pueden obtener resultados diferentes.

La **Teoría de la Probabilidad** sirve de enlace entre las dos principales ramas de la Estadística:



Espacio muestral y sucesos

Espacio muestral E: Es el conjunto de todos los posibles resultados de un experimento aleatorio.

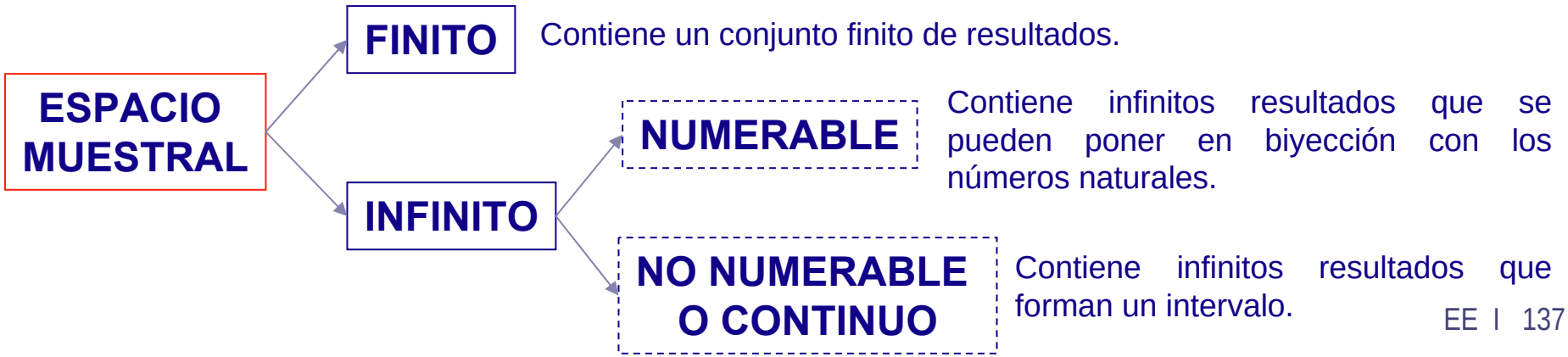
Ejemplos: Determinar el espacio muestral de los siguientes experimentos aleatorios:

(a) Lanzamiento de un dado. *(b) Lanzamiento de dos dados.*

(c) Nº de coches que entran diariamente en un garaje.

(d) Tiempo de vida de una bombilla. *(e) Temperatura diaria de un lugar.*

En función del número de resultados posibles, podemos distinguir varios tipos de espacios muestrales:



Un **suceso** es un subconjunto del **espacio muestral E**, que será **elemental** si sólo contiene un único elemento de E, o será **compuesto** si contiene varios.

<u>Ejemplo:</u> Para el experimento del lanzamiento de un dado, indicar cuáles de los siguientes sucesos son elementales y cuáles compuestos:

(a) "Salir un 2" *(b) "Salir un número par"*

(c) "Salir un número mayor que 3" *(d) "Salir un 5"*

OPERACIONES CON SUCESOS:

Dados dos sucesos A y B asociados a u experimento aleatorio:

- Se llama **unión** de A y B, $A \cup B$, al suceso que ocurre si alguno de los dos ocurre.
- Se llama **intersección** de A y B, $A \cap B$, al suceso que ocurre siempre que A y B ocurran a la vez.
- Se llama suceso **complementario** de A, $\overline{A}$, a aquel suceso que ocurre si no ocurre A.
- Se llama **diferencia** de A y B, A – B, al suceso que ocurre sí y solo sí ocurre A y no ocurre B. Se verifica que: $A - B = A \cap \overline{B}$

TIPOS DE SUCESOS: Existen distintos tipos de sucesos:

- Suceso seguro E: Es aquel suceso que ocurre siempre, coincidiendo con el espacio muestral. Dado un suceso A, siempre se cumple que: $A \cup \overline{A} = E$
- Suceso imposible $\varnothing$: Es aquel suceso que no ocurre nunca. Se cumplirá que: $\overline{\varnothing} = E \qquad \overline{E} = \varnothing$

Dados dos sucesos A y B asociados a un experimento aleatorio.

- Se dicen que son **incompatibles** o **mutuamente excluyentes** si no pueden ocurrir simultáneamente, luego se verificará que: $A \cap B = \varnothing$
- Se dice que A está **contenido** o **incluido** en B si cada vez que ocurre A, también ocurre B, denotándose por $A \subseteq B$.
- Se define el suceso A **condicionado** a B, denotado por A / B, como aquel suceso que consiste en que ocurre A sabiendo que B ha ocurrido.

Ejemplo: Sea el experimento consistente en el lanzamiento de un dado, y sean los sucesos A = "sale un número par", B = "sale un número mayor o igual que 3" y C = "sale un 1 ó un 5". Determinar los siguientes sucesos:

$$\overline{A},\ \overline{C},\ A \cup B,\ B \cup C,\ A \cap B,\ A \cap C,\ B - C,\ C / B$$

PROPIEDADES DE LA UNIÓN Y LA INTERSECCIÓN DE SUCESOS:

(a) Asociativa: $A \cup (B \cup C) = (A \cup B) \cup C \qquad A \cap (B \cap C) = (A \cap B) \cap C$

(b) Conmutativa: $A \cup B = B \cup A \qquad A \cap B = B \cap A$

(c) Elemento neutro: $A \cup \varnothing = A \qquad A \cap E = A$

(d) Distributiva: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \qquad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(e) Leyes de Morgan: $\overline{A \cup B} = \overline{A} \cap \overline{B} \qquad \overline{A \cap B} = \overline{A} \cup \overline{B}$

Ejercicios: Simplificar las siguientes expresiones:

(1) $\overline{A \cup \left(B \cap \overline{A}\right)}$

(2) $\left(A \cup \left(\overline{A \cup B}\right)\right) \cap B$

La probabilidad y sus enfoques

Ya se ha indicado que en cualquier **experimento aleatorio** es imposible predecir el resultado de antemano. Sin embargo, la **Probabilidad** intenta explicar la aparición de los distintos resultados.

El concepto de probabilidad se puede interpretar de varias maneras:

- ***Interpretación objetiva, clásica o de Laplace:*** La **probabilidad** de un suceso se obtiene como el cociente entre los casos favorables al suceso y los casos posibles totales del experimento, suponiendo que todos los sucesos elementales de E son equiprobables.

$$\text{Probabilidad} = \frac{\text{Nº de casos favorables}}{\text{Nº de casos posibles}}$$

Ejemplo: Para el experimento que consiste en extraer una carta de la baraja española, determinar la probabilidad de los siguientes sucesos:

(a) Salir una copa *(b) Salir un rey* *(c) Salir una figura*

- ***Interpretación frecuentalista:*** Se basa en la posibilidad de repetir un experimento bajo las mismas condiciones. Al aumentar el número de pruebas realizadas **n**, la frecuencia relativa **f** de un suceso A tiende a estabilizarse en torno a un valor fijo. Se entiende por **frecuencia relativa** asociada a un suceso el cociente entre el número de veces que ocurre, **m**, y el número de pruebas realizadas, **n**.

$$f(A) = \frac{m}{n} = \frac{n(A)}{n} = P(A)$$

Ejemplo: Sea el experimento que consiste en lanzar un clavo al aire, pudiendo caer de punta o de lado (no equiprobables). Suponiendo que se repite el experimento 1000 veces y que en 12 de ellas cayó el clavo de punta, determinar la probabilidad de que el clavo caiga de punta.

Propiedades de las frecuencias:

(1) $0 \le f(A) = \frac{n(A)}{n} = \frac{m}{n} \le 1$ (2) $f(E) = \frac{n(E)}{n} = \frac{n}{n} = 1$ $f(\varnothing) = \frac{n(\varnothing)}{n} = \frac{0}{n} = 0$

(3) Si A y B son sucesos incompatibles:

$$f(A \cup B) = \frac{n(A \cup B)}{n} = \frac{n(A) + n(B)}{n} = \frac{m + m'}{n} = \frac{m}{n} + \frac{m'}{n} = f(A) + f(B)$$

- ***Interpretación subjetiva o personalista:*** En este caso, la probabilidad se considera como una medida de opinión personal sobre la ocurrencia de un suceso, de manera que dos personas pueden plantear diferentes valores.

Esta interpretación se basa en la experiencia del decisor, sus creencias, su aversión al riesgo, etc.

Ejemplo: ¿Cuál es la probabilidad de que el Tenerife se mantenga en 1ª?

Definición axiomática de probabilidad

Esta definición se basa en un conjunto de axiomas que permitirán construir un modelo matemático de la probabilidad que sea capaz de explicar las regularidades observadas en los sucesos asociados a un experimento aleatorio.

Dado un espacio muestral **E** y una σ-álgebra **Å**, diremos que la siguiente función **P** es una probabilidad si verifica los tres **axiomas de Kolmogorov**.

$$\mathbf{P: Å \longrightarrow [0,1]}$$
$$A \longrightarrow P(A)$$

Una colección de sucesos **Å** es una **σ-álgebra** si verifica:

(1) $\forall$ A $\in$ Å se verifica que $\overline{A} \in$ Å

(2) Dada una sucesión infinita de sucesos de Å: A_1, A_2, ..., se verifica que:

$$\bigcup_{i=1}^{\infty} A_i \in Å$$

Nota 1: Este modelo matemático debe englobar tanto la interpretación clásica como la frecuentalista de la probabilidad.

Nota 2: A la terna (E, Å, P) se le denomina **espacio probabilístico**.

AXIOMAS DE KOLMOGOROV

$$\text{Axioma } 1: \forall A \in \text{Å} : 0 \le P(A) \le 1$$

$$\text{Axioma } 2: P(E) = 1$$

$$\text{Axioma } 3: \text{Sea } A_1, A_2, ..., A_k \in \text{Å} \text{ una sucesión de sucesos incompatibles dos a dos } (A_i \cap A_j = \varnothing, \forall i \neq j)$$

$$\text{Entonces: } P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i)$$

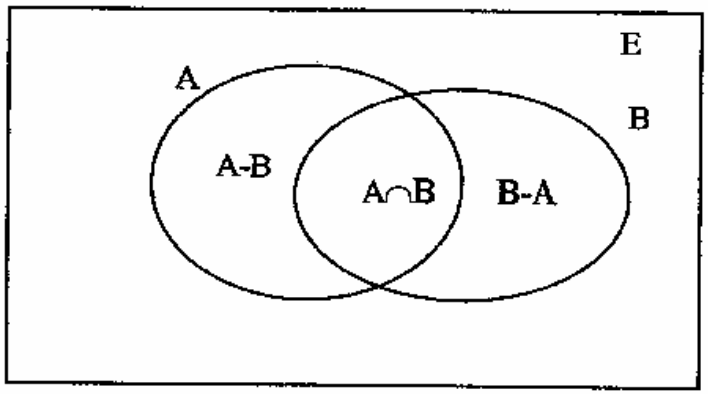
Nota: Estos tres axiomas son equivalentes a las propiedades de las frecuencias relativas.

CONSECUENCIAS DE LOS AXIOMAS:

$$\text{(a) } \forall A \in \text{Å} : P\left(\overline{A}\right) = 1 - P(A) \qquad \text{(b) } P(\varnothing) = 0$$

$$\text{(c.1) } \forall A, B \in \text{Å} : P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

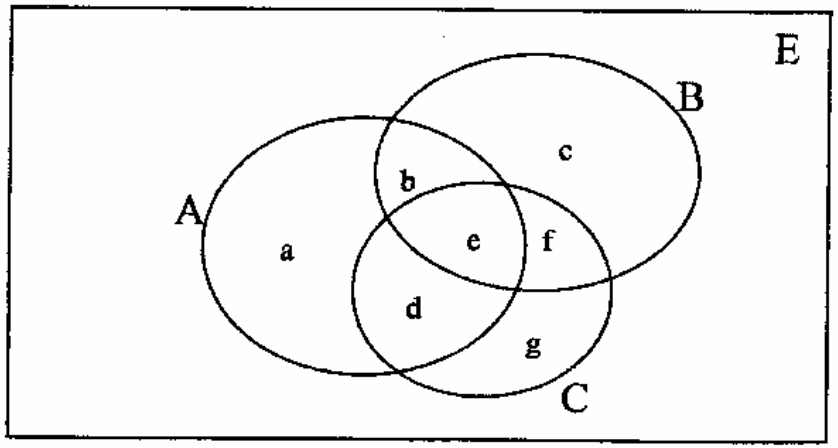
$$\text{(c.2) } \forall A, B, C \in \text{Å} : P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$



$$A \cup B = (A - B) \cup (A \cap B) \cup (B - A)$$

$$A = (A - B) \cup (A \cap B)$$

$$B = (B - A) \cup (A \cap B)$$



$$R = P(A) + P(B) + P(C)$$

$$S = P(A \cap B) + P(A \cap C) + P(B \cap C)$$

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

$$(d)\ \forall\ A, B \in Å : A \subseteq B \Rightarrow P(A) \leq P(B)\ \text{ y }\ P(B - A) = P(B) - P(A)$$

<u>Ejemplo 1:</u> Sean A y B dos sucesos tales que A ∪ B= E, P (A) = 0'8 y P(B) = 0'5. Calcular:

$$(a)\ P(A \cap B) \qquad (b)\ P(A \cup \overline{B}) \qquad (c)\ P(\overline{A} \cup B) \qquad (d)\ P(\overline{A} \cup \overline{B})$$

<u>Ejemplo 2:</u> ¿Es posible una asignación de probabilidad con P (A) = 1/2, P(B) = 1/3, P (A ∩ B) = 2/3?

Probabilidad condicionada

Anteriormente hemos introducido el concepto de probabilidad considerando que la única información disponible sobre el experimento era el espacio muestral E. Sin embargo, hay situaciones en las que se cuenta con información adicional sobre dicho experimento, lo que puede hacer cambiar la probabilidad de ocurrencia de un suceso (aumentándola o disminuyéndola) o bien no modificarla.

Ejemplo 1: Para el experimento del lanzamiento de un dado, consideramos el suceso A = "salir un 2" y B = "salir nº par". Calcular P (A) y P (A/B).

Ejemplo 2: Para el ejemplo anterior, considerando B = "salir nº impar", determinar P (A/B).

Ejemplo 3: Para el experimento consistente en lanzar dos veces un dado, se considera"n los sucesos A = "salir un 2 en el 2º lanzamiento" y A = "salir un 3 en el 1º". Calcular P (A) y P (A/B).

Sea **E** el espacio muestral asociado a un experimento aleatorio y sean A y B $\in$ **Å**, tales que P (B) > 0. Se define la **probabilidad de** A **condicionada al suceso** B como:

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

Esta definición se aceptará si verifica los tres axiomas de **Kolmogorov**, es decir, si verifica que:

$$\text{Axioma } 1: \forall A \in Å : 0 \leq P(A/B) \leq 1$$

$$\text{Axioma } 2: P(E/B) = 1$$

$$\text{Axioma } 3: \text{Sea } A_1, A_2, ..., A_k \in Å \text{ una sucesión de sucesos incompatibles dos a dos } (A_i \cap A_j = \varnothing, \forall i \neq j)$$

$$\text{Entonces: } P\left(\bigcup_{i=1}^{k} A_i / B\right) = \sum_{i=1}^{k} P(A_i / B)$$

- Sea un espacio probabilístico **(E, Å, P)**, y dos sucesos cualesquiera A y B de Å. Se dice que A y B son **estocásticamente independientes** cuando la ocurrencia de B no influye en la de A, y viceversa. En este caso, se verificará que:

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

$$\boxed{P(A \cap B) = P(A).P(B)}$$

- Dados tres sucesos A, B y C, diremos que son **globalmente independientes** si se cumple que:

$$P(A \cap B) = P(A).P(B) \quad P(A \cap C) = P(A).P(C) \quad P(B \cap C) = P(B).P(C)$$

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

- En general, **n** sucesos A_1, A_2, ..., A_n son **globalmente independientes** si se verifica que:

$$P(A_i \cap A_j) = P(A_i).P(A_j), \forall i \neq j$$

$$P(A_i \cap A_j \cap A_k) = P(A_i).P(A_j).P(A_k), \forall i \neq j \neq k$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1).P(A_2)\cdots P(A_n)$$

- Diremos que los sucesos A_1, A_2, ..., A_n son **independientes dos a dos** si cualquier par de dichos sucesos son estocásticamente independientes.

Nota: Si A_1, A_2, ..., A_n son globalmente independientes → son independientes dos a dos.

Ejemplo: Tenemos un experimento consistente en observar la descendencia de una familia seleccionada al azar. Consideremos los sucesos A = "la familia tiene como mucho una hija" y B = "la familia tiene hijos de ambos sexos". Determinar si A y B son independientes en cada una de las siguientes situaciones:

(a) La familia tiene 2 descendientes. (b) *La familia tiene 3 descendientes.*

Teoremas de la Intersección, Probabilidad total y de Bayes

TEOREMA DE LA INTERSECCIÓN:

- Dados dos sucesos A y B, se verifica que:

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

- Dados n sucesos A_1, A_2, ..., A_n, se verificará que:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1).P(A_2/A_1).P(A_3/A_1 \cap A_2)\cdots P(A_n/A_1 \cap A_2 \cap \ldots \cap A_n)$$

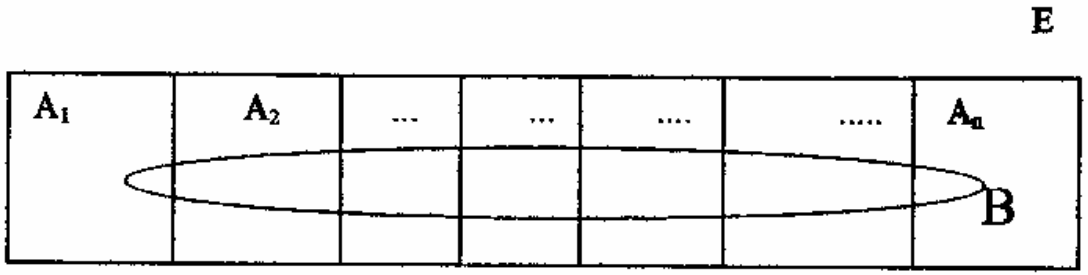
Sistema completo de sucesos: Un conjunto de n sucesos A_1, A_2, ..., A_n se dice que forman un **sistema completo de sucesos** si cumplen las dos condiciones siguientes:

$$(a)\ \bigcup_{i=1}^{n} A_i = E \qquad (b)\ A_i \cap A_j = \varnothing\ ,\ \forall\, i \neq j$$

TEOREMA DE LA PROBABILIDAD TOTAL:

Dado un sistema completo de sucesos A_1, A_2, ..., A_n, y un suceso B, entonces se verifica que:

$$P(B) = \sum_{i=1}^{n} P(A_i).P(B/A_i)$$



$$B = B \cap E = B \cap \left(\bigcup_{i=1}^{n} A_i \right)$$

<u>Ejemplo:</u> Tres máquinas de funcionamiento independiente elaboran toda la producción de una empresa: la primera, la mitad; la segunda, una quinta parte; y la tercera, el resto. Estas máquinas vienen produciendo un 2 %, 4 % y 3 % de unidades defectuosas, respectivamente.

(a) ¿Qué porcentaje de piezas defectuosas produce la empresa?

(b) Calcular la probabilidad de que, elegida una pieza al azar, haya sido producida por la primera máquina o no sea defectuosa.

<u>Probabilidades a priori y a posteriori:</u>

- Los sucesos A_i de un sistema completo de sucesos pueden interpretarse como <u>causas</u> que influyen en un suceso cualquiera B, por lo que las $P(A_i)$ reciben el nombre de **probabilidades a priori**.
- Sin embargo, estas probabilidades $P(A_i)$ pueden verse modificadas por la ocurrencia del suceso B, obteniendo las **probabilidades a posteriori**, $P(A_i / B)$.

TEOREMA DE BAYES:

Sea A_1, A_2, ..., A_n un sistema completo de sucesos y sea B un suceso cualquiera. Entonces:

$$P(A_j / B) = \frac{P(A_j).P(B/A_j)}{\sum_{i=1}^{n} P(A_i).P(B/A_i)}$$

Ejemplo: Tenemos dos urnas, una con 3 bolas blancas y 2 negras, y la otra con 2 bolas blancas y 3 negras. Se selecciona una urna al azar y extraemos una bola.

(a) ¿Cuál es la probabilidad de que la bola sea blanca?

(b) Determinar la probabilidad de que la bola seleccionada proceda de la 2ª urna, sabiendo que fue blanca.

Estadística Empresarial I

Tema 6

Estadística de Atributos

Introducción

Este tema se va a centrar en el estudio de los caracteres de los individuos de la población que no pueden medirse numéricamente, denominados **cualitativos o atributos**.

Atributos: **A, B, C, ...** Modalidades: $\mathbf{a_1, a_2, ...; b_1, b_2, ...; c_1, c_2, ...}$

Ejemplos: Sexo, profesión o nacionalidad.

El estudio de los atributos es de gran interés en campos como el Marketing o el Diseño de Encuestas, ya que en muchas ocasiones no es aconsejable hacer preguntas en las que el encuestado tenga que cuantificar.

Ejemplo:

A = "Tipo de mercancía exportada por cada empresa"

a_i	n_i	f_i
Bienes de consumo	6	0'6
Bienes de capital	3	0'3
Bienes intermedios	1	0'1
	10	

- ¿Tienen sentido las frecuencias acumuladas?
- ¿Y las principales medidas de posición: media, mediana y moda?

Tabla de contingencia

- En el caso bidimensional (A, B), podremos plantearnos el estudio del grado de **asociación** existente entre ambos atributos. Para ello, habrá que disponer los datos en una tabla de doble entrada denominada **tabla de contingencia**.

A \ B	b_1	b_2	...	b_j	...	b_k	$n_{i.}$
a_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	$n_{1.}$
a_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	$n_{2.}$
:	:	:	:	:	:	:	:
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	$n_{i.}$
:	:	:	:	:	:	:	:
a_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	$n_{h.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.k}$	N

Distribuciones marginales

a_i	$n_{i.}$
a_1	$n_{1.}$
a_2	$n_{2.}$
:	:
a_h	$n_{h.}$
	N

b_j	$n_{.j}$
b_1	$n_{.1}$
b_2	$n_{.2}$
:	:
b_k	$n_{.k}$
	N

$$\sum_{i=1}^{h} n_{i.} = \sum_{j=1}^{k} n_{.j} = \sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = N$$

Independencia

De análoga forma al caso de las variables, podemos decir que, dados dos atributos **A** y **B**:

A y **B** son independientes $\Leftrightarrow \frac{n_{ij}}{N} = \frac{n_{i.}}{N}\frac{n_{.j}}{N} \quad \forall i, j \Leftrightarrow n_{ij} = \frac{n_{i.}\, n_{.j}}{N} \quad \forall i, j$

Frecuencia observada	**Frecuencia teórica**
$F.O. = n_{ij}$	$F.T. = n'_{ij} = \frac{n_{i.}\, n_{.j}}{N}$

Así: **A** y **B** son independientes $\Leftrightarrow F.O. = F.T. \quad \forall i, j$

Se verifica, además, que: $\sum_{i=1}^{h}\sum_{j=1}^{k} n'_{ij} = N$

Tablas de contingencia 2x2

A continuación, vamos a tratar de obtener un coeficiente que cuantifique el grado de asociación entre dos atributos, en el caso en que los dos atributos presenten dos modalidades.

A \ B	b_1	b_2	$n_{i.}$
a_1	n_{11}	n_{12}	$n_{1.}$
a_2	n_{21}	n_{22}	$n_{2.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	**N**

Q de Yule: Coeficiente que permite medir la asociación entre dos modalidades de diferentes atributos, a_i y b_j.

$$Q_{ij} = \frac{N.H_{ij}}{n_{11}n_{22} + n_{12}n_{21}} \quad i = 1,2; j = 1,2$$

siendo **H = F.O. – F.T.**

- $Q_{ij} = 0$ ($H_{ij} = 0$) → Independencia entre a_i y b_j.
- $Q_{ij} > 0$ ($H_{ij} > 0$) → Existe atracción o asociación positiva entre a_i y b_j
- $Q_{ij} < 0$ ($H_{ij} < 0$) → Existe repulsión o asociación negativa entre a_i y b_j

Para medir el grado de asociación, se suele utilizar más el coeficiente **Q de Yule** que **H**, debido a que este último no se encuentra acotado y el primero sí.

$$-1 \leq Q \leq 1$$

Repulsión completa entre ambas modalidades

Atracción completa entre ambas modalidades

Además, se va a verificar que:

(1) La atracción entre $\mathbf{a_1}$ y $\mathbf{b_1}$ implica una atracción entre $\mathbf{a_2}$ y $\mathbf{b_2}$ y una repulsión entre $\mathbf{a_1}$ y $\mathbf{b_2}$, y $\mathbf{a_2}$ y $\mathbf{b_1}$.

(2) La repulsión entre $\mathbf{a_1}$ y $\mathbf{b_1}$ implica una repulsión entre $\mathbf{a_2}$ y $\mathbf{b_2}$ y una atracción entre $\mathbf{a_1}$ y $\mathbf{b_2}$, y $\mathbf{a_2}$ y $\mathbf{b_1}$.

Ejercicio: Comprobar que $\mathbf{H_{11} = H_{22} = - H_{12} = - H_{21}}$.

Ejemplo: A continuación se indica la distribución de 50 personas según sexo y su condición de fumador/no fumador. Determinar el grado de asociación entre:

Fuma/Sexo	H	M	$n_{i.}$
Sí	20	12	32
No	6	12	18
$n_{.j}$	26	24	**50**

(a) "mujer" y "no fumador".
(b) "hombre" y "no fumador".
(c) "hombre" y "fumador".
(d) "mujer" y "fumador".

Tablas de contingencia hxk

En este apartado se tratará de obtener algún coeficiente que permita medir el grado de asociación entre dos atributos **A** y **B**, con **h** y **k** modalidades, respectivamente.

Coeficiente de contingencia χ^2 de Pearson

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(n_{ij} - n'_{ij})^2}{n'_{ij}} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n'_{ij}} - N = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{F.O.^2}{F.T.} - N$$

Propiedades: (1) $\chi^2 \geq 0$ (2) χ^2 no está acotado superiormente.

Coeficiente de contingencia C de Pearson

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Propiedades:

(1) $0 \leq C \leq 1$
(2) **C = 0** → Independencia entre los atributos.
(3) **C = 1** → Perfecta asociación entre los atributos (Sólo se logra si los atributos tienen infinitas modalidades).

Coeficiente de contingencia T^2 de Tschuprow

$$T^2 = \frac{\chi^2}{N\sqrt{(h-1)(k-1)}}$$

Propiedades:

(1) **$0 \leq T^2 \leq 1$,** sea cual sea el número de modalidades de cada atributo (**h** y **k**).
(2) **$T^2 = 0$** → Independencia entre los atributos.
(3) **$T^2 = 1$** → Perfecta asociación entre los atributos.

Ejemplo: La siguiente tabla recoge la distribución de las calificaciones del primer parcial de EEI del curso 91/92 para los 398 alumnos matriculados, teniendo en cuenta el grupo al que pertenecen.

Curso / Nota	Susp.	Aprob.	Notab.	Sobre.	Total
1º A	86	21	8	3	118
1º B	73	27	7	0	107
1º C	44	19	2	0	65
1º D	61	34	9	4	108
Total	264	101	26	7	398

Discutir la asociación entre el grupo de cada alumno y la calificación obtenida.

Correlación ordinal

Existe un tipo de atributos que, aunque no se puedan medir numéricamente, son susceptibles de algún tipo de ordenación. Estaremos, pues, ante un **atributo jerarquizado**, que se caracteriza porque entre sus modalidades se puede establecer una ordenación o clasificación, según dos criterios diferentes de ordenación.

La **correlación ordinal** parte de un atributo **A** cuyas modalidades están jerarquizadas, y se centra en el estudio del grado de concordancia existente entre los dos criterios de ordenación (**X** e **Y**) establecidos sobre las modalidades de dicho atributo.

Ejemplo: Órdenes de preferencia de dos jueces X e Y sobre 5 candidatas en un concurso de belleza.

	PEPA	MARY	LOLA	PACA	ROSA
X	1	2	3	4	5
Y	3	1	4	2	5

Correlación por rangos de Spearman

Sea **A** el atributo jerarquizado según los criterios **X** e **Y.**

Coeficiente de correlación ordinal o por rangos de Spearman

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N} \longrightarrow d_i = x_i - y_i$$

Nota: La expresión de ρ se puede deducir a partir de la definición del coeficiente de correlación lineal **r.**

Así pues: $-1 \leq \rho \leq 1$

$\rho = 1 \Leftrightarrow d_i = 0, \forall i = 1, 2, ..., N \Leftrightarrow x_i = y_i, \forall i = 1, 2, ..., N \longrightarrow$ **Concordancia perfecta**

$\rho = -1 \longrightarrow$ **Disconcordancia perfecta** $\longrightarrow$ $x_1 = y_N$, $x_2 = y_{N-1}$, ..., $x_{N-1} = y_2$, $x_N = y_1$

$\rho = 0 \longrightarrow$ **Independencia**

Correlación por rangos de Kendall

Sea **A** el atributo jerarquizado según los criterios **X** e **Y.**

Coeficiente de correlación por rangos de Kendall $\tau = \dfrac{S}{\dfrac{N.(N-1)}{2}}$

Se basa en el concepto de inversión de los rangos respecto al orden natural.

Para determinar el valor de **S**, procedemos como sigue:

(1) Se ordena uno de los criterios de forma creciente, X por ejemplo, dando lugar a X* y obteniendo, a la vez, un determinado orden para el otro criterio, Y*.

(2) Se compara cada rango Y_i^* con cada rango posterior Y_j^*, obteniendo un valor f_{ij} de una función que asigna el valor +1 si $Y_i^* < Y_j^*$ y un -1 si $Y_i^* > Y_j^*$.

(3) Finalmente: $S = \sum_{i<j} f_{ij} \longrightarrow \max S = (N-1)+(N-2)+\ldots+2+1 = \dfrac{N(N-1)}{2}$

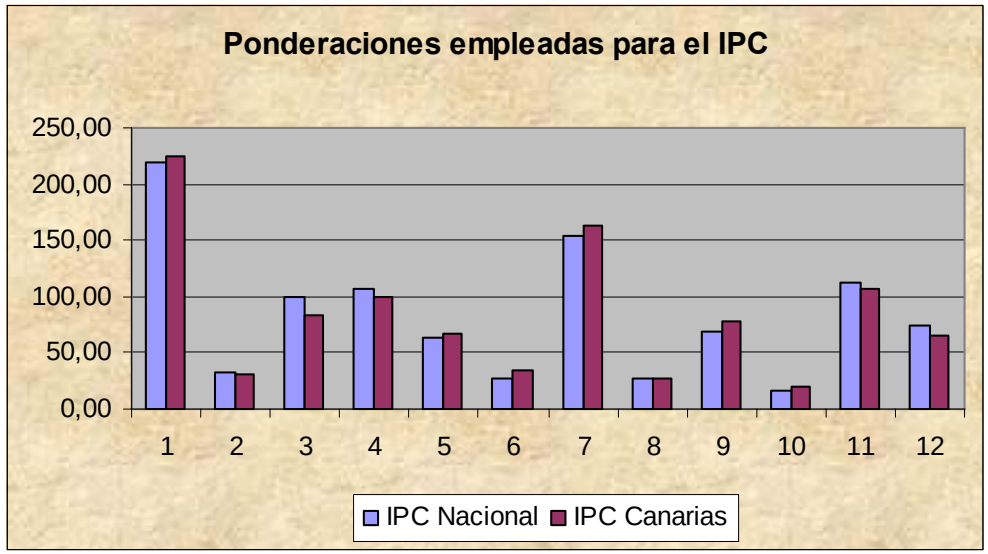
Así pues, **$-1 \leq \tau \leq 1$**. Existe concordancia si $\tau > 0$ y discordancia si $\tau < 0$.

<u>Ejemplo:</u> Para el ejemplo anterior, hallar el coeficiente τ e interpretarlo.

Información real del IPC

Ponderaciones empleadas para determinar el IPC nacional y el IPC de Canarias en el año 2003.

(1) Alimentos y bebidas no alcohólicas	219,31	224,36
(2) Bebidas alcohólicas y tabaco	31,82	29,93
(3) Vestido y calzado	98,99	83,64
(4) Vivienda	106,84	100,45
(5) Menaje	64,10	67,08
(6) Medicina	27,53	34,76
(7) Transporte	153,23	163,89
(8) Comunicaciones	27,35	27,41
(9) Ocio y cultura	68,34	77,62
(10) Enseñanza	16,75	19,18
(11) Hoteles, café y restaurante	111,81	106,17
(12) Otros	73,93	65,53
	1000	**1000**



IPC Nacional con base 2001

01/02	02/02	03/02	04/02	05/02	06/02	07/02	08/02	09/02	10/02	11/02	12/02
101,3	101,4	102,2	103,6	103,9	104	103,2	103,5	103,9	104,9	105,1	105,5

01/03	02/03	03/03	04/03	05/03	06/03	07/03	08/03	09/03	10/03
105	105,2	106	106,8	106,7	106,8	106,1	106,6	106,9	107,7