

Apuntes
de
TÉCNICAS CUANTITATIVAS 1

José Alberto Hermoso Gutiérrez
Dpto. Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada

Apuntes de TÉCNICAS CUANTITATIVAS 1

© 2013,
José Alberto Hermoso Gutiérrez
Edita e imprime: Copicentro S. L.
ISBN:978-84-15814-40-5
Depósito Legal: GR-1690/2013
Impreso en España. Printed in Spain

Quedan rigurosamente prohibidas, bajo las sanciones establecidas en las leyes, la reproducción o almacenamiento total o parcial de la presente publicación, incluyendo el diseño de la portada, así como la transmisión de la misma por cualquiera de sus medios tanto si es eléctrico, como químico, mecánico, óptico, de grabación o bien de fotocopia, sin la autorización escrita de los titulares del copyright

A Celia



Obtén el **Título Oficial de HSK1 (A1)** en un sólo curso.

3 horas semanales (2 días · clases de 1,5 h)
MATRÍCULA GRATIS | 75€/MES

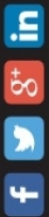
[TAMBIÉN COREANO | Profesores nativos con experiencia]

Ábrete a un mundo de oportunidades laborales, aprendiendo el idioma del futuro.

¡DIFERÉNCIATE!

T 958 998 295
granada@mandarincenters.com

C/ Torre de los Picos, 9
18008 · Granada



www.mandarincenters.com



1.- Variables estadísticas unidimensionales.

1.1 Variables estadísticas. Tablas estadísticas. Representaciones gráficas	7
1.2 Momentos centrados y no centrados	12
1.3 Medidas de posición	16
1.4 Medidas de dispersión	27
1.5 Medidas de forma	33
1.6 Medidas de concentración	35
1.7 Ejercicios resueltos	40

2.- Variables estadísticas bidimensionales.

2.1 Representaciones numéricas en dos columnas y en tablas de contingencia	66
2.2 Distribuciones marginales y condicionadas. Independencia de variables estadísticas .	68
2.3 Covarianza y coeficiente de correlación lineal	71
2.4 Recta de regresión de mínimos cuadrados	77
2.5 Ejercicios resueltos	87

3.- Números índices.

3.1 Tasas de variación	104
3.2 Índice elemental. Índice sintético	111
3.3 Índices de precios, de cantidades y de valor	114
3.4 Enlace de series de números índices con distinta base	117
3.5 Deflación de series económicas	118
3.6 Dependencia de un índice general de un grupo de productos	121
3.7 Ejercicios resueltos	123

4.- Análisis descriptivo de series cronológicas.

4.1 Definición de una serie cronológica. Representación gráfica	135
4.2 Componentes de una serie cronológica. Modelos	136
4.3 Tendencia secular: ajuste de una recta de mínimos cuadrados y medias móviles . . .	138
4.4 Variación estacional. Desestacionalización	143
4.5 Predicción	155
4.6 Ejercicios resueltos	157

5.- Probabilidad.

5.1 Definición de probabilidad. Asignación de probabilidades	172
5.2 Definición de probabilidad condicionada. Sucesos dependientes e independientes . .	179
5.3 Fórmula de la probabilidad total. Fórmula de Bayes	182
5.4 Ejercicios resueltos	184

6.- Variables aleatorias y distribuciones de probabilidad.

6.1 Concepto de variable aleatoria. Distribución de probabilidad	194
6.2 Función de distribución. Variables aleatorias discretas y variables aleatorias continuas	196
6.3 Valor esperado de una variable aleatoria. Momentos	200
6.4 Otras medidas de posición, dispersión y forma	202
6.5 Variables aleatorias bidimensionales. Independencia de variables aleatorias	206
6.6 Ejercicios resueltos	214

7.- Distribuciones discretas de probabilidad.

7.1 Distribución Uniforme discreta	238
7.2 Distribución Binomial	238
7.3 Distribución de Poisson	241
7.4 Distribución Hipergeométrica	243
7.5 Distribución Geométrica	245
7.6 Ejercicios resueltos	247

1. VARIABLES ESTADÍSTICAS UNIDIMENSIONALES.

1.1 Variables estadísticas. Tablas estadísticas. Representaciones gráficas.

El término **Estadística** procede del latín “status” debido a que sus primeras aplicaciones tuvieron que ver con la recogida y cuantificación de información referente al estado: censos de población, ejército, cosechas, impuestos, etc. El término estadística o estadísticas en muchos casos hace referencia a una cantidad de información, por ejemplo estadísticas de precios, estadísticas de producción, etc.

La descripción de esta información es el objetivo de la **Estadística Descriptiva**. Para llevar a cabo esta tarea fundamentalmente nos apoyaremos en las representaciones gráficas de dichos datos y en su síntesis mediante medidas que resuman las características más relevantes.

Podemos distinguir, fundamentalmente, dos **tipos de fenómenos**:

- *Fenómenos causales o determinísticos*: Son aquellos que presentan los mismos resultados si se realizan en idénticas condiciones (por ejemplo las reacciones químicas).
- *Fenómenos aleatorios o estadísticos*: Son los que no se puede predecir el resultado aunque sean conocidas las condiciones de realización (por ejemplo el lanzamiento de una moneda). Fenómenos de naturaleza social, económica,... en los que la incertidumbre de su comportamiento se debe también a la imposibilidad de repetirlos en las mismas condiciones, son tratados como fenómenos estadísticos.

Se denomina **población** al conjunto de elementos sobre los que se quiere realizar un estudio. En la práctica se observa un subconjunto de la población que debe ser representativo y al que llamamos **muestra**.

Entre otras muchas razones para restringirnos al estudio de muestras destacamos:

- *Rapidez*. Por ejemplo, las elecciones.
- *Evitar la destrucción de la población*. Como en el control de calidad.
- *Economía y precisión*. El estudio de una población completa puede llevar a cometer muchos errores, mientras que en una muestra se puede dedicar más atención a la calidad de los datos.

Utilizaremos el término **variable** (variable estadística) para referirnos a la característica en estudio. Se dice que una variable es cuantitativa cuando sus diversas *modalidades* pueden ser medidas numéricamente (precios, edad,...). Denominaremos **atributos** a las variables no numéricas (sexo, profesión,...).

Según los valores que toman las variables estadísticas distinguimos:

- **Variables discretas:** toman valores aislados (número de empleados, habitaciones de una vivienda,...)
- **Variables continuas:** pueden tomar todos los valores de un intervalo (estatura, peso,...)

En la práctica toda variable es discreta debido a la precisión limitada de los aparatos de medida, por ello distinguiremos entre variables con *pocas modalidades* (discretas) y variables que toman un *gran número de valores* distintos (donde se incluyen las variables continuas y otras muchas magnitudes sociales y económicas como salarios, población de ciudades,...)

Tablas estadísticas.

Una vez recogidos los datos se procede a su descripción con la finalidad de obtener el mayor conocimiento acerca del fenómeno.

El primer paso de esta descripción consiste en la ordenación, clasificación, recuento y representación de los datos. Posteriormente se procede a resumirlos en cantidades que miden características del fenómeno.

► EJEMPLO 1.1

Un estudio sobre el tipo de vivienda en construcción en una gran urbe ha aportado los siguientes datos

TIPO	
Colectiva	56
Unifamiliar	14
total	70

Número de dormitorios	
1	7
2	14
3	21
4	21
5	7
total	70

► EJEMPLO 1.2

Para conocer el salario/hora de los trabajadores de un sector se ha observado una muestra de 100. Los datos se han recogido en la siguiente tabla



Salario/hora	
0-10	25
10-20	40
20-40	20
40-50	15
total	100

A estas representaciones numéricas se denominan **tablas estadísticas**.

Estas tablas recogen la siguiente información: las **modalidades**, x_i , (valores que ha presentado el fenómeno) de forma individual o agrupadas en **intervalos** y las **frecuencias absolutas** de cada modalidad o intervalo, n_i , (número de veces que se ha observado esa modalidad o valores del intervalo). El **número total de observaciones**, n , se obtiene sumando todas las frecuencias absolutas (lo que hemos llamado total en los ejemplos 1.1 y 1.2)

$$\sum_{i=1}^k n_i = n$$

A las tablas estadísticas se le puede añadir más información, como:

La **frecuencia relativa** de la modalidad o intervalo i , f_i , es el cociente de la frecuencia absoluta sobre el total de observaciones

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

A menudo las frecuencias relativas se multiplican por 100 para expresar en tanto por ciento la medida en que se presenta cada modalidad o intervalo de valores.

La **frecuencia absoluta acumulada** de la modalidad o intervalo i , N_i , es el número de veces que se han observado valores menores o iguales que dicha modalidad o intervalo.

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

Análogamente se define la **frecuencia relativa acumulada**, F_i , a partir de las frecuencias relativas.

Se denomina **distribución de frecuencias** al conjunto de valores que presenta una variable estadística junto con sus frecuencias. Ésta se representa según el siguiente modelo donde no todas las columnas son necesarias y su posición puede cambiarse

$L_{i-1} - L_i$	x_i	n_i	f_i	N_i	F_i
$L_0 - L_1$	x_1	n_1	f_1	N_1	F_1
$L_1 - L_2$	x_2	n_2	f_2	N_2	F_2
...
$L_{k-1} - L_k$	x_k	n_k	f_k	N_k	F_k
		n	1		

► EJEMPLO 1.3

Usando los datos de los ejemplos 1.1 y 1.2.

x_i	n_i	f_i
Colectiva	56	0,80
Unifamiliar	14	0,20
total	70	1

x_i	n_i	N_i	f_i	F_i
1	7	7	0,10	0,10
2	14	21	0,20	0,30
3	21	42	0,30	0,60
4	21	63	0,30	0,90
5	7	70	0,10	1
total	70		1	

$L_{i-1} - L_i$	x_i	n_i	f_i	N_i	F_i
0-10	5	25	0,25	25	0,25
10-20	15	40	0,40	65	0,65
20-40	30	20	0,20	85	0,85
40-50	45	15	0,15	100	1
total		100	1		

Nota: en variables de tipo continuo cada intervalo de valores o **clase** está representado por su punto medio o **marca de clase**, x_i .

Cada día más, debido al uso de los **ordenadores** (hojas de cálculo, programas de estadística,...), la representación numérica de los datos se reduce a escribir en una columna los valores observados sin ordenarlos, clasificarlos o contarlos (**sin frecuencias**), tal y como aparece en la siguiente imagen.

x_i
x_1
x_2
...
x_n

	A	B	C	D	E	F
1	INGRESOS					
2	13456,23					
3	4567,75					
4	44678,30					
5	24567,90					
6	346,85					
7	23459,55					
8	4678,00					
9	567,80					
10	23567,95					
11	346,85					
12	567,80					
13	4567,75					
14	4678,00					
15	13456,23					

Representaciones gráficas.

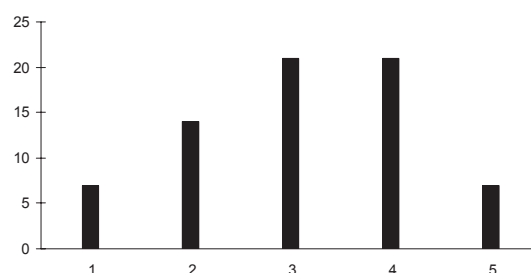
Aunque es cierto que una tabla estadística contiene toda la información observada de un fenómeno, es conveniente en ocasiones traducir toda esa información en un gráfico que nos permita realizar una rápida síntesis visual.

Entre las representaciones gráficas más utilizadas están el diagrama de barras, histograma, y diagrama de sectores.

DIAGRAMA DE BARRAS

La longitud de la barra nos informa de la frecuencia con que se ha observado cada valor de la variable.

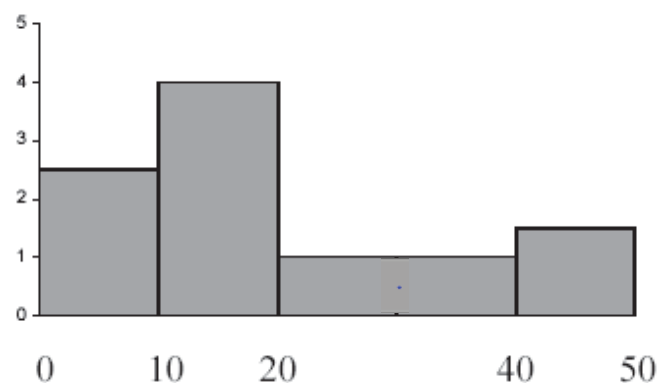
Número de dormitorios	n_i
1	7
2	14
3	21
4	21
5	7



HISTOGRAMA

En esta representación y en la siguiente es el área de las figuras utilizadas (rectángulos o sectores) la que nos indica con qué frecuencia se observa cada clase o modalidad

$L_{i-1} - L_i$	n_i	a_i	h_i
0-10	25	10	2,5
10-20	40	10	4
20-40	20	20	1
40-50	15	10	1,5



La **pirámide de población** consiste en dos histogramas colindantes, uno para la edad de los hombres y otro para la edad de las mujeres. La típica forma triangular de este gráfico le ha dado el nombre de pirámide de población.

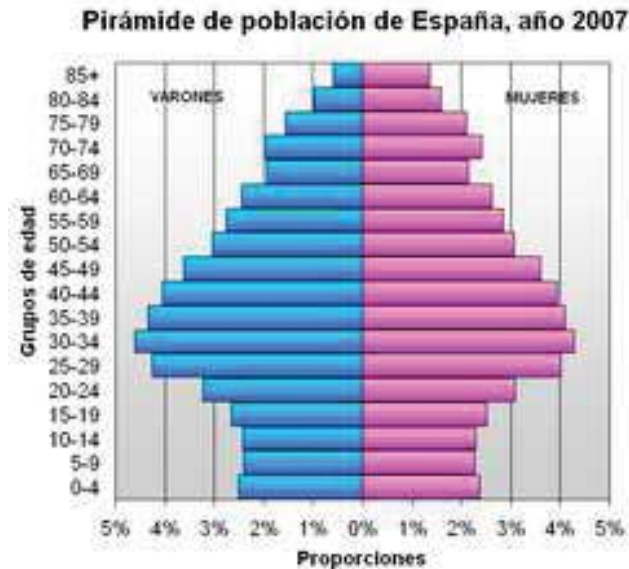
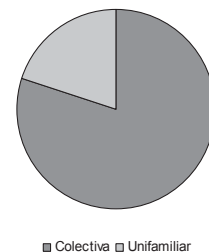


DIAGRAMA DE SECTORES

x_i	n_i	f_i	sector
Colectiva	56	0,80	$0,80 \times 360^\circ = 288^\circ$
Unifamiliar	14	0,20	$0,20 \times 360^\circ = 72^\circ$
total	70	1	360°



1.2 Momentos centrados y no centrados.

Los momentos son unos valores calculados a partir de la distribución de frecuencias que resumen la información relativa a alguna propiedad de la variable.

Como veremos más adelante, la media aritmética y la varianza son casos particulares de momentos que resumen el valor global y la dispersión de los valores que presenta la variable estadística. Otros momentos serán utilizados para medir ciertas características relativas a la forma de la distribución de frecuencias.

En la práctica utilizaremos momentos de órdenes uno a cuatro.

Momentos no centrados.

Se definen los momentos no centrados (o respecto al origen) como:

$$a_r = \frac{1}{n} \sum_{i=1}^k x_i^r n_i = \sum_{i=1}^k x_i^r f_i \quad (\text{para tablas con frecuencias})$$

$$a_r = \frac{1}{n} \sum_{i=1}^n x_i^r \quad (\text{para tablas sin frecuencias})$$

Propiedades:

- El momento no centrado a_1 se conoce también como media (aritmética) y se suele notar como \bar{x} .
- $a_0 = 1$.

Momentos centrados.

Se definen los momentos centrados (o respecto a la media) como:

$$m_r = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^r n_i = \sum_{i=1}^k (x_i - \bar{x})^r f_i \quad (\text{para tablas con frecuencias})$$

$$m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r \quad (\text{para tablas sin frecuencias})$$

Propiedades:

- El momento centrado m_2 se conoce también como varianza y se suele notar como S^2 .
- $m_0 = 1, \quad m_1 = 0$.

$$\begin{aligned} m_2 &= \frac{\sum_{i=1}^k (x_i - \bar{x})^2 n_i}{n} = \frac{\sum_{i=1}^k x_i^2 n_i}{n} + \frac{\sum_{i=1}^k \bar{x}^2 n_i}{n} - \frac{\sum_{i=1}^k 2\bar{x}x_i n_i}{n} = \frac{\sum_{i=1}^k x_i^2 n_i}{n} + \frac{\bar{x}^2 \sum_{i=1}^k n_i}{n} - \frac{2\bar{x} \sum_{i=1}^k x_i n_i}{n} = \\ &= a_2 + \bar{x}^2 - 2\bar{x}a_1 = a_2 + a_1^2 - 2a_1a_1 = a_2 - a_1^2 \end{aligned}$$

- Análogamente, desarrollando el correspondiente binomio elevado a r , cualquier momento centrado puede escribirse en función de los momentos no centrados. Señalamos por su importancia los casos $r=2$ y $r=3$.

$$m_3 = a_3 - 3a_2a_1 + 2a_1^3 \qquad m_4 = a_4 - 4a_3a_1 + 6a_2a_1^2 - 3a_1^4$$

Cálculo de los momentos.

Para facilitar la obtención de los momentos, los cálculos se disponen en una tabla como sigue:

► **EJEMPLO 1.4 (momentos no centrados)**

Tabla sin frecuencias.

x_i	x_i^2	x_i^3	x_i^4
3	9	27	81
3	9	27	81
2	4	8	16
2	4	8	16
3	9	27	81
5	25	125	625
18	60	222	900

$$\bar{x} = a_1 = \frac{1}{n} \sum_{i=1}^n x_i = \frac{18}{6} = 3 \quad a_2 = \frac{1}{n} \sum_{i=1}^n x_i^2 = \frac{60}{6} = 10$$

$$a_3 = \frac{1}{n} \sum_{i=1}^n x_i^3 = \frac{222}{6} = 37 \quad a_4 = \frac{1}{n} \sum_{i=1}^n x_i^4 = \frac{900}{6} = 150$$

Tabla con frecuencias. Variable discreta.

x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$x_i^3 n_i$	$x_i^4 n_i$
1	7	7	7	7	7
2	14	28	56	112	224
3	21	63	189	567	1701
4	21	84	336	1344	5376
5	7	35	175	875	4375
total	$n=70$	217	763	2905	11683

$$\bar{x} = a_1 = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{217}{70} = 3'1 \quad a_2 = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i = \frac{763}{70} = 10'9$$

$$a_3 = \frac{1}{n} \sum_{i=1}^k x_i^3 n_i = \frac{2905}{70} = 41'5 \quad a_4 = \frac{1}{n} \sum_{i=1}^k x_i^4 n_i = \frac{11683}{70} = 166'9$$

Tabla con frecuencias. Variable continua.

$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$x_i^3 n_i$	$x_i^4 n_i$
0-10	5	25	125	625	3125	15625
10-20	15	40	600	9000	135000	2025000
20-40	30	20	600	18000	540000	16200000
40-50	45	15	675	30375	1366875	61509375
total		$n=100$	2000	58000	2045000	79750000

$$\bar{x} = a_1 = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{2000}{100} = 20 \quad a_2 = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i = \frac{58000}{100} = 580$$

$$a_3 = \frac{1}{n} \sum_{i=1}^k x_i^3 n_i = \frac{2045000}{100} = 20450 \quad a_4 = \frac{1}{n} \sum_{i=1}^k x_i^4 n_i = \frac{79750000}{100} = 797500$$

► **EJEMPLO 1.5 (momentos centrados)**

Tabla sin frecuencias.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
3	0	0	0	0
3	0	0	0	0
2	-1	1	-1	1
2	-1	1	-1	1
3	0	0	0	0
5	2	4	8	16
total	0	6	6	18

$$m_1 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{0}{6} = 0 \quad m_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{6}{6} = 1$$

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 = \frac{6}{6} = 1 \quad m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4 = \frac{18}{6} = 3$$

También podríamos haberlos calculado en función de los momentos no centrados que hemos obtenido en el ejemplo 1.4. El momento centrado de orden dos (varianza) se suele calcular así

$$S^2 = m_2 = a_2 - a_1^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = 10 - 3^2 = 1$$

Dado que los cálculos para tablas con frecuencias son análogos para variables discretas y continuas, incluimos un solo ejemplo de momentos centrados con frecuencias.

Tabla con frecuencias. Variable continua.

$L_{i-1} - L_i$	x_i	n_i	$(x_i - \bar{x})$	$(x_i - \bar{x}) n_i$	$(x_i - \bar{x})^2 n_i$	$(x_i - \bar{x})^3 n_i$	$(x_i - \bar{x})^4 n_i$
0-10	5	25	-15	-375	5625	-84375	1265625
10-20	15	40	-5	-200	1000	-5000	25000
20-40	30	20	10	200	2000	20000	200000
40-50	45	15	25	375	9375	234375	5859375
total		$n=100$		0	18000	165000	7350000

$$m_1 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}) n_i = \frac{0}{100} = 0 \quad m_2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_i = \frac{18000}{100} = 180$$

$$m_3 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^3 n_i = \frac{165000}{100} = 1650 \quad m_4 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^4 n_i = \frac{7350000}{100} = 73500$$

También podríamos haberlos calculado en función de los momentos no centrados que hemos obtenido en el ejemplo 1.4. El momento centrado de orden dos (varianza) se suele calcular así

$$S^2 = m_2 = a_2 - a_1^2 = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = 580 - 20^2 = 580 - 400 = 180$$

Cálculos como los anteriores, dispuestos en forma de **tabla**, son **muy fáciles** de hacer con la ayuda de una **hoja de cálculo**.

1.3 Medidas de posición.

Las distintas medias, la moda y la mediana tratan de representar mediante un solo valor a un conjunto de datos y suelen tomar una posición central respecto de los mismos, motivo por el que son conocidas como **medidas de posición central**.

Media aritmética.

Es el promedio más familiar y utilizado en los más diversos ámbitos, aunque no es el único ni el más adecuado en todas las ocasiones. Se define como

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \sum_{i=1}^k x_i f_i \quad (\text{para tablas con frecuencias})$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (\text{para tablas sin frecuencias})$$

Coincide con el momento no centrado $a_1 = \bar{x}$.

Las **calculadoras científicas** nos permiten obtener fácilmente algunos valores estadísticos con la opción **SD**, entre ellos la media aritmética \bar{x} .

► EJEMPLO 1.6.

x_i	n_i	$x_i n_i$
1	7	7
2	14	28
3	21	63
4	21	84
5	7	35
total	70	217

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{217}{70} = 3,1$$

$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$
0-10	5	25	125
10-20	15	40	600
20-40	30	20	600
40-50	45	15	675
total		100	2000

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{2000}{100} = 20$$

Propiedades de la media aritmética:

- Consideramos n observaciones agrupadas en s conjuntos de datos con n_1, n_2, \dots, n_s observaciones cada uno y con medias $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_s$ respectivamente, entonces la media \bar{x} de las n observaciones es:

$$\bar{x} = \frac{\bar{x}_1 n_1 + \dots + \bar{x}_s n_s}{n_1 + \dots + n_s} = \frac{1}{n} \sum_{i=1}^s \bar{x}_i n_i$$

- Con frecuencia se dividen o multiplican los valores de la variable por una constante, ex_i (**cambio de escala**), por ejemplo cuando decidimos expresar los valores en millones en lugar de en euros (en \$ en lugar de €,...). En otras ocasiones se suma o resta una constante a los valores de la variable, $x_i + c$ (**cambio de origen**). Si realizamos una o ambas transformaciones sobre la variable original obtenemos una nueva variable, $y_i = ex_i + c$, cuya media está relacionada con la media de la variable de partida según:

$$\bar{y} = e\bar{x} + c$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^k y_i n_i = \frac{1}{n} \sum_{i=1}^k (ex_i + c) n_i = \frac{1}{n} \sum_{i=1}^k ex_i n_i + \frac{1}{n} \sum_{i=1}^k cn_i = e \frac{1}{n} \sum_{i=1}^k x_i n_i + c \frac{1}{n} \sum_{i=1}^k n_i = e\bar{x} + c$$

Media geométrica.

$$G = \sqrt[n]{\prod_{i=1}^k x_i^{n_i}} = \sqrt[n]{x_1^{n_1} \dots x_k^{n_k}} \quad (\text{para tablas con frecuencias})$$

$$G = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \dots x_n} \quad (\text{para tablas sin frecuencias})$$

La media geométrica se utiliza para promediar porcentajes, tasas, índices de precios,... es decir, en aquellos casos en los que la variable representa **variaciones acumulativas**. La media geométrica es menor que la media aritmética calculada sobre los mismos datos.

► EJEMPLO 1.7.

El valor de la vivienda ha sufrido en los últimos 5 años los siguientes incrementos

incremento
6%
5%
17%
20%
14%

Obtenga el incremento anual medio del valor de la vivienda en estos cinco años.

Solución:

Una vivienda cuyo valor fuese V_0 al comienzo de estos cinco años alcanzaría un valor

$$V_1 = V_0 + \frac{6}{100}V_0 = 1,06V_0$$

al final del primer año, que se transformaría en

$$V_2 = V_1 + \frac{5}{100}V_1 = 1,05V_1 = 1,05 \times 1,06 \times V_0$$

después de transcurridos dos años.

Así sucesivamente, al final de los cinco años

$$V_5 = 1,14 \times 1,20 \times 1,17 \times 1,05 \times 1,06 \times V_0$$

$$V_5 = 1,7814V_0 \quad \left(\frac{V_5}{V_0} = 1,7814 \right)$$

Luego ha habido un incremento del precio de la vivienda en los últimos cinco años del 78,14% (0,7814 por uno).

El incremento medio anual será aquel valor r (en tanto por uno) tal que si se hubiera observado durante todo el periodo (últimos 5 años) ese incremento constante, el resultado final habría sido el mismo

$$1,14 \times 1,20 \times 1,17 \times 1,05 \times 1,06 \times V_0 = (1+r) \times (1+r) \times (1+r) \times (1+r) \times (1+r) \times V_0$$

por tanto

$$(1+r)^5 = 1,14 \times 1,20 \times 1,17 \times 1,05 \times 1,06 = 1,7814$$

$$1+r = \sqrt[5]{1,14 \times 1,20 \times 1,17 \times 1,05 \times 1,06} = \sqrt[5]{1,7814} = 1,1224$$

$$r = 0,1224 \Rightarrow \text{en tanto por ciento } r\% = 12,24\%$$

$1+r$ es la media geométrica de los valores $1+r_i$, donde r_i es el incremento en cada año expresado en tanto por uno.

Teniendo en cuenta que $1,14 \times 1,20 \times 1,17 \times 1,05 \times 1,06 = 1,7814 = \frac{V_5}{V_0}$, otra forma de expresar r es

$$1+r = \sqrt[5]{\frac{V_5}{V_0}}$$

O en general para n años

$$1+r = \sqrt[n]{\frac{V_n}{V_0}}$$

Utilizaremos esto último cuando se desconozcan los incrementos anuales (6%, 5%,...) pero se conozcan los valores inicial (V_0) y final (V_n) del bien que se esté analizando (como en el ejemplo 1.8)

La **media aritmética no es adecuada en este contexto** como puede verse

$$\bar{r} = \frac{0,06 + 0,05 + 0,17 + 0,20 + 0,14}{5} = 0,124$$

(en tanto por ciento) $\bar{r} = \frac{6\% + 5\% + 17\% + 20\% + 14\%}{5} = 12,4\% \quad (12,4\% \neq 12,24\%)$

Según la media aritmética habría un incremento del valor de la vivienda en los 5 años de $(1,124)^5 = 1,794 \Rightarrow 79,4\%$, que no se corresponde con la realidad del ejemplo (78,14%).

► EJEMPLO 1.8.

Una vivienda que en el año 2000 se compró por 125.000€ se ha vendido en el año 2007 por 500.000€. Otra vivienda que se compró en 1995 por 100.000€ se vendió en el 2006 por 700.000€. ¿Cuál de las viviendas incrementó más su valor?

Solución:

Responderemos apoyándonos en el incremento anual medio observado en el valor de cada vivienda.

$$\sqrt[7]{\frac{V_{2007}}{V_{2000}}} = \sqrt[7]{\frac{500.000}{125.000}} = \sqrt[7]{4} = 1,219 \Rightarrow \text{incremento anual medio del } 21,9\% \text{ (primera vivienda)}$$

$$\sqrt[11]{\frac{V_{2006}}{V_{1995}}} = \sqrt[11]{\frac{700.000}{100.000}} = \sqrt[11]{7} = 1,1935 \Rightarrow \text{incremento anual medio del } 19,35\% \text{ (segunda vivienda)}$$

La primera vivienda experimentó un incremento anual medio del 21,9% en su valor mientras que la segunda experimentó un incremento menor, 19,35%.

Media armónica.

$$H = \frac{n}{\sum_{i=1}^k \frac{n_i}{x_i}} \quad (\text{para tablas con frecuencias})$$

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad (\text{para tablas sin frecuencias})$$

Se utiliza para promediar velocidades, precios por unidad, cambios de divisas,...

La media armónica es menor que la media geométrica y esta última menor que la media aritmética.

► EJEMPLO 1.9.

Si se sube en bicicleta a Sierra Nevada a una velocidad de 10 Km/h y se baja a una velocidad de 60 Km/h, a qué velocidad media se ha hecho el recorrido completo de ida y vuelta. Suponga que la distancia desde la casa del ciclista a Sierra Nevada es de 30 Km.

Solución:

Si calculamos la **media aritmética**, obtenemos $\bar{x} = \frac{10+60}{2} = 35 \text{ Km/h}$, que **no es la solución**.

Entendiendo por velocidad media aquella velocidad constante a la que si se realizara el recorrido se tardaría lo mismo, claramente 35 Km/h no lo cumple.

$\frac{30+30}{35} = 1,7143h$ mientras que hemos tardado $\frac{30}{10} = 3h$ en subir y $\frac{30}{60} = 0,5h$ en bajar, en total 3 horas y media.

La solución sería la media armónica $H = \frac{2}{\sum_{i=1}^2 \frac{1}{x_i}} = \frac{2}{\frac{1}{10} + \frac{1}{60}} = \frac{2}{0,10 + 0,0167} = 17,1429 \text{ km/h}$

(Obsérvese que la distancia, 30 Km., no interviene en el cálculo)

A esa velocidad media tardaríamos $\frac{30+30}{17,1429} = 3,49999h$, es decir 3,5 h (la diferencia que se

observa es debida a los errores de redondeo), lo que realmente se ha tardado en la subida y bajada.

Nota:

Utilizando el valor de las distancias recorridas, la media se hubiera calculado como

$$H = \frac{2}{\sum_{i=1}^2 \frac{1}{x_i}} = \frac{2}{\frac{1}{10} + \frac{1}{60}} = \frac{60}{\frac{30}{10} + \frac{30}{60}}$$

y en general

$$H = \frac{D}{\sum_{i=1}^k \frac{d_i}{v_i}} = \frac{D}{\frac{d_1}{v_1} + \dots + \frac{d_k}{v_k}} = \frac{\text{distancia total}}{\text{tiempo total como suma de los tiempos parciales}}$$



► EJEMPLO 1.10.

Una agencia inmobiliaria ha vendido las siguientes viviendas

Precio	Superficie	precio / m ²
240.000	60	4.000
180.000	90	2.000
420.000	100	4.200

Cual ha sido el precio medio al que ha vendido el metro cuadrado.

Solución:

n_i	$\frac{n_i}{x_i}$	x_i
240.000	60	4.000
180.000	90	2.000
420.000	100	4.200
840.000	250	

$$H = \frac{n}{\sum_{i=1}^k \frac{n_i}{x_i}} = \frac{840.000}{250} = \frac{\text{Precio total}}{\text{Superficie total}} = 3.360 \text{ €/m}^2$$

Todos aquellos datos que representen **precios por unidad** como cambio de divisas (p.e., \$/€, €/€), precios de alimentación, combustibles,... (p.e., €/Kg, €/l), precio de la vivienda por m² (€/m²), etc., **se promedian utilizando la media armónica**. ◀

Hay más tipos de medias, como la **media cuadrática**, ..., que no estudiaremos en esta asignatura.

Moda.

Es el valor que se presenta con más frecuencia. Se nota **Mo**. Puede haber varias modas.

Para **variables discretas** y **atributos** su cálculo es inmediato. En las **variables continuas** la mayor o menor frecuencia de las observaciones en un intervalo depende en parte de su amplitud, por lo que para calcular la moda consideraremos las frecuencias observadas en conjuntos de igual amplitud (amplitud unidad).

Dentro del **intervalo modal** (el de mayor frecuencia por unidad de amplitud, mayor altura en el histograma) hay que seleccionar un punto como moda, para lo cual no hay un único criterio. Señalaremos tres criterios diferentes que notaremos como $Mo(I)$, $Mo(II)$ y $Mo(III)$.

Uno de ellos consiste en tomar el punto medio o marca de clase:

$$Mo(I) = \frac{L_{i-1} + L_i}{2} = x_i$$

Otro criterio sitúa la moda a una distancia de los extremos del intervalo proporcional a las alturas de los intervalos anterior y posterior al intervalo modal:

$$Mo(II) = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i$$

Y el tercero tiene en cuenta un procedimiento gráfico para situar la moda dentro del intervalo modal:

$$Mo(III) = L_{i-1} + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} a_i$$

En las expresiones anteriores L_{i-1} es el extremo inferior del intervalo modal, a_i la amplitud del intervalo modal y h_i la altura del intervalo modal (h_{i-1} la altura del intervalo anterior, h_{i+1} la altura del intervalo posterior).

► EJEMPLO 1.11.

x_i	n_i
1	7
2	14
3	21
4	21
5	7
	$n = 70$

Mo=3

Mo=4

TIPO	n_i
Colectiva	56
Unifamiliar	14
total	70

Mo=Colectiva

$L_{i-1} - L_i$	n_i	a_i	h_i
0-10	25	10	2,5
10-20	30	10	3
20-40	40	20	2
40-50	15	10	1,5
	$n = 100$		

Intervalo modal: 10-20.

Dentro del intervalo modal hay que seleccionar un punto como moda. Hay diversos criterios:

$$Mo(I) = \frac{L_{i-1} + L_i}{2} = \frac{10 + 20}{2} = 15$$

$$Mo(II) = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i = 10 + \frac{2}{2,5 + 2} 10 = 14,44$$

$$Mo(III) = L_{i-1} + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} a_i = 10 + \frac{3 - 2,5}{(3 - 2,5) + (3 - 2)} 10 = 13,33$$

Mediana.

Es aquel valor, **Me**, que divide a la muestra ordenada en dos partes iguales, es decir, hay el mismo número de datos menores que la mediana como mayores que ella.

Si hay un número impar de observaciones, la mediana es el único valor central

$$5, 10, 30, 45, 50 \Rightarrow Me=30$$

Si hay un número par de observaciones, la mediana es el punto medio de los dos valores centrales

$$5, 10, 30, 45 \Rightarrow Me = (10+30)/2 = 40/2 = 20$$

Si tenemos los datos representados en una tabla estadística la mediana se calcula buscando el valor

que deja por debajo de él una frecuencia acumulada igual a $\frac{n}{2}$

► EJEMPLO 1.12.

Para **variables discretas** buscamos en la columna de frecuencias absolutas acumuladas el valor $\frac{n}{2}$,

$$\left(\frac{n}{2} = \frac{70}{2} = 35 \right)$$

x_i	n_i	N_i
1	7	7
2	14	21
3	14	35
4	28	63
5	7	70
$n = 70$		

pudiendo ocurrir que hay un $N_i = \frac{n}{2}$, en cuyo caso la mediana

$$\text{es: } Me = \frac{x_i + x_{i+1}}{2} = \frac{3+4}{2} = 3,5,$$

x_i	n_i	N_i
1	7	7
2	14	21
3	21	42
4	21	63
5	7	70
$n = 70$		

o bien, como en este otro ejemplo, todos los $N_i \neq \frac{n}{2}$. Entonces, se

busca el primer $N_i > \frac{n}{2}$, siendo la modalidad x_i asociada a esa

frecuencia acumulada el valor que se toma como mediana, $Me = 3$.

En **variables continuas** se distinguen las mismas dos posibilidades:

$L_{i-1} - L_i$	n_i	N_i
0-10	20	20
10-20	30	50
20-40	35	85
40-50	15	100
$n = 100$		

$N_i = \frac{n}{2}$. La mediana es el extremo superior del intervalo

donde se alcanza la mitad de las observaciones.

$$Me = L_i = 20$$

$L_{i-1} - L_i$	a_i	n_i	N_i
100-110	10	25	25
110-120	10	40	65
120-140	20	20	85
140-150	10	15	100
$n = 100$			

Todos los $N_i \neq \frac{n}{2}$. La mediana está en el intervalo donde

por primera vez $N_i > \frac{n}{2}$ y se calcula mediante:

$$Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i = 110 + \frac{50 - 25}{40} 10 = 116,25$$

que es el resultado de repartir homogéneamente las 40 observaciones sobre la amplitud 10 del intervalo. Según la tabla, hasta 110 hay una frecuencia acumulada de 25, hasta 120 hay una frecuencia acumulada de 65 (por tanto 40 observaciones entre 110 y 120), luego el valor **Me** hasta el que hay una frecuencia acumulada de $\frac{n}{2} = 50$ estará entre 110 y 120. Si las observaciones se reparten *homogéneamente* en el intervalo, la distancia entre la mediana y el extremo inferior del intervalo será proporcional al número de observaciones entre dichos valores

$$\frac{120 - 110}{Me - 110} = \frac{40}{50 - 25} \Rightarrow \frac{120 - 110}{40} \times (50 - 25) = Me - 110$$

$$\frac{120 - 110}{40} = \text{Amplitud entre cada una de las 40 observaciones del intervalo 110-120.}$$

$$(50 - 25) = \text{Número de observaciones en el intervalo 110-Me.}$$

Despejando el valor **Me** se obtiene la anterior expresión para la mediana

$$\frac{10}{Me - 110} = \frac{40}{50 - 25} \Rightarrow \frac{Me - 110}{10} = \frac{50 - 25}{40} \Rightarrow Me - 110 = \frac{50 - 25}{40} 10 \Rightarrow Me = 110 + \frac{50 - 25}{40} 10$$



Percentiles.

Estas medidas (P_1, \dots, P_{99} o C_1, \dots, C_{99}) dividen a la muestra ordenada en 100 conjuntos con igual número de observaciones, $\frac{n}{100}$, habiendo por tanto $\alpha \frac{n}{100}$ observaciones menores que P_α . La mediana coincide con P_{50} . Salvo este percentil, el resto de percentiles ocupan una posición no central respecto de los datos de la muestra, propiedad por la que reciben la denominación de **medidas de posición no central**.

Otros casos particulares de percentiles son los denominados **cuartiles** ($Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$) que dividen a la muestra en 4 conjuntos con igual número de observaciones, $\frac{n}{4}$, y los deciles ($D_1 = P_{10}$, $D_2 = P_{20}$, ..., $D_9 = P_{90}$) que dividen a la muestra en 10 conjuntos con igual número de observaciones, $\frac{n}{10}$.

El cálculo de los percentiles es similar al de la mediana, cambiando $\frac{n}{2} = 50 \frac{n}{100}$ por $\alpha \frac{n}{100}$ dependiendo del percentil, P_α , que se quiera calcular.

En las distribuciones de variables discretas **no hay consenso sobre la forma de calcular los percentiles**, existiendo en la literatura científica nueve métodos diferentes que conducen a resultados diferentes. Por ello, al calcular cualquier percentil por medio de software o manualmente, es básico saber e indicar el método utilizado. **Excel no utiliza el mismo método que en estos apuntes.**

► EJEMPLO 1.13.

Calcule sobre las siguientes tablas los percentiles 30 y 85

x_i	n_i	N_i	$30 \frac{n}{100} = 21 \Rightarrow P_{30} = \frac{x_i + x_{i+1}}{2} = \frac{2+3}{2} = 2,5$ $85 \frac{n}{100} = 59,5 \Rightarrow P_{85} = 4$
1	7	7	
2	14	21	
3	14	35	
4	28	63	
5	7	70	
$n = 70$			

$L_{i-1} - L_i$	a_i	n_i	N_i	$85 \frac{n}{100} = 85 \Rightarrow P_{85} = 140$
100-110	10	25	25	
110-120	10	40	65	
120-140	20	20	85	
140-150	10	15	100	
		$n = 100$		

$$30 \frac{n}{100} = 30 \Rightarrow N_i \neq \alpha \frac{n}{100} \Rightarrow P_\alpha \text{ está en el intervalo donde por primera vez } N_i > \alpha \frac{n}{100}$$

se calcula con una expresión similar a la de la mediana, sustituyendo $\frac{n}{2}$ por $\alpha \frac{n}{100}$, que es el resultado de repartir homogéneamente las observaciones sobre la amplitud del intervalo

$$P_\alpha = L_{i-1} + \frac{\alpha \frac{n}{100} - N_{i-1}}{n_i} a_i$$

$$P_{30} = L_{i-1} + \frac{30 \frac{n}{100} - N_{i-1}}{n_i} a_i = 110 + \frac{30 - 25}{40} 10 = 111,25 \quad \blacktriangleleft$$

► EJEMPLO 1.14.

Los saldos de las cuentas abiertas por los clientes de una sucursal bancaria se distribuyen de acuerdo a la siguiente tabla

SALDOS	NUMERO DE CLIENTES
0-200	100
200-1.000	400
1.000-5.000	300
5.000-30.000	50

Se consideran clientes preferentes al 10% de los clientes con mayores saldos, ¿cuál ha de ser el saldo para que un cliente sea considerado como tal?

¿Qué porcentaje de clientes tienen un saldo superior a 900€?

Solución:

SALDOS	NUMERO DE CLIENTES	N_i
0-200	100	100
200-1.000	400	500
1.000-5.000	300	800
5.000-30.000	50	850

$$n = 850 \Rightarrow 90 \frac{850}{100} = 765$$

$$P_{90} = L_{i-1} + \frac{90 \frac{n}{100} - N_{i-1}}{n_i} a_i = 1000 + \frac{765 - 500}{300} 4000 = 4533,33€$$

Se considerarán clientes preferentes a los que tienen un saldo superior a 4533,33€.

$$P_{\alpha} = L_{i-1} + \frac{\alpha \frac{n}{100} - N_{i-1}}{n_i} a_i = 900 = 200 + \frac{\alpha \frac{850}{100} - 100}{400} 800 \Rightarrow \alpha = 52,94$$

El 52,94% de los clientes tienen un saldo inferior a 900€, por tanto un 47,06% = (100-52.94)% de clientes tienen un saldo superior a 900€.

Considerando el reparto homogéneo de las observaciones en cada intervalo, las anteriores cuestiones se podrían haber resuelto también de la siguiente manera

$$\frac{5000 - 1000}{P_{90} - 1000} = \frac{800 - 500}{765 - 500} \Rightarrow \frac{4000}{P_{90} - 1000} = \frac{300}{265} \Rightarrow P_{90} - 1000 = \frac{4000 \times 265}{300} = 3533,33 \Rightarrow P_{90} = 4533,33€$$

$$\frac{1000 - 200}{900 - 200} = \frac{500 - 100}{x - 100} \Rightarrow \frac{800}{700} = \frac{400}{x - 100} \Rightarrow x - 100 = \frac{400 \times 700}{800} = 350 \Rightarrow x = 450 . \text{ Hay 450 clientes}$$

con un saldo inferior a 900€ que representan un $\frac{450}{850} 100 = 52,94\%$.



1.4 Medidas de dispersión.

Las **medidas de dispersión** cuantifican la variabilidad o esparcimiento de los datos. Cuando esta dispersión se mide respecto de alguna medida de posición central (por ejemplo, la media) nos indica la mayor o menor **representatividad** de dicha medida.

Recorridos

El **recorrido o rango R** se define como la diferencia entre los valores extremos.

$$R = \text{máximo} - \text{mínimo}$$

Es la medida de dispersión más fácil de calcular pero tiene el inconveniente de que sólo utiliza dos valores (estando sujeta a posibles datos erróneos) por lo que no nos da una medida precisa de la dispersión de todos los datos.

El recorrido intercuartílico R_I se define como la diferencia entre el tercer y primer cuartil,

$$R_I = Q_3 - Q_1 = P_{75} - P_{25}$$

representa la amplitud del intervalo donde se encuentra el 50% de las observaciones centrales de la muestra. Con esta medida se evita la fuerte influencia que tienen los valores extremos en el recorrido R .

Con la misma idea se pueden definir distintos recorridos utilizando otros percentiles.

Varianza

$$S^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_i \quad (\text{para tablas con frecuencias})$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{para tablas sin frecuencias})$$

La varianza coincide con el momento centrado de orden 2, $S^2 = m_2$.

Mide la dispersión o distancia de los datos, x_i , respecto de la media aritmética, \bar{x} . Esta medida está expresada en las unidades de los datos al cuadrado (p.e., €, hab.²,...) por lo que no tiene una interpretación fácil. Con el objeto de tener una medida de dispersión expresada en las mismas unidades que los datos en estudio, se define la desviación típica como la raíz cuadrada positiva de la varianza.

Con frecuencia se dividen o multiplican los valores de la variable por una constante, ex_i (**cambio de escala**), por ejemplo cuando decidimos expresar los valores en millones en lugar de en euros (en \$ en lugar de en €,...). En otras ocasiones se suma o resta una constante a los valores de la variable, $x_i + c$ (**cambio de origen**). Si realizamos una o ambas transformaciones sobre la variable original obtenemos una nueva variable, $y_i = ex_i + c$, cuya varianza está relacionada con la varianza de la variable de partida mediante:

$$S_y^2 = e^2 S_x^2$$

(los cambios de origen en los valores de la variable no afectan al valor de la varianza, pero sí los cambios de escala).

$$\begin{aligned} S_y^2 &= \frac{1}{n} \sum_{i=1}^k (y_i - \bar{y})^2 n_i = \frac{1}{n} \sum_{i=1}^k (ex_i + c - e\bar{x} - c)^2 n_i = \frac{1}{n} \sum_{i=1}^k (ex_i - e\bar{x})^2 n_i = \frac{1}{n} \sum_{i=1}^k (e(x_i - \bar{x}))^2 n_i = \\ &= \frac{1}{n} \sum_{i=1}^k e^2 (x_i - \bar{x})^2 n_i = e^2 \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_i = e^2 S_x^2 \end{aligned}$$

En general, para cualquier momento centrado se tiene que

$$m_r(y) = e^r m_r(x)$$

Desviación típica.

La **medida de dispersión absoluta más utilizada** es la desviación típica, S .

$$S = \sqrt{S^2}$$

Las calculadoras científicas nos permiten obtener fácilmente algunos valores estadísticos con la opción **SD**, entre ellos la desviación típica $S = \sigma_n$ y elevándola al cuadrado obtenemos la varianza S^2 .

Los **cambios de origen** en los valores de la variable no afectan al valor de la desviación típica, pero sí los **cambios de escala** según la expresión $S_y = e S_x$.

► EJEMPLO 1.15.

Calcule el rango, la varianza y la desviación típica para la siguiente distribución de frecuencias.

$L_{i-1} - L_i$	n_i
0-10	1
10-20	2
20-30	3
30-40	4
	10

Solución:

$L_{i-1} - L_i$	n_i	x_i	$x_i n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
0-10	1	5	5	400	400
10-20	2	15	30	100	200
20-30	3	25	75	0	0
30-40	4	35	140	100	400
	10		250		1000

$$R=40-0=40$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{250}{10} = 25$$

$$S^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_i = \frac{1000}{10} = 100$$

$$S = \sqrt{S^2} = \sqrt{100} = 10$$

Desigualdad de Tchebycheff

Esta desigualdad nos permite **entender mejor el significado de la desviación típica** como medida de dispersión.

Dado un conjunto de datos con media \bar{x} y desviación típica S , la proporción de datos en el intervalo $(\bar{x} - kS, \bar{x} + kS)$ es mayor o igual que $1 - \frac{1}{k^2}$ ($k > 1$).

$$p[x_i \in (\bar{x} - kS, \bar{x} + kS)] \geq 1 - \frac{1}{k^2}$$

Otra forma equivalente de expresar el mismo resultado, haciendo $t = kS \Leftrightarrow k = \frac{t}{S} \Leftrightarrow k^2 = \frac{t^2}{S^2}$, dice que la proporción de datos en el intervalo $(\bar{x} - t, \bar{x} + t)$ es mayor o igual que $1 - \frac{S^2}{t^2}$.

Por ejemplo, para $k=2$, $(\bar{x} - 2S, \bar{x} + 2S)$, tenemos que alejándonos de la media dos desviaciones típicas, a la derecha e izquierda, abarcaríamos más del 75% de las observaciones, $1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0,75$.

► EJEMPLO 1.16.

Se sabe que el número medio de unidades diarias de un determinado producto que vende un supermercado es $\bar{x} = 100$ y la desviación típica $S = 40$. Si cada día el supermercado repone hasta completar 200 unidades del producto en sus estanterías:

¿Cuántos días al año la demanda será mayor que su oferta?

¿Cuánto habría que reponer para asegurar que no va a faltar producto en las estanterías el 95% de los días?

Solución:

$\bar{x} + t = 200 \Rightarrow t = 100 \Rightarrow$ la proporción de observaciones dentro del intervalo $(\bar{x} - t, \bar{x} + t) = (0, 200)$ es mayor o igual que $1 - \frac{S^2}{t^2} = 1 - \frac{1600}{10000} = 0,84$ y por tanto fuera de dicho intervalo (demanda inferior a 0 o superior a 200) la proporción de observaciones es menor o igual a $0,16 = 1 - 0,84$. Es decir, en menos del 16% de los días la demanda es superior a 200 unidades.

$$1 - \frac{S^2}{t^2} = 1 - \frac{1600}{t^2} = 0,95 \Rightarrow \frac{1600}{0,05} = 32000 = t^2 \Rightarrow t = 178,9 \Rightarrow \bar{x} + t \approx 279$$

Reponiendo hasta completar 279 unidades en las estanterías aseguramos que más del 95% de los días no faltará producto en el supermercado.



Variable tipificada

Se define la variable X tipificada como la nueva variable obtenida al realizar el siguiente cambio

$$Z = \frac{X - \bar{x}}{S}$$

Esta nueva variable se caracteriza por tener media cero y desviación típica 1.

La tipificación de variables **se puede utilizar para establecer comparaciones** entre valores de dos variables (por ejemplo, las calificaciones de dos alumnos en dos centros diferentes).

► EJEMPLO 1.17.

Se quiere comparar los precios de dos viviendas con las mismas características, una en Madrid y otra en Granada. El precio medio de las viviendas del tipo considerado es 200.000€ en Madrid y 140.000€ en Granada, las desviaciones típicas son respectivamente 20.000€ y 15.000€. Las dos viviendas a comparar tienen unos precios de 260.000€ (Madrid) y 190.000€ (Granada). ¿cuál de las dos viviendas está alcanzando un mayor valor en su mercado?

Solución:

$$\frac{260.000 - 200.000}{20.000} = \frac{60.000}{20.000} = 3 \quad \frac{190.000 - 140.000}{15.000} = \frac{50.000}{15.000} = 3,33$$

$3,33 > 3 \Rightarrow$ por tanto la vivienda de Granada está alcanzando mayor valor en su mercado que la de Madrid.

Una vivienda con las características de la de Granada en Madrid tendría un precio de unos 266.600€

$$\frac{x - 200.000}{20.000} = 3,33 \Rightarrow x = (20.000 \times 3,33) + 200.000 = 266.600$$



Los recorridos, la varianza y la desviación típica son medidas de **dispersión absoluta** que dependen de la unidad de medida de la variable (y por tanto les afecta el *cambio de escala*), lo anterior conlleva que no se puedan comparar estas medidas en dos variables con distinta unidad de medida. En la práctica para evitar este problema se prefiere trabajar con otras medidas de dispersión, obtenidas de las anteriores, que se denominan medidas de **dispersión relativa**. Las medidas de dispersión relativa son **adimensionales** (*no dependen de la unidad de medida de la variable estadística y por tanto no les afectan los cambios de escala*)

Coeficiente de Variación.

Es la **medida de dispersión relativa más utilizada**. Se define como el cociente de la desviación típica sobre la media aritmética

$$CV = \frac{S}{x}$$

Esta medida es invariante frente a cambios de escala pero le afecta los cambios de origen.

► EJEMPLO 1.18.

La distribución de los salarios mensuales de 10 trabajadores con igual cualificación profesional es

Salarios en cientos de €	n_i
0-10	1
10-20	2
20-30	3
30-40	4
	10

El horario de trabajo no es único para todos, siendo 6 el número medio de horas trabajadas cada día y 1 hora la desviación típica. ¿Es coherente la distribución de los salarios con la de las horas trabajadas?

Razonamiento:

Si todos los empleados trabajan las mismas horas, lo coherente es que reciban el mismo salario (la variación del salario sería cero y también la del número de horas). Quien trabaje más debe recibir más salario y quien trabaje menos debe recibir menos, es decir, debe haber la misma variabilidad en las horas trabajadas que en los salarios percibidos.

Solución:

La media y desviación típica de los salarios según hemos calculado en el ejemplo 1.15 sobre estos mismos datos son:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{250}{10} = 25 \quad S = \sqrt{S^2} = \sqrt{100} = 10 \quad \Rightarrow \quad CV(\text{salarios}) = \frac{10}{25} = 0,4$$

Mientras que $CV(\text{horarios}) = \frac{1}{6} = 0,167$. La dispersión en los salarios es más del doble que en las horas trabajadas, luego no es coherente la distribución de salarios en relación con las horas de trabajo.

Otra forma de razonar:

$$x = \text{horas trabajadas}, \quad y = \text{salario}$$

Si $y = kx$, donde $k = \text{salario/hora}$, debido a la invariabilidad del coeficiente de variación frente a cambios de escala $\Rightarrow CV_x = CV_y$, pero si $CV_x \neq CV_y \Rightarrow y \neq kx$. Los salarios no son proporcionales a las horas trabajadas. ◀

► EJEMPLO 1.19.

Las subvenciones, en millones de euros, a las pequeñas empresas en 2006 y 2007 según comunidades autónómicas se recogen en la siguiente tabla

	2006	2007
ANDALUCIA	42,95	69,15
ARAGON	5,91	19,25
ASTURIAS	2,23	28,5
BALEARES	1,24	12,5
CANARIAS	3,29	25,8
CANTABRIA	7,20	16,25
CASTILLA LA MANCHA	11,02	19,5
CASTILLA LEON	11,15	26,05
CATALUÑA	25,89	58,2
COMUNIDAD VALENCIANA	17,26	58,35
EXTREMADURA	6,08	41,5
GALICIA	6,75	29
MADRID	8,15	13,55
MURCIA	9,78	25,35
NAVARRA	1,84	15,5
PAIS VASCO	7,88	34,35
LA RIOJA	1,38	17,2

¿Se han mantenido las diferencias entre las subvenciones recibidas en 2006 y 2007? Si han variado, indique en qué sentido.

Solución:

	2006 x_i	$(x_i - \bar{x})^2$	2007 y_i	$(y_i - \bar{y})^2$
ANDALUCIA	42,95	1085,7025	69,15	1532,7225
ARAGON	5,91	16,7281	19,25	115,5625
ASTURIAS	2,23	60,3729	28,5	2,25
BALEARES	1,24	76,7376	12,5	306,25
CANARIAS	3,29	45,0241	25,8	17,64
CANTABRIA	7,20	7,84	16,25	189,0625
CASTILLA LA MANCHA	11,02	1,0404	19,5	110,25
CASTILLA LEON	11,15	1,3225	26,05	15,6025
CATALUÑA	25,89	252,4921	58,2	795,24
COMUNIDAD VALENCIANA	17,26	52,7076	58,35	803,7225
EXTREMADURA	6,08	15,3664	41,5	132,25
GALICIA	6,75	10,5625	29	1
MADRID	8,15	3,4225	13,55	270,6025
MURCIA	9,78	0,0484	25,35	21,6225
NAVARRA	1,84	66,5856	15,5	210,25
PAIS VASCO	7,88	4,4944	34,35	18,9225
LA RIOJA	1,38	74,3044	17,2	163,84
TOTAL	170,00	1774,75	510,00	4706,79

$$n = 17 \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{170}{17} = 10 \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{510}{17} = 30$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1774,75}{17} = 104,40 \quad S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{4706,79}{17} = 276,87$$

$$S_x = \sqrt{S_x^2} = 10,22 \quad S_y = \sqrt{S_y^2} = 16,64$$

$$CV_{2006} = \frac{S_x}{\bar{x}} = \frac{10,22}{10} = 1,022 \quad CV_{2007} = \frac{S_y}{\bar{y}} = \frac{16,64}{30} = 0,555$$

Las diferencias entre las subvenciones recibidas por las distintas comunidades autonómicas han disminuido en 2007 en relación al año anterior. ◀

1.5 Medidas de forma.

Momentos como la media y la varianza nos aportan información sobre la posición y dispersión de los datos. En este apartado, las denominadas **medidas de forma** cuantifican características observables en la forma de la representación gráfica que nos proporcionan más información sobre el fenómeno en estudio.

Hay diversas medidas de forma, las más utilizadas se basan en los momentos.

Coeficiente de asimetría de Fisher.

$$g_1 = \frac{m_3}{S^3}$$

Se basa en que en distribuciones simétricas por cada observación a la derecha de la media hay otra a igual distancia a la izquierda, por tanto la expresión

$$\sum_{i=1}^n (x_i - \bar{x})^3 = 0$$

Si la gráfica es asimétrica a la izquierda se rompe el anterior equilibrio entre sumandos positivos y negativos, resultando la suma negativa. Lo mismo ocurre cuando la distribución de frecuencias es asimétrica a la derecha, resultando la suma positiva. El signo del denominador de g_1 es siempre positivo, por tanto:

- Si la distribución es simétrica $\Rightarrow g_1 = 0$
- Si la distribución es asimétrica a la izquierda $\Rightarrow g_1 < 0$
- Si la distribución es asimétrica a la derecha $\Rightarrow g_1 > 0$

El signo de g_1 lo aporta m_3 , se define dividiendo por S^3 para conseguir que el coeficiente sea una **medida adimensional** que pueda compararse con la asimetría de otras distribuciones, además también se consigue así que sea **independiente de cambios de origen y escala**.

Coeficiente de curtosis (o apuntamiento) de Fisher.

Las medidas de curtosis se utilizan en distribuciones unimodales simétricas o levemente asimétricas para cuantificar la mayor o menor frecuencia de observaciones en torno a la media. La mayor frecuencia de observaciones próximas a la media dará lugar a una representación gráfica más apuntada, la menor frecuencia de observaciones próximas a la media dará lugar a una representación más aplanada. El perfil de apuntamiento que se toma como referencia es el de la conocida *campana de Gauss* o *curva normal*.

La medida de apuntamiento más utilizada es el coeficiente de curtosis de Fisher

$$g_2 = \frac{m_4}{S^4} - 3$$

Al igual que el coeficiente de asimetría de Fisher es **adimensional** e **independiente de cambios de origen y escala**.

Los valores del coeficiente de curtosis de Fisher se interpretan de la siguiente manera:

- Si la distribución tiene un apuntamiento *normal* (mesocúrtica) $\Rightarrow g_2 = 0$

- Si la distribución es más aplanada que la *curva normal* (platicúrtica) $\Rightarrow g_2 < 0$
- Si la distribución es más apuntada que la *curva normal* (leptocúrtica) $\Rightarrow g_2 > 0$

1.6 Medidas de concentración.

Las medidas de concentración miden la mayor o menor igualdad en el reparto de una cantidad (por ejemplo, la masa salarial total de una empresa, ...). Ante este problema eminentemente económico, medidas estadísticas como la media, la varianza, ..., no son significativas, por lo que es necesario construir unos indicadores específicos. Debido a la naturaleza de los fenómenos que aquí se consideran, las variables tomarán sólo valores positivos (por éste y otros motivos, no deben hacerse cambios de origen).

La característica que se va a estudiar puede presentar las siguientes situaciones límite:

- *Máxima concentración*: Cuando un solo individuo recibe la cantidad total a repartir y el resto nada.
- *Equidistribución (mínima concentración)*: Todos los individuos reciben la misma cantidad.

Entre ambas situaciones extremas hay infinidad de situaciones intermedias que trataremos de cuantificar con las siguientes medidas de concentración:

Curva de concentración de Lorenz.

Ilustraremos la construcción de la **curva de Lorenz** con un ejemplo.

► EJEMPLO 1.20.

Estudiar la concentración de los salarios/hora de 25 trabajadores recogidos en la siguiente tabla

$L_{i-1} - L_i$	n_i
500-1500	3
1500-2500	7
2500-3500	8
3500-4500	4
4500-5500	2
5500-6500	1

Hacemos los siguientes cálculos

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
500	1500	1000	3	3000	3	3000	12	4,1
1500	2500	2000	7	14000	10	17000	40	23,3
2500	3500	3000	8	24000	18	41000	72	56,2
3500	4500	4000	4	16000	22	57000	88	78,1
4500	5500	5000	2	10000	24	67000	96	91,8
5500	6500	6000	1	6000	25	73000	100	100,0
total			$n = 25$	$\sum_{j=1}^k x_j n_j = 73000$			408	353,5

Donde n_i son los trabajadores en cada intervalo, $x_i n_i$ es la suma de los salarios de los trabajadores en cada intervalo, acumulando los anteriores valores obtenemos respectivamente N_i y u_i .

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

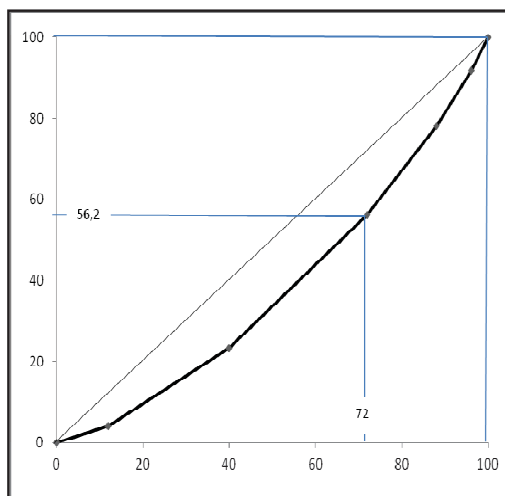
(representan el número de trabajadores y la suma de los salarios de todos los trabajadores hasta el intervalo i)

Por último, p_i y q_i son los valores N_i y u_i expresados en porcentajes.

$$p_i = \frac{N_i}{n} 100 \quad q_i = \frac{u_i}{\sum_{j=1}^k x_j n_j} 100$$

(representan el porcentaje de trabajadores y el porcentaje de los salarios de todos los trabajadores hasta el intervalo i)

La **curva de Lorenz** es la representación gráfica de los puntos con coordenadas (p_i, q_i) , $i = 1, \dots, k$, a los que se añade el punto $(0, 0)$. Como puede verse en el gráfico, la curva de Lorenz siempre parte del punto $(0, 0)$ y termina en el punto $(100, 100)$.



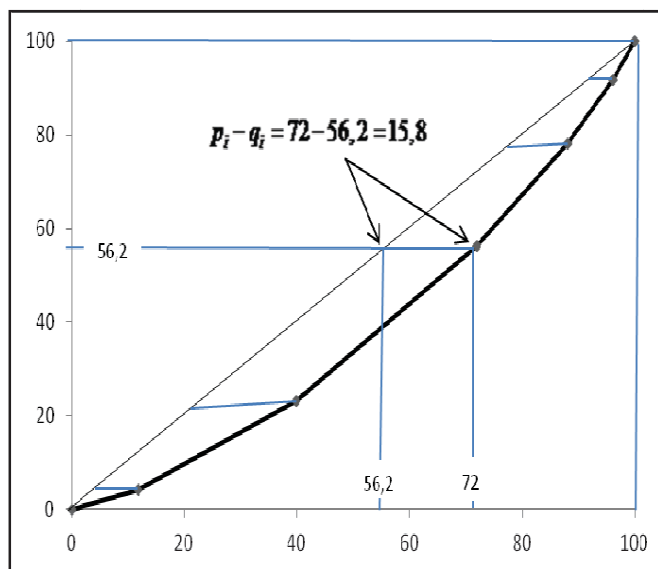
Interpretación de la curva de Lorenz.

Si hay un reparto equitativo (todos reciben lo mismo), a un porcentaje p_i de trabajadores le corresponde un porcentaje q_i de la cantidad total repartida igual a p_i . Es decir los valores p_i son iguales a los valores q_i y la curva de Lorenz coincide con la línea $y = x$ (bisectriz del primer cuadrante). Cuanto más próximos estemos de esta situación de reparto equitativo más próxima estará la curva de Lorenz de la recta $y = x$, cuanto mayor sea la concentración del reparto más se alejará la curva de Lorenz de dicha línea.

Índice de Gini.

El índice de Gini cuantifica la anterior propiedad de la curva de Lorenz basándose en la distancia de los puntos (p_i, q_i) , $i = 1, \dots, k-1$, a la recta $y = x$. [No se tienen en cuenta los puntos $(0,0)$ y $(p_k, q_k) = (100, 100)$].

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



Interpretación del índice de Gini.

- Si hay un reparto equitativo (equidistribución) $\Rightarrow p_i = q_i \quad i = 1, \dots, k-1 \Rightarrow I_G = 0$
- Si hay concentración máxima (todos reciben nada, salvo uno que recibe todo)
 $\Rightarrow q_i = 0 \quad i = 1, \dots, k-1, \quad q_k = 100 \Rightarrow I_G = 1$

Luego el índice de Gini es un valor entre 0 y 1 que mide la concentración, indicando mayor concentración cuanto mayor sea y mayor equidistribución cuanto menor sea.

En el ejemplo anterior el índice de Gini vale $I_G = 1 - \frac{253,5}{308} = 0,1769$.

(Obsérvese que a la suma de los p_i y q_i se ha restado 100 puesto que $p_k = 100$ y $q_k = 100$ no se incluyen en el cálculo del índice de Gini)

Mediala.

La mediala o valor medial, **MI**, es aquel valor tal que **la suma de las observaciones** menores que él es igual a la suma de las observaciones mayores que él.

Se trata pues de una mediana sobre los valores $x_i n_i$ en lugar de sobre las frecuencias n_i . Su cálculo se realiza de forma análoga, buscando el valor de la variable asociado a la mitad de la cantidad total

repartida, $\frac{\sum_{j=1}^k x_j n_j}{2}$, o equivalentemente asociado a $q_i = 50\%$.

Según lo anterior, en variables continuas (valores agrupados en intervalos) la mediala se obtiene de

$$MI = L_{i-1} + \frac{50 - q_{i-1}}{(n_i x_i)\%} a_i = L_{i-1} + \frac{50 - q_{i-1}}{q_i - q_{i-1}} a_i$$

Donde $(n_i x_i)\%$ representa el porcentaje recibido en el reparto por el intervalo i (intervalo donde se encuentra la mediala), L_{i-1} es el extremo inferior del intervalo medial (intervalo donde por primera vez $q_i > 50$) y a_i es la amplitud del intervalo donde está la mediala.

La mediala de los datos del ejemplo anterior es:

$$MI = 2500 + \frac{50 - 23,3}{56,2 - 23,3} 1000 = 3311,55$$

¿De qué forma nos ayuda la mediala a medir la concentración? Si hay equidistribución la mediana y mediala coinciden, separándose más cuanto mayor sea la concentración. Por tanto la respuesta es

$$\Delta M = MI - Me$$

En el ejemplo anterior $Me = 2500 + \frac{12,5 - 10}{8} 1000 = 2812,5$ $\Delta M = MI - Me = 499,05$

El valor ΔM , que depende de la unidad de medida de la variable, se suele relativizar en comparación con el rango, $\Delta M / R$.

En nuestro ejemplo, $R = 6500 - 500 = 6000$ $\frac{\Delta M}{R} = 0,0832$, que confirma de nuevo lo que ya conocíamos por la curva de Lorenz e índice de Gini de que la concentración es débil.

En el anterior ejemplo hemos visto cómo sobre una variable continua se calculan las medidas de concentración, en el siguiente ejemplo lo veremos sobre una variable discreta.

► EJEMPLO 1.21.

Dos familias con 4 y 5 hijos respectivamente deciden repartir parte de sus patrimonios entre ellos de la siguiente forma.

Familia A
300000
500000
200000
150000

Familia B
1500000
1000000
2000000
1000000
2000000

¿Cuál de los dos repartos es más equitativo?

Solución:

De estas tablas pasamos los datos a tablas estadísticas con frecuencias para variables discretas.

Familia A:

x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
150000	1	150000	1	150000	25	13,0
200000	1	200000	2	350000	50	30,4
300000	1	300000	3	650000	75	56,5
500000	1	500000	4	1150000	100	100,0
total	4	1150000			250	199,9

$$I_G(A) = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{99,9}{150} = 0,334$$

Para obtener la mediana, sencillamente buscamos en la columna de los q_i dónde se supera por primera vez el valor 50%, $q_3 = 56,5 > 50 \Rightarrow MI(A) = x_3 = 300000$.

(Si hay un $q_i = 50 \Rightarrow ML = \frac{x_i + x_{i+1}}{2}$, análogamente a como se hace en el cálculo de la mediana)

Familia B:

x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
1000000	2	2000000	2	2000000	40	26,7
1500000	1	1500000	3	3500000	60	46,7
2000000	2	4000000	5	7500000	100	100,0
total	5	7500000			200	173,4

$$I_G(B) = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{73,4}{100} = 0,266$$

Para obtener la mediana, sencillamente buscamos en la columna de los q_i dónde se supera por primera vez el valor 50%, $q_3 = 100 > 50 \Rightarrow ML(B) = x_3 = 2000000$.

Comparando ambos índices de Gini se observa que es más equitativo el reparto de la familia B

$$I_G(A) = 0,334 > I_G(B) = 0,266$$



1.7 Ejercicios resueltos.

- La compañía de telefonía móvil *Noteoigo* está considerando cambiar sus tarifas. Para ello ha observado la duración en segundos de 1000 llamadas realizadas por sus abonados:

Duración de las llamadas	Número de llamadas
0"-20"	15
20"-60"	180
60"-90"	195
90"-180"	405
180"-300"	205

Obtenga:

- Duración media de las llamadas.
- Coste medio de las llamadas según las siguientes tarifas:
 - 10 céntimos el establecimiento de llamada, más 6 céntimos por minuto (proporcionalmente las fracciones de minuto).

b.2) 8 céntimos por minuto (proporcionalmente las fracciones de minuto), sin coste de establecimiento de llamada.

c) Porcentaje de llamadas que superan los 2 minutos de duración.

Solución:

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	N_i	p_i
0	20	10	15	150	15	1,5
20	60	40	180	7200	195	19,5
60	90	75	195	14625	390	39
90	180	135	405	54675	795	79,5
180	300	240	205	49200	1000	100
			1000	125850		

$$a) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{125850}{1000} = 125,85 \text{ segundos}$$

b) Utilizaremos como afecta a la media un cambio de origen y/o escala:

$$Y = eX + c \quad \Rightarrow \quad \bar{y} = e\bar{x} + c$$

X =duración de las llamadas en segundos.

Y =duración de las llamadas en minutos.

Z_1 = coste de la llamada según la opción b.1

Z_2 = coste de la llamada según la opción b.2

$$Y = \frac{1}{60} X \quad \Rightarrow \quad \bar{y} = \frac{1}{60} \bar{x} = \frac{125,85}{60} = 2,0975$$

$$b.1) \quad Z_1 = 10 + 6Y \quad \Rightarrow \quad \bar{z}_1 = 10 + (6 \times \bar{y}) = 10 + (6 \times 2,0975) = 22,585 \text{ céntimos.}$$

$$b.2) \quad Z_2 = 8Y \quad \Rightarrow \quad \bar{z}_2 = 8 \times \bar{y} = 8 \times 2,0975 = 16,78 \text{ céntimos.}$$

c) Calculamos, interpolando, el porcentaje de llamadas que duran menos de 120 segundos, y se lo restamos a 100%.

90	39
120	x
180	79,5

$$\frac{180 - 90}{79,5 - 39} = \frac{120 - 90}{x - 39} \quad \Rightarrow \quad x = 52,5\% \quad \Rightarrow \quad 100 - 52,5 = 47,5\%$$

El 47,5% de las llamadas superan los dos minutos.

2. Se dispone de la siguiente información sobre los salarios anuales brutos de los empleados de una empresa (en miles de euros):

Salarios	0-20	20-60	60-70	70-90
nº empleados	10	45	30	15

a) Obtenga el coeficiente de variación de los salarios.

b) ¿Qué salario es superado por el 60% de los empleados?

c) ¿Qué tanto por ciento de empleados tienen un salario superior a 63000 euros?

Solución:

$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	N_i	p_i
0-20	10	10	100	1000	10	10
20-60	40	45	1800	72000	55	55
60-70	65	30	1950	126750	85	85
70-90	80	15	1200	96000	100	100
		n=100	5050	295750		

$$a) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i n_i = \frac{1}{100} 5050 = 50,5$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 n_i - \bar{x}^2 = \frac{295750}{100} - 50,5^2 = 407,25 \quad S = \sqrt{407,25} = 20,18$$

$$CV = \frac{S}{\bar{x}} = \frac{20,18}{50,5} = 0,3996$$

b)

L_i	p_i
20	10
x	40
60	55

$$\frac{60-20}{55-10} = \frac{x-20}{40-10} \quad \frac{1200}{45} + 20 = x \quad x = 46,6$$

46666 euros no son superados por el 40% de los empleados

46666 euros son superados por el 60% de los empleados

c)

L_i	p_i
60	55
63	x
70	85

$$\frac{70-60}{85-55} = \frac{63-60}{x-55} \quad x-55 = \frac{3 \times 30}{10} \quad x = 64$$

64% de los empleados tienen salario inferior a 63000 euros

100 - x = 36% de los empleados tienen salario superior a 63000 euros

3. Se tienen datos sobre los beneficios en millones de euros (ya deflactados) obtenidos por las empresas de tres sectores productivos en los años 2008 y 2012:

Sector	2008		2012	
	Beneficio medio (millones de euros)	Número de empresas	Beneficio medio (millones de euros)	Número de empresas
A	40	800	35	735
B	65	1200	60	1165
C	50	1000	35	900

Calcule para los tres sectores en conjunto:

- a) El beneficio medio en 2008 y 2012.
b) La disminución media anual (en %) de los beneficios en el periodo 2008-2012.

Solución:

$$a) \quad \bar{x}_{2008} = \frac{1}{n} \sum_{i=1}^s \bar{x}_i n_i = \frac{1}{3000} ((40 \times 800) + (65 \times 1200) + (50 \times 1000)) = \frac{160000}{3000} = 53,33$$

$$\bar{x}_{2012} = \frac{1}{n} \sum_{i=1}^s \bar{x}_i n_i = \frac{1}{2800} ((35 \times 735) + (60 \times 1165) + (35 \times 900)) = \frac{127125}{2800} = 45,4$$

- b) $1+r = \sqrt[4]{\frac{45,4}{53,33}} = 0,96 \Rightarrow r = -0,04 \text{ } (-4\%)$. Se ha producido una disminución media anual del 4% en los beneficios del sector.

4. El salario medio, la desviación típica de los salarios y el número de empleados en tres empresas filiales de G.E.C.O.S.A. son

	Salario medio	Desviación típica	Número de empleados
A	1365	105	100
B	1300	100	300
C	1560	120	170

Se decide subir el salario un 5% en la filial A, un 6% en la filial B y 50 euros a los empleados de la filial C.

- a) ¿En cuál de las tres filiales son más homogéneos los salarios antes de la subida?
b) ¿En cuál de las tres filiales son más homogéneos los salarios tras la subida?

Solución:

- a) Antes de la subida la homogeneidad de los salarios es la misma en las tres filiales

$$CV_A = \frac{105}{1365} = CV_B = \frac{100}{1300} = CV_C = \frac{120}{1560} = 0,0769$$

- b) Después de la subida la homogeneidad no cambia en las filiales A y B pues el coeficiente de variación es invariante frente a cambios de escala.

$$Y = eX \Rightarrow CV_Y = \frac{\bar{y}}{S_y} = \frac{e\bar{x}}{eS_x} = \frac{\bar{x}}{S_x} = CV_X$$

Sin embargo, si cambia en la filial C al cambiar la media aunque no la desviación típica.

Si llamamos X al salario antes de la subida e Y al salario después de la subida:

$$\text{Filial A} \quad Y = X + 0,05X = 1,05X$$

$$\text{Filial B} \quad Y = X + 0,06X = 1,06X$$

$$\text{Filial C} \quad Y = X + 50$$

$$CV_A = CV_B = 0,0769 \quad CV_C = \frac{120}{1560 + 50} = \frac{120}{1610} = 0,0745$$

Por tanto, después de la subida la filial C es la de salarios más homogéneos (aunque C tiene la mayor desviación típica, recuerde que para comparar la dispersión de dos variables estadísticas se ha de usar una medida de dispersión relativa)

5. Sea una distribución de frecuencias con media 300, varianza 36 y $n=5000$. ¿Cuántas observaciones contiene el conjunto $(-\infty, 288] \cup [312, \infty)$?

Solución:

La desigualdad de Tchebycheff afirma que la proporción de datos en el intervalo $(\bar{x} - kS, \bar{x} + kS)$ es mayor o igual que $1 - \frac{1}{k^2}$ ($k > 1$).

$$p\left[x_i \in (\bar{x} - kS, \bar{x} + kS)\right] \geq 1 - \frac{1}{k^2}$$

En nuestro caso: $288 = 300 - 12 = \bar{x} - 2S$ $312 = 300 + 12 = \bar{x} + 2S$

Por tanto, la proporción de observaciones en el intervalo (288, 312) es mayor o igual que

$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = \frac{3}{4} = 0,75$, de modo que en el conjunto $(-\infty, 288] \cup [312, \infty)$ habrá menos de un

25% de las observaciones, es decir, menos de 1250 observaciones $\left(\frac{25}{100} 5000 = 1250\right)$.

6. En un fenómeno se han observado 2000 individuos con media 1000 y varianza 100. ¿Cuántas observaciones, como mínimo, son mayores que 970 y menores que 1030?

Solución:

De nuevo, de acuerdo a la desigualdad de Tchebycheff:

$970 = 1000 - 30 = \bar{x} - 3S$, $1030 = 1000 + 30 = \bar{x} + 3S$, en el intervalo $(\bar{x} - 3S, \bar{x} + 3S)$ hay una

proporción de observaciones mayor o igual que $1 - \frac{1}{3^2} = \frac{8}{9} = 0,8889$. El 88,89% de las 2000

observaciones son 1777,8, luego como mínimo hay 1778 observaciones entre 970 y 1030.

7. Para asignar los puestos de trabajo en una cadena de montaje se realiza un test a los 90 empleados; 12 de ellos realizarán un trabajo tipo A (los que obtengan mejor puntuación), otros tantos un trabajo tipo C (los que saquen puntuación más baja), y el resto realizarán labores tipo B. El resultado del test fue:

Puntuación	0-30	30-50	50-70	70-100	100-120	120-150
n_i	10	15	20	20	20	5

¿Cuál fue la puntuación en el test para los que desempeñarán un trabajo tipo B?

Solución:

L_{i-1}	L_i	n_i	N_i
0	30	10	10
30	50	15	25
50	70	20	45
70	100	20	65
100	120	20	85
120	150	5	90

Si los 12 mejores harán el trabajo A, los $78=90-12$ restantes con la puntuación más baja no lo harán. Buscamos 78 en la columna de frecuencias acumuladas, como no aparece interpolaremos entre los siguientes valores

100	65
x	78
120	85

$$\frac{120-100}{85-65} = \frac{x-100}{78-65} \Rightarrow x=113$$

Los que obtengan más de 113 harán el trabajo A.

Los 12 con puntuación más baja harán el trabajo C. Buscamos 12 en la columna de frecuencias acumuladas e interpolamos entre los valores más próximos

30	10
x	12
50	25

$$\frac{50-30}{25-10} = \frac{x-30}{12-10} \Rightarrow x=32,67$$

Los que obtengan menos de 32,67 puntos harán el trabajo C.

Por tanto harán el trabajo B los que obtengan una puntuación entre 32,67 y 113 puntos.

8. En la siguiente tabla se recogen los salarios anuales, en miles de euros, de los empleados de dos empresas del mismo sector

salarios	Número de empleados	
	Empresa A	Empresa B
5-15	2	5
15-25	15	5
25-35	28	10
35-45	45	50
45-55	10	30

- ¿Qué empresa constituye un grupo más homogéneo de empleados en cuanto a salarios se refiere?
- ¿Qué empresa presenta mayor concentración en sus salarios?
- ¿Qué porcentaje de la masa salarial de la empresa A perciben los trabajadores cuyo salario anual está comprendido entre 22500 y 30000 euros?
- ¿Qué salario anual percibe un empleado de la empresa B que se encuentra dentro del 25% de los mejor pagados en dicha empresa?

Solución:**Empresa A**

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	N_i	u_i	p_i	q_i
5	15	10	2	20	200	2	20	2	0,58
15	25	20	15	300	6000	17	320	17	9,25
25	35	30	28	840	25200	45	1160	45	33,53
35	45	40	45	1800	72000	90	2960	90	85,55
45	55	50	10	500	25000	100	3460	100	100,00
			100	3460	128400			254	228,90

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{3460}{100} = 34,6$$

$$S^2 = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \frac{128400}{100} - 34,6^2 = 86,84 \Rightarrow S = \sqrt{86,84} = 9,3188 \Rightarrow CV(A) = \frac{S}{\bar{x}} = 0,2693$$

$$I_G(A) = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{118,90}{154} = 0,163$$

Empresa B

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	N_i	u_i	p_i	q_i
5	15	10	5	50	500	5	50	5	1,27
15	25	20	5	100	2000	10	150	10	3,80
25	35	30	10	300	9000	20	450	20	11,39
35	45	40	50	2000	80000	70	2450	70	62,03
45	55	50	30	1500	75000	100	3950	100	100,00
			100	3950	166500			205	178,48

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{3950}{100} = 39,5$$

$$S^2 = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \frac{166500}{100} - 39,5^2 = 104,75 \Rightarrow S = \sqrt{104,75} = 10,2347 \Rightarrow CV(B) = \frac{S}{\bar{x}} = 0,2591$$

$$I_G(B) = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{78,48}{105} = 0,2526$$

- a) Para ver donde son más homogéneos los salarios estudiamos la variabilidad o dispersión de esta variable. Para poder comparar la dispersión necesitamos una medida de dispersión relativa como el coeficiente de variación. Se observa una dispersión similar en ambas empresas, si bien algo mayor en la empresa A

$$CV(A) = 0,2693 > CV(B) = 0,2591$$

Observe que si nos hubiéramos apoyado en los valores de las varianzas o de las desviaciones típicas la conclusión hubiese sido la opuesta.

- b) El índice de Gini en la empresa A es menor que en la empresa B, luego hay menos concentración en los salarios de la empresa A.

$$I_G(A) = 0,163 < I_G(B) = 0,2526$$

- c) Buscamos en la tabla de la empresa A el tanto por ciento de masa salarial acumulada asociada a los valores 22,5 y 30. Como ninguno aparece en la tabla interpolaremos entre los valores más próximos:

15	0,58
22,5	x
25	9,25

$$\frac{25-15}{9,25-0,58} = \frac{22,5-15}{x-0,58} \Rightarrow x = 7,08$$

25	9,25
30	x
33,53	33,53

$$\frac{35-25}{33,53-9,25} = \frac{30-25}{x-9,25} \Rightarrow x = 21,39$$

Los trabajadores con un salario inferior a 30000 euros reciben el 21,39% de la masa salarial. Los que su salario es inferior a 22500 euros reciben el 7,08% de la masa salarial. Por tanto los que su salario está comprendido entre 22500 y 30000 euros recibirán el 14,31% de la masa salarial ($21,39-7,08=14,31$).

- d) El salario que es superado por el 25% de los empleados de la empresa B mejor pagados es el mismo salario que no es alcanzado por el 75% restante. Buscamos en la columna de los p_i el valor 75, como no aparece, interpolamos entre los valores más próximos

45	70
x	75
55	100

$$\frac{55-45}{100-70} = \frac{x-45}{75-70} \Rightarrow x = 46,667 \Rightarrow 46667 \text{ euros}$$

9. El sueldo mensual, en euros, correspondiente a los empleados de dos factorías de una misma empresa es

Sueldo mensual	Número de empleados. Factoría A	Número de empleados. Factoría B
600-1000	20	16
1000-1400	40	20
1400-2000	30	32
2000-3000	10	32

- ¿Qué sueldo corresponde al 60% de los empleados de la empresa?
- Calcule el sueldo del 25% de los empleados de la factoría B con menor salario.
- ¿Qué sueldo puede ser considerado moda de la factoría A?
- Halle el sueldo medio: de la factoría A, de la factoría B y de la empresa.
- ¿En cuál de las dos factorías los sueldos son más homogéneos?

- f) ¿En cuál de las dos factorías los sueldos tienen una concentración mayor?
- g) Calcule la mediana y la media para todos los empleados de la empresa.
- h) ¿Qué porcentaje de empleados de la empresa tienen un sueldo superior a 2300 euros?
- i) ¿Qué sueldo corresponde al 30% de los empleados de la empresa con mayor sueldo?
- j) Calcule el porcentaje de empleados de la empresa, con menor sueldo, que reciben el 43% de la nómina.
- k) ¿Cuántos empleados de la empresa, con mayor sueldo, reciben el 25% de la nómina?

Solución:

Sumando el número de empleados en las dos factorías de la empresa obtenemos la distribución de frecuencias para todos los empleados de la empresa.

Sueldo mensual	Número de empleados. Factoría A	Número de empleados. Factoría B	Número de empleados. EMPRESA
600-1000	20	16	36
1000-1400	40	20	60
1400-2000	30	32	62
2000-3000	10	32	42

En primer lugar realizamos en las siguientes tablas los cálculos necesarios para responder a todos los apartados:

Factoría A

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	a_i	h_i	N_i	u_i	p_i	q_i
600	1000	800	20	16000	12800000	400	0,05	20	16000	20	11,43
1000	1400	1200	40	48000	57600000	400	0,1	60	64000	60	45,71
1400	2000	1700	30	51000	86700000	600	0,05	90	115000	90	82,14
2000	3000	2500	10	25000	62500000	1000	0,01	100	140000	100	100,00
			100	140000	219600000				270	239,29	

Factoría B

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	N_i	u_i	p_i	q_i	
600	1000	800	16	12800	10240000	16	12800	16	7,48	
1000	1400	1200	20	24000	28800000	36	36800	36	21,50	
1400	2000	1700	32	54400	92480000	68	91200	68	53,27	
2000	3000	2500	32	80000	200000000	100	171200	100	100,00	
			100	171200	331520000				220	182,24

Empresa

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	N_i	u_i	p_i	q_i	
600	1000	800	36	28800	23040000	36	28800	18	9,25	
1000	1400	1200	60	72000	86400000	96	100800	48	32,39	
1400	2000	1700	62	105400	179180000	158	206200	79	66,26	
2000	3000	2500	42	105000	262500000	200	311200	100	100,00	
			200	311200	551120000				245	207,90

- a) Esta pregunta puede tener diferentes interpretaciones: el 60% de los empleados de la empresa con mayores sueldos, el 60% con menores sueldos, el 60% de los sueldos intermedios o centrales,...

Para su resolución aquí supondremos el 60% de los empleados con menores sueldos.

60 no aparece en la columna de los p_i de la tabla “**Empresa**”, por lo que interpolaremos entre los valores más próximos a 60 en la tabla.

1400	48
x	60
2000	79

$$\frac{2000-1400}{79-48} = \frac{x-1400}{60-48} \Rightarrow x = 1632,26 \text{ euros}$$

El 60% de los empleados con menores sueldos tienen un salario inferior a 1632,26 €.

- b) 25 no aparece en la columna de los p_i de la tabla “**Factoría B**”, por lo que interpolaremos entre los valores más próximos a 25 en la tabla.

1000	16
x	25
1400	36

$$\frac{1400-1000}{36-16} = \frac{x-1000}{25-16} \Rightarrow x = 1180 \text{ euros}$$

El 25% de los empleados de la factoría B con menores salarios tienen un salario por debajo de 1180 €.

- c) La moda es el valor más frecuente de la variable. En variables continuas se sitúa en el intervalo de mayor altura (que no siempre coincide con el de mayor frecuencia, aunque sí en este caso). Calculamos las alturas dividiendo la frecuencia absoluta, n_i , entre la amplitud del intervalo, a_i . Dentro del intervalo de mayor altura ($h_i = 0,1$, 1000-1400) hay diferentes criterios para situar la moda, el más sencillo es el del punto medio del intervalo. Según dicho criterio, 1200 euros puede considerarse la moda en la factoría A.

$$d) \bar{x}(A) = \frac{140000}{100} = 1400 \quad \bar{x}(B) = \frac{171200}{100} = 1712 \quad \bar{x}(\text{empresa}) = \frac{311200}{200} = 1556$$

$$e) S^2(A) = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \frac{219600000}{100} - 1400^2 = 236000 \Rightarrow S(A) = \sqrt{236000} = 485,8$$

$$CV(A) = \frac{S(A)}{\bar{x}(A)} = 0,347$$

$$S^2(B) = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \frac{331520000}{100} - 1712^2 = 384256 \Rightarrow S(B) = \sqrt{384256} = 619,88$$

$$CV(B) = \frac{S(B)}{\bar{x}(B)} = 0,362$$

Hay menos dispersión, por tanto son más homogéneos los sueldos, en la factoría A, aunque como puede observarse las diferencias no son importantes.

$$f) \quad I_G(A) = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{139,29}{170} = 0,181 \quad I_G(B) = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{82,24}{120} = 0,315$$

Hay una mayor concentración en el reparto de los sueldos en la factoría B.

$$g) \quad Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i = 1400 + \frac{100 - 96}{62} 600 = 1438,71$$

$$Ml = L_{i-1} + \frac{50 - q_{i-1}}{q_i - q_{i-1}} a_i = 1400 + \frac{50 - 32,39}{66,26 - 32,39} 600 = 1711,96$$

- h) En la tabla **“Empresa”** buscamos el valor 2300. Éste no está, interpolamos entre los valores más próximos a 2300:

2000	79	$\frac{3000 - 2000}{100 - 79} = \frac{2300 - 2000}{x - 79} \Rightarrow x = 85,3\% \Rightarrow 100 - 85,3 = 14,7\%$
2300	x	
3000	100	

Luego el 14,7% de empleados en la empresa tienen un sueldo superior a 2300 €.

- i) El mismo sueldo que es superado sólo por el 30% de los empleados de la empresa con mayor sueldo no es alcanzado por el 70% restante. Buscamos en la tabla **“Empresa”** el valor 70 en la columna p_i e interpolamos entre los valores más próximos:

1400	48	$\frac{2000 - 1400}{79 - 48} = \frac{x - 1400}{70 - 48} \Rightarrow x = 1825,8 \text{ euros}$
x	70	
2000	79	

El 30% de los empleados de la empresa con mayor sueldo tienen un sueldo superior a 1825,8 €.

- j) Buscamos en la columna q_i de la tabla **“Empresa”** el valor 43 e interpolamos:

48	32,39	$\frac{79 - 48}{66,26 - 32,39} = \frac{x - 48}{43 - 32,39} \Rightarrow x = 57,71\%$
x	43	
79	66,26	

- k) Buscamos en la tabla **“Empresa”** el porcentaje de empleados, con menor sueldo, que reciben el 75% de la nómina

79	66,26	$\frac{100 - 79}{100 - 66,26} = \frac{x - 79}{75 - 66,26} \Rightarrow x = 84,44\% \Rightarrow 100 - 84,44 = 15,56\%$
x	75	
100	100	

Por tanto, el 15,56% restante de empleados recibirán el 25% restante de la nómina. El 15,56% de los 200 empleados de la empresa son $31,12 \cong 31$ empleados.

10. Se conoce el salario medio, la desviación típica de los salarios y el número de empleados de dos empresas filiales:

Filial	Sueldo medio (€)	Desviación típica (€)	Número de empleados
A	853,4	60	25
B	795	52	35

Se decide subir el sueldo un 5% a los empleados de la filial A y subir 138 euros a cada uno de los de la filial B. Justifique en cuál de las dos filiales los sueldos serán más heterogéneos después de la subida.

Solución:

Llamemos X =sueldo antes de la subida e Y =sueldo tras la subida.

En la filial A la relación entre X e Y es un sencillo cambio de escala:

$$Y=X+0,05X=1,05X$$

En la filial B la relación entre X e Y es un cambio de origen

$$Y=X+138$$

En la filial A la media y desviación típica para la nueva variable Y son:

$$\bar{y}(A) = 1,05 \times \bar{x}(A) = 1,05 \times 853,4 = 897,12$$

$$S_y(A) = 1,05 \times S_x(A) = 1,05 \times 60 = 63$$

En la filial B la media y desviación típica para la nueva variable Y son:

$$\bar{y}(B) = \bar{x}(B) + 138 = 795 + 138 = 933$$

La desviación típica de X e Y es la misma en la filial B puesto que los cambios de origen no afectan a su valor.

Para comparar la variabilidad de los sueldos después de la subida utilizamos el coeficiente de variación

$$CV_A = \frac{63}{897,12} = 0,0702 \qquad CV_B = \frac{52}{933} = 0,0557$$

Luego, después de la subida, son más heterogéneos los sueldo en la filial A.

11. En una empresa de embalaje se conoce el número de cajas que hacen los 75 empleados de esa sección al final de una semana:

Número de cajas	350-400	400-450	450-500	500-550
% de empleados	26,67	22,67	36	14,66

- Calcule la varianza.
- Calcule el número de cajas que con más frecuencia hace un empleado.
- Para incentivar la productividad se decide aumentar el salario 70 euros a la tercera parte de los empleados que más cajas hace. ¿Cuál ha de ser el número mínimo de cajas que debe hacer un empleado para que le suban el sueldo?

Solución:

L_{i-1}	L_i	x_i	f_i	$x_i f_i$	$x_i^2 f_i$	a_i	h_i	F_i	p_i
350	400	375	0,2667	100,0125	37504,6875	50	0,005334	0,2667	26,67
400	450	425	0,2267	96,3475	40947,6875	50	0,004534	0,4934	49,34
450	500	475	0,36	171	81225	50	0,0072	0,8534	85,34
500	550	525	0,1466	76,965	40406,625	50	0,002932	1	100

1 444,325 200084 261,35

- a) En este ejemplo nos dan la distribución de frecuencias en tantos por ciento. Estos tantos por ciento son las frecuencias relativas multiplicadas por 100.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \sum_{i=1}^k x_i f_i = 444,325$$

$$S^2 = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 f_i \right) - \bar{x}^2 = 200084 - 444,325^2 = 2659,29$$

- b) Al ser todos los intervalos de igual amplitud, el intervalo de mayor altura coincide con el de mayor frecuencia relativa (y absoluta), 450-500. Según los distintos criterios de obtención de la moda obtenemos:

$$Mo(I) = \frac{L_{i-1} + L_i}{2} = \frac{450 + 500}{2} = 475$$

$$Mo(II) = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i = 450 + \frac{0,002932}{0,004534 + 0,002932} 50 = 469,64$$

$$Mo(III) = L_{i-1} + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} a_i = 450 + \frac{0,0072 - 0,004534}{(0,0072 - 0,004534) + (0,0072 - 0,002932)} 50 = 469,22$$

- c) $1/3=0,3333$, al 33,33% que más paquetes hacen se le sube el salario, al 66,67% restante que menos paquetes hacen no se le subirá. Calcularemos el percentil 66,67 buscando este valor en la columna p_i de la anterior tabla. Los valores p_i se han obtenido a partir de las frecuencias relativas acumuladas, F_i , expresando éstas en tanto por ciento. Interpolando entre los valores más próximos a 66,67:

450	49,34
x	66,67
500	85,34

$$\frac{500 - 450}{85,34 - 49,34} = \frac{x - 450}{66,67 - 49,34} \Rightarrow x = 474,07 \text{ cajas}$$

474 cajas no son superadas por dos tercios (66,67%) de los empleados, por tanto serán superadas por el tercio de empleados restantes.

12. Se dispone de la siguiente información sobre una empresa:

Sueldos/ hora (€)	% de empleados	q_i
0-6	21,43	3,38
6-20	28,57	22,93
20-30	35,71	69,92
30-50	14,29	100

- Calcule el índice de Gini e interprételo.
- Calcule la media y la mediana.
- ¿Qué porcentaje de empleados gana más de 27 euros/hora?

Solución:

A partir de los % de empleados, dividiendo por 100 obtenemos las frecuencias relativas y acumulándolos obtenemos los p_i

L_{i-1}	L_i	x_i	f_i	$x_i f_i$	p_i	q_i
0	6	3	0,2143	0,6429	21,43	3,38
6	20	13	0,2857	3,7141	50	22,93
20	30	25	0,3571	8,9275	85,71	69,92
30	50	40	0,1429	5,716	100	100,00
			1	19,0005	257,14	196,23

$$a) I_G = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{96,23}{157,14} = 0,3876$$

El índice de Gini es un valor entre 0 y 1, cuanto más próximo esté a 1 indica una mayor concentración en el reparto. En este caso su valor nos indica una débil concentración.

$$b) \bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \sum_{i=1}^k x_i f_i = 19,0005$$

Sencillamente observando la tabla encontramos que el 50% de las observaciones son menores o iguales a 20, por tanto $Me=20$.

- Calcularemos, interpolando, el porcentaje de empleados que ganan menos de 27 €/hora, el resto de empleados hasta completar el 100 ganarán más de 27 €/hora.

20	50
27	x
30	85,71

$$\frac{30-20}{85,71-50} = \frac{27-20}{x-50} \Rightarrow x = 75\% \Rightarrow 100-75 = 25\%$$

- Se sabe que una cuenta de ahorro ha producido anualmente, en los últimos cinco años, los siguientes intereses: 4%, 7%, 6%, 5% y 8%. Calcule la rentabilidad anual media para el período en cada uno de los siguientes casos:

- Se mantienen en la cuenta los intereses.
- Se retiran anualmente los intereses.

Solución:

- a) Si partimos de un capital C_0 , al final del primer año se habrá transformado en

$$C_1 = C_0 + \frac{4}{100} C_0 = 1,04 C_0$$

Al final del segundo año C_1 se habrá transformado en

$$C_2 = C_1 + \frac{7}{100} C_1 = 1,07 C_1 = 1,07(1,04 C_0)$$

Así, al final del quinto año tendremos un capital:

$$C_5 = C_4 + \frac{8}{100} C_4 = 1,08 C_4 = 1,08 \times 1,05 \times 1,06 \times 1,07 \times 1,04 \times C_0$$

La rentabilidad media sería aquella rentabilidad r (en tanto por uno) que puede sustituir a las rentabilidades observadas (4%, 7%, 6%, 5% y 8%) obteniéndose el mismo capital al final, es decir

$$C_5 = 1,08 \times 1,05 \times 1,06 \times 1,07 \times 1,04 \times C_0 = (1+r) \times (1+r) \times (1+r) \times (1+r) \times (1+r) \times C_0 = (1+r)^5 C_0$$

$$1+r = \sqrt[5]{1,08 \times 1,05 \times 1,06 \times 1,07 \times 1,04} = 1,0599 \quad \Rightarrow \quad r = 0,0599 \quad (5,99\%)$$

- b) Si retiramos anualmente los intereses, al comienzo de cada año hay siempre el mismo capital C_0 y al cabo de los 5 años tendremos C_0 más los intereses producidos en cada año.

$$C_5 = C_0 + \frac{4}{100} C_0 + \frac{7}{100} C_0 + \frac{6}{100} C_0 + \frac{5}{100} C_0 + \frac{8}{100} C_0$$

En este caso la rentabilidad media r cumpliría que

$$C_5 = C_0 + rC_0 + rC_0 + rC_0 + rC_0 + rC_0$$

Por tanto

$$\left(\frac{4}{100} + \frac{7}{100} + \frac{6}{100} + \frac{5}{100} + \frac{8}{100} \right) C_0 = 5rC_0$$

$$r = \frac{0,04 + 0,07 + 0,06 + 0,05 + 0,08}{5} = 0,06$$

Cuando los valores de la variable, como en el apartado a), representan incrementos acumulativos el promedio adecuado es la media geométrica. Si como en el apartado b) los incrementos no son acumulativos se debe usar como promedio la media aritmética.

14. Un inversor mantuvo el mismo tipo de inversión durante cuatro años consecutivos. Si las respectivas rentabilidades fueron del 11%, 8%, 7% y 9%, justifique cual es el promedio más adecuado para obtener el rendimiento medio anual y calcúlelo.

Solución:

El promedio más adecuado es la media geométrica, tal y como se justifica en el apartado a) del anterior ejercicio.

$$1+r = \sqrt[4]{1,11 \times 1,08 \times 1,07 \times 1,09} = 1,0874 \quad \Rightarrow \quad r = 0,0874 \quad (8,74\%)$$

15. Para comparar los rendimientos entre empresas europeas y estadounidenses pertenecientes al mismo sector, se han seleccionado 25 empresas de características semejantes en cada lugar:

Europa		Estados Unidos	
Beneficio en millones de €	n_i	Beneficios en millones de \$	n_i
1,5	5	3	4
3	8	3,5	2
4,2	6	4	6
5	4	4,5	4
6	2	5	4
		6	5

- ¿Dónde hay mayor homogeneidad en las empresas del sector?
- Estudie la concentración de los beneficios en cada lugar y compárelas.
- ¿Qué relación hay entre las conclusiones de los apartados a y b?

Solución:**Europa**

x_i	n_i	$x_i n_i$	$x_i^2 n_i$	N_i	u_i	p_i	q_i
1,5	5	7,5	11,25	5	7,5	20	8,46
3	8	24	72	13	31,5	52	35,51
4,2	6	25,2	105,84	19	56,7	76	63,92
5	4	20	100	23	76,7	92	86,47
6	2	12	72	25	88,7	100	100,00
	25	88,7	361,09			340	294,36

Estados Unidos

x_i	n_i	$x_i n_i$	$x_i^2 n_i$	N_i	u_i	p_i	q_i
3	4	12	36	4	12	16	10,81
3,5	2	7	24,5	6	19	24	17,12
4	6	24	96	12	43	48	38,74
4,5	4	18	81	16	61	64	54,95
5	4	20	100	20	81	80	72,97
6	5	30	180	25	111	100	100
	25	111	517,5			332	294,59

$$a) \quad \bar{x}(\text{Europa}) = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{88,7}{25} = 3,548$$

$$S^2(\text{Europa}) = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \frac{361,09}{25} - 3,548^2 = 1,855$$

$$S(Europa) = \sqrt{1,855} = 1,362$$

$$CV(Europa) = \frac{S(Europa)}{\bar{x}(Europa)} = \frac{1,362}{3,548} = 0,3839$$

$$\bar{x}(Estados Unidos) = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{111}{25} = 4,44$$

$$S^2(Estados Unidos) = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \frac{517,5}{25} - 4,44^2 = 0,9864$$

$$S(Estados Unidos) = \sqrt{0,9864} = 0,993$$

$$CV(Estados Unidos) = \frac{S(Estados Unidos)}{\bar{x}(Estados Unidos)} = \frac{0,993}{4,44} = 0,224$$

Hay menos dispersión, por tanto son más homogéneos los beneficios, en las empresas de Estados Unidos.

$$b) \quad I_G(Europa) = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{194,36}{240} = 0,19$$

$$I_G(Estados Unidos) = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{194,59}{232} = 0,161$$

Hay algo más de concentración en los beneficios de las empresas de Europa donde obtenemos un mayor valor del índice de Gini.

- c) No hay ninguna relación entre la concentración y la dispersión, son características diferentes de los datos que se miden con diferentes coeficientes.

16. Los saldos de las cuentas abiertas por los clientes de una sucursal bancaria se distribuyen de acuerdo a la siguiente tabla

Saldos (€)	Número de clientes
0-200	175
200-1000	500
1000-5000	300
5000-30000	25

Nota: llamaremos saldo total de la sucursal a la suma de los saldos de todas las cuentas abiertas en la misma.

- a) ¿Qué porcentaje del saldo total de la sucursal pertenece a los clientes cuyo saldo está comprendido entre 6000 y 10000 euros?

- b) Si se consideran clientes preferentes al 10% de los clientes con mayores saldos, ¿cuál ha de ser el saldo para que un cliente sea considerado como tal?
- c) ¿Qué porcentaje del saldo total de la sucursal pertenece al 30% de los clientes con mayor saldo? ¿Cuánto dinero representa ese porcentaje?
- d) ¿Qué saldo mínimo tienen los clientes con mayores saldos que reúnen el 20% del saldo total de la sucursal?
- e) ¿Qué porcentaje de clientes superan los 2000 euros de saldo?
- f) ¿Qué tanto por ciento de clientes con mayor saldo reúnen el 30% del saldo total de la sucursal?

Solución:

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
0	200	100	175	17500	175	17500	17,5	1,06
200	1000	600	500	300000	675	317500	67,5	19,18
1000	5000	3000	300	900000	975	1217500	97,5	73,56
5000	30000	17500	25	437500	1000	1655000	100	100
				1000	1655000	282,5	193,81	

- a) Buscamos el porcentaje del saldo total de la sucursal, q_i , correspondiente a clientes con menos de 6000 euros de saldo y clientes con menos de 10000 euros de saldo. Como dichos valores no aparecen directamente en la anterior tabla, interpolamos entre los valores más próximos:

5000	73,56
6000	x
30000	100

$$\frac{30000 - 5000}{100 - 73,56} = \frac{6000 - 5000}{x - 73,56} \Rightarrow x = 74,62$$

5000	73,56
10000	x
30000	100

$$\frac{30000 - 5000}{100 - 73,56} = \frac{10000 - 5000}{x - 73,56} \Rightarrow x = 78,85$$

A los clientes con menos de 10000 euros de saldo les corresponde el 78,85% del saldo total de la sucursal, a los clientes con menos de 6000 euros de saldo les corresponde el 74,62% del saldo total de la sucursal, por tanto a los clientes con un saldo comprendido entre 6000 y 10000 euros les corresponde el 4,23% del saldo total de la sucursal ($4,23 = 78,85 - 74,62$).

- b) Buscamos el saldo que no es superado por el 90% de los restantes clientes con menores saldos:

1000	67,5
x	90
5000	97,5

$$\frac{5000 - 1000}{97,5 - 67,5} = \frac{x - 1000}{90 - 67,5} \Rightarrow x = 4000$$

- c) Buscamos el porcentaje del saldo total de la sucursal que pertenece al 70% de los clientes con menor saldo:

67,5	19,18	$\frac{97,5-67,5}{73,56-19,18} = \frac{70-67,5}{x-19,18} \Rightarrow x = 23,7\% \Rightarrow 100-23,7 = 76,3\%$
70	x	
97,5	73,56	

Al 70% de los clientes con menos saldo les pertenece el 23,7% del saldo total de la sucursal, por tanto al 30% restante de clientes con más saldo les pertenece el 76,3% del saldo total de la sucursal.

Estimamos el saldo total de la sucursal por $\sum_{i=1}^k x_i n_i = 1655000$, el 76,3% del saldo total de la

sucursal es: $\frac{76,3}{100} 1655000 = 1262765$

(Utilizamos la expresión “*estimamos*” para el valor del saldo total de la sucursal puesto que hemos supuesto que todas las cuentas dentro de cada intervalo de la tabla tienen un saldo igual al punto central de dicho intervalo)

- d) Ese saldo mínimo será igual al saldo máximo del resto de clientes que reunirán el 80% del saldo total de la sucursal (80=100-20):

5000	73,56	$\frac{30000-5000}{100-73,56} = \frac{x-5000}{80-73,56} \Rightarrow x = 11089,25$
x	80	
30000	100	

- e) Calcularemos el porcentaje de clientes que no superan los 2000 euros de saldo:

1000	67,5	$\frac{5000-1000}{97,5-67,5} = \frac{2000-1000}{x-67,5} \Rightarrow x = 75\% \Rightarrow 100-75 = 25\%$
2000	x	
5000	97,5	

El 25% de los clientes superan los 2000 euros de saldo.

- f) Buscamos el tanto por ciento de clientes con menor saldo que representan el 70% del saldo total de la sucursal:

67,5	19,18	$\frac{97,5-67,5}{73,56-19,18} = \frac{x-67,5}{70-19,18} \Rightarrow x = 95,5\% \Rightarrow 100-95,5 = 4,5\%$
x	70	
97,5	73,56	

Por tanto, el 4,5% de clientes con mayor saldo representan el 30% restante del saldo total de la sucursal.

17. El montante de las pólizas correspondientes a los agentes visitantes de una compañía de seguros de vida se distribuye como sigue:

Montante de pólizas (millones de euros)	Número de agentes
0-5	8
5-10	10
10-20	16
20-40	16

- a) Calcule el índice de Gini. Comente el resultado.

- b) Calcule la mediana y la mediala. Comente los resultados.
- c) ¿Por debajo de qué montante de pólizas se encuentra el 40% de los agentes que menos cartera de seguros tienen?
- d) ¿Qué tanto por ciento de agentes tienen un montante de pólizas superior a 30 millones de euros?

Solución:

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
0	5	2,5	8	20	8	20	16	2,45
5	10	7,5	10	75	18	95	36	11,66
10	20	15	16	240	34	335	68	41,10
20	40	30	16	480	50	815	100	100
			50	815			220	155,21

$$a) I_G = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{55,21}{120} = 0,54$$

Este índice toma valores entre 0 y 1, siendo la concentración del reparto más fuerte cuanto más próximo esté a 1. En este caso el grado de concentración es intermedio entre los casos extremos como muestra el valor del índice.

- b) Estas medidas se pueden calcular mediante las expresiones:

$$Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i \qquad \qquad \qquad Ml = L_{i-1} + \frac{50 - q_{i-1}}{q_i - q_{i-1}} a_i$$

O bien, interpolando el valor 50 en las columnas p_i y q_i respectivamente (método que utilizaremos):

10	36
x	50
20	68

$$\frac{20-10}{68-36} = \frac{x-10}{50-36} \quad \Rightarrow \quad x = Me = 14,375$$

20	41,1
x	50
40	100

$$\frac{40-20}{100-41,1} = \frac{x-20}{50-41,1} \quad \Rightarrow \quad x = Ml = 23,02$$

La mayor o menor diferencia entre ellas nos da una medida del grado de concentración. Cuanto mayor sea la concentración mayor será la distancia entre la mediana y la mediala. Para tener una medida relativa de dicha distancia, ésta se suele comparar por cociente con el rango de la variable (diferencia entre el mayor y menor valor que toma la variable).

$$\frac{Ml - Me}{R} = \frac{23,02 - 14,375}{40 - 0} = 0,216$$

La anterior medida es un valor entre 0 y 1 que indica una mayor concentración cuanto más próximo esté de 1.

- c) Interpolamos 40 entre los valores más próximos de la tabla:

10	36
x	40
20	68

$$\frac{20-10}{68-36} = \frac{x-10}{40-36} \Rightarrow x = 11,25$$

- d) Calculamos el tanto por ciento de agentes que tienen un montante de pólizas inferior a 30 millones de euros

20	68
30	x
40	100

$$\frac{40-20}{100-68} = \frac{30-20}{x-68} \Rightarrow x = 84\% \Rightarrow 100-84 = 16\%$$

El 84% de los agentes tienen un montante de pólizas inferior a 30 millones de euros, el 16% restante de los agentes tendrán un montante de pólizas superior a 30 millones de euros.

18. La siguiente tabla de frecuencias muestra los salarios mensuales de los empleados de una empresa

Salarios (€)	Número de empleados
600-1000	30
1000-1400	110
1400-2000	40
2000-3000	20

- Calcule la mediana y la mediana.
- Analice la concentración en base a dichas medidas.
- Calcule el índice de Gini.
- ¿Qué porcentaje de empleados tienen un salario superior a 2400 €?
- ¿Qué porcentaje de la masa salarial perciben aquellos trabajadores cuyo salario está comprendido entre 1200 y 1500 euros?
- Calcule el salario de aquellos empleados mejor pagados que, en su conjunto, perciben el 30% de la masa salarial pagada en dicha empresa al mes.
- ¿Entre qué valores se sitúa el salario del 30% de los empleados con mayor retribución?
- ¿Qué porcentaje de masa salarial se dedica a pagar el salario del 40% de los empleados que menos ganan?

Solución:

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
600	1000	800	30	24000	30	24000	15	8,76
1000	1400	1200	110	132000	140	156000	70	56,93
1400	2000	1700	40	68000	180	224000	90	81,75
2000	3000	2500	20	50000	200	274000	100	100
			200	274000				
						275	247,45	

- a) Estas medidas se pueden calcular mediante las expresiones:

$$Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i = 1000 + \frac{\frac{200}{2} - 30}{110} 400 = 1254,545$$

$$Ml = L_{i-1} + \frac{50 - q_{i-1}}{q_i - q_{i-1}} a_i = 1000 + \frac{50 - 8,76}{56,93 - 8,76} 400 = 1342,45$$

- b) La mayor o menor diferencia entre ellas nos da una medida del grado de concentración. Cuanto mayor sea la concentración mayor será la distancia entre la mediana y la mediana. Para tener una medida relativa de dicha distancia, ésta se suele comparar por cociente con el rango de la variable (diferencia entre el mayor y menor valor que toma la variable).

$$\frac{Ml - Me}{R} = \frac{1342,45 - 1254,545}{3000 - 600} = 0,0366$$

La anterior medida es un valor entre 0 y 1 que indica una mayor concentración cuanto más próximo esté de 1. Como puede observarse hay poca concentración.

$$c) I_G = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{147,45}{175} = 0,157$$

- d) Calculamos el tanto por ciento de empleados con un salario inferior a 2400 €:

2000	90
2400	x
3000	100

$$\frac{3000 - 2000}{100 - 90} = \frac{2400 - 2000}{x - 90} \Rightarrow x = 94\% \Rightarrow 100 - 94 = 6\%$$

El 94% de los empleados tienen un salario inferior a 2400 €, por tanto el 6% restante tendrán un salario superior a 2400 €.

- e) Calculamos el tanto por ciento de masa salarial que perciben los empleados con un salario inferior a 1200 € y 1500 € respectivamente:

1000	8,76
1200	x
1400	56,93

$$\frac{1400 - 1000}{56,93 - 8,76} = \frac{1200 - 1000}{x - 8,76} \Rightarrow x = 32,85\%$$

1400	56,93
1500	x
2000	81,75

$$\frac{2000 - 1400}{81,75 - 56,93} = \frac{1500 - 1400}{x - 56,93} \Rightarrow x = 61,07\%$$

Los empleados con un salario inferior a 1500 € reciben el 61,07% de la masa salarial, los empleados con un salario inferior a 1200 € reciben el 32,85% de la masa salarial, por tanto los empleados con un salario comprendido entre 1200 y 1500 euros reciben el 28,22% de la masa salarial (61,07-32,85=28,22).

- f) El salario tal que los salarios que lo superan suponen en su conjunto el 30% de la masa salarial es el mismo salario tal que los salarios que no lo superan suponen en su conjunto el 70% restante de la masa salarial.

1400	56,93
x	70
2000	81,75

$$\frac{2000-1400}{81,75-56,93} = \frac{x-1400}{70-56,93} \Rightarrow x = 1715,9$$

- g) Calculamos el salario por debajo del cual se encuentran el 70% de los empleados peor pagados. Este valor se encuentra en la tabla, por lo que no es necesario interpolar. Si el 70% de los empleados peor pagados tienen un salario por debajo de 1400 €, el 30% restante de empleados mejor pagados tendrán un salario por encima de 1400 €, es decir entre 1400 y 3000 euros (donde 3000 es el máximo valor que puede tomar la variable según la tabla).
- h) Interpolamos el valor 40 en la columna de tantos por ciento acumulados de empleados:

15	8,76
40	x
70	56,93

$$\frac{70-15}{56,93-8,76} = \frac{40-15}{x-8,76} \Rightarrow x = 30,66\%$$

19. La siguiente tabla muestra intervalos de renta, porcentajes acumulados de población y porcentajes acumulados de renta

Intervalos de renta (€)	% de población	% de renta
0-800	10	1,87
800-1500	40	18,03
1500-2500	60	36,77
2500-3000	80	62,53
3000-5000	100	100

- a) Calcule la mediana.
- b) Calcule la mediana.
- c) Calcule el porcentaje de renta que corresponde al 30% de la población mejor pagada.

Solución:

L_{i-1}	L_i	p_i	q_i
0	800	10	1,87
800	1500	40	18,03
1500	2500	60	36,77
2500	3000	80	62,53
3000	5000	100	100

290 219,20

- a) Interpolamos el valor 50 en la columna del % acumulado de población:

1500	40
x	50
2500	60

$$\frac{2500-1500}{60-40} = \frac{x-1500}{50-40} \Rightarrow x = Me = 2000$$

- b) Interpolamos el valor 50 en la columna del % acumulado de renta

2500	36,77
x	50
3000	62,53

$$\frac{3000 - 2500}{62,53 - 36,77} = \frac{x - 2500}{50 - 36,77} \Rightarrow x = Ml = 2756,8$$

- c) Buscaremos en la tabla el porcentaje de renta que corresponde al 70% de la población peor pagada:

60	36,77
70	x
80	62,53

$$\frac{80 - 60}{62,53 - 36,77} = \frac{70 - 60}{x - 36,77} \Rightarrow x = 49,65\% \Rightarrow 100 - 49,65 = 50,35\%$$

El 70% de la población peor pagada recibe el 49,65% de toda la renta, por tanto el 30% restante de la población mejor pagada recibe el 50,35% restante de la renta.

20. Con los datos de la siguiente tabla:

Ingresos (€)	0-600	600-800	800-1000	1000-1500	1500-2000
Número de familias	15	30	55	70	42

- Calcule el índice de Gini.
- Obtenga la mediana y la mediana.
- Calcule el porcentaje de familias con mayores ingresos que obtienen en su conjunto el 33% de los ingresos totales.

Solución:

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
0	600	300	15	4500	15	4500	7,08	1,91
600	800	700	30	21000	45	25500	21,23	10,81
800	1000	900	55	49500	100	75000	47,17	31,78
1000	1500	1250	70	87500	170	162500	80,19	68,86
1500	2000	1750	42	73500	212	236000	100	100
			212	236000			255,67	213,36

$$a) I_G = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{113,36}{255,67} = 0,2718$$

- b) Interpolamos el valor 50 en la columna p_i para obtener la mediana:

1000	47,17
x	50
1500	80,19

$$\frac{1500 - 1000}{80,19 - 47,17} = \frac{x - 1000}{50 - 47,17} \Rightarrow x = Me = 1042,9$$

Interpolamos el valor 50 en la columna q_i para obtener la mediana:

1000	31,78
x	50
1500	68,86

$$\frac{1500 - 1000}{68,86 - 31,78} = \frac{x - 1000}{50 - 31,78} \Rightarrow x = Ml = 1245,7$$

- c) Calculamos en primer lugar el porcentaje de familias con menores ingresos que obtienen en su conjunto el 67% de los ingresos totales (67=100-33):

47,17	31,78
x	67
80,19	68,86

$$\frac{80,19 - 47,17}{68,86 - 31,78} = \frac{x - 47,17}{67 - 31,78} \Rightarrow x = 78,5\% \Rightarrow 100 - 78,5 = 21,5\%$$

El 78,5% de las familias con menos ingresos obtienen el 67% de los ingresos totales, por tanto el 21,5% restante de familias con mayores ingresos obtienen el 33% restante de los ingresos totales.

21. El servicio Central de Correos realiza una encuesta sobre el franqueo medio de las cartas que diariamente tiene que distribuir. La información recibida sobre 500 cartas es la siguiente:

Franqueo	0,30	0,40	0,50	0,70	1	1,20	1,80	2	2,50
Cartas	145	132	84	50	48	22	10	8	1

- Determine el franqueo medio en la muestra y verifique si es representativo.
- Si la muestra anterior es representativa del total de las cartas que diariamente se reparten, calcule si el servicio es rentable, teniendo en cuenta que se reparten 350000 cartas al día y que el costo medio diario es de 200000 euros.

Solución:

x_i	n_i	$x_i n_i$	$x_i^2 n_i$
0,30	145	43,5	13,05
0,40	132	52,8	21,12
0,50	84	42	21
0,70	50	35	24,5
1,00	48	48	48
1,20	22	26,4	31,68
1,80	10	18	32,4
2,00	8	16	32
2,50	1	2,5	6,25
	500	284,2	230

- Para verificar la representatividad de la media aritmética podemos usar cualquier medida de dispersión apoyada en este promedio, como la varianza, la desviación típica o el coeficiente de variación. Cuanto menor sea el valor de estas medidas de dispersión mayor será la representatividad de la media aritmética.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{284,2}{500} = 0,5684$$

$$S^2 = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 n_i \right) - \bar{x}^2 = \frac{230}{500} - 0,5684^2 = 0,1369 \quad S = \sqrt{0,1369} = 0,37$$

$$CV = \frac{S}{\bar{x}} = \frac{0,37}{0,5684} = 0,651$$

- b) Si se reparten 350000 cartas al día y el franqueo medio de cada carta es 0,5684 euros, se estima un franqueo total diario de $350000 \times 0,5684 = 198940$ euros que no alcanza el costo medio diario de 200000 euros, luego el servicio tiene pérdidas, no es rentable.

2. VARIABLES ESTADÍSTICAS BIDIMENSIONALES.

En los dos temas anteriores se ha estudiado sobre cada individuo de la muestra sólo el valor que presenta un carácter. En este tema se medirá sobre cada individuo dos caracteres (por ejemplo: salario y edad), estos dos caracteres los notaremos como X e Y . A la variable que representa dicho estudio se denomina *variable estadística bidimensional*.

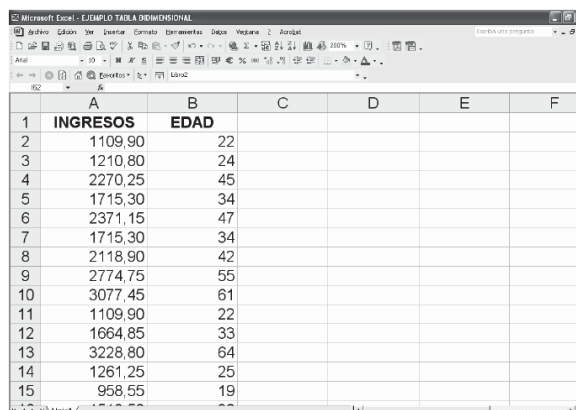
Con lo desarrollado en el tema 1 se pueden estudiar separadamente la variable X o la variable Y , calculando sus medidas de posición, dispersión, ... Pero lo **más importante** que se introduce en el estudio de *variables estadísticas bidimensionales* es el **análisis conjunto** de ambas variables para conocer la **relación entre ellas**.

2.1 Representaciones numéricas en dos columnas y en tablas de contingencia.

Dos columnas.

La tabla más sencilla para representar los valores de una variable estadística bidimensional consiste en colocar en dos columnas los pares de valores observados. Es la forma habitual de representarlos en hojas de cálculo y programas de estadística

x_i	y_i
x_1	y_1
x_2	y_2
...	...
x_n	y_n



	A	B	C	D	E	F
1	INGRESOS	EDAD				
2	1109,90	22				
3	1210,80	24				
4	2270,25	45				
5	1715,30	34				
6	2371,15	47				
7	1715,30	34				
8	2118,90	42				
9	2774,75	55				
10	3077,45	61				
11	1109,90	22				
12	1664,85	33				
13	3228,80	64				
14	1261,25	25				
15	958,55	19				

Tablas de contingencia.

También llamada tabla de correlación, es una tabla de doble entrada en la que se recogen ordenadas y clasificadas las observaciones presentadas por una variable estadística bidimensional.

Para ilustrar su forma nos apoyamos en el ejemplo siguiente:

► EJEMPLO 2.1.

La distribución según salarios (Y) y edades (X) de un grupo de 100 jóvenes es

X/Y	500-1000	1000-1500	1500-2000	$n_{i\bullet}$
20	10	3	2	15
21	5	15	5	25
22	2	20	15	37
23	0	13	10	23
$n_{\bullet j}$	17	51	32	$n=100$

En general la notación para este tipo de tablas es la siguiente

X/Y	B_1 $* \quad L_0^* - L_1^*$ y_1	...	B_j $L_{j-1}^* - L_j^*$ y_j	...	B_p $L_{p-1}^* - L_p^*$ y_p	$n_{i\bullet}$
$* \quad A_1 \quad L_0 - L_1 \quad x_1$	n_{11}	...	n_{1j}	...	n_{1p}	$n_{1\bullet}$
...
$A_i \quad L_{i-1} - L_i \quad x_i$	n_{i1}	...	n_{ij}	...	n_{ip}	$n_{i\bullet}$
...
$A_k \quad L_{k-1} - L_k \quad x_k$	n_{k1}	...	n_{kj}	...	n_{kp}	$n_{k\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$...	$n_{\bullet j}$...	$n_{\bullet p}$	n

* Pueden utilizarse para atributos, variables continuas o discretas.

El número de veces que se ha observado el par de valores (x_i, y_j) lo notamos como n_{ij} y se denomina **frecuencia absoluta conjunta** de dicho par.

La suma de todas las frecuencias absolutas es igual al número de observaciones, $\sum_{i=1}^k \sum_{j=1}^p n_{ij} = n$.

La **frecuencia relativa conjunta** se define a partir de la frecuencia absoluta como $f_{ij} = \frac{n_{ij}}{n}$, donde

n es el número total de observaciones. Su suma es igual a la unidad, $\sum_{i=1}^k \sum_{j=1}^p f_{ij} = 1$.

Se denomina **distribución bidimensional de frecuencias** al conjunto de valores $((x_i, y_j), n_{ij})$, $i=1, \dots, k$, $j=1, \dots, p$. Si utilizamos las frecuencias relativas se denomina **distribución bidimensional de frecuencias relativas**.

2.2 Distribuciones marginales y condicionadas. Independencia de variables estadísticas.

Distribuciones marginales.

Otros valores importantes que aparecen en la tabla de contingencia son $n_{i\bullet}$, $i=1,\dots,k$ y $n_{\bullet j}$, $j=1,\dots,p$, conocidos como **frecuencias absolutas marginales**.

$$n_{i\bullet} = \sum_{j=1}^p n_{ij} \quad n_{\bullet j} = \sum_{i=1}^k n_{ij} \quad \sum_{i=1}^k n_{i\bullet} = \sum_{j=1}^p n_{\bullet j} = n$$

Las **frecuencias relativas marginales** se definen a partir de las frecuencias absolutas como

$$f_{i\bullet} = \frac{n_{i\bullet}}{n} \quad f_{\bullet j} = \frac{n_{\bullet j}}{n} \quad \sum_{i=1}^k f_{i\bullet} = \sum_{j=1}^p f_{\bullet j} = 1$$

Utilizaremos estas frecuencias que aparecen en la última fila y columna de la *tabla de contingencia* para estudiar separadamente una variable de la otra. A estas distribuciones de frecuencias se les denominan **distribuciones marginales**.

Para estudiar separadamente una variable de la otra cuando tenemos los datos en *dos columnas* basta con quedarnos sólo con la columna de datos de dicha variable.

► EJEMPLO 2.2.

La distribución marginal de las edades en el ejemplo 2.1 es

x_i	$n_{i\bullet}$
20	15
21	25
22	37
23	23
	100

La distribución marginal de los salarios en el ejemplo 2.1 es

y_j	$n_{\bullet j}$
500-1000	17
1000-1500	51
1500-2000	32
	100

Sobre cada una de estas distribuciones marginales (unidimensionales) se puede estudiar todo lo que sobre distribuciones unidimensionales hemos visto (medias, percentiles, dispersión,...), permitiéndonos un conocimiento individual de cada una de estas variables, por ejemplo,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i\bullet}$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_{i\bullet}$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^p y_j n_{\bullet j}$$

$$S_y^2 = \frac{1}{n} \sum_{j=1}^p (y_j - \bar{y})^2 n_{\bullet j}$$

pero si queremos conocer la *relación entre dichas variables* hemos de analizar la **distribución conjunta (bidimensional) de ambas variables X e Y** (como se hará en los apartados 2.3 y 2.4).

Distribuciones condicionadas.

A partir de la representación conjunta de X e Y en la tabla de contingencia también se pueden estudiar las denominadas **distribuciones condicionadas** cuyo significado se expone con el siguiente ejemplo.

► EJEMPLO 2.3.

Distribución de los salarios condicionada a que el joven tiene 22 años

y_j	n_j
500-1000	2
1000-1500	20
1500-2000	15
	37

Distribución de las edades de los jóvenes que tienen un salario mayor que 1000€ pero menor que 1500€

x_i	n_i
20	3
21	15
22	20
23	13
	51



Cada columna de frecuencias de la tabla de contingencia (salvo la marginal) determina una *distribución condicionada de X* . Cada fila de frecuencias de la tabla de contingencia (salvo la marginal) determina una *distribución condicionada de Y* .

A partir de los datos representados en dos columnas no es tan fácil ni inmediata la determinación de las distribuciones condicionadas como en las tablas de contingencia.

Independencia de variables estadísticas.

Se dice que X e Y son independientes estadísticamente si todas las distribuciones condicionadas coinciden, en otras palabras, las frecuencias relativas de las distribuciones condicionadas son iguales (además también coinciden con las frecuencias relativas marginales). Esto es lo mismo que decir que el comportamiento de X no depende del valor de Y y que el comportamiento de Y tampoco depende del valor de X.

► EJEMPLO 2.4.

Comprobar si hay independencia estadística entre el salario, X, y el sexo, Y, en el grupo de 49 trabajadores representados en la siguiente tabla

X/Y	hombre	mujer	TOTAL
500-1000	20	8	28
1000-1500	10	4	14
1500-2000	5	2	7

Solución:

X/Y	hombre	mujer	TOTAL
500-1000	$\frac{20}{35} = 0,5714$	$\frac{8}{14} = 0,5714$	$\frac{28}{49} = 0,5714$
1000-1500	$\frac{10}{35} = 0,2857$	$\frac{4}{14} = 0,2857$	$\frac{14}{49} = 0,2857$
1500-2000	$\frac{5}{35} = 0,1429$	$\frac{2}{14} = 0,1429$	$\frac{7}{49} = 0,1429$

Luego hay independencia, el salario no depende del sexo, es decir, los salarios de los hombres no son distintos de los de las mujeres. ◀

Más cómodo, pero equivalente a comprobar que las frecuencias relativas de las distintas distribuciones condicionadas son iguales, es ver si se cumple

$$n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \quad \forall i, j$$

O equivalentemente

$$n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \Leftrightarrow \frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} \Leftrightarrow f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \quad \forall i, j$$

► **EJEMPLO 2.5.**

X/Y	hombre	mujer	$n_{i\cdot}$
500-1000	$20 = \frac{28 \times 35}{49}$	$8 = \frac{28 \times 14}{49}$	28
1000-1500	$10 = \frac{14 \times 35}{49}$	$4 = \frac{14 \times 14}{49}$	14
1500-2000	$5 = \frac{7 \times 35}{49}$	$2 = \frac{7 \times 14}{49}$	7
$n_{\cdot j}$	35	14	$n = 49$

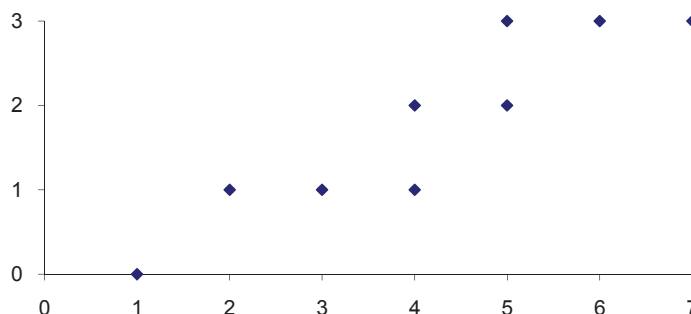
Usando esta caracterización de independencia, al trabajar con número enteros, evitamos los posibles errores de redondeo en los decimales de las frecuencias relativas. ◀

2.3 Covarianza y coeficiente de correlación lineal.

La representación gráfica más utilizada en el caso bidimensional es la **nube de puntos** o **diagrama de dispersión**. Consiste en representar un punto por cada observación, donde las coordenadas del punto coinciden con los valores X e Y .

► **EJEMPLO 2.6.**

X	Y
3	1
5	2
4	1
6	3
7	3
5	3
4	2
2	1
1	0



La **forma de la nube de puntos** nos informa sobre la **relación entre X e Y**. Esta información se cuantifica con el cálculo de la covarianza.

Covarianza

Se define para datos en *tablas de contingencia* como

$$S_{xy} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x})(y_j - \bar{y}) n_{ij}$$

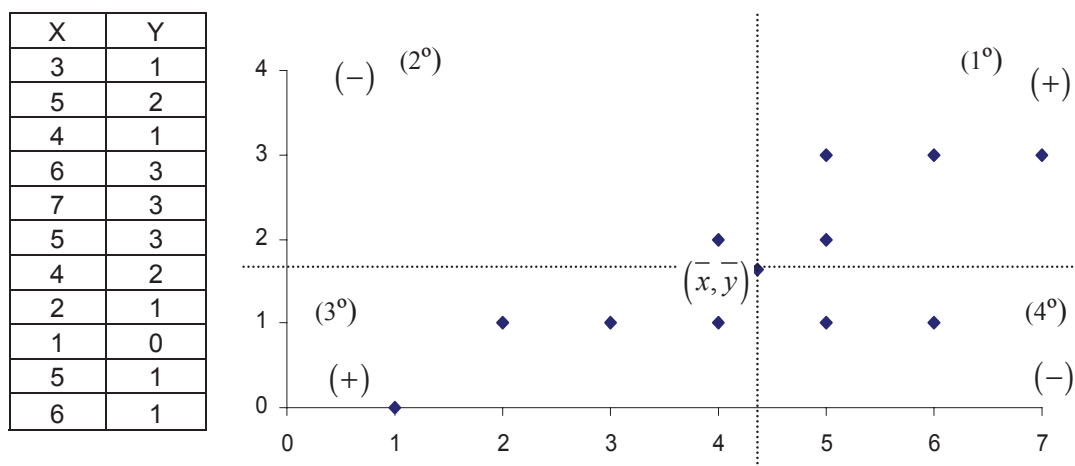
y para datos en *dos columnas* como

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

La covarianza nos da una medida de la variabilidad conjunta y por tanto de la asociación o relación entre las variables X e Y. El signo de la covarianza indica en qué sentido varían conjuntamente las variables. Si la **covarianza es positiva** las dos variables cambian, en general, en el mismo sentido (si una variable aumenta de valor la otra también aumenta, análogamente si una disminuye la otra hará lo mismo, en este caso decimos que la relación entre las variables es positiva o directa). Si la **covarianza es negativa** las dos variables cambian, en general, en sentidos opuestos (si una aumenta la otra disminuye, en este caso decimos que la relación entre las variables es negativa o inversa).

► EJEMPLO 2.7.

Calcule la covarianza para la siguiente tabla de datos e interprete su valor de acuerdo a la representación gráfica.



Solución:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
3	1	-1,36	-0,64	0,87
5	2	0,64	0,36	0,23
4	1	-0,36	-0,64	0,23
6	3	1,64	1,36	2,23
7	3	2,64	1,36	3,60
5	3	0,64	1,36	0,87
4	2	-0,36	0,36	-0,13
2	1	-2,36	-0,64	1,50
1	0	-3,36	-1,64	5,50
5	1	0,64	-0,64	-0,40
6	1	1,64	-0,64	-1,04
totales	48	18		13,46

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{48}{11} = 4,36 \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{18}{11} = 1,64 \quad S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{13,46}{11} = 1,22$$

Observemos los cuatro cuadrantes centrados en (\bar{x}, \bar{y}) . Las observaciones en los cuadrantes 1° y 3° dan lugar a sumandos positivos en la expresión de la covarianza,

$$\left[(1^\circ): x_i > \bar{x}, y_i > \bar{y} \right] \left[(3^\circ): x_i < \bar{x}, y_i < \bar{y} \right] \Rightarrow (x_i - \bar{x})(y_i - \bar{y}) > 0,$$

debido a que ambas diferencias son de igual signo. Mientras que los puntos en los cuadrantes 2° y 4° (sombreados en gris en la tabla anterior) dan lugar a sumandos negativos al tener dichas diferencias signos contrarios.

Cuando la relación entre las variables es directa predominan los puntos en los cuadrantes 1° y 3°, por tanto pesan más los sumandos positivos en la expresión de la covarianza, arrojando ésta un valor positivo (como el caso que nos ocupa). Cuando la relación entre las variables es inversa predominan los puntos en los cuadrantes 2° y 4° provocando un valor negativo de la covarianza.



Una forma más cómoda para el cálculo de la covarianza se obtiene al desarrollar la expresión que la define

$$\begin{aligned} S_{xy} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) = \\ &= \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \left(\bar{y} \frac{1}{n} \sum_{i=1}^n x_i \right) - \left(\bar{x} \frac{1}{n} \sum_{i=1}^n y_i \right) + \left(\bar{x} \bar{y} \frac{1}{n} \sum_{i=1}^n 1 \right) = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y} = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y} \end{aligned}$$

$$S_{xy} = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y} \quad (\text{datos en dos columnas})$$

$$S_{xy} = \left(\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} \right) - \bar{x} \bar{y} \quad (\text{datos en tablas de contingencia})$$

► EJEMPLO 2.8.

Obtenga la covarianza utilizando esta última expresión

x_i	y_i	$x_i y_i$
3	1	3
5	2	10
4	1	4
6	3	18
7	3	21
5	3	15
4	2	8
2	1	2
1	0	0
5	1	5
6	1	6
48	18	92

$$n = 11 \quad \bar{x} = \frac{48}{11} = 4,36 \quad \bar{y} = \frac{18}{11} = 1,64$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{92}{11} - (4,36 \times 1,64) = 1,21$$

(La diferencia de una centésima con el ejemplo 2.7 en el cálculo de la covarianza es debida a los errores de redondeo)

Si realizamos un **cambio de origen y escala** sobre las variables originales X e Y

$$X^* = eX + c \quad Y^* = hY + d$$

la covarianza de las nuevas variables es

$$S_{x^* y^*} = eh S_{xy}$$

(al valor de la covarianza no le afecta los cambios de origen pero sí los cambios de escala).

$$\begin{aligned} S_{x^* y^*} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p (x_i^* - \bar{x}^*) (y_j^* - \bar{y}^*) n_{ij} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p (ex_i + c - e\bar{x} - c) (hy_j + d - h\bar{y} - d) n_{ij} = \\ &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p (ex_i - e\bar{x}) (hy_j - h\bar{y}) n_{ij} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p e(x_i - \bar{x}) h(y_j - \bar{y}) n_{ij} = eh \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x}) (y_j - \bar{y}) n_{ij} = eh S_{xy} \end{aligned}$$

En caso de **independencia** estadística $\Rightarrow S_{xy} = 0$ (sin embargo, $S_{xy} = 0$ no implica necesariamente que se dé la condición de independencia estadística). Para demostrarlo nos apoyamos en la caracterización de la independencia estadística

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

$$\begin{aligned} S_{xy} &= \left(\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} \right) - \bar{x} \bar{y} = \left(\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p x_i y_j \frac{n_{i\bullet} n_{\bullet j}}{n} \right) - \bar{x} \bar{y} = \left(\frac{1}{n} \sum_{i=1}^k x_i n_{i\bullet} \sum_{j=1}^p y_j \frac{n_{\bullet j}}{n} \right) - \bar{x} \bar{y} = \\ &= \left(\frac{1}{n} \sum_{i=1}^k x_i n_{i\bullet} \cdot \frac{1}{n} \sum_{j=1}^p y_j n_{\bullet j} \right) - \bar{x} \bar{y} = (\bar{x} \bar{y}) - \bar{x} \bar{y} = 0 \end{aligned}$$

La covarianza no está acotada de forma uniforme para todas las variables ni está expresada en las mismas unidades, por lo que no es fácil interpretar su valor ni hacer comparaciones. Este problema se soluciona con el coeficiente de correlación lineal.

Coefficiente de correlación lineal.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Su **signo** es el **mismo** que el de la **covarianza**, además este coeficiente toma valores entre -1 y 1, $(-1 \leq r_{xy} \leq 1)$, midiendo objetivamente el grado de variación conjunta que tienen las variables X e Y . Cuando las variables están tipificadas el coeficiente de correlación lineal coincide con la covarianza.

El **cambio de origen y escala** no afecta al valor de este coeficiente. Sabemos que

$$X^* = eX + c \quad Y^* = hY + d \quad \Rightarrow \quad S_{x^*y^*} = ehS_{xy} \quad S_{x^*} = eS_x \quad S_{y^*} = hS_y$$

por tanto,

$$r_{x^*y^*} = \frac{S_{x^*y^*}}{S_{x^*}S_{y^*}} = \frac{ehS_{xy}}{eS_x hS_y} = \frac{S_{xy}}{S_x S_y} = r_{xy}$$

► EJEMPLO 2.9.

Calcule el coeficiente de correlación lineal con los datos del ejemplo 2.7.

Solución:

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
3	1	3	9	1
5	2	10	25	4
4	1	4	16	1
6	3	18	36	9
7	3	21	49	9
5	3	15	25	9
4	2	8	16	4
2	1	2	4	1
1	0	0	1	0
5	1	5	25	1
6	1	6	36	1
48	18	92	242	40

$$n = 11 \quad \bar{x} = \frac{48}{11} = 4,364 \quad \bar{y} = \frac{18}{11} = 1,636 \quad S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{92}{11} - (4,364 \times 1,636) = 1,22$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{242}{11} - 4,364^2 = 2,9555 \Rightarrow S_x = 1,72$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{40}{11} - 1,636^2 = 0,9599 \Rightarrow S_y = 0,98$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{1,22}{1,72 \times 0,98} = 0,7238$$

En el ejemplo anterior se recogen todos los cálculos necesarios en el estudio de variables estadísticas bidimensionales para datos en dos columnas. Veamos en el siguiente ejemplo como se llevan a cabo dichos cálculos cuando los datos están en una tabla de contingencia.

► **EJEMPLO 2.10.**

Calcule el coeficiente de correlación lineal con los datos del ejemplo 2.1.

X/Y	500-1000	1000-1500	1500-2000	$n_{i\bullet}$
20	10	3	2	15
21	5	15	5	25
22	2	20	15	37
23	0	13	10	23
$n_{\bullet j}$	17	51	32	$n=100$

Solución:

Si alguna de las variables es continua (datos agrupados en intervalos), como en este caso los salarios, tomaremos la marca de clase de cada intervalo, y_j , como valor representativo para hacer los cálculos. Junto a cada frecuencia n_{ij} calcularemos los productos $x_i y_j n_{ij}$ que incluiremos dentro de un recuadro para no confundirlos con los valores de las frecuencias.

X/Y	500-1000 750	1000-1500 1250	1500-2000 1750	$n_{i\bullet}$	$x_i n_{i\bullet}$	$x_i^2 n_{i\bullet}$
20	10 150000	3 75000	2 70000	15	300	6000
21	5 78750	15 393750	5 183750	25	525	11025
22	2 33000	20 550000	15 577500	37	814	17908
23	0 0	13 373750	10 402500	23	529	12167
$n_{\bullet j}$	17	51	32	$n=100$	2168	47100
$y_j n_{\bullet j}$	12750	63750	56000	132500		
$y_j^2 n_{\bullet j}$	9562500	79687500	98000000	187250000		

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i\bullet} = \frac{2168}{100} = 21,68$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^p y_j n_{\bullet j} = \frac{132500}{100} = 1325$$

$$S_{xy} = \left(\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} \right) - \bar{x} \bar{y} = \frac{150000 + \dots + 402500}{100} - (21,68 \times 1325) = \frac{2888000}{100} - 28726 = 154$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_{i\bullet} = \frac{1}{n} \sum_{i=1}^k x_i^2 n_{i\bullet} - \bar{x}^2 = \frac{47100}{100} - 21,68^2 = 0,9776 \quad \Rightarrow \quad S_x = 0,9887$$

$$S_y^2 = \frac{1}{n} \sum_{j=1}^p (y_j - \bar{y})^2 n_{\bullet j} = \frac{1}{n} \sum_{j=1}^p y_j^2 n_{\bullet j} - \bar{y}^2 = \frac{187250000}{100} - 1325^2 = 116875 \quad \Rightarrow \quad S_y = 341,87$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{154}{0,9887 \times 341,87} = 0,4556$$



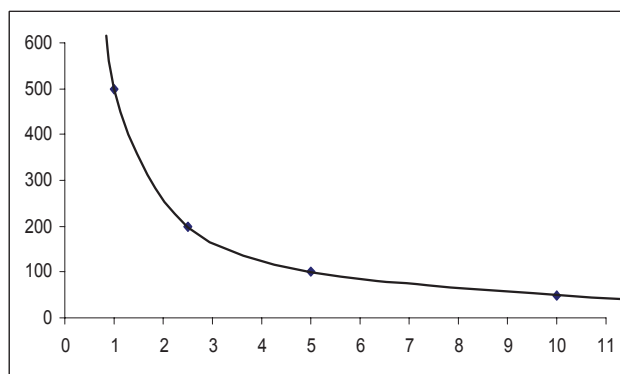
2.4 Recta de regresión de mínimos cuadrados.

Uno de los objetivos en toda ciencia es encontrar relaciones entre los hechos que estudia. Esas relaciones normalmente se intentan traducir en expresiones matemáticas. Así, si se observa el tiempo empleado por un móvil para recorrer 500 Km. cuando éste se mueve con velocidad uniforme, pueden obtenerse los siguientes valores

t (horas)	v (Km./h)
5	100
10	50
2.5	200
1	500

Estos valores están claramente relacionados y esa relación puede expresarse mediante la función $vt=500$. Los datos observados son valores particulares que verifican dicha ecuación.

Representando gráficamente dicha función $v = \frac{500}{t}$ y los puntos observados (t_i, v_i) , se tiene



Existen otro tipo de variables como el consumo y la renta, la oferta y la demanda,... donde no cabe duda de que hay una relación, pero donde es imposible definir sobre ellas una función matemática que verifiquen rigurosamente. A este tipo de dependencia entre variables la denominaremos **dependencia estadística** frente a la **dependencia funcional** que es el término utilizado para referirse a relaciones estrictas entre las variables, como la observada entre v y t .

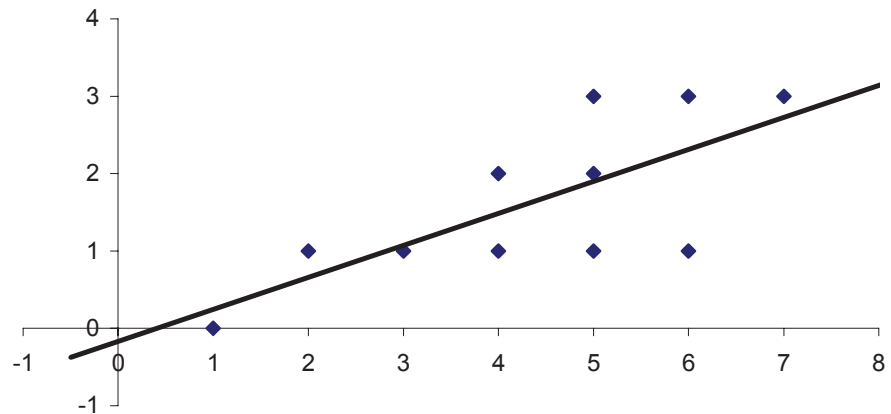
Es destacable el hecho de que dependencias de tipo estadístico son muy frecuentes en Economía y, en general, en todas las Ciencias Sociales.

Si dos variables presentan una dependencia estadística, no funcional, no es posible encontrar una ecuación tal que los valores de las variables la verifiquen. Esto equivale gráficamente al hecho de

que no es posible encontrar una función tal que su gráfica pase por todos los puntos del diagrama de dispersión asociado a las variables observadas

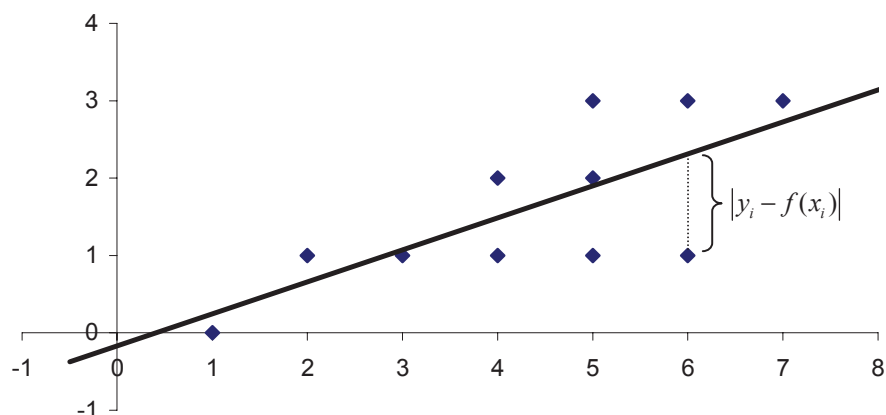
► **EJEMPLO 2.11.**

X	Y
3	1
5	2
4	1
6	3
7	3
5	3
4	2
2	1
1	0
5	1
6	1



Ante la imposibilidad de encontrar una gráfica que pase por todos los puntos de la nube hay que aceptar el razonamiento lógico de que **la función cuya gráfica más se aproxime a los datos observados expresará mejor la relación entre los mismos.**

Hacer **regresión** consiste en ajustar lo mejor posible una función a una serie de valores observados, gráficamente equivale a encontrar la curva que esté lo más próxima que se pueda a los puntos del diagrama de dispersión.



El procedimiento analítico para obtener dicha función se basa en hacer mínima la suma de las distancias (al cuadrado) entre los puntos de la nube, (x_i, y_i) , y la gráfica de la función, $\hat{y}_i = f(x_i)$,

es decir, trata de minimizar $\sum_{i=1}^n (y_i - f(x_i))^2$. Por este motivo se le conoce como **ajuste por mínimos cuadrados**.

Una vez obtenida dicha función $y = f(x)$, para un valor dado $x = x_i$ **estimaríamos el valor de y** mediante $\hat{y}_i = f(x_i)$. Probablemente el verdadero valor de la variable Y asociado a x_i no coincidirá con \hat{y}_i pero el ajuste por mínimos cuadrados nos garantiza que dicha diferencia $(y_i - \hat{y}_i)$ será mínima en la medida de lo posible.

► EJEMPLO 2.12.

La recta del ejemplo 2.11 se ha ajustado por mínimos cuadrado y su expresión es $y = 0,41x - 0,17$, cuando $x_i = 3$, $\hat{y}_i = (0,41 \times 3) - 0,17 = 1,06$. Como podemos observar en la tabla y gráfica del ejemplo 2.11, cuando $x_i = 3$ la variable y toma el valor $y_i = 1 \neq 1,06$, siendo en este caso el error de la estimación muy pequeño. En otros puntos como $x_i = 6$ los errores que se cometen son mayores, $y_i = 1$ o $y_i = 3$, $\hat{y}_i = (0,41 \times 6) - 0,17 = 2,29$. ◀

¿Qué aplicaciones puede tener la regresión en Economía y otras Ciencias Sociales? Las aplicaciones son varias aunque quizás la más interesante sea la de **predecir**, es decir, conociendo el valor de una variable estimar el valor que presentará otra variable relacionada con la primera.

A la variable que se quiere predecir se denomina **dependiente, endógena o explicada**. La variable cuyo conocimiento se utiliza para hacer la predicción se llama **independiente, exógena o explicativa**.

Utilizaremos la expresión **regresión simple** cuando se usa una sola variable independiente y **regresión múltiple** cuando se use más de una variable independiente.

Cuando se hace un ajuste por mínimos cuadrados hay que especificar el tipo de función que se quiere ajustar, recta $y = a + bx$, hipérbola $y = a + \frac{b}{x}$, potencia $y = ax^b$,... Esto dependerá de la forma en que se relacionan las variables, forma que está reflejada en el diagrama de dispersión o nube de puntos. Vamos a considerar el caso más sencillo, además de ser el más utilizado, el de la

línea recta, que se conoce como **regresión lineal simple, recta de regresión, recta de mínimos cuadrados o recta de regresión de mínimos cuadrados**.

Regresión lineal.

Aplicando el ajuste por mínimos cuadrados a funciones de la forma $y = a + bx$ obtenemos que la recta que mejor se ajusta a un conjunto de datos, $(x_i, y_i) \quad i = 1, \dots, n$, tiene por coeficientes

$$b = \frac{S_{xy}}{S_x^2} \quad a = \bar{y} - b\bar{x} = \bar{y} - \frac{S_{xy}}{S_x^2} \bar{x}$$

A los **coeficientes** de la recta, $y = a + bx$, se les conocen como: **a=ordenada en el origen** (punto donde corta al eje Y) y **b=pendiente de la recta**. Ambos tienen un significado propio que se ilustra en el ejemplo 2.13.

Otra forma de escribir una recta es la conocida **expresión punto-pendiente**, $y - y_0 = b(x - x_0)$, donde además de la pendiente se indica un punto, (x_0, y_0) , por el que la recta pasa.

Sustituyendo a y b en la expresión $y = a + bx$ y reordenando convenientemente se obtiene la expresión punto-pendiente

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x})$$

que se puede escribir también en función del coeficiente de correlación lineal como

$$y - \bar{y} = r \frac{S_y}{S_x} (x - \bar{x})$$

Según las anteriores expresiones, la recta de regresión siempre pasa por el punto (\bar{x}, \bar{y}) .

La recta obtenida se denomina **recta de regresión de Y/X**, siendo una de sus aplicaciones predecir la variable Y , conocido el valor de la variable X . Si queremos predecir el valor de X , conocido el valor de Y , no se debe sustituir el valor conocido de Y en la anterior recta y despejar el valor de X . Para ello debemos utilizar la **recta de regresión de X/Y** que en general no coincide con la anterior, siendo su expresión

$$x - \bar{x} = \frac{S_{xy}}{S_y^2} (y - \bar{y}) \quad x - \bar{x} = r \frac{S_x}{S_y} (y - \bar{y})$$

Que habitualmente se simplifica reduciéndola a su forma más simple: $x = a' + b'y$.

► EJEMPLO 2.13.

Se han observado precio y superficie de un conjunto de viviendas. Claramente esas variables están relacionadas. Se ha ajustado por mínimos cuadrados la siguiente recta

$$p = 50000 + 3000m \quad p = \text{precio} \quad m = \text{superficie en m}^2$$

La **ordenada en el origen** nos dice lo que vale la variable dependiente cuando la independiente es cero. Esta cantidad es difícil de interpretar intuitivamente en algunos casos.

La **pendiente de la recta nos indica lo que aumenta la variable dependiente por cada unidad que aumenta la independiente**. En este caso lo que se encarece el valor de la vivienda por cada metro más de superficie.

La ordenada en el origen distinta de cero puede justificar situaciones como la siguiente: una vivienda de 50 m² tiene un precio de 200000€ mientras que una vivienda con el doble de superficie, 100 m², tiene un precio de 350000€

$$200000 = 50000 + (3000 \times 50) \quad 350000 = 50000 + (3000 \times 100) \quad \blacktriangleleft$$

Analizando la forma punto-pendiente de las rectas de regresión observamos que ambas se cortan en el punto (\bar{x}, \bar{y}) y **coincidirán** dichas rectas si además tienen la misma pendiente (tomamos como referencia el eje X para la pendiente).

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}) \quad x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y}) \Rightarrow \frac{S_y^2}{S_{xy}}(x - \bar{x}) = y - \bar{y}$$
$$\frac{S_{xy}}{S_x^2} = \frac{S_y^2}{S_{xy}} \Leftrightarrow \frac{S_{xy} S_{xy}}{S_x^2 S_y^2} = 1 \Leftrightarrow \frac{S_{xy}^2}{S_x^2 S_y^2} = 1 \Leftrightarrow r^2 = 1$$

Es decir, coinciden cuando $r^2 = 1$ (lo que significa, como veremos más adelante, que la relación lineal entre las variables es perfecta y la recta pasa por todos los puntos del diagrama de dispersión).

► EJEMPLO 2.14.

Obtenga para los datos del ejemplo 2.11 los coeficientes de la recta de regresión de Y/X.

Solución:

Utilizaremos la expresión $y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$. Por tanto, todo se reduce a calcular las dos medias, la covarianza y la varianza de X.

En el ejemplo 2.9 se calcularon todas esas cantidades

$$\bar{x} = 4,364 \quad \bar{y} = 1,636 \quad S_{xy} = 1,22 \quad S_x^2 = 2,9555$$

y sustituyendo obtenemos

$$y - 1,636 = \frac{1,22}{2,9555}(x - 4,364) \Leftrightarrow y - 1,636 = 0,4128(x - 4,364) \Leftrightarrow y = 0,4128x - 0,1655$$

(Nota: la pequeña diferencia que se observa con los coeficientes del ejemplo 2.12 es debida a errores de redondeo) ◀

Correlación lineal.

Tan importante es conocer la forma en que se relacionan las variables como su **grado de relación o asociación**. De esto último se ocupa la **correlación**.

El grado de asociación entre las variables indica en qué medida se puede encontrar una expresión que explica una variable en función de otra. Así en el ejemplo de la velocidad y el tiempo, donde el grado de asociación es *completo*, la expresión $vt=500$ explica perfectamente el valor de una variable conociendo la otra. En otros casos la asociación entre las variables no será tan fuerte y la expresión hallada no explicará exactamente a la variable dependiente en función de la independiente.

Todo esto equivale gráficamente al hecho de que la gráfica ajustada pase por los puntos observados (caso de $vt=500$) o bien pase más o menos próxima a ellos. Según lo anterior, el estudio de la correlación equivale al estudio de la **bondad del ajuste** de una curva a un conjunto de puntos.

Para medir la bondad del ajuste nos fijaremos en la distancia entre la gráfica de la función y los puntos de la nube o diagrama de dispersión. Estas distancias se conocen como errores o *residuos*.

Cuando este análisis se aplica sobre la recta de regresión hablamos de **correlación lineal**.

Varianza residual. Coeficiente de determinación.

La media de los residuos o errores al cuadrado que se cometen en un ajuste se conoce como **varianza residual** y es una medida de la bondad del ajuste (observe los residuos en el gráfico del ejemplo 2.11)

$$S_{ry}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

Esta medida tiene el inconveniente de no estar acotada entre unos valores fijos por lo que tenemos el problema de no saber con precisión si está tomando valores significativamente grandes o pequeños. Para superar dicho problema se define a partir de ella el **coeficiente de determinación** como

$$R^2 = 1 - \frac{S_{ry}^2}{S_y^2}, \text{ siendo } 0 \leq R^2 \leq 1$$

En el caso de la recta, este coeficiente coincide con el coeficiente de correlación lineal al cuadrado, $R^2 = r^2$.

Si $R^2 = 0$ tenemos el peor ajuste que se puede hacer por mínimos cuadrados.

$$\text{En el caso de la recta, ésta se reduce a } y - \bar{y} = r \frac{S_y}{S_x} (x - \bar{x}), \quad r=0 \Rightarrow y - \bar{y} = 0$$

$$y = \bar{y}$$

(análogamente la recta de X/Y es $x - \bar{x} = 0 \Leftrightarrow x = \bar{x}$, y las dos rectas de regresión son perpendiculares)

Al no relacionarse (linealmente) la variable independiente con la variable dependiente y por tanto no aportar ninguna información, la variable independiente no aparece en la expresión de la recta de regresión que se reduce a estimar los valores de la variable Y mediante \bar{y} sea quien sea el valor de X .

Si $R^2 = 1$ estamos ante un ajuste por mínimos cuadrados perfecto, la gráfica de la función pasa por todos los puntos de la nube (como en el ejemplo de $vt=500$). En el caso de la recta, como se dijo anteriormente, ambas rectas de regresión coinciden y pasan por todos los puntos.

Para valores intermedios, $0 < R^2 < 1$, según esté R^2 más próximo a un extremo u otro *nos indicará un peor o mejor ajuste*. Además, en el caso de la recta, cuanto más se aproxime a 1 más se cierra el ángulo que forman las rectas de regresión y cuanto más se aproxime a cero mayor será dicho ángulo.

Otra **interpretación** de este coeficiente de determinación, R^2 , es *la proporción de la varianza de Y (comportamiento de la variable Y) que puede atribuirse a su relación con X* .

► EJEMPLO 2.15.

Dé una medida de la bondad del ajuste de la recta del ejemplo 2.14. Obtenga la varianza residual.

Solución:

En el ejemplo 2.9 se calculó el coeficiente de correlación lineal para dichos datos

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{1,22}{1,72 \times 0,98} = 0,7238 \Rightarrow r_{xy}^2 = 0,5239$$

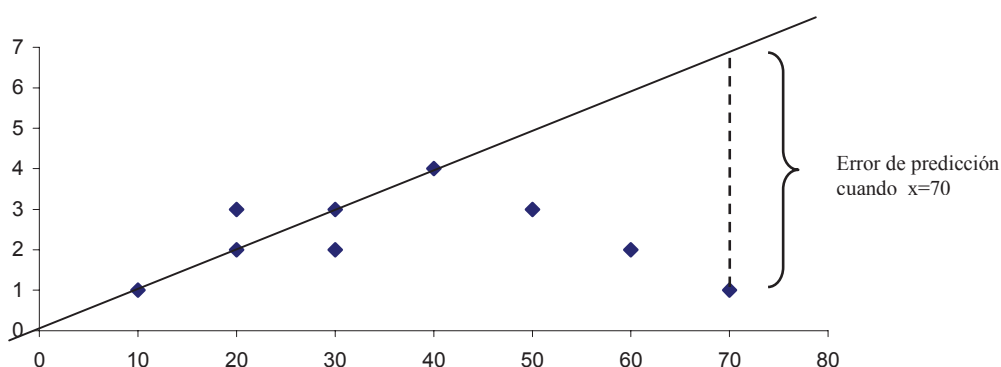
Varianza residual:

$$r^2 = 1 - \frac{S_{ry}^2}{S_y^2} \Rightarrow S_{ry}^2 = (1 - r^2) S_y^2 = (1 - 0,5239) 0,9599 = 0,457$$

Predicciones.

Predecir consiste en determinar a partir del modelo ajustado el valor de la variable dependiente para un valor dado de la variable independiente. Cuando la estimación se hace para un valor de la variable independiente situado fuera de su rango de variación se denomina **extrapolación**, si este valor se encuentra dentro del rango de valores analizados se denomina **interpolación**.

Para la interpolación la fiabilidad de los valores pronosticados será tanto mayor cuanto mejor sea el ajuste, sin embargo cuando hacemos predicciones para valores muy alejados del rango de variación de la variable independiente se corre el riesgo adicional de que el modelo ajustado no sea válido para dichos valores en la medida dada por R^2 . Este hecho es fácil de entender gráficamente



Se ha ajustado la recta para valores de x comprendidos entre 10 y 40, donde la recta refleja la tendencia creciente de los valores de y en función de x . En este rango de valores las predicciones que se hicieran no estarían sujetas a graves errores. Supongamos que desconocemos cómo se comporta el fenómeno para valores de x mayores que 40 y decidiéramos hacer una predicción (extrapolación) para $x=70$, en tal caso cometeríamos un grave error de predicción.

► EJEMPLO 2.16.

Se conoce que el consumo de pan en la dieta de los países desarrollados está relacionado con la renta per cápita. Ajuste una recta que explique dicha relación sobre los siguientes datos, interprete los coeficientes y halle la bondad del ajuste.

PAIS	Kg./habitante	Renta per cápita
Alemania	40	14,8
Bélgica	41	15,1
Francia	57	13,5
Grecia	71	7,3
Holanda	61	12
Italia	42	14
Portugal	69	7,2

Solución:

x_i	y_i	$x_i y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
14,8	40	592	7,92	208,18
15,1	41	619,1	9,70	180,33
13,5	57	769,5	2,29	6,61
7,3	71	518,3	21,96	274,61
12	61	732	0,00	43,18
14	42	588	4,06	154,47
7,2	69	496,8	22,90	212,33
83,9	381	4315,7	68,83	1079,71

En el ejemplo 2.9 se utilizó para el cálculo de la varianza $S_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$

En este ejemplo, sin embargo, vamos a usar la expresión $S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

(Compare ambos métodos y decida cual es más fácil de utilizar)

$$\begin{aligned}\bar{x} &= \frac{83,9}{7} = 11,9857 & \bar{y} &= \frac{381}{7} = 54,4286 \\ S_x^2 &= \frac{68,83}{7} = 9,833 & S_y^2 &= \frac{1079,71}{7} = 154,244 \\ S_{xy} &= \frac{4315,7}{7} - \bar{x}\bar{y} = -35,836\end{aligned}$$

Sustituyendo en la ecuación de la recta de Y/X, $y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x})$, obtenemos

$$\begin{aligned}y - 54,4286 &= \frac{-35,836}{9,833} (x - 11,9857) \Leftrightarrow y - 54,4286 = -3,64 (x - 11,9857) \\ y &= -3,64x + 98,1\end{aligned}$$

La pendiente negativa (-3.64) indica que a mayor renta per cápita menor es el consumo de pan en la dieta.

Para hallar la bondad del ajuste calculamos el coeficiente de correlación lineal al cuadrado

$$r_{xy}^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{(-35,836)^2}{9,833 \times 154,244} = 0,85$$

► EJEMPLO 2.17.

Una fábrica de cerveza ha tomado al azar 8 semanas del año, observando la temperatura media en cada una de ellas y los miles de litros de cerveza demandados

Temperatura	Cerveza
10	26
28	82
30	98
35	103
20	60
12	35
27	68
25	71

¿Puede la fábrica planificar la cantidad de producción en función de la temperatura esperada? ¿De qué forma?

Solución:

En primer lugar estudiaremos el grado de asociación que hay entre las variables, si éste es importante hallaremos la expresión según la cual se relacionan.

x_i	y_i	$x_i y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
10	26	260	178,89	1753,52
28	82	2296	21,39	199,52
30	98	2940	43,89	907,52
35	103	3605	135,14	1233,77
20	60	1200	11,39	62,02
12	35	420	129,39	1080,77
27	68	1836	13,14	0,02
25	71	1775	2,64	9,77
187	543	14332	535,88	5246,88

$$\bar{x} = \frac{187}{8} = 23,375 \quad \bar{y} = \frac{543}{8} = 67,875$$

$$S_x^2 = \frac{535,88}{8} = 66,98 \quad S_y^2 = \frac{5246,88}{8} = 655,86$$

$$S_{xy} = \frac{14332}{8} - \bar{x}\bar{y} = 204,92$$

$$r_{xy}^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{(204,92)^2}{66,98 \times 655,86} = 0,956$$

Podemos decir que existe una fuerte relación lineal entre las variables, $r_{xy}^2 \cong 1$, de modo que la fábrica podrá planificar con mucha fiabilidad la cantidad de producción en función de la

temperatura utilizando la recta de regresión de Y/X: $y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x})$

$$y - 67,875 = \frac{204,92}{66,98} (x - 23,375) \Leftrightarrow y - 67,875 = 3,06 (x - 23,375)$$

$$y = 3,06x - 3,65$$

Por ejemplo para una temperatura media de 33^0 se esperará una demanda de

$$y = (3,06 \times 33) - 3,65 = 97,33 \quad (97.330 \text{ litros})$$



2.5 Ejercicios resueltos.

- Una fábrica de cerveza ha tomado al azar 10 semanas del año, observando la temperatura media en cada una de ellas y los miles de litros de cerveza demandados. Los datos recogidos se muestran en la siguiente tabla:

<i>Temperatura</i>	<i>Cerveza</i>	<i>Temperatura</i> ²	<i>Cerveza</i> ²	<i>Temperatura</i> × <i>Cerveza</i>
12	33	144	1089	396
25	66	625	4356	1650
10	26	100	676	260
28	82	784	6724	2296
30	98	900	9604	2940
35	103	1225	10609	3605
20	60	400	3600	1200
12	35	144	1225	420
27	68	729	4624	1836
25	71	625	5041	1775
TOTAL:	224	642	5676	47548

Utilizando los totales de la tabla anterior, responda a las siguientes cuestiones:

- Obtenga la recta de regresión que permita explicar la demanda de cerveza en función de la temperatura.
- ¿En qué proporción el consumo de cerveza puede atribuirse a la temperatura media semanal?
- ¿Qué demanda se espera para una semana con una temperatura media de 30 grados?

Solución: $X = \text{temperatura}$ $Y = \text{demanda de cerveza}$

$$n = 10 \quad \bar{x} = \frac{224}{10} = 22,4 \quad \bar{y} = \frac{642}{10} = 64,2$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{5676}{10} - 22,4^2 = 65,84 \Rightarrow S_x = \sqrt{65,84} = 8,1142$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{47548}{10} - 64,2^2 = 633,16 \Rightarrow S_y = \sqrt{633,16} = 25,1627$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{16378}{10} - (22,4 \times 64,2) = 199,72$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{199,72}{65,84} = 3,0334$$

$$y - \bar{y} = b(x - \bar{x}) \Rightarrow y - 64,2 = 3,0334(x - 22,4) \Rightarrow y = 3,0334x - 3,748$$

$$a) \quad r = \frac{S_{xy}}{S_x S_y} = 0,9782 \Rightarrow r^2 = 0,9568 \Rightarrow 95,68\%$$

$$b) \quad \hat{y}_{x=30} = (3,0334 \times 30) - 3,748 = 87,25 \text{ miles de litros de cerveza.}$$

2. Una empresa tiene a sus trabajadores clasificados según el salario (en miles de euros anuales) y la antigüedad (en años) en el puesto de trabajo.

	<i>Antigüedad</i>			
<i>Salario</i>	0 – 10	10 – 20	20 – 30	30 – 50
10 – 20	4	2	1	0
20 – 25	1	0	3	0
25 – 30	5	2	1	10
30 – 40	0	1	2	5
40 – 50	0	1	4	6

- Calcule el índice de Gini de concentración de los salarios de todos los empleados de la empresa.
- ¿Qué porcentaje de empleados hay en la empresa con un salario superior a 30.000 euros al año?
- ¿Qué sueldo anual corresponde al 30% de los empleados mejor remunerados?
- Calcule el salario medio de los empleados que no superan los 20 años de antigüedad en el puesto de trabajo.
- ¿Respecto de cuál de las dos variables son los trabajadores más homogéneos?
- Calcule el coeficiente de correlación lineal.
- Estime el sueldo para un empleado con una antigüedad de 12 años en el puesto de trabajo.

Solución:

Notaremos X =salario e Y =antigüedad. Aunque la tabla del enunciado es una tabla de contingencia de una variable estadística bidimensional, en los apartados a, b, c, d y e nos preguntan sólo sobre variables unidimensionales representadas en la tabla. Los apartados a, b y c se refieren a los salarios de todos los empleados, es decir a la distribución marginal de los salarios. El apartado e compara las dos distribuciones marginales, salarios y antigüedad. Mientras que el apartado d se refiere a una distribución condicionada de los salarios.

a)

L_{i-1}	L_i	x_i	$n_{i\bullet}$	$x_i n_{i\bullet}$	$x_i^2 n_{i\bullet}$	N_i	u_i	p_i	q_i
10	20	15	7	105	1575	7	105	14,58	7,17
20	25	22,5	4	90	2025	11	195	22,92	13,31
25	30	27,5	18	495	13612,5	29	690	60,42	47,10
30	40	35	8	280	9800	37	970	77,08	66,21
40	50	45	11	495	22275	48	1465	100,00	100,00
			48	1465	49287,5			275,00	233,79

$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{133,79}{175} = 0,2355$$

- b) Según la anterior tabla hay un 60,42% de empleados con un salario inferior a 30000 euros/año luego habrá $100 - 60,42 = 39,58\%$ de empleados con un salario superior a 30000 euros/año.
- c) El sueldo que es superado por el 30% de los empleados es el mismo que no es superado por el 70% restante de empleados. Buscamos en la anterior tabla $p_i = 0,70$, al no aparecer interpolamos entre los valores más próximos

30	60,42
x	70
40	77,08

$$\frac{40 - 30}{77,08 - 60,42} = \frac{x - 30}{70 - 60,42} \Rightarrow x = 35,75$$

El 30% de los empleados mejor remunerados tienen un sueldo por encima de 35750 euros/año.

- d) Para la distribución condicionada sobre la que nos preguntan su media, sumamos las dos primeras columnas de frecuencias de la tabla.

L_{i-1}	L_i	x_i	n_i	$x_i n_i$
10	20	15	6	90
20	25	22,5	1	22,5
25	30	27,5	7	192,5
30	40	35	1	35
40	50	45	1	45
			16	385

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{385}{16} = 24,0625$$

- e) Para comparar la homogeneidad de los trabajadores en las dos variables utilizamos una medida de dispersión relativa, el coeficiente de variación.

Con los datos de la tabla del apartado a) calculamos el coeficiente de variación de los salarios:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i\bullet} = \frac{1465}{48} = 30,5208$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_{i\cdot} = \frac{1}{n} \sum_{i=1}^k x_i^2 n_{i\cdot} - \bar{x}^2 = \frac{49287,5}{48} - 30,5208^2 = 95,3 \quad \Rightarrow \quad S_x = 9,762$$

$$CV_x = \frac{S_x}{\bar{x}} = \frac{9,762}{30,5208} = 0,32$$

A partir de la distribución marginal de la antigüedad obtenemos su coeficiente de variación:

L_{i-1}	L_i	y_j	$n_{\cdot j}$	$y_j n_{\cdot j}$	$y_j^2 n_{\cdot j}$
0	10	5	10	50	250
10	20	15	6	90	1350
20	30	25	11	275	6875
30	50	40	21	840	33600
			48	1255	42075

$$\bar{y} = \frac{1}{n} \sum_{j=1}^p y_j n_{\cdot j} = \frac{1255}{48} = 26,1458$$

$$S_y^2 = \frac{1}{n} \sum_{j=1}^p (y_j - \bar{y})^2 n_{\cdot j} = \frac{1}{n} \sum_{j=1}^p y_j^2 n_{\cdot j} - \bar{y}^2 = \frac{42075}{48} - 26,1458^2 = 192,96 \quad \Rightarrow \quad S_y = 13,891$$

$$CV_y = \frac{S_y}{\bar{y}} = \frac{13,891}{26,1458} = 0,53$$

Luego los trabajadores son más homogéneos respecto del salario que de la antigüedad.

- f) En el apartado anterior se han calculado las medias y desviaciones típicas marginales, sólo nos falta calcular la covarianza para lo cual construimos la siguiente tabla (representamos cada intervalo por su punto medio o marca de clase):

$x_i y_j n_{ij}$	5	15	25	40
15	300	450	375	0
22,5	112,5	0	1687,5	0
27,5	687,5	825	687,5	11000
35	0	525	1750	7000
45	0	675	4500	10800

$$\sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} = 41375$$

$$S_{xy} = \left(\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} \right) - \bar{x} \bar{y} = \frac{41375}{48} - (30,5208 \times 26,1458) = 63,99$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{63,99}{9,762 \times 13,891} = 0,472$$

- g) En la recta de regresión de X/Y , $x - \bar{x} = \frac{S_{xy}}{S_y^2} (y - \bar{y})$, sustituiremos $y=12$.

$$x - \bar{x} = \frac{S_{xy}}{S_y^2} (y - \bar{y}) \quad \Rightarrow \quad x - 30,5208 = \frac{63,99}{192,96} (y - 26,1458) \quad \Rightarrow \quad x = 21,85 + 0,3316y$$

$$\hat{x}_{y=12} = 21,85 + (0,3316 \times 12) = 25,83 \quad \Rightarrow \quad 25830 \text{ euros anuales}$$

3. Una oficina donde se alquilan apartamentos ha observado que durante el mes de julio el número de apartamentos alquilados varía según el precio de los mismos. Del año anterior se tiene la siguiente información:

Precios por día (€)	Número de apartamentos alquilados
15	934
40	512
60	364
85	180
150	80

Calcule:

- Precio medio por día de los apartamentos alquilados.
- ¿En qué medida podemos considerar que el número de apartamentos alquilados depende de los precios?

Si se mantienen el comportamiento de los clientes y los precios:

- ¿Cuántos apartamentos se alquilarían con un precio de 90 € por día?
- ¿Cuál será el precio máximo al que se podrán ofrecer apartamentos en alquiler?

Solución:

a)

x_i	n_i	$x_i n_i$
15	934	14010
40	512	20480
60	364	21840
85	180	15300
150	80	12000
	2070	83630

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{83630}{2070} = 40,4 \text{ euros por día}$$

En los siguientes apartados el análisis de los datos es completamente diferente al del apartado a. Trataremos de estudiar la relación que existe entre los precios y la frecuencia con la que se alquilan apartamentos a dichos precios.

X =precio al que se ofertan los apartamentos.

Y =frecuencia de ocupación de los apartamentos.

- Calcularemos el coeficiente de correlación lineal al cuadrado, que nos indica en qué medida una variable depende linealmente de la otra.

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
15	934	14010	225	872356
40	512	20480	1600	262144
60	364	21840	3600	132496
85	180	15300	7225	32400
150	80	12000	22500	6400
335	2070	83630	35150	1305796

$$n = 5 \quad \bar{x} = \frac{335}{5} = 67 \quad \bar{y} = \frac{2070}{5} = 414$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{83630}{5} - (67 \times 414) = -11012 \quad \Rightarrow \quad S_{xy}^2 = 121264144$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{35150}{5} - 67^2 = 2541$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{1305796}{5} - 414^2 = 89763,2$$

$$r_{xy}^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{121264144}{2541 \times 89763,2} = 0,5317$$

c) Ajustamos la recta de regresión de Y/X : $y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x})$

$$y - 414 = \frac{-11012}{2541} (x - 67) \Leftrightarrow y - 414 = -4,3337 (x - 67) \Leftrightarrow y = -4,3337x + 704,3579$$

Sustituyendo $x=90$ en la recta de regresión se obtiene la estimación del número de apartamentos que se alquilarían a ese precio

$$\hat{y}_{x=90} = (-4,3337 \times 90) + 704,3579 = 314,3249$$

d) El precio máximo al que se pueden ofrecer apartamentos es aquel para el que la demanda sería nula. ¿Cuál será el valor de X para el que estimamos que Y será nulo?

$$0 = -4,3337x + 704,3579 \quad \Rightarrow \quad x = \frac{-704,3579}{-4,3337} = 162,53 \text{ euros por día}$$

4. Se consideran 50 establecimientos de alimentación, atendiendo a dos factores: tiempo que llevan funcionando (X , en años) y beneficio anual (Y , en cientos de miles de euros).

X/Y	0-1	1-3	3-4	$n_{i\cdot}$
0-5	0	2	5	7
5-10	2	4	4	10
10-15	8	4	6	18
15-20	10	3	2	15
$n_{\cdot j}$	20	13	17	$n=50$

a) De una estimación del beneficio anual para un establecimiento con 12 años de antigüedad.

b) Calcule la varianza residual.

Solución:

a) Ajustamos la recta de regresión de Y/X : $y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$ y sustituimos en ella $x=12$.

Representamos cada intervalo por su punto medio o marca de clase y efectuamos en la siguiente tabla los cálculos necesarios para obtener las medias, varianzas y covarianza.

$x_i y_j n_{ij}$	0,5	2	3,5	$n_{i\bullet}$	$x_i n_{i\bullet}$	$x_i^2 n_{i\bullet}$
2,5	0	10	43,75	7	17,5	43,75
7,5	7,5	60	105	10	75	562,5
12,5	50	100	262,5	18	225	2812,5
17,5	87,5	105	122,5	15	262,5	4593,75
$n_{\bullet j}$	20	13	17	$n=50$	580	8012,5
$y_j n_{\bullet j}$	10	26	59,5	95,5	$\sum_i \sum_j x_i y_j n_{ij} = 953,75$	
$y_j^2 n_{\bullet j}$	5	52	208,25	265,25		

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i\bullet} = \frac{580}{50} = 11,6 \quad \bar{y} = \frac{1}{n} \sum_{j=1}^p y_j n_{\bullet j} = \frac{95,5}{50} = 1,91$$

$$S_{xy} = \left(\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} \right) - \bar{x} \bar{y} = \frac{953,75}{50} - (11,6 \times 1,91) = -3,081$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_{i\bullet} = \frac{1}{n} \sum_{i=1}^k x_i^2 n_{i\bullet} - \bar{x}^2 = \frac{8012,5}{50} - 11,6^2 = 25,69$$

$$S_y^2 = \frac{1}{n} \sum_{j=1}^p (y_j - \bar{y})^2 n_{\bullet j} = \frac{1}{n} \sum_{j=1}^p y_j^2 n_{\bullet j} - \bar{y}^2 = \frac{265,25}{50} - 1,91^2 = 1,6569$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}) \Rightarrow y - 1,91 = \frac{-3,081}{25,69}(x - 11,6) \Rightarrow y = 3,3012 - 0,1199x$$

$$\hat{y}_{x=12} = 3,3012 - (0,1199 \times 12) = 1,862 \text{ cientos de miles de euros}$$

$$b) \quad r_{xy}^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{(-3,081)^2}{25,69 \times 1,6569} = 0,223$$

$$r^2 = 1 - \frac{S_{ry}^2}{S_y^2} \Rightarrow S_{ry}^2 = (1 - r^2) S_y^2 = (1 - 0,223) \times 1,6569 = 1,2874$$

5. El valor promedio, en euros, de las acciones en un grupo de inmobiliarias y en un grupo de bancos en los años 2001-2010 aparecen en la siguiente tabla:

AÑO	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
X	35	40	42	43	40	53	54	49	40	55
Y	102	101	97	98	98	100	97	91	95	95

X =promedio del precio de las acciones de las inmobiliarias.

Y =promedio del precio de las acciones de los bancos.

- Calcule el coeficiente de correlación lineal.
- ¿Qué cabe esperar que le ocurrirá al valor de las acciones de las inmobiliarias si se sabe que las acciones de los bancos aumentarán su valor?

Solución:

a)

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
35	102	1225	10404	3570
40	101	1600	10201	4040
42	97	1764	9409	4074
43	98	1849	9604	4214
40	98	1600	9604	3920
53	100	2809	10000	5300
54	97	2916	9409	5238
49	91	2401	8281	4459
40	95	1600	9025	3800
55	95	3025	9025	5225
451	974	20789	94962	43840

$$n=10 \quad \bar{x} = \frac{451}{10} = 45,1 \quad \bar{y} = \frac{974}{10} = 97,4 \quad S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{43840}{10} - (45,1 \times 97,4) = -8,74$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{20789}{10} - 45,1^2 = 44,89 \Rightarrow S_x = 6,7$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{94962}{10} - 97,4^2 = 9,44 \Rightarrow S_y = 3,0725$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-8,74}{6,7 \times 3,0725} = -0,4246$$

- b) El coeficiente de la variable Y en la recta de regresión de X/Y :

$$x - \bar{x} = \frac{S_{xy}}{S_y^2} (y - \bar{y})$$

es negativo como la covarianza, luego dicha expresión nos indica que al aumentar Y disminuirá el valor esperado para X . La anterior afirmación es fiable en la medida en que la recta de regresión exprese adecuadamente la relación entre Y y X . El coeficiente de correlación lineal al cuadrado es quien determina si dicha recta se ajusta bien a los valores de las variables. En este caso dicho coeficiente al cuadrado es 0,18 que no es un valor próximo a 1 y por tanto la afirmación hecha no ofrece grandes garantías.

6. En un estudio sobre consumo de tabaco se ha preguntado a unos jóvenes sobre su edad, X , y el número de cigarrillos que fuman al día, Y , obteniendo los siguientes resultados:

X/Y	0-4	4-8	8-14	$n_{i\cdot}$
15-20	3	5	10	18
20-24	6	7	5	18
24-28	7	6	1	14
$n_{\cdot j}$	16	18	16	$n=50$

- Calcule la edad más frecuente de aquellos jóvenes que fuman más de 4 cigarrillos al día.
- Calcule la edad media y varianza de los jóvenes encuestados.
- ¿Qué media es más representativa, la edad media de los encuestados o la edad media de los que fuman más de 4 cigarrillos diarios?
- ¿Cuántos cigarrillos fuman el 20% de jóvenes más fumadores?

Solución:

Aunque en el enunciado tenemos una distribución de frecuencias bidimensional, todas las cuestiones se plantean sobre variables estadísticas unidimensionales en la anterior tabla.

Sobre la **distribución condicionada de frecuencias de los 34 jóvenes que fuman más de 4 cigarrillos** calcularemos la moda (apartado a), media y desviación típica (apartado c)

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	h_i
15	20	17,5	15	262,5	4593,75	3
20	24	22	12	264	5808	3
24	28	26	7	182	4732	1,75
			34	708,5	15133,75	

- La edad más frecuente o moda de los jóvenes que fuman más de 4 cigarrillos se encuentra en el intervalo de mayor altura. En este caso la mayor altura se alcanza en dos intervalos, por tanto hay dos modas. Vamos a calcularlas según los tres procedimientos estudiados:

La altura anterior al primer intervalo se toma igual a cero (análogamente se hace con la altura posterior al último intervalo, aunque en este ejemplo no se necesita).

$$Mo(I) = \frac{L_{i-1} + L_i}{2} = \frac{15 + 20}{2} = 17,5$$

$$Mo(II) = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i = 15 + \frac{3}{0 + 3} 5 = 20$$

$$Mo(III) = L_{i-1} + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} a_i = 15 + \frac{3 - 0}{(3 - 0) + (3 - 3)} 5 = 20$$

$$Mo(I) = \frac{L_{i-1} + L_i}{2} = \frac{20 + 24}{2} = 22$$

$$Mo(II) = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i = 20 + \frac{1,75}{3+1,75} 4 = 21,47$$

$$Mo(III) = L_{i-1} + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} a_i = 20 + \frac{3-3}{(3-3) + (3-1,75)} 4 = 20$$

b) Sobre la **distribución marginal de X** obtenemos su media y varianza.

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
15	20	17,5	18	315	5512,5
20	24	22	18	396	8712
24	28	26	14	364	9464
			50	1075	23688,5

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i n_i = \frac{1075}{50} = 21,5$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 n_i - \bar{x}^2 = \frac{23688,5}{50} - 21,5^2 = 11,52$$

c) Para estudiar lo representativa que es una medida de posición central se utiliza una medida de dispersión calculada sobre dicha medida, en este caso podríamos utilizar la varianza. Si se quiere comparar la representatividad de la media en dos variables utilizaremos una medida de dispersión relativa calculada sobre la media, concretamente el coeficiente de variación.

A partir del apartado anterior es inmediato obtener el coeficiente de variación para todos los encuestados:

$$S = \sqrt{11,52} = 3,3941$$

$$CV = \frac{S}{\bar{x}} = \frac{3,3941}{21,5} = 0,1579$$

Y de la tabla utilizada en el apartado a se obtiene el coeficiente de variación para los que fuman más de 4 cigarrillos al día:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i n_i = \frac{708,5}{34} = 20,8382$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 n_i - \bar{x}^2 = \frac{15133,75}{34} - 20,8382^2 = 10,88 \quad S = \sqrt{10,88} = 3,30$$

$$CV = \frac{S}{\bar{x}} = \frac{3,30}{20,8382} = 0,158$$

Los coeficientes de variación en ambos casos casi coinciden, luego son igualmente representativas las medias de las dos distribuciones.

d)

L_{i-1}	L_i	n_i	N_i	p_i
0	4	16	16	32
4	8	18	34	68
8	14	16	50	100

50

El número de cigarrillos que superan el 20% de jóvenes que más fuman es el mismo número de cigarrillos que no alcanzan el 80% restante de fumadores. Interpolaremos el valor 80 entre los valores más próximos de la columna p_i .

8	68
x	80
14	100

$$\frac{14-8}{100-68} = \frac{x-8}{80-68} \Rightarrow x = 10,25 \text{ cigarrillos diarios}$$

7. Una empresa de importación dispone de una cuota de mercado del 4% del sector. En los seis últimos años el volumen de importación y la producción de los sectores que han absorbido dichas importaciones son (en millones de euros):

Importación	Producción
220	1050
330	1200
450	1250
500	1300
650	1400
670	1540

- Cuál será el volumen de importación de esa empresa en un año en que la producción industrial estimada es de 2000 millones de euros (suponiendo que la relación industrial se mantenga en dicho año)
- ¿Qué fiabilidad tiene dicha estimación?
- Calcule la varianza residual.

Solución: $X = \text{Importaciones}$ $Y = \text{Producción}$

- Ajustaremos la recta de regresión de X/Y : $x - \bar{x} = \frac{S_{xy}}{S_y^2} (y - \bar{y})$

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
220	1050	48400	1102500	231000
330	1200	108900	1440000	396000
450	1250	202500	1562500	562500
500	1300	250000	1690000	650000
650	1400	422500	1960000	910000
670	1540	448900	2371600	1031800
2820	7740	1481200	10126600	3781300

$$n = 6 \quad \bar{x} = \frac{2820}{6} = 470 \quad \bar{y} = \frac{7740}{6} = 1290$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{10126600}{6} - 1290^2 = 23666,67$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{3781300}{6} - (470 \times 1290) = 23916,67$$

$$x - \bar{x} = \frac{S_{xy}}{S_y^2} (y - \bar{y}) \Rightarrow x - 470 = \frac{23916,67}{23666,67} (y - 1290) \Rightarrow x = -833,62 + 1,01056y$$

$$\hat{x}_{y=2000} = -833,62 + (1,01056 \times 2000) = 1187,5$$

- b) La fiabilidad de la anterior estimación depende de la bondad del ajuste de la recta utilizada. Dicha bondad la mide el coeficiente de correlación lineal al cuadrado.

$$S_y^2 = 23666,67 \Rightarrow S_y = \sqrt{23666,67} = 153,84$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{1481200}{6} - 470^2 = 25966,67 \Rightarrow S_x = 161,14$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{23916,67}{161,14 \times 153,84} = 0,965 \Rightarrow r_{xy}^2 = 0,93$$

- c) La varianza residual para la recta de regresión de X/Y es:

$$S_{rx}^2 = (1 - r^2) S_x^2 = (1 - 0,93) \times 25966,67 = 1817,67$$

8. En ocho empresas los beneficios y los gastos en investigación el año pasado fueron los siguientes (en miles de euros):

Beneficios	550	600	400	500	300	400	500	500
Gastos	400	400	350	500	400	450	350	550

- Obtenga el coeficiente de correlación lineal y comente el resultado.
- Estime el beneficio que obtendría una empresa si dedicara al gasto en investigación 5000000€.

Solución:a) $X = \text{Beneficios}$ $Y = \text{Gastos}$

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
550	400	302500	160000	220000
600	400	360000	160000	240000
400	350	160000	122500	140000
500	500	250000	250000	250000
300	400	90000	160000	120000
400	450	160000	202500	180000
500	350	250000	122500	175000
500	550	250000	302500	275000
3750	3400	1822500	1480000	1600000

$$n = 8 \quad \bar{x} = \frac{3750}{8} = 468,75 \quad \bar{y} = \frac{3400}{8} = 425$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{1822500}{8} - 468,75^2 = 8085,9375 \Rightarrow S_x = 89,9218$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{1480000}{8} - 425^2 = 4375 \Rightarrow S_y = 66,1438$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{1600000}{8} - (468,75 \times 425) = 781,25$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{781,25}{89,9218 \times 66,1438} = 0,13135 \Rightarrow r_{xy}^2 = 0,01725$$

Como puede observarse, el coeficiente de correlación lineal toma un valor pequeño (próximo a cero) por lo que hay una relación muy débil entre ambas variables. La estimación que hagamos en el próximo apartado tendrá poca fiabilidad.

b) Para estimar $X = \text{Beneficios}$ en función de $Y = \text{Gastos}$ ajustaremos la recta de regresión de X/Y :

$$x - \bar{x} = \frac{S_{xy}}{S_y^2} (y - \bar{y}) \Rightarrow x - 468,75 = \frac{781,25}{4375} (y - 425) \Rightarrow x = 392,86 + 0,17857y$$

$$\hat{x}_{y=500} = 392,86 + (0,17857 \times 500) = 482,14$$

9. En la siguiente tabla se clasifican los trabajadores de una empresa según su edad y antigüedad en el puesto de trabajo (ambas variables en años)

Antigüedad	Edad			
	18-30	30-40	40-50	50-65
0-3	4	2	2	0
3-6	1	0	3	1
6-10	5	2	1	10
10-15	0	1	2	5
Más de 15	0	1	4	6

¿Qué distribución es más homogénea, la de los trabajadores con menos de 6 años de antigüedad o la de los que superan esa cifra?

Solución:

Tenemos que comparar la dispersión en dos distribuciones condicionadas, la de trabajadores con menos de 6 años de antigüedad y la de trabajadores con más de 6 años de antigüedad. Para ello usamos el coeficiente de variación.

Trabajadores con menos de 6 años de antigüedad

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
18	30	24	5	120	2880
30	40	35	2	70	2450
40	50	45	5	225	10125
50	65	57,5	1	57,5	3306,25
			13	472,5	18761,25

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i n_i = \frac{472,5}{13} = 36,346$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 n_i - \bar{x}^2 = \frac{18761,25}{13} - 36,346^2 = 122,1 \quad S = \sqrt{122,1} = 11,05$$

$$CV = \frac{S}{\bar{x}} = \frac{11,05}{36,346} = 0,304$$

Trabajadores con más de 6 años de antigüedad

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
18	30	24	5	120	2880
30	40	35	4	140	4900
40	50	45	7	315	14175
50	65	57,5	21	1207,5	69431,25
			37	1782,5	91386,25

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i n_i = \frac{1782,5}{37} = 48,1757$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 n_i - \bar{x}^2 = \frac{91386,25}{37} - 48,1757^2 = 149 \quad S = \sqrt{149} = 12,2066$$

$$CV = \frac{S}{\bar{x}} = \frac{12,2066}{48,1757} = 0,2534$$

$0,2534 < 0,304$, luego es más homogénea la distribución de los trabajadores con más de seis años de antigüedad.

10. La distribución de frecuencias absolutas de la variable bidimensional (*Número de hijos, Gasto anual en miles de euros*) para 20 familias objeto de un estudio es:

Nº de hijos	Gasto anual		
	0-10	10-20	20-30
0	6	0	0
1	5	2	1
3	1	3	0
4	0	1	1

Si el comportamiento del resto de las familias en la ciudad fuese similar al descrito en la tabla, ¿cuál sería el gasto anual para una familia con dos hijos?

Solución: $X = \text{Número de hijos}$ $Y = \text{Gasto anual}$

Ajustamos la recta de regresión de Y/X : $y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$ y sustituimos en ella $x=2$.

Representamos cada intervalo por su punto medio o marca de clase y efectuamos en la siguiente tabla los cálculos necesarios para obtener las medias, varianza de X y covarianza.

$x_i y_j n_{ij}$	5	15	25	$n_{i\bullet}$	$x_i n_{i\bullet}$	$x_i^2 n_{i\bullet}$
0	0	0	0	6	0	0
1	25	30	25	8	8	8
3	15	135	0	4	12	36
4	0	60	100	2	8	32
$n_{\bullet j}$	12	6	2	$n=20$	28	76
$y_j n_{\bullet j}$	60	90	50	200	$\sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} = 390$	

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i\bullet} = \frac{28}{20} = 1,4 \quad \bar{y} = \frac{1}{n} \sum_{j=1}^p y_j n_{\bullet j} = \frac{200}{20} = 10$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_{i\bullet} = \frac{1}{n} \sum_{i=1}^k x_i^2 n_{i\bullet} - \bar{x}^2 = \frac{76}{20} - 1,4^2 = 1,84$$

$$S_{xy} = \left(\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p x_i y_j n_{ij} \right) - \bar{x} \bar{y} = \frac{390}{20} - (1,4 \times 10) = 5,5$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}) \Rightarrow y - 10 = \frac{5,5}{1,84}(x - 1,4) \Rightarrow y = 5,815 + 2,989x$$

$$\hat{y}_{x=2} = 5,815 + (2,989 \times 2) = 11,793 \text{ miles de euros} \Rightarrow 11793 \text{ euros}$$

11. Los beneficios anuales (en miles de euros) y el número de empleados de las 50 empresas de un determinado polígono industrial son:

Beneficios	Nº de empleados		
	0-10	10-20	20-40
0-20	8	2	0
20-40	10	4	2
40-60	4	6	4
60-100	0	0	10

- a) Calcule el índice de Gini de la concentración de los beneficios.
b) ¿Cuál es el beneficio mínimo del 40% de las empresas más rentables?
c) ¿Cuál es el porcentaje de empresas más rentables que obtienen el 25% del total de los beneficios anuales de dicho polígono?

Solución: $X = \text{Beneficios}$ $Y = \text{Número de empleados}$

En los tres apartados nos piden que calculemos medidas sobre la **distribución marginal de X**.

a)

L_{i-1}	L_i	x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
0	20	10	10	100	10	100	20	4,81
20	40	30	16	480	26	580	52	27,88
40	60	50	14	700	40	1280	80	61,54
60	100	80	10	800	50	2080	100	100,00
			50	2080			252	194,23

$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{94,23}{152} = 0,38$$

- b) El beneficio mínimo del 40% de las empresas más rentables es igual al beneficio máximo del 60% restante de empresas menos rentables. Interpolamos 60 en la columna de los p_i

40	52
x	60
60	80

$$\frac{60-40}{80-52} = \frac{x-40}{60-52} \Rightarrow x = 45,714 \text{ miles de euros} \Rightarrow 45714 \text{ euros}$$

- c) Calculamos las empresas menos rentables que obtienen el 75% restante del total de los beneficios anuales de dicho polígono. Interpolamos 75 en la columna de los q_i

80	61,54
x	75
100	100

$$\frac{100-80}{100-61,54} = \frac{x-80}{75-61,54} \Rightarrow x = 87\% \Rightarrow 100-87 = 13\%$$

Si el 75% de las empresas menos rentables obtienen el 87% del total de los beneficios, el 25% restante de las empresas más rentables obtienen el 13% restante de los beneficios.

12. En la siguiente tabla se han recogido los datos sobre ingresos y ahorro de una muestra de 5 familias

Ingresos	1630	2450	3210	3320	4250
Ahorro	250	650	700	720	980

¿Cuál sería el ahorro de una familia con 3000 € de ingresos? ¿Es fiable la anterior estimación?

Solución: $X = \text{Ingresos}$ $Y = \text{Ahorro}$

Ajustaremos la recta de regresión de Y/X y sustituiremos $x=3000$ para estimar el ahorro en función de los ingresos. Calcularemos el coeficiente de determinación asociado a la recta (coeficiente de correlación lineal al cuadrado) para valorar la fiabilidad de la estimación.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1630	250	2656900	62500	407500
2450	650	6002500	422500	1592500
3210	700	10304100	490000	2247000
3320	720	11022400	518400	2390400
4250	980	18062500	960400	4165000
14860	3300	48048400	2453800	10802400

$$n = 5 \quad \bar{x} = \frac{14860}{5} = 2972 \quad \bar{y} = \frac{3300}{5} = 660$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{48048400}{5} - 2972^2 = 776896$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{2453800}{5} - 660^2 = 55160$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{10802400}{5} - (2972 \times 660) = 198960 \Rightarrow S_{xy}^2 = 39585081600$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Rightarrow y - 660 = \frac{198960}{776896} (x - 2972) \Rightarrow y = -101,1 + 0,2561x$$

$$\hat{y}_{x=3000} = -101,1 + (0,2561 \times 3000) = 667,2 \text{ €}$$

$$r_{xy}^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{39585081600}{776896 \times 55160} = 0,9237$$

Dado que el valor del coeficiente de determinación es próximo a 1 y que la predicción se está haciendo para un valor de X dentro del rango de valores utilizados en el ajuste de la recta, podemos decir que la anterior predicción es bastante fiable.

3. NÚMEROS ÍNDICES.

3.1 Tasas de variación.

Las tasas de variación son medidas que utilizamos para cuantificar la variación temporal (o espacial) de una variable. A lo largo del tema consideraremos siempre una variación temporal pero todo es igualmente aplicable al caso espacial.

Variación absoluta.

Sea una serie de observaciones de una variable X ordenadas en el tiempo, obtenidas en períodos temporales de la misma duración:

$$x_0, x_1, x_2, \dots, x_{t-1}, x_t, x_{t+1}, \dots, x_n$$

(supondremos para simplificar la exposición y por ser uno de los casos más frecuentes, que disponemos de observaciones mensuales. Lo mismo sería aplicable a otros períodos de tiempo).

Se define la **variación absoluta** en el mes t respecto del mes anterior como la diferencia:

$$\Delta_1(x_t) = x_t - x_{t-1}$$

Y en general, la variación absoluta en el mes t respecto de k meses antes como la diferencia:

$$\Delta_k(x_t) = x_t - x_{t-k}$$

Su signo positivo o negativo nos indica una evolución creciente o decreciente respectivamente. Estas diferencias están expresadas en las mismas unidades que la serie original, lo que dificulta hacer comparaciones. Incluso cuando se usa la misma unidad de medida, el valor de la variación absoluta no es fácil de interpretar. Veámoslo en el siguiente ejemplo.

► EJEMPLO 3.1.

Supongamos que el valor de una serie de datos, X, en el momento t es igual a 315€ y en el momento t-1 era igual a 300€, la variación absoluta sería

$$\Delta_1(x_t) = x_t - x_{t-1} = 315 - 300 = 15 \text{ €}$$

Consideremos otra serie, Y, que en momento t-1 era igual a 1000€ y en el momento t igual a 1015€, su variación absoluta sería de igual signo que la anterior (ambas series tienen una evolución creciente) y de igual cuantía

$$\Delta_1(y_t) = y_t - y_{t-1} = 1015 - 1000 = 15 \text{ €}$$

Sin embargo, es evidente que la evolución en ambos casos no tiene la misma importancia. ◀

Para medir de forma **más precisa** estas variaciones, de forma que sean **comparables**, hay que observarlas en relación a los datos que las provocan, es decir, en **términos relativos**.

Variaciones relativas: Tasas de variación.

La **variación relativa** de una variable en el período t , o **tasa de variación**, se define como el cociente (ratio o razón) de la variación absoluta sobre el valor previo de la variable en dicha variación absoluta:

$$T_1(t) = T_1(x_t) = \frac{\Delta_1(x_t)}{x_{t-1}} = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

y en general para cualquier número de meses k , como:

$$T_k(t) = T_k(x_t) = \frac{\Delta_k(x_t)}{x_{t-k}} = \frac{x_t - x_{t-k}}{x_{t-k}} = \frac{x_t}{x_{t-k}} - 1$$

Dependiendo del intervalo de tiempo considerado ($x_{t-1} \rightarrow x_t$, $x_{t-k} \rightarrow x_t$), hablaremos de tasas de variación mensual, trimestral, anual,...

La tasa de variación es una proporción o tanto por uno. Es habitual multiplicarla por 100 para expresarla en tanto por ciento o porcentaje.

Se trata de un **valor adimensional** por lo que se puede utilizar para **comparar** la evolución de series de observaciones, aunque éstas estén expresadas en diferentes unidades.

Estas tasas se calculan sólo sobre variables que toman valores positivos en todo período t ($x_t > 0$).

► EJEMPLO 3.2.

En el ejemplo anterior la tasa de variación para la serie de valores, X , sería:

$$T_1(t) = \frac{\Delta_1(x_t)}{x_{t-1}} = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1 = \frac{315}{300} - 1 = 1,05 - 1 = 0,05 \quad \Rightarrow \quad 5\%$$

Para la otra serie, Y , valdría:

$$T_1(t) = \frac{y_t}{y_{t-1}} - 1 = \frac{1015}{1000} - 1 = 1,015 - 1 = 0,015 \quad \Rightarrow \quad 1,5\%$$

Lo que muestra que aunque en ambos casos la variación absoluta es la misma, en la primera serie supone un 5% sobre el valor inicial (300) y en la segunda serie sólo un 1,5% sobre el valor inicial (1000), señalando la mayor importancia de la variación en la primera serie comparada con la segunda.

La tasa de variación toma el mismo valor cuando las cantidades que se comparan son proporcionales. Por ejemplo: $315 \times 5 = 1575$, $300 \times 5 = 1500$

$$T_1(t) = \frac{\Delta_1(z_t)}{z_{t-1}} = \frac{z_t - z_{t-1}}{z_{t-1}} = \frac{z_t}{z_{t-1}} - 1 = \frac{1575}{1500} - 1 = 1,05 - 1 = 0,05 \Rightarrow 5\%$$



Si la variación absoluta es positiva también lo será la correspondiente tasa de variación. Lo mismo vale para valores negativos. Por tanto, una tasa de variación positiva indica una evolución creciente (como en estos casos del ejemplo). Una tasa de variación negativa indica una evolución decreciente (por ejemplo, $T_1(t) = -0,03$, indica que la serie ha disminuido su valor en el período t un 3% en relación a su valor en el período t-1).

Relacionado con la **tasa de variación** está el **factor de variación unitaria** que se define como

$$1 + T_1(t) = 1 + \frac{x_t}{x_{t-1}} - 1 = \frac{x_t}{x_{t-1}}$$

y en general para cualquier número de meses k, como:

$$1 + T_k(t) = 1 + \frac{x_t}{x_{t-k}} - 1 = \frac{x_t}{x_{t-k}}$$

Donde como puede observarse, se comparan por cociente el valor actual de la serie, x_t , con un valor anterior, x_{t-1} (o x_{t-k}).

Se denomina así porque expresa en lo que se transforma una unidad entre los períodos de tiempo considerados.

► EJEMPLO 3.3.

En nuestro ejemplo, el factor de variación unitaria para la serie de valores, X, sería:

$$1 + T_1(t) = \frac{x_t}{x_{t-1}} = \frac{315}{300} = 1,05$$

Cada unidad en t-1 se transforma en 1,05 unidades en t, así las 300 unidades en t-1 se transforman en $300 \times 1,05 = 315$ unidades en t, 500 unidades se transformarían en $500 \times 1,05 = 525$, etc.

Para la otra serie, Y, el factor de variación unitaria valdría:

$$1 + T_1(t) = \frac{y_t}{y_{t-1}} = \frac{1015}{1000} = 1,015$$

una unidad en $t-1$ se transformaría en 1,015 unidades en t , así las 1000 unidades en $t-1$ se transforman en $1000 \times 1,015 = 1015$ unidades en t , 300 unidades se transformarían en $300 \times 1,015 = 304,5$, etc. ◀

Equivalencia entre tasas de variación de diferentes períodos.

No sólo se utilizan tasas de variación mensuales, $T_1(t)$, también se trabaja frecuentemente con tasas de variación referidas a otros períodos como trimestres, $T_3(t)$, semestres, $T_6(t)$, o años, $T_{12}(t), \dots$, donde el subíndice 1, 3, 6 o 12 refleja la separación, en meses, entre los períodos que se comparan en las respectivas tasas.

Si la tasa de variación mensual de un determinado mes se mantuviera en los once meses siguientes, ¿cuál sería la tasa de variación anual (*equivalente*)? La respuesta es que la ***tasa de variación anual equivalente*** sería aquella que nos conduce a la misma variación absoluta:

Un valor inicial, x_0 , que sufriera una variación mensual, $T_1(t)$, durante el primer mes y los once siguientes se transformaría en,

$$x_1 = (1 + T_1(t)) x_0 \quad (\text{al cabo del primer mes})$$

$$x_2 = (1 + T_1(t)) x_1 = (1 + T_1(t))(1 + T_1(t)) x_0 = (1 + T_1(t))^2 x_0 \quad (\text{al cabo de dos meses})$$

y sucesivamente en

$$x_{12} = (1 + T_1(t))^{12} x_0 \quad (\text{al cabo de un año})$$

La tasa de variación anual equivalente produciría la misma variación absoluta, pasando en un año del valor inicial x_0 al valor final x_{12} , es decir

$$x_{12} = (1 + \hat{T}_{12}) x_0$$

Igualando las dos últimas expresiones obtenemos

$$(1 + T_1(t))^{12} = (1 + \hat{T}_{12}) \quad \Leftrightarrow \quad \hat{T}_{12} = (1 + T_1(t))^{12} - 1$$

La primera expresión indica que el producto del factor de variación unitaria mensual por sí mismo, las doce veces que se repite, nos da el factor de variación unitaria anual equivalente.

Nota: lo notamos \hat{T}_{12} (en lugar de T_{12}) para indicar que es un valor estimado bajo el supuesto de que la tasa de variación mensual es la misma para todos los meses, $T_1(t)$.

T_{12} se obtendría a partir de la tasa de variación mensual de cada uno de los doce meses que forman el año como

$$(1 + T_1(t))(1 + T_1(t+1)) \dots (1 + T_1(t+11)) = (1 + T_{12}) \Leftrightarrow T_{12} = [(1 + T_1(t))(1 + T_1(t+1)) \dots (1 + T_1(t+11))] - 1$$

La primera igualdad expresa que el producto de los factores de variación unitaria de todos los meses que componen el año es igual al factor de variación unitaria anual.

Para indicar que esta tasa anual equivalente se obtiene a partir de una tasa mensual, la notaremos como

$$\hat{T}_{12}^{(1)} = (1 + T_1(t))^{12} - 1$$

Si la tasa anual equivalente se obtiene a partir del primer trimestre del año, se nota

$$\hat{T}_{12}^{(3)} = (1 + T_3(t))^4 - 1$$

Y si se ha obtenido a partir del primer semestre del año

$$\hat{T}_{12}^{(6)} = (1 + T_6(t))^2 - 1$$

Análogamente si se ha obtenido a partir de otros períodos en los que se pueda dividir el año.

Leeremos $\hat{T}_{12}^{(1)}$ como *tasa de variación mensual elevada a anual*, $\hat{T}_{12}^{(3)}$ como *tasa de variación trimestral elevada a anual*, etc.

Todo lo anterior es la base teórica que se utiliza cuando para estimar la inflación anual sólo se dispone de la evolución de los precios registrada por el IPC en el primer mes del año (primer trimestre, ...)

► EJEMPLO 3.4.

En el primer semestre de 2012 los precios han subido un 1,17%. ¿Cuál será la inflación anual estimada para el año 2012 si en el segundo semestre se mantiene la misma tendencia observada en el primero?

Solución:

Expresamos 1,17% en tanto por uno dividiendo por 100 y sustituimos en la expresión de la tasa de variación semestral elevada a anual

$$\hat{T}_{12}^{(6)} = (1 + T_6(t))^2 - 1 = (1 + 0,0117)^2 - 1 = 0,02354 \Rightarrow 2,354\%$$

Si en el segundo semestre de 2012 sigue la misma tendencia en la subida de precios que en el primer semestre, al cabo del año los precios habrán aumentado un 2,354%



Tasas medias de variación.

Sea la siguiente serie mensual de observaciones

$$x_0, x_1, x_2, \dots, x_{12}$$

con las siguientes tasas mensuales de variación asociadas

$$T_1(1) = \frac{x_1}{x_0} - 1 \quad T_1(2) = \frac{x_2}{x_1} - 1 \quad \dots \quad T_1(12) = \frac{x_{12}}{x_{11}} - 1$$

¿Cuál sería el valor de la tasa mensual de variación que aplicada repetidamente a lo largo de los doce meses nos llevaría al mismo valor final del período anual, x_{12} , partiendo del valor inicial, x_0 ? Es decir, cuál sería la tasa mensual media de variación TM_1 .

Con los factores de variación unitaria observados se tiene que

$$x_{12} = x_0 (1 + T_1(1))(1 + T_1(2)) \dots (1 + T_1(12))$$

Si se repite una tasa mensual TM_1 durante los doce meses, transformándose x_0 en x_{12}

$$x_{12} = x_0 (1 + TM_1)(1 + TM_1) \dots (1 + TM_1) = x_0 (1 + TM_1)^{12}$$

Igualando ambas expresiones

$$(1 + T_1(1))(1 + T_1(2)) \dots (1 + T_1(12)) = (1 + TM_1)^{12} \quad \Leftrightarrow \quad \sqrt[12]{(1 + T_1(1))(1 + T_1(2)) \dots (1 + T_1(12))} = (1 + TM_1)$$

Esta última expresión indica que el factor de variación unitaria asociado a la tasa media de variación mensual es la **media geométrica** de los factores de variación unitaria asociados a las diferentes tasas de variación mensual. Despejando la tasa media de variación mensual se obtiene

$$TM_1 = \sqrt[12]{(1 + T_1(1))(1 + T_1(2)) \dots (1 + T_1(12))} - 1$$

Por otra parte, es posible calcular la tasa media de variación a partir de los valores inicial y final del período como sigue

$$x_{12} = x_0 (1 + T_1(1))(1 + T_1(2)) \dots (1 + T_1(12)) \quad \Leftrightarrow \quad \frac{x_{12}}{x_0} = (1 + T_1(1))(1 + T_1(2)) \dots (1 + T_1(12))$$

$$TM_1 = \sqrt[12]{(1 + T_1(1))(1 + T_1(2)) \dots (1 + T_1(12))} - 1 = \sqrt[12]{\frac{x_{12}}{x_0}} - 1$$

De forma similar se calcularía la tasa media de variación si en lugar de meses nos referimos a otros períodos o si en lugar de 12 datos se dispone de un número diferente.

► EJEMPLO 3.5.

El consumo de agua, en m^3 , de una comunidad de regantes fue de diciembre de 2010 a junio de 2011

t	$consumo$
2010 diciembre	15400
2011 enero	17300
2011 febrero	25600
2011 marzo	30200
2011 abril	27000
2011 mayo	26400
2011 junio	24800

Calcule:

- La variación absoluta trimestral.
- Tasas mensuales de variación.
- Tasa semestral de variación.
- Tasa mensual media de variación en todo el período.

Solución:

a)

t	$\Delta_3(x_t)$
2011 marzo	$30200-15400=14800$
2011 abril	$27000-17300=9700$
2011 mayo	$26400-25600=800$
2011 junio	$24800-30200=-5400$

b)

t	$T_1(t)$
2011 enero	$\frac{17300}{15400} - 1 = 0,1234$
2011 febrero	$\frac{25600}{17300} - 1 = 0,4798$
2011 marzo	$\frac{30200}{25600} - 1 = 0,1797$
2011 abril	$\frac{27000}{30200} - 1 = -0,1060$
2011 mayo	$\frac{26400}{27000} - 1 = -0,0222$
2011 junio	$\frac{24800}{26400} - 1 = -0,0606$

c) Dos formas:

$$T_6(junio\ 11) = \frac{24800}{15400} - 1 = 0,6104$$

O bien como

$$1 + T_6(junio\ 11) = (1 + T_1(enero\ 11)) \times (1 + T_1(febrero\ 11)) \times \dots \times (1 + T_1(junio\ 11)) = \\ 1,1234 \times 1,4798 \times 1,1797 \times 0,8940 \times 0,9778 \times 0,9394 = 1,6104$$

$$T_6(junio\ 11)\% = 61,04\%$$

d) Varias formas:

$$1 + TM_1 = \sqrt[6]{\frac{24800}{15400}} = \sqrt[6]{(1 + T_1(\text{enero } 11)) \times (1 + T_1(\text{febrero } 11)) \times \dots \times (1 + T_1(\text{junio } 11))} = \\ = \sqrt[6]{1 + T_6(\text{junio } 11)} = 1,269 \quad \Rightarrow \quad TM_1 \% = 26,9\%$$

3.2 Índice elemental. Índice sintético.

Índice elemental.

Sea la evolución temporal de una magnitud X : $x_0, x_1, \dots, x_t, \dots$, se denomina **índice elemental** (o índice simple) de la magnitud X en el período t respecto al periodo 0 al cociente

$$I_{t/0}(X) = \frac{x_t}{x_0}$$

Nota: observe la relación del **índice elemental** con la tasa de variación y el factor de variación

unitaria definidos anteriormente: $1 + T_1(t) = 1 + \frac{x_t}{x_{t-1}} - 1 = \frac{x_t}{x_{t-1}}$, $1 + T_t(t) = 1 + \frac{x_t}{x_{t-t}} - 1 = \frac{x_t}{x_0}$, es

decir, **el índice elemental es el factor de variación unitaria entre los períodos 0 y t.**

El período 0 utilizado como período de referencia se denomina **período base** y el período t que se compara con el anterior es el **período corriente** (o período actual).

Este índice expresa, en tanto por uno, la evolución que ha experimentado la magnitud X desde el período 0 hasta el período t . Valores mayores que uno indican que X ha aumentado, menores que uno expresan una disminución de la magnitud X e igual a uno que se ha mantenido su valor. Habitualmente los números índices se multiplican por 100 para expresarlos en porcentajes.

► EJEMPLO 3.6.

El precio de la gasolina ha pasado de 1,125€/litro a 1,328€/litro. Calculando el correspondiente índice elemental podemos ver que

$$I_{1/0}(X) = \frac{x_1}{x_0} = \frac{1,328}{1,125} = 1,18$$

la gasolina es ahora un 18% más cara que antes, o bien

$$I_{0/1}(X) = \frac{x_0}{x_1} = \frac{1,125}{1,328} = 0,847 \quad \Rightarrow \quad 1 - 0,847 = 0,153$$

antes su valor era el 84,7% de su valor actual, es decir, un 15,3% más barata que actualmente. ◀

De la misma definición de índice elemental resulta la siguiente propiedad (**propiedad circular**) que nos va a permitir referirnos a períodos base distintos del período 0

$$\frac{x_t}{x_0} = \frac{x_t}{x_{t'}} \cdot \frac{x_{t'}}{x_0} \quad \Leftrightarrow \quad I_{t/0}(X) = I_{t/t'}(X) I_{t'/0}(X)$$

de donde

$$I_{t/t'}(X) = \frac{I_{t/0}(X)}{I_{t'/0}(X)}$$

► EJEMPLO 3.7.

El año pasado el precio de la gasolina ha experimentado entre los meses de enero y junio una subida, según el índice elemental, de 1,185 y entre los meses de enero y diciembre una subida, según el mismo índice, de 1,225. ¿Qué subida ha experimentado el precio de la gasolina entre los meses de junio y diciembre?

Solución:

Etiquetamos por comodidad como 0, 1 y 2 a los tres meses de enero, junio y diciembre respectivamente.

$$I_{2/1}(X) = \frac{I_{2/0}(X)}{I_{1/0}(X)} = \frac{1,225}{1,185} = 1,0338$$

De junio a diciembre ha aumentado 0,0338 en tanto por uno (3,38%).

Observe que no sería válido el siguiente razonamiento: de enero a junio ha aumentado un 18,5%, de enero a diciembre ha aumentado un 22,5%, por tanto de junio a diciembre ha aumentado un 22,5-18,5=4%. ◀

Índice sintético.

Consideremos ahora un conjunto X de magnitudes simples, $X_1, X_2, \dots, X_i, \dots, X_n$ (por ejemplo, precios de distintos artículos). Los índices elementales de las magnitudes simples X_i se definen como

$$I_{t/0}(X_i) = \frac{x_{it}}{x_{i0}}$$

Queremos resumir en un único índice (que llamaremos **índice sintético, compuesto o complejo**) los índices elementales de las magnitudes simples X_i .

Hay distintas soluciones, todas ellas se basan en el cálculo de **medias sobre dichos índices elementales**.

Distinguiremos entre **índices sintéticos sin ponderar** (a todas las magnitudes simples se le da la misma importancia) e **índices sintéticos ponderados** (a cada magnitud simple se le asigna una importancia diferente).

Índices sintéticos sin ponderar.

Índice de Sauerbeck.

Es la media aritmética de los índices elementales

$$S_{t/0}(X) = \frac{1}{n} \sum_{i=1}^n \frac{x_{it}}{x_{i0}}$$

Índice de Bradstreet y Dudot.

Se define como la media agregativa

$$D_{t/0}(X) = \frac{\sum_{i=1}^n x_{it}}{\sum_{i=1}^n x_{i0}}$$

Índices sintéticos ponderados.

Para el cálculo de estos índices se utilizan coeficientes de ponderación que cuantifican la importancia relativa de cada magnitud:

- u_{i0} , importancia relativa de la magnitud simple X_i en el período base 0. $\left(\sum_{i=1}^n u_{i0} = 1 \right)$
- u_{it} , importancia relativa de la magnitud simple X_i en el período actual t.
 $\left(\sum_{i=1}^n u_{it} = 1 \right)$

Índice de Laspeyres.

Es la media aritmética de los índices elementales con ponderaciones del período base

$$L_{t/0}(X) = \sum_{i=1}^n \frac{x_{it}}{x_{i0}} u_{i0}$$

Índice de Paasche.

Es la media armónica de los índices elementales con ponderaciones del período actual

$$P_{t/0}(X) = \frac{1}{\sum_{i=1}^n \frac{x_{i0}}{x_{it}} u_{it}}$$

Índice de Fisher.

Es la media geométrica de los índices de Laspeyres y Paasche

$$F_{t/0}(X) = \sqrt{L_{t/0}(X) P_{t/0}(X)}$$

3.3 Índices de precios, de cantidades y de valor.

Se tiene información sobre precios y cantidades de n artículos:

Período base		Período actual	
precios	cantidades	precios	cantidades
p_{10}	q_{10}	p_{1t}	q_{1t}
p_{20}	q_{20}	p_{2t}	q_{2t}
\cdot	\cdot	\cdot	\cdot
p_{i0}	q_{i0}	p_{it}	q_{it}
\cdot	\cdot	\cdot	\cdot
p_{n0}	q_{n0}	p_{nt}	q_{nt}

Se puede obtener fácilmente el valor total de cada artículo para el período base, $p_{i0}q_{i0}$, y para el período actual, $p_{it}q_{it}$. Por tanto el valor global de todos los artículos en el período base será

$$\sum_{i=1}^n p_{i0}q_{i0} \text{ y el valor de todos los artículos en el período actual } \sum_{i=1}^n p_{it}q_{it}.$$

Se suele tomar como factores de ponderación el valor relativo de cada artículo en relación al valor global

$$u_{i0} = \frac{p_{i0}q_{i0}}{\sum_{j=1}^n p_{j0}q_{j0}} \quad u_{it} = \frac{p_{it}q_{it}}{\sum_{j=1}^n p_{jt}q_{jt}}$$

Índices de precios.

Para hallar los índices sintéticos de precios utilizaremos como índices elementales $\frac{p_{it}}{p_{i0}}$

$$L_{t/0}^p = \sum_{i=1}^n \frac{p_{it}}{p_{i0}} u_{i0} = \sum_{i=1}^n \frac{p_{it}}{p_{i0}} \frac{p_{i0} q_{i0}}{\sum_{j=1}^n p_{j0} q_{j0}} = \sum_{i=1}^n \frac{p_{it} q_{i0}}{\sum_{j=1}^n p_{j0} q_{j0}} = \frac{\sum_{i=1}^n p_{it} q_{i0}}{\sum_{i=1}^n p_{i0} q_{i0}}$$

Es decir, el índice de precios de Laspeyres consiste en calcular el valor de las cantidades del período base a precios del período base y período actual y compararlos por cociente.

Mientras que en el índice de precios de Paasche se consideran las cantidades del período actual, se determina su valor a precios del período base y actual y se comparan por cociente.

$$P_{t/0}^p = \frac{1}{\sum_{i=1}^n \frac{p_{i0}}{p_{it}} u_{it}} = \frac{1}{\sum_{i=1}^n \frac{p_{i0}}{p_{it}} \frac{p_{it} q_{it}}{\sum_{j=1}^n p_{jt} q_{jt}}} = \frac{1}{\sum_{i=1}^n \frac{p_{i0} q_{it}}{\sum_{j=1}^n p_{jt} q_{jt}}} = \frac{\sum_{i=1}^n p_{it} q_{it}}{\sum_{i=1}^n p_{i0} q_{it}}$$

En el caso de que las cantidades del período base y actual coincidan también coincidirán los dos índices anteriores, $L_{t/0}^p = P_{t/0}^p$.

Se llama **cesta de la compra** a un conjunto de artículos (y cantidades), obtenidos a partir de la encuesta de presupuestos familiares, representativos de un determinado nivel de vida. El **índice de precios de consumo, I.P.C.**, es el **índice de precios de Laspeyres** sobre la **cesta de la compra**.

Índices de cantidades.

Ahora los índices elementales son los cocientes $\frac{q_{it}}{q_{i0}}$

$$L_{t/0}^q = \sum_{i=1}^n \frac{q_{it}}{q_{i0}} u_{i0} = \frac{\sum_{i=1}^n p_{i0} q_{it}}{\sum_{i=1}^n p_{i0} q_{i0}}$$

El índice de cantidades de Laspeyres compara el valor de las cantidades de los períodos base y actual con precios del período base. Mientras que el índice de cantidades de Paasche compara el valor de las cantidades de los períodos base y actual con precios del período actual.

$$P_{t/0}^q = \frac{\sum_{i=1}^n p_{it} q_{it}}{\sum_{i=1}^n p_{it} q_{i0}}$$

Tanto para precios como para cantidades se define el índice de Fisher como la media geométrica de los índices de Laspeyres y Paasche (por tanto su valor está comprendido entre ambos).

Índices de Marshall-Edgeworth.

Adopta una posición (valor) intermedia entre los índices de Laspeyres y Paasche considerando la media de las cantidades en los períodos base y actual para hallar el **índice de precios**

$$E_{t/0}^p = \frac{\sum_{i=1}^n p_{it} \frac{(q_{i0} + q_{it})}{2}}{\sum_{i=1}^n p_{i0} \frac{(q_{i0} + q_{it})}{2}} = \frac{\sum_{i=1}^n p_{it} (q_{i0} + q_{it})}{\sum_{i=1}^n p_{i0} (q_{i0} + q_{it})}$$

Para hallar el **índice de cantidades** considera la media de los precios en los períodos base y actual

$$E_{t/0}^q = \frac{\sum_{i=1}^n q_{it} \frac{(p_{i0} + p_{it})}{2}}{\sum_{i=1}^n q_{i0} \frac{(p_{i0} + p_{it})}{2}} = \frac{\sum_{i=1}^n q_{it} (p_{i0} + p_{it})}{\sum_{i=1}^n q_{i0} (p_{i0} + p_{it})}$$

Índices de valor.

Supongamos un fenómeno caracterizado por un conjunto de cantidades y precios que varían a lo largo del tiempo (p.e. la producción). Nos planteamos estudiar la variación en el valor agregado de dicho fenómeno.

En el año base $V_0 = \sum_{i=1}^n p_{i0} q_{i0}$, y en el año actual $V_t = \sum_{i=1}^n p_{it} q_{it}$.

Denominamos **índice de valor agregado** al cociente

$$I_{t/0}^V = \frac{V_t}{V_0} = \frac{\sum_{i=1}^n p_{it} q_{it}}{\sum_{i=1}^n p_{i0} q_{i0}}$$

Se puede demostrar fácilmente que el índice anterior puede obtenerse como el producto de un índice de precios por un índice de cantidades, según:

$$I_{t/0}^V = L_{t/0}^p P_{t/0}^q = L_{t/0}^q P_{t/0}^p.$$

3.4 Enlace de series de números índices con distinta base.

Supongamos que tenemos un número índice referido al período base t_1 , I_{t/t_1} , y queremos expresarlo con período base t_2 , I_{t/t_2} , además conocemos el índice que relaciona los períodos t_1 y t_2 , I_{t_2/t_1} . La forma de obtener I_{t/t_2} a partir de I_{t/t_1} consiste en aplicar la propiedad circular. Esta propiedad no la cumplen los índices sintéticos más usuales, a pesar de ello el procedimiento práctico actúa como si se verificase tal propiedad.

$$I_{t/t_2} = \frac{I_{t/t_1}}{I_{t_2/t_1}}$$

Si se quiere expresar el índice I_{t/t_2} en la base t_1 , haremos uso de la misma propiedad escrita como

$$I_{t/t_1} = I_{t/t_2} I_{t_2/t_1}.$$

► EJEMPLO 3.8.

Dadas las siguientes series de números índices de precios al consumo obtenga una sola serie para todo el periodo con base 2001 y otra con base en 2003

Año	IPC (base 1990)	IPC (base 2001)	IPC (base2004)
1997	115		
1998	123		
1999	129		
2000	132		
2001	135	100	
2002		103	
2003		107	
2004		113	100
2005			110
2006			120
2007			123

Solución:

Para aplicar la propiedad circular y cambiar de base, los números índices deben estar expresados en tanto por uno, para lo cual los dividimos primero por 100.

Año	IPC (base 2001)	IPC (base 2001) en %
1997	$1,15/1,35=0,8518$	85,18
1998	$1,23/1,35=0,9111$	91,11
1999	$1,29/1,35=0,9555$	95,55
2000	$1,32/1,35=0,9777$	97,77
2001	1,00	100
2002	1,03	103
2003	1,07	107
2004	1,13	113
2005	$1,10 \times 1,13=1,243$	124,3
2006	$1,20 \times 1,13=1,356$	135,6
2007	$1,23 \times 1,13=1,3899$	138,99

Año	IPC (base 2003)	IPC (base 2003) en %
1997	$0,8518/1,07=0,7961$	79,61
1998	$0,9111/1,07=0,8515$	85,15
1999	$0,9555/1,07=0,893$	89,3
2000	$0,9777/1,07=0,9137$	91,37
2001	$1,00/1,07=0,9346$	93,46
2002	$1,03/1,07=0,9626$	96,26
2003	$1,07/1,07=1,00$	100
2004	$1,13/1,07=1,0561$	105,61
2005	$1,243/1,07=1,1617$	116,17
2006	$1,356/1,07=1,2673$	126,73
2007	$1,3899/1,07=1,299$	129,9

3.5 Deflación de series económicas.

Un problema frecuente en estudios económicos consiste en el análisis de una sucesión de valores expresados en **moneda corriente** de cada año (también se utilizan las expresiones precios corrientes o nominales). Para poder comparar dichas cantidades es necesario homogeneizarlas en el sentido de que los valores estén expresados en los “*mismos precios*” (**precios constantes o reales**, también se utilizan las expresiones moneda constante o términos reales). La homogenización mencionada se denomina **deflación** y consiste en dividir los valores de la serie económica por un índice adecuado, conocido como **deflactor**. Cada fenómeno concreto exige de un deflactor adecuado, por ejemplo la capacidad de consumo tendría por deflactor el índice de precios al consumo.

EJEMPLO 3.9.

En los últimos siete años el gasto en enseñanza, en millones de euros corrientes, y los índices de precios al consumo han sido

Año	Gastos en millones €	IPC (base 2001)	IPC (base 2004)
2001	150	100	
2002	230	115	
2003	240	135	
2004	290	180	100
2005	300		105
2006	330		120
2007	400		150

Calcúlese el porcentaje, en términos reales, en que ha variado el gasto en enseñanza durante los siete años considerados.

Solución:

A **precios corrientes o valores nominales** la evolución del gasto en enseñanza ha sido según la tasa de variación:

$$\frac{400-150}{150} = 1,6667 \Rightarrow 166,67\%$$

Para obtener los gastos en enseñanza en **términos reales** hemos de deflactar los valores corrientes, transformando los **euros corrientes** de cada año en **euros constantes** del año base.

En este ejemplo los IPC son tales que si procedemos a deflactar directamente tendríamos hasta el 2004 los gastos expresados en euros de 2001 y a partir de 2004 en euros de este año, por lo que las cantidades no serían comparables. Por ello en primer lugar vamos a expresar todos los índices en la misma base, usando $I_{t/2001} = I_{t/2004} I_{2004/2001}$ (índices expresados en tanto por uno)

IPC (base 2001)	IPC (base 2001) en %
1,00	100
1,15	115
1,35	135
1,80	180
1,89=1,05x1,80	189
2,16=1,20x1,80	216
2,70=1,50x1,80	270

O bien, usando $I_{t/2004} = \frac{I_{t/2001}}{I_{2004/2001}}$ (índices expresados en tanto por uno)

IPC (base 2004)	IPC (base 2004) en %
$0,5555 = \frac{1,00}{1,80}$	55,55
$0,6389 = \frac{1,15}{1,80}$	63,89
$0,75 = \frac{1,35}{1,80}$	75
1,00	100
1,05	105
1,20	120
1,50	150

Si expresamos los gastos en términos reales, en euros de 2001, tenemos:

Gastos en euros de 2001
150/1=150
230/1,15=200
240/1,35=177,78
290/1,80=161,11
300/1,89=158,73
330/2,16=152,77
400/2,70=148,15

$148,15 - 150 = -1,85$, ha habido una disminución absoluta en términos reales de 1,85 millones de euros de 2001 lo que supone una disminución relativa (según tasa de variación)

$$\frac{148,15 - 150}{150} = \frac{-1,85}{150} = -0,0123 \Rightarrow -0,0123 \times 100 = -1,23\%$$

Si deflactamos tomando como referencia los euros de 2004, se tiene

Gastos en euros de 2004
150/0,5555=270
230/0,6389=359,94
240/0,75=320
290/1=290
300/1,05=285,71
330/1,20=275
400/1,50=266,67

$266,67 - 270 = -3,33$, ha habido una disminución absoluta del gasto en enseñanza durante el período 2001-2007 de 3,33 millones de euros de 2004, por tanto una disminución relativa de

$$\frac{266,67 - 270}{270} = \frac{-3,33}{270} = -0,0123 \Rightarrow -0,0123 \times 100 = -1,23\%$$

Obsérvese que se llega a la misma variación relativa en términos reales, tanto si se toma un año base como otro. ◀

► EJEMPLO 3.10.

El salario medio en los últimos cinco años en una empresa siderometalúrgica y los índices de precio de consumo han sido

Años	Salarios	IPC
2008	1760	107
2009	1830	113
2010	1950	124,3
2011	2100	135,6
2012	2200	139

Estudie el valor real de los salarios en euros constantes del 2012. Calcule la tasa de crecimiento medio anual de los salarios en términos reales.

Solución:

Años	Salarios corrientes	IPC (base 2012)	Salarios en euros constantes (euros 2012)
2008	1760	$1,07/1,39=0,7698$	$1760/0,7698=2286,3$
2009	1830	$1,13/1,39=0,8129$	$1830/0,8129=2251,2$
2010	1950	$1,243/1,39=0,8942$	$1950/0,8942=2180,7$
2011	2100	$1,356/1,39=0,9755$	$2100/0,9755=2152,7$
2012	2200	$1,39/1,39=1,00$	$2200/1=2200$

$$\sqrt[4]{\frac{S_4}{S_0}} - 1 = \sqrt[4]{\frac{2200}{2286,3}} - 1 = 0,99 - 1 = -0,01$$

Ha habido una disminución media anual del 1% ($-0,01 \times 100 = -1\%$) en el período de tiempo considerado. ◀

► EJEMPLO 3.11.

El propietario de una vivienda alquiló esta en 500€ mensuales a cuatro estudiantes de la Facultad de CC. Económicas y Empresariales en Octubre de 2010. Este había pactado revisar el precio del alquiler en Enero de cada año de acuerdo al incremento de precios al consumo del último año. El valor del IPC se recoge en la siguiente tabla

Años	IPC
2009	113
2010	124,3
2011	135,6
2012	139

¿A cuánto ascenderá el precio del alquiler a partir de Enero de 2013?

Solución:

Años	IPC %	$IPC_{\text{año } t / \text{año } t-1}$	Alquiler
2009	113		
2010	124,3	$124,3/113=1,10$	500
2011	135,6	$135,6/124,3=1,0909$	$500 \times 1,10=550$
2012	139	$139/135,6=1,0251$	$550 \times 1,0909=599,99 \approx 600$
2013			$600 \times 1,0251=615,06 \approx 615$

3.6 Dependencia de un índice general respecto de un grupo de productos.

Vamos a analizar cómo afecta a un índice general la variación en uno de los productos (artículos) o grupo de productos considerados en su construcción.

Muchos índices, entre ellos el IPC español, están basados en el índice de Laspeyres.

Los índices de Laspeyres y Paasche poseen la **propiedad de agregación**: El índice de Laspeyres sobre un conjunto de productos es igual al índice de Laspeyres calculado sobre los índices de Laspeyres de cada subconjunto de productos. Lo mismo ocurre con el índice de Paasche.

Escribiremos la propiedad de agregación con la siguiente notación general (donde el índice $I_{t/0}$ representa a $L_{t/0}$ en el caso del IPC, X_i representa al grupo i)

$$I_{t/0}(X) = \sum_{i=1}^n I_{t/0}(X_i) u_{i0}$$

La variación absoluta del índice entre dos períodos t y t' será

$$\Delta I(X) = I_{t'/0}(X) - I_{t/0}(X) = \sum_{i=1}^n (I_{t'/0}(X_i) - I_{t/0}(X_i)) u_{i0} = \sum_{i=1}^n \Delta I(X_i) u_{i0}$$

► EJEMPLO 3.12.

Las ponderaciones de los siguientes grupos en el IPC para el año base son

Grupo	Ponderación en %
Alimentos	35
Vestido	9
Vivienda	17
Menaje y hogar	7
Salud	3
Transporte	12
Ocio	5
Enseñanza	8
Otros	4
Total	100

Si el IPC de este año ha sido 115,96, calcule el IPC para el próximo año en los siguientes casos:

- El índice de la vivienda se incrementa 15 puntos permaneciendo igual el resto de los índices.
- Los índices de vivienda y transporte se incrementan 10 y 5 puntos respectivamente, los índices de vestido y enseñanza disminuyen 2 y 1 punto respectivamente, no variando el resto de los índices.

Solución:

$$a) \Delta I(X) = \sum_{i=1}^n \Delta I(X_i) u_i = 0,15 \times 0,17 = 0,0255$$

$$\text{IPC(próx. año)} = \text{IPC(año actual)} + \Delta \text{IPC} = 115,96 + 2,55 = 118,51$$

$$b) \Delta I(X) = \sum_{i=1}^n \Delta I(X_i) u_i = (0,10 \times 0,17) + (0,05 \times 0,12) + (-0,02 \times 0,09) + (-0,01 \times 0,08) = 0,0204$$

$$\text{IPC(próx. año)} = \text{IPC(año actual)} + \Delta \text{IPC} = 115,96 + 2,04 = 118,00$$



3.7 Ejercicios resueltos.

1. La tabla muestra los beneficios anuales, en millones de euros, de una empresa y los índices de precios (IPC) para el período 2005-2010:

Año	2005	2006	2007	2008	2009	2010
Beneficios	9,3	9,5	9	10	11,4	13
IPC	105	110	112			
IPC			100	103	108	114

- Enlace las series de números índice tomando como base el año 2010.
- Obtenga los beneficios anuales a precios del 2010.
- Calcule las tasas de variación anual (en %) de los beneficios reales.
- Calcule la tasa media de variación anual (en %) de los beneficios reales.

Solución:

AÑO	2005	2006	2007	2008	2009	2010
BENEFICIO	9,3	9,5	9	10	11,4	13
IPC	105	110	112	115,36	120,96	127,68
IPC	93,75	98,21	100	103	108	114
IPC(base 2010)	82,24	86,15	87,72	90,35	94,74	100,00
BENEFICIOS (precios 2010)	11,308	11,027	10,26	11,068	12,033	13
TASAS DE VARIACION ANUAL (%)		-2,48	-6,96	7,88	8,72	8,04

- Las dos series de números índices (IPC) pueden enlazarse de las dos formas que aparecen en la tabla (dividiendo la primera serie por 1,12 o multiplicando la segunda serie por 1,12). Posteriormente, haciendo que el IPC para el año 2010 sea 100 (dividiendo por 1,2768 la primera serie de IPC o por 1,14 la segunda serie de IPC) se obtiene la serie de números índices (IPC) enlazada con base el año 2010. Debido al redondeo puede haber en algún caso una pequeña diferencia entre ambos métodos.
- Dividiendo los beneficios corrientes entre los IPC (base 2010) se obtienen los beneficios a precios del 2010 que aparecen en la tabla. Para estos cálculos se expresan los IPC en tantos por uno.

$$\frac{9,3}{0,8224} = 11,308 \quad \dots \quad \frac{13}{1} = 13$$

- A partir de los cocientes entre los beneficios reales (a precios del 2010) de años consecutivos se obtienen las tasas de variación anual. Se multiplican por 100 para expresarlas en %.

$$\frac{11,027}{11,308} - 1 = -0,0248 \quad \dots \quad \frac{13}{12,033} - 1 = 0,0804$$

- Se puede calcular de dos formas:

- Como media geométrica sobre las tasas de variación anual (sumándole 1) de los beneficios reales

$$\sqrt[5]{(1-0,0248)(1-0,0696)(1+0,0788)(1+0,0872)(1+0,0804)} - 1 = 1,0283 - 1 = 0,0283$$

- A partir del cociente entre los beneficios reales en 2010 y 2005

$$\sqrt[5]{\frac{13}{11,308}} - 1 = 0,0283$$

De ambas formas se obtiene que la tasa media de variación anual de los beneficios reales es del 2,83%.

2. Los beneficios, en millones de euros, obtenidos por una empresa y los valores del IPC se muestran en la tabla siguiente:

Año	2005	2006	2007	2008	2009	2010	2011
Beneficios	2,3	3,7	4,5	5,3	6	6,2	6,4
IPC	100	105	108	115			
IPC				100	102	104	107

- Obtenga los beneficios de los distintos años en millones de euros de 2011.
- Calcule la tasa de variación anual media de los beneficios en términos reales.

Solución:

- Multiplicando la segunda serie de IPC por 1,15 se enlaza con la primera serie de IPC. Dividiendo la serie enlazada por 1,2305 se cambia de base a 2011. Dividiendo los beneficios nominales o corrientes por esta última serie de IPC (en tanto por 1) se obtienen los beneficios reales en euros de 2011.

Año	2005	2006	2007	2008	2009	2010	2011
Beneficio	2,3	3,7	4,5	5,3	6	6,2	6,4
IPC	100	105	108	115	117,3	119,6	123,05
IPC				100	102	104	107
IPC(2011)	81,27	85,33	87,77	93,46	95,33	97,20	100,00
Beneficio (2011)	2,830	4,336	5,127	5,671	6,294	6,379	6,400

b)

$$\sqrt[6]{\frac{6,4}{2,83}} - 1 = 0,1457 \Rightarrow \text{incremento anual medio de los beneficios en términos reales de un } 14,57\%$$

3. Los beneficios (en miles de euros) obtenidos en una pequeña empresa y la evolución del IPC se muestran en la tabla siguiente:

Años	2006	2007	2008	2009	2010	2011
Beneficios	25,8	28,3	30,1	33,9	39	42
IPC	125	128	130			
IPC			100	104	107	109

- Obtenga los beneficios en euros constantes del año 2011.
- Obtenga las tasas de variación que, de año en año, han sufrido los beneficios reales, así como la tasa media anual de variación de los beneficios reales.

Solución:

- En primer lugar enlazamos las dos series de números índices (multiplicando por 1,30 la segunda serie de IPC), a continuación cambiamos de base al año 2011 (dividiendo la serie enlazada de IPC por 1,417). Dividiendo los beneficios por esta última serie de números índices (en tanto por uno) se obtienen los beneficios en euros constantes del año 2011.

Años	2006	2007	2008	2009	2010	2011
Beneficios	25,8	28,3	30,1	33,9	39	42
IPC	125	128	130	135,2	139,1	141,7
IPC			100	104	107	109
IPC (base 2011)	88,21	90,33	91,74	95,41	98,17	100,00
Beneficios en miles de euros de 2011	29,25	31,33	32,81	35,53	39,73	42,00

- Dividimos los beneficios reales (beneficios en miles de euros de 2011) de cada año sobre los del año anterior y le restamos 1 para obtener las tasas de variación anual. Si quisiéramos expresarlas en tantos por ciento se multiplicarían por 100.

Años	2006	2007	2008	2009	2010	2011
Tasas de variación anual real		0,0712	0,0472	0,0829	0,1182	0,0572
Tasas de variación anual real en %		7,12	4,72	8,29	11,82	5,72

La tasa media anual de variación de los beneficios reales se puede obtener de dos formas:

A partir de las tasas de variación anual reales o a partir de los valores inicial y final de los beneficios reales.

$$TM = \sqrt[5]{(1+T_1(2007))(1+T_1(2008))...(1+T_1(2011))} - 1 = \sqrt[5]{\frac{B_{2011}}{B_{2006}}} - 1$$

Según la primera:

$$TM = \sqrt[5]{1,0712 \times 1,0472 \times 1,0829 \times 1,1182 \times 1,0572} - 1 = \sqrt[5]{1,436} - 1 = 0,075$$

Según la segunda:

$$\sqrt[5]{\frac{B_{2011}}{B_{2006}}} - 1 = \sqrt[5]{\frac{42}{29,25}} - 1 = \sqrt[5]{1,436} - 1 = 0,075$$

Luego la tasa media anual de variación de los beneficios reales ha sido del 7,5%.

- El beneficio anual, en millones de euros corrientes, de una empresa y los índices de precios al consumo han sido:

Año	Millones de euros corrientes	IPC	IPC
2007	1,40	110	-
2008	1,45	115	-
2009	1,52	118	110
2010	1,60	-	105
2011	1,64	-	107
2012	1,80	-	110

- a) Calcule el beneficio total, para el periodo 2007-2012, en euros constantes del 2012.
b) La tasa media de variación anual del beneficio en términos reales.

Solución:

- a) Para enlazar las dos series de IPC, dividimos la primera serie por 1,18 y la multiplicamos por 1,10.

Año	IPC	IPC
2007	110	102,5424
2008	115	107,2034
2009	118	110
2010	-	105
2011	-	107
2012	-	110

Cambiamos de base la serie de IPC enlazada dividiéndola por 1,10.

Año	IPC (base 2012)
2007	93,2203
2008	97,4576
2009	100
2010	95,4545
2011	97,2727
2012	100

Dividimos los beneficios en euros corrientes entre los IPC con base 2012 (expresados en tanto por 1), obteniendo los beneficios en euros constantes del 2012 que suman un total para el período 2007-2012 de 9,6718 millones de euros.

Año	Millones de euros constantes (2012)
2007	1,5018
2008	1,4878
2009	1,5200
2010	1,6762
2011	1,6860
2012	1,8000

- b) A partir de la tabla de los beneficios en euros constantes o términos reales obtenemos la tasa de variación anual media como

$$\sqrt[5]{\frac{B_{2012}}{B_{2007}}} - 1 = \sqrt[5]{\frac{1,8}{1,5018}} - 1 = \sqrt[5]{1,19856} - 1 = 0,0369 \Rightarrow 3,69\%$$

5. El gasto en medicamentos de una clínica en euros corrientes del período 2008-2012 y el índice de precios al consumo han sido:

Año	2008	2009	2010	2011	2012
Gasto en miles de euros	18	21	20	25	27

Año	2008	2009	2010	2011	2012
IPC	125	130			
IPC		100	105	120	124

- Obtenga el gasto total en medicamentos en euros de 2012.
- Calcule la tasa de crecimiento medio anual del gasto en medicamentos en términos reales.

Solución:

- Enlazamos las dos series de IPC dividiendo la primera por 1,30.

Año	2008	2009	2010	2011	2012
IPC	96,1538	100	105	120	124

Cambiamos los IPC a la base 2012 dividiendo la serie enlazada de IPC por 1,24.

Año	2008	2009	2010	2011	2012
IPC (base 2012)	77,5434	80,6452	84,6774	96,7742	100

Obtenemos el gasto en medicamentos en euros de 2012 dividiendo el gasto corriente o nominal entre los valores de la serie de IPC con base 2012 (expresada en tantos por 1).

Año	2008	2009	2010	2011	2012
Gasto real en miles de euros de 2012	23,213	26,040	23,619	25,833	27

Lo que nos da un gasto real total en medicamentos durante el período 2008-2012 de 125705 euros de 2012.

- A partir de la tabla del gasto real obtenemos la tasa de crecimiento anual medio como

$$\sqrt[4]{\frac{G_{2012}}{G_{2008}}} - 1 = \sqrt[4]{\frac{27}{23,213}} - 1 = \sqrt[4]{1,1631} - 1 = 0,0385 \Rightarrow 3,85\%$$

6. Los beneficios distribuidos por una sociedad anónima y los índices de precios en el período 2008-2012 fueron:

Año	2008	2009	2010	2011	2012
Beneficio (€)	4875	6742	9524	8635	7421
IPC (2005)	132	135	137		
IPC (2010)			100	103	108

- Enlace en una serie única el IPC con base en 2008.
- Obtenga la evolución real de los beneficios en euros constantes del 2012.
- Calcule la tasa de crecimiento medio anual de los beneficios distribuidos en el período 2008-2012 en euros corrientes o nominales y en euros constantes o reales.

Solución:

- Enlazamos ambas series con base en 2010 dividiendo la primera serie de IPC por 1,37. A continuación cambiamos a la base 2008 dividiendo esta última serie por 0,963504

Año	2008	2009	2010	2011	2012
IPC (2010)	96,3504	98,5401	100	103	108
IPC (2008)	100	102,2727	103,7878	106,9015	112,0909

- b) Obtenemos en primer lugar la serie de IPC con base en 2012 dividiendo la serie IPC (2010) por 1,08. Posteriormente se dividen los beneficios nominales entre los valores de esta última serie de IPC (2012), expresada en tantos por 1, y se obtienen los beneficios reales en euros constantes de 2012.

Año	2008	2009	2010	2011	2012
IPC (2012)	89,2133	91,2409	92,5926	95,3704	100
Beneficio en euros constantes de 2012	5464,43	7389,23	10285,92	9054,17	7421

- c) A partir de la tabla del enunciado, con los **beneficios** corrientes o **nominales** obtenemos

$$\sqrt[4]{\frac{BN_{2012}}{BN_{2008}}} - 1 = \sqrt[4]{\frac{7421}{4875}} - 1 = \sqrt[4]{1,5223} - 1 = 0,1108 \Rightarrow 11,08\%$$

A partir de la tabla obtenida en el apartado b, con los **beneficios reales** en euros constantes de 2012 obtenemos

$$\sqrt[4]{\frac{BR_{2012}}{BR_{2008}}} - 1 = \sqrt[4]{\frac{7421}{5464,43}} - 1 = \sqrt[4]{1,3581} - 1 = 0,0795 \Rightarrow 7,95\%$$

Aunque en valores nominales se observe un crecimiento medio anual de los beneficios del 11,08%, el crecimiento medio anual real ha sido del 7,95%.

7. Dadas las siguientes series de números índices y salario medio de una región en miles de euros corrientes

Año	IPC (2005)	IPC (2009)
2008	107	
2009	113	
2010		110
2011		120
2012		123

Salario medio
1,1
1,7
1,9
2,1
2,2

- a) Complete ambas series de números índices.
b) Estudie el valor real de los salarios medios en dicho período, en euros constantes de 2012.
c) Calcule la tasa de crecimiento medio anual de los salarios en términos reales.

Solución:

- a) En la segunda serie de IPC añadimos el valor 100 para el año 2009 dado que ese es el año base.

Multiplicando todos los índices de la segunda serie por 1,13 completamos la primera serie de IPC y dividiendo por 1,13 los índices de la primera serie completamos la segunda serie de IPC.

Año	IPC (2005)	IPC (2009)
2008	107	94,69
2009	113	100
2010	124,3	110
2011	135,6	120
2012	138,99	123

- b) El valor real de los salarios medios en euros constantes de 2012 se obtiene dividiendo el salario medio nominal del enunciado entre la serie de IPC con base en 2012 (que obtenemos de la serie IPC con base en 2009 dividiéndola por 1,23), expresada en tantos por uno.

Año	IPC (2012)	Salarios medios reales
2008	76,9837	1,43
2009	81,3008	2,09
2010	89,4309	2,12
2011	97,5610	2,15
2012	100	2,20

- c) A partir de la tabla de salarios medios reales obtenemos la tasa de crecimiento anual medio como

$$\sqrt[4]{\frac{S_{2012}}{S_{2008}}} - 1 = \sqrt[4]{\frac{2,2}{1,43}} - 1 = \sqrt[4]{1,5385} - 1 = 0,1137 \Rightarrow 11,37\%$$

8. La recaudación anual, en euros, de un ayuntamiento en el período 2008-2012 junto con los índices de precios correspondientes han sido:

Años	2008	2009	2010	2011	2012
Recaudación	20000	21500	22200	21900	22800
IPC	100	102,5	104		
IPC			100	103	105

- a) Obtenga una serie única para el IPC con año base 2009.
b) Obtenga la recaudación total del ayuntamiento en el período 2008-2012 expresada en euros constantes de 2012.
c) ¿Cuál ha sido la tasa de crecimiento medio anual de la recaudación del ayuntamiento a precios corrientes y a precios constantes?
d) Calcule las tasas anuales de variación real y nominal en la recaudación del ayuntamiento.

Solución:

- a) Dividiendo la primera serie de IPC por 1,04 la enlazamos con la segunda, posteriormente dividiendo esta por 0,985577 se obtiene la serie de IPC con base en 2009.

Años	2008	2009	2010	2011	2012
IPC	96,1538	98,5577	100	103	105
IPC (base 2009)	97,5610	100	101,4634	104,5073	106,5366

- b) Cambiamos la base de la última serie de IPC (2009) al año 2012 dividiéndola por 1,065366. Seguidamente dividimos la recaudación nominal por esta serie de IPC (2012), expresada en

tantos por uno, obteniendo así la recaudación en euros de 2012, siendo el total para el período 2008-2012 de 113180,63 euros de 2012

Años	2008	2009	2010	2011	2012
IPC (2012)	91,5751	93,8644	95,2381	98,0952	100
Recaudación en euros de 2012	21840	22905,37	23310,01	22325,25	22800

- c) A partir de la tabla del enunciado, con la **recaudación** corriente o **nominal** obtenemos

$$\sqrt[4]{\frac{RN_{2012}}{RN_{2008}}} - 1 = \sqrt[4]{\frac{22800}{20000}} - 1 = \sqrt[4]{1,14} - 1 = 0,0333 \Rightarrow 3,33\%$$

A partir de la tabla obtenida en el apartado b, con la **recaudación real** en euros constantes de 2012 obtenemos

$$\sqrt[4]{\frac{RR_{2012}}{RR_{2008}}} - 1 = \sqrt[4]{\frac{22800}{21840}} - 1 = \sqrt[4]{1,043956} - 1 = 0,0108 \Rightarrow 1,08\%$$

Aunque en valores nominales se observe un crecimiento medio anual de la recaudación del 3,33%, el crecimiento medio anual real ha sido sólo del 1,08%.

- d) Las tasas anuales de variación se obtienen comparando por cociente un año con el anterior y restándole 1 (multiplicando por 100 se expresa en tantos por ciento):

$$T_1(t) = \frac{x_t}{x_{t-1}} - 1$$

Años	2008	2009	2010	2011	2012
Recaudación nominal	20000	21500	22200	21900	22800
Tasas de variación anual nominal (%)	-	7,50	3,26	-1,35	4,11
Recaudación en euros de 2012	21840	22905,37	23310,01	22325,25	22800
Tasas de variación anual real (%)	-	4,88	1,77	-4,22	2,13

Las tasas de variación anual real no dependen del año al que estén referidos los euros constantes.

9. La siguiente tabla recoge el IPC de los seis últimos meses del año 2011 y las ventas nominales de un establecimiento en miles de euros:

	julio	agosto	septiembre	octubre	noviembre	diciembre
IPC enero 2005	113,9	114,2	114,6	114,6		
IPC				100	99,91	99,82
Ventas	120	126	114	132	140	136

- a) Obtenga una sola serie de números índices con base en el mes de agosto.
b) Calcule el total de ventas del período julio-diciembre en euros constantes del mes de diciembre.

Solución:

- a) Enlazamos las dos series de IPC multiplicando la segunda por 1,146 y cambiamos a la base agosto 2011 dividiendo la serie enlazada por 1,142.

	julio	agosto	septiembre	octubre	noviembre	diciembre
IPC enero 2005	113,9	114,2	114,6	114,6	114,4969	114,3937
IPC agosto 2011	99,74	100	100,35	100,35	100,26	100,17

- b) Para expresar las ventas en euro constantes de diciembre 2011 hallamos el IPC con base en diciembre de 2011 (dividiendo cualquiera de las dos series del apartado a por 1,143937 o por 1,0017 respectivamente) y dividimos por dichos valores (expresados en tantos por uno) las ventas nominales.

	julio	agosto	septiembre	octubre	noviembre	diciembre
IPC diciembre 2011	99,57	99,83	100,18	100,18	100,09	100,00
Ventas en euros de diciembre 2011	120,52	126,21	113,80	131,76	139,87	136

El total de ventas del período julio-diciembre en euros constantes del mes de diciembre es la suma de esta última fila (768,16 miles de euros, 768160 €).

10. La siguiente tabla recoge el IPC y las ventas anuales de una panadería en miles de euros corrientes

Año	IPC (2007)	IPC (2010)	Ventas
2008	102,7		120
2009	104,1		145
2010	106,9	100	160
2011		104,1	165
2012		107,0	210

- a) Deflacte la serie de ventas anuales, tomando como base el año 2008. Interprete el resultado.
b) Obtenga la tasa media de variación anual de las ventas en euros constantes. Interprete el resultado.

Solución:

- a) Enlazamos ambas series de IPC multiplicando la segunda por 1,069 y a continuación cambiamos a la base 2008 dividiendo todos sus valores por 1,027. Por último se dividen las ventas corrientes sobre la última serie (IPC 2008) expresada en tantos por uno.

Año	IPC (2007)	IPC (2008)	Ventas deflactadas (miles de euros constantes 2008)
2008	102,7	100	120
2009	104,1	101,3632	143,050
2010	106,9	104,0896	153,714
2011	111,2829	108,3573	152,274
2012	114,383	111,3759	188,551

Las ventas deflactadas nos indican el valor de estas en cada uno de los años si los precios del año 2008 hubieran permanecido constantes el resto de los años.

- b) En este apartado no se especifica el año al que se refieren los euros constantes, pero no importa pues la tasa media de variación anual de las ventas es la misma sea cual sea el año al que estén referidos dichos euros constantes. Usaremos los valores en euros constantes del año 2008 obtenidos en el apartado anterior.

$$\sqrt[4]{\frac{V_{2012}}{V_{2008}}} - 1 = \sqrt[4]{\frac{188,551}{120}} - 1 = \sqrt[4]{1,57126} - 1 = 0,1196 \Rightarrow 11,96\%$$

El valor real de las ventas ha pasado de 120000€ a 188551€ en el período 2008-2012, lo que equivale a un aumento anual medio del 11,96% de su valor.

11. Los salarios medios de una empresa y los índices de precios al consumo en los años 2007-2012 fueron:

Años	2007	2008	2009	2010	2011	2012
Salarios (€)	1520	1580	1600	1630	1640	1840
Índices	140	162	175	190	200	205

- Estudie el valor real del salario en euros de 2011.
- Calcule la variación porcentual anual del salario nominal y del salario real
- ¿Cuál ha sido la tasa media anual acumulativa de variación de los salarios en términos reales?

Solución:

- Cambiamos de base en la serie de índices al año 2011 dividiendo la serie del enunciado por 2,00 y seguidamente dividimos los salarios medios en euros corrientes por la nueva serie de índices (expresada en tantos por uno) para obtener los salarios en euros constantes de 2011.

Años	2007	2008	2009	2010	2011	2012
Índices (2011)	70	81	87,5	95	100	102,5
Salarios en euros constantes de 2011	2171,43	1950,62	1828,57	1715,79	1640,00	1795,12

- Las tasas anuales de variación se obtienen comparando por cociente un año con el anterior y restándole 1 (multiplicando por 100 se expresa en porcentaje):

$$T_1(t) = \frac{x_t}{x_{t-1}} - 1$$

Años	2007	2008	2009	2010	2011	2012
Salario nominal	1520	1580	1600	1630	1640	1840
Tasas de variación anual nominal (%)	-	3,95	1,27	1,88	0,61	12,20
Salario real en euros de 2011	2171,43	1950,62	1828,57	1715,79	1640,00	1795,12
Tasas de variación anual real (%)	-	-10,17	-6,26	-6,17	-4,42	9,46

Las tasas de variación anual real no dependen del año al que estén referidos los euros constantes.

- La tasa de variación media anual real es la misma sea cual sea el año al que estén referidos los euros constantes. Consideraremos los valores en euros constantes del año 2011 obtenidos en el apartado a

$$\sqrt[5]{\frac{SR_{2012}}{SR_{2007}}} - 1 = \sqrt[5]{\frac{1795,12}{2171,43}} - 1 = \sqrt[5]{0,8267} - 1 = -0,0373 \Rightarrow -3,73\%$$

El valor real de los salarios (en euros de 2011) ha pasado de 2171,43€ a 1795,12€ en el período 2007-2012, lo que equivale a una disminución anual media del 3,73% de su valor.

12. La siguiente tabla recoge el IPC en los meses de enero y diciembre de 2011 para cada uno de los grupos que constituyen la cesta de la compra.

Grupo	Alimentación	Vestido	Vivienda	Menaje	Salud	Transporte	Cultura	Otros
Enero	175,4	179,7	163	160,4	167,9	161,1	161,1	192,7
Diciembre	181,1	188,4	170,6	167,3	178,2	166,5	171	205,3

El índice global de precios se calcula como una media ponderada de los índices de precios de cada grupo, donde las ponderaciones en tantos por ciento son:

Grupo	Alimentación	Vestido	Vivienda	Menaje	Salud	Transporte	Cultura	Otros
Ponderación (%)	33	8,7	18,6	7,4	2,4	14,4	7	8,5

Obtenga la tasa de variación del IPC global en el período enero 2011-diciembre 2011.

Solución:

Calculamos el índice global en enero:

$$\frac{1}{100}((175,4 \times 33) + (179,7 \times 8,7) + (163 \times 18,6) + (160,4 \times 7,4) + (167,9 \times 2,4) + (161,1 \times 14,4) + (161,1 \times 7) + (192,7 \times 8,5)) = 170,59$$

Calculamos el índice global en diciembre:

$$\frac{1}{100}((181,1 \times 33) + (188,4 \times 8,7) + (170,6 \times 18,6) + (167,3 \times 7,4) + (178,2 \times 2,4) + (166,5 \times 14,4) + (171 \times 7) + (205,3 \times 8,5)) = 177,94$$

La tasa de variación del IPC global en el período enero 2011-diciembre 2011 es

$$\frac{177,94}{170,59} - 1 = 0,0431 \Rightarrow 4,31\%$$

13. La siguiente tabla muestra información sobre precios y cantidades correspondientes a los artículos que se indican

Año	Leche		Pan		Carne		Pescado	
	p_i	q_i	p_i	q_i	p_i	q_i	p_i	q_i
2009	0,60	1442	0,27	21315	7,50	19432	5	8470
2010	0,68	1818	0,32	29997	8,25	21013	5,58	12220
2011	0,70	1925	0,35	22721	9,50	27947	5,63	10513
2012	0,75	1980	0,40	26346	10	31084	6,19	7758

Empleando el índice de Paasche para estos productos, se pide:

- Si una familia en 2009 ha destinado 5000 unidades monetarias (u.m.) para la compra de estos artículos y 7500 u.m. en 2012, ¿puede decirse que ha habido incremento real en dicho presupuesto? ¿De qué porcentaje?
- Si en 2009 una cierta cantidad de carne costaba 17,50 u.m. ¿cuánto costaría comprar esa misma cantidad en 2012?

Solución: Calculamos el índice de precios de Paasche

$$P_{t/0}^p = \frac{\sum_{i=1}^n p_{it} q_{it}}{\sum_{i=1}^n p_{i0} q_{it}} = \frac{370885,42}{280221,42} = 1,3235$$

	2009		2012			
	p_{i0}	q_{i0}	p_{it}	q_{it}	$p_{i0}q_{it}$	$p_{it}q_{it}$
Leche	0,6	1442	0,75	1980	1188	1485
Pan	0,27	21315	0,4	26346	7113,42	10538,4
Carne	7,5	19432	10	31084	233130	310840
Pescado	5	8470	6,19	7758	38790	48022,02
					280221,42	370885,42

- a) Con el índice de precios de Paasche deflactamos las 7500 u.m. de 2012 a u.m. de 2009,

$$\frac{7500}{1,3235} = 5666,79 \text{ u.m. de 2009}$$

Ha habido un incremento real en dicho presupuesto de 666,79 u.m. de 2009 lo que representa un incremento porcentual del 13,34%

$$\frac{5666,79}{5000} - 1 = 0,1334 \Rightarrow 13,34\%$$

- b) Para la magnitud simple “*precio de la carne*” el índice simple o elemental en el año 2012 respecto del año 2009 es $\frac{10}{7,5} = 1,3333$, lo que indica que dicho precio ha aumentado su valor un 33,33% en el citado período. Lo que costaba 17,50 u.m. en 2009 costaría $17,50 \times 1,3333 = 23,33$ u.m. en 2012.

4. ANÁLISIS DESCRIPTIVO DE SERIES CRONOLÓGICAS.

4.1 Definición de una serie cronológica. Representaciones numérica y gráfica.

Una **serie cronológica o temporal** es una sucesión de observaciones de una variable Y , ordenadas en el tiempo, habitualmente obtenidas en períodos de tiempo de la misma duración.

Gran parte de los datos publicados sobre actividades económicas tienen esta forma.

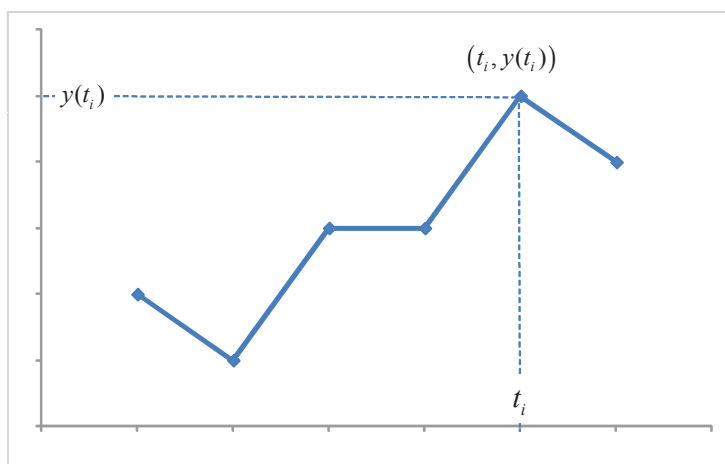
Es frecuente representar dichos datos en una tabla como la que sigue

Tiempo	Observaciones
t_1	$y(t_1) = y_1$
t_2	$y(t_2) = y_2$
...	...
t_i	$y(t_i) = y_i$
...	...
t_n	$y(t_n) = y_n$

Utilizaremos la notación $y(t_i)$ o más abreviada y_i , indistintamente, según se quiera o no destacar explícitamente la dependencia de la variable Y respecto del tiempo.

Una serie cronológica puede considerarse un tipo especial de variable estadística bidimensional. Por tanto se le pueden aplicar técnicas propias de variables bidimensionales.

Por ejemplo, su representación gráfica es similar a la nube de puntos o diagrama de dispersión, dibujaremos un punto con coordenadas $(t_i, y(t_i))$ por cada observación. Para marcar la evolución (creciente o decreciente) a lo largo del tiempo de las observaciones, se unirán dichos puntos mediante segmentos.



Si dentro de cada período de tiempo t_i se tienen varias observaciones, los datos se representan en una tabla como la siguiente (por ejemplo, observaciones mensuales, trimestrales, ... en cada año)

Año/estaciones	I	...	j	...	s
t_1	y_{11}	...	y_{1j}	...	y_{1s}
...
t_i	y_{i1}	...	y_{ij}	...	y_{is}
...
t_n	y_{n1}	...	y_{nj}	...	y_{ns}

► EJEMPLO 4.1

Las denuncias en las Oficinas de Información al Consumidor en los últimos años han sido (expresadas en decenas de miles)

	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre
2007	5	3,5	7
2008	8	5	9
2009	10	7	10,5
2010	11,5	10	13

4.2 Componentes de una serie cronológica. Modelos.

Para el estudio de una serie cronológica, ésta se descompone en cuatro componentes:

- **Tendencia secular**, $\tau(t)$: Es el movimiento de la serie a largo plazo, es decir, refleja el comportamiento general de la serie. Por ejemplo, la tendencia creciente del IPC.
- **Variación estacional**, $E(t)$: Representa fluctuaciones de la serie en períodos de tiempo que se repiten con una periodicidad conocida. Por ejemplo, los crecimientos y disminuciones en la serie por el hecho de estar en una determinada estación del año.
- **Variación cíclica**, $C(t)$: Representa el comportamiento de la serie de carácter periódico, con períodos de duración diferente, desconocida y en general superior a un año. Por ejemplo, los ciclos económicos con etapas de prosperidad, recesión y recuperación.
- **Variación irregular, residual o aleatoria**, $\varepsilon(t)$: Refleja hechos impredecibles que ocurren aleatoriamente y que normalmente suponen ligeras desviaciones de los valores de la variable respecto de las componentes anteriores, aunque en otras ocasiones no es así (catástrofes como el terremoto de Japón, ...)

El primer problema que se nos plantea es la construcción de un modelo que reuniendo las anteriores componentes explique el comportamiento de la serie cronológica. Básicamente consideramos dos modelos.

- **Modelo aditivo:** Supone que las observaciones se generan como suma de las cuatro componentes

$$Y(t) = \tau(t) + E(t) + C(t) + \varepsilon(t)$$

En este modelo cada componente se expresa en la misma unidad que las observaciones.

La variación irregular es independiente de las demás componentes, es decir, la magnitud de sus valores no depende de las otras componentes.

- **Modelo multiplicativo:** Las observaciones están generadas por el producto de las componentes (modelo multiplicativo puro)

$$Y(t) = \tau(t)E(t)C(t)\varepsilon(t)$$

En este modelo la tendencia secular se expresa en la misma unidad que las observaciones y las demás componentes en tantos por uno. Aquí no se cumple la hipótesis de independencia de la variación irregular respecto de las demás componentes sino que es proporcional al resto de componentes.

En términos generales es más adecuado el modelo multiplicativo que el aditivo para la descripción de fenómenos económicos (por ejemplo, los factores estacionales y cíclicos no afectarán de la misma manera, en términos absolutos, a las ventas de un pequeño comercio y de un hipermercado, sino proporcionalmente al volumen de ventas de cada uno)

Existen varios procedimientos para determinar el tipo de modelo al que responde una serie cronológica. La idea en todos consiste en poner de manifiesto si las fluctuaciones de la serie son aproximadamente constantes o proporcionales al valor de la tendencia. Uno de ellos es:

Análisis de la variabilidad de las diferencias y cocientes estacionales.

Calculamos las diferencias y cocientes estacionales: Para cada estación j se comparan los datos en años consecutivos, $(i-1)$, i , mediante la diferencia y cociente de ambos.

$$d_{ij} = y_{ij} - y_{(i-1)j} \qquad k_{ij} = \frac{y_{ij}}{y_{(i-1)j}}$$

A continuación calculamos los coeficientes de variación sobre las diferencias y sobre los cocientes. Si $CV(d) < CV(k)$ se elegirá el modelo aditivo, en caso contrario se optará por el modelo multiplicativo.

► EJEMPLO 4.2

Estudiemos la conveniencia del modelo aditivo o multiplicativo sobre los datos del ejemplo 4.1.

	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre
2007	5	3,5	7
2008	8	5	9
2009	10	7	10,5
2010	11,5	10	13

Solución:

$d_{ij} = y_{ij} - y_{(i-1)j}$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre
2007	---	---	---
2008	$8 - 5 = 3$	$5 - 3,5 = 1,5$	$9 - 7 = 2$
2009	2	2	1,5
2010	1,5	3	2,5

$k_{ij} = \frac{y_{ij}}{y_{(i-1)j}}$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre
2007	---	---	---
2008	$\frac{8}{5} = 1,6$	$\frac{5}{3,5} = 1,43$	$\frac{9}{7} = 1,29$
2009	1,25	1,4	1,17
2010	1,15	1,43	1,24

Con los datos de cada tabla calculamos la media, varianza y desviación típica para finalmente hallar el coeficiente de variación:

$$\bar{d} = 2,111 \quad S_d^2 = 0,321 \quad S_d = 0,567 \quad \Rightarrow \quad CV(d) = \frac{S_d}{\bar{d}} = 0,268$$

$$\bar{k} = 1,328 \quad S_k^2 = 0,019 \quad S_k = 0,138 \quad \Rightarrow \quad CV(k) = \frac{S_k}{\bar{k}} = 0,104$$

$CV(d) > CV(k)$, por tanto sería más adecuado el modelo multiplicativo. ◀

4.3 Tendencia secular: ajuste de una recta de mínimos cuadrados y medias móviles.

Estudiaremos dos procedimientos para la determinación de la tendencia secular en una serie cronológica:

- El primer método tiene un **enfoque global**, consiste en el ajuste de una **recta de mínimos cuadrados** al conjunto de todas las observaciones (podría considerarse el ajuste de otra función si las características de la serie así lo indicaran).

- El segundo método tiene un **enfoque local**, sólo se utilizan algunas observaciones para el cálculo de la tendencia en cada período mediante la media de dichas observaciones (**medias móviles**).

Es aconsejable utilizar métodos locales en las previsiones a corto plazo porque se adaptan mejor y más rápidamente a las circunstancias cambiantes. Sin embargo, en las previsiones a largo plazo donde nos apoyamos en aspectos permanentes de la evolución del fenómeno es mejor usar métodos globales.

Método del ajuste de una recta de mínimos cuadrados.

Considerando una serie cronológica como un caso particular de variable estadística bidimensional, ajustaremos la **recta de regresión de mínimos cuadrados de Y/X** tal y como hemos visto en el tema 2.

► EJEMPLO 4.3

De nuevo utilizaremos los datos del ejemplo 4.1 para obtener la **tendencia secular** mediante el **ajuste de una recta por mínimos cuadrados**.

Para que los datos sobre los que vamos a obtener la tendencia secular contengan fundamentalmente a ésta, se eliminan previamente las oscilaciones debidas a factores estacionales calculando los valores medios anuales (por estación). Y sobre dichos valores realizamos el ajuste.

	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	media anual
2007	5	3,5	7	5,17
2008	8	5	9	7,33
2009	10	7	10,5	9,17
2010	11,5	10	13	11,50

$$\bar{y}_1 = \frac{5 + 3,5 + 7}{3} = 5,167 \quad \dots \quad \bar{y}_4 = \frac{11,5 + 10 + 13}{3} = 11,5$$

t_i	$x_i = t_i - 2006$	$y_i = \text{media anual}$	x_i^2	$x_i y_i$
2007	1	5,167	1	5,167
2008	2	7,333	4	14,667
2009	3	9,167	9	27,5
2010	4	11,50	16	46
totales	10	33,167	30	93,334

Cuando se trabaja con años, un cambio de origen del tipo $x_i = t_i - 2006$ facilita notablemente los cálculos.

Exactamente igual que en el tema 2, obtenemos la **recta de regresión de Y/X**:

$$n = 4 \quad \bar{x} = \frac{10}{4} = 2,5 \quad \bar{y} = \frac{33,17}{4} = 8,2917$$

$$S_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{30}{4} - 2,5^2 = 1,25$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{93,334}{4} - (2,5 \times 8,2917) = 2,604$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Leftrightarrow y - 8,2917 = \frac{2,604}{1,25} (x - 2,5) \Leftrightarrow y = 3,08 + 2,083x$$

Finalmente deshacemos el cambio de origen $x_i = t_i - 2006$ y expresamos y como la tendencia secular $\tau(t)$ a la que representa:

$$y = 3,08 + 2,083x \Leftrightarrow \tau(t) = 3,08 + 2,083(t - 2006) \Leftrightarrow \tau(t) = -4175,42 + 2,083t \quad \blacktriangleleft$$

Método de las medias móviles.

En este método no se supone una forma funcional para la tendencia (ni una recta ni cualquier otro tipo de curva).

El método de las medias móviles es un método de **suavizamiento** de la serie cronológica que transforma las observaciones originales (con una representación gráfica típica de dientes de sierra) en unos valores con menores fluctuaciones. Para la aplicación de este método se van calculando sucesivamente medias aritméticas sobre subconjuntos de datos originales, en cada nueva media se elimina la observación más antigua e introduce la siguiente observación, avanzando así desde la primera hasta la última observación (de ahí su calificativo de **móviles**).

Las medias móviles que utilizan igual ponderación para cada una de las observaciones se denominan *medias móviles no ponderadas*, otra posibilidad consiste en usar ponderaciones distintas para darle más importancia a las observaciones centrales en el cálculo de cada media (*medias móviles ponderadas*).

Se denomina **media móvil de amplitud h** a la que se calcula sobre h observaciones.

Cuando este método se utiliza en la obtención de la tendencia secular, el conjunto de datos sobre los que se calcula cada media móvil debe contener a todas las estaciones para eliminar los altibajos debidos a factores estacionales (h debe ser igual al número de estaciones o un múltiplo de éste).

El valor de cada media móvil se asocia al período central de los períodos sobre los que se ha calculado. Si h es impar el período central está clara y unívocamente determinado (ejemplo 4.4), sin embargo, cuando h es par hay dos períodos centrales y la media móvil se asocia al punto intermedio entre ambos. Para que las medias móviles siempre estén referidas a los mismos períodos que las observaciones originales, en este último caso es necesario proceder a *centrar las medias móviles*.

para lo que se volverán a calcular medias móviles de amplitud 2 sobre las medias móviles de amplitud h (ejemplo 4.5).

► EJEMPLO 4.4

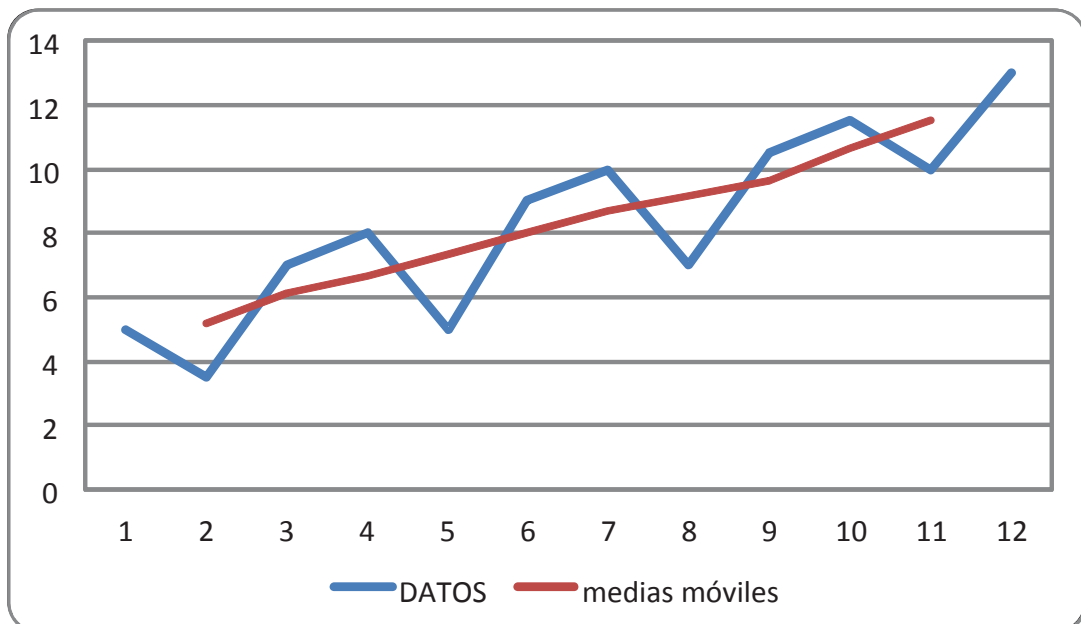
Utilizamos de nuevo los datos del ejemplo 4.1 para obtener la **tendencia secular** mediante **medias móviles**.

	1° cuatrimestre	2° cuatrimestre	3° cuatrimestre
2007	5	3,5	7
2008	8	5	9
2009	10	7	10,5
2010	11,5	10	13

Calculando medias móviles de amplitud 3 incluiremos en su obtención una observación de cada cuatrimestre, compensando (suavizando como puede verse en el gráfico) así los mayores y menores valores que presentan las distintas estaciones.

$$\frac{5+3,5+7}{3}=5,167 \quad \frac{3,5+7+8}{3}=6,167 \quad \frac{7+8+5}{3}=6,667 \quad \dots$$

medias móviles	1° cuatrimestre	2° cuatrimestre	3° cuatrimestre
2007	---	5,167	6,167
2008	6,667	7,333	8
2009	8,667	9,167	9,667
2010	10,667	11,5	---



► EJEMPLO 4.5

Obtenga la *tendencia secular* mediante *medias móviles*.

	1º trimestre	2º trimestre	3º trimestre	4º trimestre
2006	7	5	2,4	1,1
2007	10	7,1	3,6	1,4
2008	13,5	8,9	4,5	1,6
2009	17	11,2	5,3	2
2010	18,1	12,7	6,6	2,4

Calculamos las medias móviles de amplitud 4 que estarán asociadas al punto central de los 4 trimestres sobre los que se calcula:

media móvil amplitud 4	1º trimestre	2º trimestre	3º trimestre	4º trimestre	1º trimestre (año siguiente)
2006		3,875	4,625	5,15	
2007	5,45	5,525	6,4	6,85	
2008	7,075	7,125	8,0	8,575	
2009	8,775	8,875	9,15	9,525	
2010	9,85	9,95			

donde las medias se han calculado de la siguiente forma:

$$3,875 = \frac{7 + 5 + 2,4 + 1,1}{4} \quad 4,625 = \frac{5 + 2,4 + 1,1 + 10}{4} \quad 5,15 = \frac{2,4 + 1,1 + 10 + 1,7}{4} \quad \dots$$

A continuación procedemos a centrar las medias móviles en los mismos períodos que las observaciones de la serie, tomando medias móviles de amplitud 2 sobre las anteriores medias móviles de amplitud 4:

$$4,25 = \frac{3,875 + 4,625}{2} \quad 4,8875 = \frac{4,625 + 5,15}{2} \quad 5,3 = \frac{5,15 + 5,45}{2} \quad \dots$$

medias móviles amplitud 2	1º trimestre	2º trimestre	3º trimestre	4º trimestre
2006	---	---	4,25	4,8875
2007	5,3	5,4875	5,9625	6,625
2008	6,9625	7,1	7,5625	8,2875
2009	8,675	8,825	9,0125	9,3375
2010	9,6875	9,9	---	---

4.4 Variación estacional. Desestacionalización.

En el *modelo multiplicativo* la componente estacional de una serie cronológica se mide con un índice adimensional denominado **índice de variación estacional**, éste se expresa en porcentajes e indica la fluctuación del valor de la serie en dicha estación respecto del valor de la tendencia. Por ejemplo, un índice de variación estacional del 90% indica que en esa estación hay una disminución del 10% en relación al valor de la tendencia.

En el *modelo aditivo* la componente estacional indica en términos absolutos (expresada en las mismas unidades que la variable observada) la cantidad en que se ha superado (si es positiva) o no se ha alcanzado (si es negativa) la tendencia. Por ejemplo, un valor de la componente estacional de -150€ indica que a esa estación le corresponde un valor 150€ por debajo del valor de la tendencia.

Vamos a estudiar **tres procedimientos** para la obtención de la **variación estacional**:

Método de las medias simples.

Denominado también *método de las relaciones de las medias estacionales respecto a la tendencia*.

► EJEMPLO 4.6

Explicamos el método con los datos del ejemplo 4.1.

	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	media anual
2007	5	3,5	7	5,167
2008	8	5	9	7,333
2009	10	7	10,5	9,167
2010	11,5	10	13	11,50

1. Sobre los valores medios anuales ajustamos la recta de regresión de mínimos cuadrados como se hizo en el ejemplo 4.3.

$$y = 3,08 + 2,083x \Leftrightarrow \tau(t) = 3,08 + 2,083(t - 2006) \Leftrightarrow \tau(t) = -4175,42 + 2,083t$$

La pendiente b de la recta estima lo que varia la tendencia por unidad de tiempo (un año en nuestro caso), por tanto la tendencia variará $\frac{b}{s}$ por cada estación que transcurra (s =número de estaciones).

En este ejemplo la tendencia crece $\frac{2,083}{3} = 0,694$ cada cuatrimestre.

2. Calculamos los valores medios en cada estación

$$\frac{5+8+10+11,5}{4} = 8,625 \quad \dots \quad \frac{7+9+10,5+13}{4} = 9,875$$

	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre
2007	5	3,5	7
2008	8	5	9
2009	10	7	10,5
2010	11,5	10	13
media por estación	8,625	6,375	9,875
media corregida	8,625	5,681	8,487

Seguidamente eliminamos la tendencia de los anteriores valores medios por estación, restando tantas veces $\frac{b}{s} = \frac{2,083}{3} = 0,694$ como estaciones del año han pasado, obteniendo las medias corregidas:

$$8,625 - \left(0 \times \frac{b}{s}\right) = 8,625 \quad 6,375 - \left(1 \times \frac{b}{s}\right) = 5,681 \quad 9,875 - \left(2 \times \frac{b}{s}\right) = 8,487$$

3. Calculamos la media de las medias corregidas (media global corregida):

$$7,5977 = \frac{8,625 + 5,681 + 8,487}{3}$$

Comparando las medias corregidas con su promedio (por cociente o diferencia según el modelo sea multiplicativo o aditivo) obtenemos la variación estacional.

					Media global corregida
	media corregida	8,625	5,681	8,487	7,5977
I.V.E.	modelo multiplicativo	113,52%	74,77%	111,70%	
	modelo aditivo	1,0273	-1,9167	0,8893	

$$\frac{8,625}{7,5977} 100 = 113,52 \quad \dots \quad \frac{8,487}{7,5977} 100 = 111,70$$

$$8,625 - 7,5977 = 1,0273 \quad \dots \quad 8,487 - 7,5977 = 0,8893$$

En el modelo multiplicativo la variación estacional se expresa mediante los *índices de variación estacional* (IVE) que son valores adimensionales con la propiedad de que su media es 100 (1 si los IVE están expresados en tantos por uno). En el primer cuatrimestre el número de denuncias es un 13,52% mayor que la tendencia, en el segundo cuatrimestre las denuncias no llegan a ser el 100% de la tendencia sino que es un 25,23% menor ($100 - 74,77 = 25,23$), ...

En el modelo aditivo la variación estacional está expresada en las mismas unidades que la variable observada (decenas de miles de denuncias en este ejemplo) y su media es cero ($1,0273 - 1.9167 + 0,8893 = -0,0001 \neq 0$ debido a pequeños errores de redondeo). En el primer cuatrimestre hay 1,0273 decenas de miles de denuncias más que la tendencia (10273 denuncias más), en el segundo cuatrimestre hay 19167 denuncias menos que la tendencia, ... ◀

► EJEMPLO 4.7

Con el método de las medias simples obtenga la variación estacional de los datos del ejemplo 4.5

1. En primer lugar ajustamos la recta de tendencia sobre las medias anuales:

t_i	$x_i = t_i - 2005$	$y_i = \text{media anual}$	x_i^2	$x_i y_i$
2006	1	3,875	1	3,875
2007	2	5,525	4	11,05
2008	3	7,125	9	21,375
2009	4	8,875	16	35,5
2010	5	9,950	25	49,75
totales	15	35,35	55	121,55

$$n = 5 \quad \bar{x} = \frac{15}{5} = 3 \quad \bar{y} = \frac{35,35}{5} = 7,07$$

$$S_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{55}{5} - 3^2 = 2$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{121,55}{5} - (3 \times 7,07) = 3,1$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Leftrightarrow y - 7,07 = \frac{3,1}{2} (x - 3) \Leftrightarrow y = 2,42 + 1,55x$$

deshacemos el cambio de origen $x_i = t_i - 2005$ y expresamos y como la tendencia secular $\tau(t)$:

$$y = 2,42 + 1,55x \Leftrightarrow \tau(t) = 2,42 + 1,55(t - 2005) \Leftrightarrow \tau(t) = -3105,33 + 1,55t$$

2. Seguidamente eliminamos la tendencia de los valores medios por estación, restando tantas veces $\frac{b}{s} = \frac{1,55}{4} = 0,3875$ como estaciones del año han pasado, obteniendo las medias corregidas:

$$13,12 - \left(0 \times \frac{b}{s} \right) = 13,12 \quad \dots \quad 1,7 - \left(3 \times \frac{b}{s} \right) = 0,5375$$

3. Calculamos la media de las medias corregidas (media global corregida):

$$6,48875 = \frac{13,12 + 8,5925 + 3,705 + 0,5375}{4}$$

Comparando las medias corregidas con su promedio (por cociente o diferencia según el modelo sea multiplicativo o aditivo) obtenemos la variación estacional.

$$\begin{aligned} \frac{13,12}{6,48875} 100 &= 202,196 & \dots & \quad \frac{0,5375}{6,48875} 100 = 8,284 \\ 13,12 - 6,48875 &= 6,631 & \dots & \quad 0,5375 - 6,48875 = -5,951 \end{aligned}$$

En la siguiente tabla se recogen todos los cálculos:

	1º trimestre	2º trimestre	3º trimestre	4º trimestre	media anual
2006	7	5	2,4	1,1	3,875
2007	10	7,1	3,6	1,4	5,525
2008	13,5	8,9	4,5	1,6	7,125
2009	17	11,2	5,3	2	8,875
2010	18,1	12,7	6,6	2,4	9,950
media por estación	13,12	8,98	4,48	1,7	media global corregida
media corregida	13,12	8,5925	3,705	0,5375	6,48875
modelo multiplicativo	202,196%	132,421%	57,099%	8,284%	
modelo aditivo	6,631	2,104	-2,784	-5,951	

En los dos métodos siguientes la idea básica es eliminar de las observaciones el valor de la tendencia secular, para ello es fundamental conocer el modelo que se adapta a la serie cronológica.

En el modelo multiplicativo eliminaremos la tendencia mediante cociente de las observaciones sobre la tendencia

$$Y(t) = \tau(t)E(t)C(t)\varepsilon(t) \Rightarrow \frac{Y(t)}{\tau(t)} = E(t)C(t)\varepsilon(t)$$

Mientras que en el modelo aditivo lo haremos por diferencia

$$Y(t) = \tau(t) + E(t) + C(t) + \varepsilon(t) \Rightarrow Y(t) - \tau(t) = E(t) + C(t) + \varepsilon(t)$$

Posteriormente se promedian estos valores ($E(t)C(t)\varepsilon(t)$ ó $E(t) + C(t) + \varepsilon(t)$) sobre cada una de las estaciones, de esta forma los efectos unas veces positivos, otras negativos, de las componentes $C(t)$ y $\varepsilon(t)$ se anulan, recogiendo en dicho promedio fundamentalmente el valor de la componente estacional.

Método de la razón (o diferencia) a la tendencia.

Explicamos el método con los mismos datos del ejemplo 4.1 en el siguiente ejemplo.

► EJEMPLO 4.8

	1° cuatrimestre	2° cuatrimestre	3° cuatrimestre	media anual
2007	5	3,5	7	5,167
2008	8	5	9	7,333
2009	10	7	10,5	9,167
2010	11,5	10	13	11,50

1. Sobre los valores medios anuales ajustamos la recta de regresión de mínimos cuadrados como se hizo en el ejemplo 4.3.

$$y = 3,08 + 2,083x \Leftrightarrow \tau(t) = 3,08 + 2,083(t - 2006) \Leftrightarrow \tau(t) = -4175,42 + 2,083t$$

2. Estimamos la tendencia secular para cada una de las estaciones de cada año, teniendo presente que la tendencia $\tau(t) = -4175,42 + 2,083t$ para un año t se asigna al período central del año (2° cuatrimestre en este ejemplo) y que la tendencia varía $\frac{b}{s} = \frac{2,083}{3} = 0,694$ cada estación que pasa.

$$\tau(2007) = -4175,42 + (2,083 \times 2007) = 5,161 \rightarrow 2^\circ \text{ cuatrimestre de 2007}$$

...

$$\tau(2010) = -4175,42 + (2,083 \times 2010) = 11,41 \rightarrow 2^\circ \text{ cuatrimestre de 2010}$$

En el 3° cuatrimestre de 2007 la tendencia será $\frac{b}{s} = \frac{2,083}{3} = 0,694$ más que en el 2° cuatrimestre:

$$5,161 + 0,694 = 5,855$$

En el 1° cuatrimestre de 2007 la tendencia será $\frac{b}{s} = \frac{2,083}{3} = 0,694$ menos que en el 2° cuatrimestre:

$$5,161 - 0,694 = 4,467$$

...

Análogamente en el resto de los años.

También podríamos haber estimado la tendencia para todas las estaciones del primer año y sumando a éstas b se obtendría la tendencia para las estaciones del segundo año y así sucesivamente para todos los años.

El método que se sigue en este ejemplo evita la acumulación de errores de redondeo en la estimación de la tendencia.

3. Se elimina de las observaciones, $Y(t)$, el valor de la tendencia, $\tau(t)$, realizando los cocientes $\frac{Y(t)}{\tau(t)}$ si suponemos el modelo multiplicativo y mediante las diferencias $Y(t) - \tau(t)$ si el modelo es aditivo.

En lo que sigue supondremos el **modelo multiplicativo** (más adelante se repetirá todo para el modelo aditivo).

$$\frac{5}{4,467} = 1,119 \quad \dots \quad \frac{13}{12,104} = 1,074$$

4. Calculamos la media por estación de los anteriores valores y la media de dichas medias (que es igual a la media global de todos los cocientes $\frac{Y(t)}{\tau(t)}$). Los índices de variación estacional (I.V.E.) se obtienen como cociente de las medias por estación sobre la *media global* (el resultado se multiplica por 100 para expresarlo en tantos por ciento)

$$\frac{1,119 + 1,221 + 1,158 + 1,073}{4} = 1,1428 \quad \dots \quad \frac{1,196 + 1,134 + 1,048 + 1,074}{4} = 1,1130$$

$$\frac{1,1428 + 0,7488 + 1,1130}{3} = 1,001533$$

$$\frac{1,1428}{1,001533} 100 = 114,11\% \quad \dots \quad \frac{1,1130}{1,001533} 100 = 111,13\%$$

En la siguiente tabla se recogen todos los **cálculos (modelo multiplicativo)**:

$Y(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	5	3,5	7	
2008	8	5	9	
2009	10	7	10,5	
2010	11,5	10	13	
$\tau(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	4,467	5,161	5,855	
2008	6,550	7,244	7,938	
2009	8,633	9,327	10,021	
2010	10,716	11,410	12,104	
$\frac{Y(t)}{\tau(t)}$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	1,119	0,678	1,196	
2008	1,221	0,690	1,134	
2009	1,158	0,751	1,048	
2010	1,073	0,876	1,074	media global
media por estación	1,1428	0,7488	1,1130	1,001533
I.V.E.	114,11%	74,77%	111,13%	

Repetimos los pasos 3 y 4 suponiendo el **modelo aditivo**

3. Se elimina de las observaciones, $Y(t)$, el valor de la tendencia, $\tau(t)$, realizando las diferencias $Y(t) - \tau(t)$.

$$5 - 4,467 = 0,533 \quad \dots \quad 13 - 12,104 = 0,896$$

4. Calculamos la media por estación de los anteriores valores y la media de dichas medias (que es igual a la media global de todas las diferencias $Y(t) - \tau(t)$). La variación estacional (V.E.) se obtiene como diferencia de la media por estación menos la *media global*

$$\frac{0,533 + 1,450 + 1,367 + 0,784}{4} = 1,0335 \quad \dots \quad \frac{1,145 + 1,062 + 0,479 + 0,896}{4} = 0,8955$$

$$\frac{1,0335 - 1,9105 + 0,8955}{3} = 0,0062$$

$$1,0335 - 0,0062 = 1,0273 \quad \dots \quad 0,8955 - 0,0062 = 0,8893$$

En la siguiente tabla se recogen todos los **cálculos (modelo aditivo)**:

$Y(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	5	3,5	7	
2008	8	5	9	
2009	10	7	10,5	
2010	11,5	10	13	
$\tau(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	4,467	5,161	5,855	
2008	6,550	7,244	7,938	
2009	8,633	9,327	10,021	
2010	10,716	11,410	12,104	
$Y(t) - \tau(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	0,533	-1,661	1,145	
2008	1,450	-2,244	1,062	
2009	1,367	-2,327	0,479	
2010	0,784	-1,410	0,896	media global
media por estación	1,0335	-1,9105	0,8955	0,0062
V.E.	1,0273	-1,9167	0,8893	

Repetimos este método tanto para el modelo multiplicativo como aditivo en un ejemplo con un número par de estaciones. La única diferencia con el ejemplo anterior está en la etapa 2 donde el valor de la tendencia en un año t estará asignada al punto medio del año, es decir, entre las dos estaciones centrales. Para referir los valores de la tendencia a los mismos períodos que las

observaciones habrá que centrar las estimaciones de la tendencia en dichos períodos como se expone en el siguiente ejemplo.

► EJEMPLO 4.9.

Utilizamos los datos del ejemplo 4.5.

1. Sobre los valores medios anuales ajustamos la recta de regresión de mínimos cuadrados como se hizo en el ejemplo 4.7

$$y = 2,42 + 1,55x \Leftrightarrow \tau(t) = 2,42 + 1,55(t - 2005) \Leftrightarrow \tau(t) = -3105,33 + 1,55t$$

2. Estimamos la tendencia secular para cada una de las estaciones de cada año, teniendo presente que la tendencia $\tau(t) = -3105,33 + 1,55t$ para un año t se asigna al punto central del año (punto entre el 2º y 3º trimestre, en este ejemplo), que la tendencia varía $\frac{b}{s} = \frac{1,55}{4} = 0,3875$ cada estación que pasa y la mitad $\left(\frac{0,3875}{2} = 0,19375\right)$ si sólo ha transcurrido media estación. Así:

$$\tau(2006) = -3105,33 + (1,55 \times 2006) = 3,97 \rightarrow \text{punto entre 2º y 3º trimestres de 2006}$$

...

$$\tau(2010) = -3105,33 + (1,55 \times 2010) = 10,17 \rightarrow \text{punto entre 2º y 3º trimestres de 2010}$$

La tendencia centrada en el 3º trimestre de 2006 será: $3,97 + 0,19375 = 4,16375$

...

La tendencia centrada en el 3º trimestre de 2010 será: $10,17 + 0,19375 = 10,36375$

En el 4º trimestre de 2006 la tendencia será $\frac{b}{s} = \frac{1,55}{4} = 0,3875$ más que en el 3º trimestre:

$$4,16375 + 0,3875 = 4,55125$$

En el 2º trimestre de 2006 la tendencia será $\frac{b}{s} = \frac{1,55}{4} = 0,3875$ menos que en el 3º trimestre:

$$4,16375 - 0,3875 = 3,77625$$

Y en el 1º trimestre de 2006 la tendencia será $\frac{b}{s} = \frac{1,55}{4} = 0,3875$ menos que en el 2º trimestre:

$$3,77625 - 0,3875 = 3,38875$$

Análogamente en el resto de los años.

Las etapas 3 y 4 son idénticas a las del ejemplo anterior.

En la siguiente tabla se recogen todos los **cálculos (modelo multiplicativo)**:

$Y(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	media anual
2006	7	5	2,4	1,1	3,875
2007	10	7,1	3,6	1,4	5,525
2008	13,5	8,9	4,5	1,6	7,125
2009	17	11,2	5,3	2	8,875
2010	18,1	12,7	6,6	2,4	9,950
$\tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	3,38875	3,77625	4,16375	4,55125	
2007	4,93875	5,32625	5,71375	6,10125	
2008	6,48875	6,87625	7,26375	7,65125	
2009	8,03875	8,42625	8,81375	9,20125	
2010	9,58875	9,97625	10,36375	10,75125	
$\frac{Y(t)}{\tau(t)}$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	2,06566	1,32406	0,57640	0,24169	
2007	2,02480	1,33302	0,63006	0,22946	
2008	2,08052	1,29431	0,61951	0,20912	
2009	2,11476	1,32918	0,60133	0,21736	
2010	1,88763	1,27302	0,63684	0,22323	media global
media por estación	2,03467	1,31072	0,61283	0,22417	1,0456
I.V.E.	194,59%	125,36%	58,61%	21,44%	

En la siguiente tabla se recogen todos los **cálculos (modelo aditivo)**:

$Y(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	media anual
2006	7	5	2,4	1,1	3,875
2007	10	7,1	3,6	1,4	5,525
2008	13,5	8,9	4,5	1,6	7,125
2009	17	11,2	5,3	2	8,875
2010	18,1	12,7	6,6	2,4	9,950
$\tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	3,38875	3,77625	4,16375	4,55125	
2007	4,93875	5,32625	5,71375	6,10125	
2008	6,48875	6,87625	7,26375	7,65125	
2009	8,03875	8,42625	8,81375	9,20125	
2010	9,58875	9,97625	10,36375	10,75125	
$Y(t) - \tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	3,61125	1,22375	-1,76375	-3,45125	
2007	5,06125	1,77375	-2,11375	-4,70125	
2008	7,01125	2,02375	-2,76375	-6,05125	
2009	8,96125	2,77375	-3,51375	-7,20125	
2010	8,51125	2,72375	-3,76375	-8,35125	media global
media por estación	6,63125	2,10375	-2,78375	-5,95125	0
V.E.	6,63125	2,10375	-2,78375	-5,95125	

El **método de las medias simples (modelo aditivo)** y el **método de la diferencia a la tendencia coinciden**, siempre nos conducen a los mismos valores para la variación estacional. (Compárense los resultados obtenidos en los ejemplos 4.6, 4.7, 4.8 y 4.9. Las pequeñas diferencias que se aprecian en este ejemplo en relación al ejemplo 4.7 son debidas a que aquí hemos trabajado con 5 decimales).

Método de la razón (o diferencia) a las medias móviles.

Este método es igual que el anterior con la única diferencia de que aquí se estima la tendencia para cada período (estación) mediante la técnica de suavizamiento de las medias móviles tal y como vimos en los ejemplos 4.4 y 4.5 (según el número de estaciones sea impar o par).

► EJEMPLO 4.10.

Utilizamos los datos del ejemplo 4.4 donde ya se calcularon las medias móviles para estimar la tendencia en cada estación.

Modelo multiplicativo:

$Y(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	5	3,5	7	
2008	8	5	9	
2009	10	7	10,5	
2010	11,5	10	13	
$\tau(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	---	5,167	6,167	
2008	6,667	7,333	8	
2009	8,667	9,167	9,667	
2010	10,667	11,5	---	
$\frac{Y(t)}{\tau(t)}$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	---	0,67738	1,13507	
2008	1,19994	0,68185	1,12500	
2009	1,15380	0,76361	1,08617	
2010	1,07809	0,86957	---	media global
media por estación	1,14394	0,74810	1,11541	1,0025
I.V.E.	114,11%	74,62%	111,26%	

Modelo aditivo:

$Y(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	5	3,5	7	
2008	8	5	9	
2009	10	7	10,5	
2010	11,5	10	13	
$\tau(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	---	5,167	6,167	
2008	6,667	7,333	8	
2009	8,667	9,167	9,667	
2010	10,667	11,5	---	
$Y(t) - \tau(t)$	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	
2007	---	-1,667	0,833	
2008	1,333	-2,333	1	
2009	1,333	-2,167	0,833	
2010	0,833	-1,5	---	media global
media por estación	1,1663	-1,9168	0,8887	0,0461
V.E.	1,1202	-1,9629	0,8426	◀

► EJEMPLO 4.11.

Usamos los datos del ejemplo 4.5 donde se calcularon las medias móviles centradas en cada estación.

Modelo multiplicativo:

$Y(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	7	5	2,4	1,1	
2007	10	7,1	3,6	1,4	
2008	13,5	8,9	4,5	1,6	
2009	17	11,2	5,3	2	
2010	18,1	12,7	6,6	2,4	
$\tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	---	---	4,25	4,8875	
2007	5,3	5,4875	5,9625	6,625	
2008	6,9625	7,1	7,5625	8,2875	
2009	8,675	8,825	9,0125	9,3375	
2010	9,6875	9,9	---	---	
$\frac{Y(t)}{\tau(t)}$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	---	---	0,56471	0,22506	
2007	1,88679	1,29385	0,60377	0,21132	
2008	1,93896	1,25352	0,59504	0,19306	
2009	1,95965	1,26912	0,58807	0,21419	
2010	1,86839	1,28283	---	---	media global
media por estación	1,91345	1,27483	0,58790	0,21091	0,99677
I.V.E.	191,96%	127,90%	58,98%	21,16%	

Modelo aditivo:

$Y(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	7	5	2,4	1,1	
2007	10	7,1	3,6	1,4	
2008	13,5	8,9	4,5	1,6	
2009	17	11,2	5,3	2	
2010	18,1	12,7	6,6	2,4	
$\tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	---	---	4,25	4,8875	
2007	5,3	5,4875	5,9625	6,625	
2008	6,9625	7,1	7,5625	8,2875	
2009	8,675	8,825	9,0125	9,3375	
2010	9,6875	9,9	---	---	
$Y(t) - \tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2006	---	---	-1,85	-3,7875	
2007	4,7	1,6125	-2,3625	-5,225	
2008	6,5375	1,8	-3,0625	-6,6875	
2009	8,325	2,375	-3,7125	-7,3375	
2010	8,4125	2,8	---	---	
media por estación	6,99375	2,14688	-2,74688	-5,75938	media global 0,15859
V.E.	6,83516	1,98828	-2,90547	-5,91797	

Desestacionalización.

En ocasiones interesa conocer las variaciones estacionales y eliminarlas de las observaciones de la serie para poder ver mejor el comportamiento de ésta ajeno a causas estacionales. La eliminación de la componente estacional se conoce como **desestacionalización de la serie** y permite entre otras cosas comparar valores observados en estaciones distintas que están influidos, con toda seguridad, por este hecho. Así, si las ventas de juguetes presentan una variación estacional según un modelo multiplicativo del 48% en el mes de marzo y del 156% en el mes de diciembre no podremos comparar directamente las ventas de 13500€ y 41500€ respectivamente en los meses de marzo y diciembre del año pasado. Suponiendo el modelo multiplicativo, como se ha indicado, y eliminando el efecto estacional mediante cociente (mediante diferencia si el modelo es aditivo), obtendríamos el valor de la serie si ésta no se hubiera visto afectada por factores estacionales

$$Y(t) = \tau(t)E(t)C(t)\varepsilon(t) \Rightarrow \frac{Y(t)}{E(t)} = \tau(t)C(t)\varepsilon(t)$$

$$\text{marzo: } \frac{13500}{0,48} = 28125$$

$$\text{diciembre: } \frac{41500}{1,56} = 26602,56$$

Donde se observa que el mes de marzo tuvo **relativamente** un mejor comportamiento que diciembre en cuanto a las ventas (aunque en **términos absolutos** éstas fueron claramente inferiores, $13500 < 41500$).

► EJEMPLO 4.12.

Desestacionalice la serie cronológica del ejemplo 4.5, considerando el modelo multiplicativo y los I.V.E. obtenidos con el método de la razón a las medias móviles.

Solución:

Desestacionalizar una serie consiste en eliminar los altibajos observados en la misma debidos a factores estacionales. Para ello, en el **modelo multiplicativo**, dividiremos las observaciones de cada estación por su I.V.E., expresado este último en tantos por uno (en el modelo aditivo restaremos a las observaciones de cada estación el valor de su variación estacional).

$$\frac{7}{1,9196} = 3,6466 \quad \frac{5}{1,279} = 3,9093 \quad \dots$$

$Y(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre
2006	7	5	2,4	1,1
2007	10	7,1	3,6	1,4
2008	13,5	8,9	4,5	1,6
2009	17	11,2	5,3	2
2010	18,1	12,7	6,6	2,4
I.V.E.	191,96%	127,90%	58,98%	21,16%
$\frac{Y(t)}{E(t)}$	1º trimestre	2º trimestre	3º trimestre	4º trimestre
2006	3,6466	3,9093	4,0692	5,1985
2007	5,2094	5,5512	6,1038	6,6163
2008	7,0327	6,9586	7,6297	7,5614
2009	8,8560	8,7568	8,9861	9,4518
2010	9,4290	9,9296	11,1902	11,3422

En la serie desestacionalizada se aprecia un aumento casi continuo de la variable observada sin las fluctuaciones causadas por los factores estacionales. ◀

4.5 Predicción.

Conociendo las cuatro *componentes* de una serie cronológica así como el *modelo* según el cual se relacionan podríamos conocer el valor de la serie en cualquier período. La *variación irregular* es por naturaleza desconocida, del resto de componentes podemos tener a lo sumo una estimación de las mismas y el modelo es sencillamente un esquema artificial impuesto para facilitar el estudio de la

serie. Por tanto, no se podrá conocer el valor de la serie para un período futuro, a lo sumo se podrá hacer una *predicción*.

En este curso no se ha estudiado la *variación cíclica* por lo que la predicción se hará en base a la *tendencia secular* y *variación estacional*.

Si el modelo es aditivo, $Y(t) = \tau(t) + E(t) + C(t) + \varepsilon(t)$, la estimación del valor de la serie, $\hat{Y}(t)$, para un período futuro t estará dada por: $\hat{Y}(t) = \tau(t) + E(t)$.

Si la serie sigue un modelo multiplicativo, $Y(t) = \tau(t)E(t)C(t)\varepsilon(t)$, la predicción de su valor para un período futuro se hará mediante: $\hat{Y}(t) = \tau(t)E(t)$.

En ambos casos se está suponiendo que la *variación cíclica* no tiene una influencia fuerte en los valores de la serie.

► EJEMPLO 4.13.

¿Cuál sería el valor estimado de la serie del ejemplo 4.1 en el primer cuatrimestre de 2012?. Utilice los diferentes métodos y modelos.

Solución:

Según la recta de tendencia ajustada, $\tau(t) = -4175,42 + 2,083t$, el valor de la tendencia en el punto medio del año 2012 (punto medio del 2º cuatrimestre) se estima por:

$$\tau(2012) = -4175,42 + (2,083 \times 2012) = 15,576$$

Cada cuatrimestre la tendencia aumenta $\frac{2,083}{3} = 0,694$, por tanto el valor de la tendencia en el primer cuatrimestre del año 2012 se estima en $15,576 - 0,694 = 14,882$.

El valor estimado de la serie para el primer cuatrimestre de 2012 diferirá de este valor según la variación estacional para dicho cuatrimestre:

Modelo aditivo	
<i>Método de las media simples</i>	<i>Método de la diferencia a la tendencia</i>
$\hat{Y}(t) = \tau(t) + E(t) = 14,882 + 1,0273 = 15,9093$	$\hat{Y}(t) = \tau(t) + E(t) = 14,882 + 1,0273 = 15,9093$

Como se indicó anteriormente, estos dos métodos coinciden sobre el modelo aditivo (si se observa alguna diferencia es debida a errores de redondeo).

Modelo multiplicativo	
<i>Método de las media simples</i>	<i>Método de la razón a la tendencia</i>
$\hat{Y}(t) = \tau(t)E(t) = 14,882 \times \frac{113,52}{100} = 16,894$	$\hat{Y}(t) = \tau(t)E(t) = 14,882 \times \frac{114,11}{100} = 16,982$

En el *método de la razón (diferencia) a las medias móviles* la tendencia no se estima mediante el ajuste de una recta por mínimos cuadrados sobre las medias anuales sino mediante medias móviles,

por lo que tendríamos que calcular el valor de la tendencia a partir de ellas. Haría falta ajustar una recta a las medias móviles, esta recta no suele diferir mucho de la anterior recta de mínimos cuadrados sobre las medias anuales. Por este motivo, en la práctica, también para este último método se procede como en los dos anteriores: hallamos el valor estimado de la tendencia para las estaciones del año utilizando la recta de mínimos cuadrados sobre las media anuales y posteriormente se modifica de acuerdo al modelo y a la variación estacional estimada con el método de la razón (diferencia) a las medias móviles.

<i>Método de la razón a las medias móviles</i>	<i>Método de la diferencia a las medias móviles</i>
$\hat{Y}(t) = \tau(t)E(t) = 14,882 \times \frac{114,11}{100} = 16,982$	$\hat{Y}(t) = \tau(t) + E(t) = 14,882 + 1,1202 = 16,0022$

4.6 Ejercicios resueltos.

1. El número de llamadas (expresado en millones) de los abonados de la compañía *Noteoigo* en cada trimestre de los últimos años ha sido:

<i>t</i>	1º trimestre	2º trimestre	3º trimestre	4º trimestre	Total anual	Media anual
2008	6,4	7,2	5,6	8	27,2	6,8
2009	8,8	9,6	6,4	8,8	33,6	8,4
2010	9,6	10,4	8,8	11,2	40	10
2011	12	12,8	9,6	13,6	48	12
2012	14,4	14,4	11,2	16	56	14
Media por estación	10,24	10,88	8,32	11,52		

Obtenga:

- a) Tasa de variación media anual en el período 2008-2012 para el total de llamadas anuales.
- b) Índices de variación estacional (método de las medias simples, modelo multiplicativo).

Nota: $\tau(t) = 4,84 + 1,8(t - 2007) = -3607,76 + 1,8t$

- c) Estimación del número de llamadas para el 2º trimestre de 2013.

Solución:

a) $\sqrt[4]{\frac{T_{2012}}{T_{2008}}} - 1 = \sqrt[4]{\frac{56}{27,2}} - 1 = 0,1979 \Rightarrow 19,79\%$

- b) Eliminamos la tendencia de los valores medios por estación, restando tantas veces

$\frac{b}{s} = \frac{1,8}{4} = 0,45$ como estaciones del año han pasado, obteniendo las medias corregidas:

$$10,24 - \left(0 \times \frac{b}{s}\right) = 10,24 \quad 10,88 - \left(1 \times \frac{b}{s}\right) = 10,43 \quad 8,32 - \left(2 \times \frac{b}{s}\right) = 7,42 \quad 11,52 - \left(3 \times \frac{b}{s}\right) = 10,17$$

Calculamos la media de las medias corregidas (media global corregida):

$$9,565 = \frac{10,24 + 10,43 + 7,42 + 10,17}{4}$$

Comparando las medias corregidas con su promedio (por cociente) y multiplicando por 100 obtenemos los índices de variación estacional expresados en porcentajes.

$$\frac{10,24}{9,565} 100 = 107,057 \quad \dots \quad \frac{10,17}{9,565} 100 = 106,325$$

En la siguiente tabla se recogen todos los cálculos:

	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2008	6,4	7,2	5,6	8	
2009	8,8	9,6	6,4	8,8	
2010	9,6	10,4	8,8	11,2	
2011	12	12,8	9,6	13,6	
2012	14,4	14,4	11,2	16	
media por estación	10,24	10,88	8,32	11,52	media global corregida
media corregida	10,24	10,43	7,42	10,17	
IVE (%)	107,057%	109,043%	77,574%	106,325%	9,565

- c) Según la recta de tendencia ajustada, $\tau(t) = 4,84 + 1,8(t - 2007)$, el valor de la tendencia en el punto medio del año 2013 (punto entre el 2º y 3º trimestre) se estima por:

$$\tau(2013) = 4,84 + (1,8 \times 6) = 15,64$$

Cada trimestre la tendencia aumenta $\frac{1,8}{4} = 0,45$, por tanto en medio trimestre aumenta

$$\frac{0,45}{2} = 0,225.$$

El valor de la tendencia en el segundo trimestre del año 2013 se estima en $15,64 - 0,225 = 15,415$.

El valor estimado de la serie para el segundo trimestre de 2013 diferirá de este valor según la variación estacional para dicho trimestre: $15,415 \times 1,09043 = 16,81$ millones de llamadas.

2. La tendencia de la serie cronológica de ventas trimestrales de automóviles en una provincia (en miles) es:

$$\tau(t) = 135 + 16(t - 2008)$$

- Estime el valor de la tendencia para cada uno de los cuatro trimestres de 2012.
- Estime, según la tendencia, las ventas para todo el año 2012.

Solución:

- a) El valor de la tendencia en el punto central del año 2012 se estima mediante

$$\tau(2012) = 135 + 16(2012 - 2008) = 135 + 64 = 199$$

El valor de la tendencia aumenta 16 cada año, aumenta $\frac{16}{4} = 4$ cada trimestre y aumenta

$$\frac{4}{2} = 2 \text{ cada medio trimestre. Según lo anterior:}$$

$$\tau(3^{\circ} \text{ trimestre de } 2012) = 199 + 2 = 201 \text{ miles de automóviles}$$

$$\tau(2^{\circ} \text{ trimestre de } 2012) = 201 - 4 = 197 \text{ miles de automóviles}$$

$$\tau(1^{\circ} \text{ trimestre de } 2012) = 197 - 4 = 193 \text{ miles de automóviles}$$

$$\tau(4^{\circ} \text{ trimestre de } 2012) = 201 + 4 = 205 \text{ miles de automóviles}$$

Como puede comprobarse el valor de la tendencia en el punto central del año 2012 es igual a la media de la tendencia en los cuatro trimestres:

$$\frac{193 + 197 + 201 + 205}{4} = 199$$

- b) Las ventas totales en el año 2012 se pueden obtener como suma de las ventas en los cuatro trimestres: $193 + 197 + 201 + 205 = 796$. O a partir de la tendencia en el punto central del año 2012: $4 \times 199 = 796$.

3. Las ventas de motocicletas (en miles) en un país han sido las siguientes:

	Año				
Cuatrimestre	2002	2003	2004	2005	2006
1°	26	26	25	25	24
2°	52	53	53	52	51
3°	22	23	23	23	24

Calcule los índices de variación estacional según el método de las medias simples.

Solución:

Al mencionar el término índice de variación estacional nos están indicando que debemos utilizar el modelo multiplicativo.

En primer lugar ajustamos la recta de tendencia sobre las medias anuales

$$\frac{26 + 52 + 22}{3} = 33,3333 \quad \dots \quad \frac{24 + 51 + 24}{3} = 33$$

t_i	$x_i = t_i - 2001$	$y_i = \text{media anual}$	x_i^2	$x_i y_i$
2002	1	33,3333	1	33,3333
2003	2	34	4	68
2004	3	33,6667	9	101,0001
2005	4	33,3333	16	133,3332
2006	5	33	25	165
totales	15	167,3333	55	500,6666

$$n = 5 \quad \bar{x} = \frac{15}{5} = 3 \quad \bar{y} = \frac{167,3333}{5} = 33,4667$$

$$S_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{55}{5} - 3^2 = 2$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{500,6666}{5} - (3 \times 33,4667) = -0,26678$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Leftrightarrow y - 33,4667 = \frac{-0,26678}{2} (x - 3) \Leftrightarrow y = 33,86687 - 0,13339x$$

Deshacemos el cambio de origen $x_i = t_i - 2001$ y expresamos y como la tendencia secular $\tau(t)$

$$y = 33,86687 - 0,13339x \Leftrightarrow \tau(t) = 33,86687 - 0,13339(t - 2001) \Leftrightarrow \tau(t) = 300,78 - 0,13339t$$

Seguidamente eliminamos la tendencia de los valores medios por estación, restando tantas veces

$$\frac{b}{s} = \frac{-0,13339}{3} = -0,04446 \text{ como estaciones del año han transcurrido desde el comienzo del mismo,}$$

obteniendo las medias corregidas:

$$52,2 - \left(\frac{b}{s} \right) = 52,24446 \quad 23 - \left(2 \times \frac{b}{s} \right) = 23,08892$$

Calculamos la media de las medias corregidas (media global corregida):

$$33,51113 = \frac{25,2 + 52,24446 + 23,08892}{3}$$

Y comparando las medias corregidas con su promedio, por cociente, obtenemos los índices de variación estacional.

$$\frac{25,2}{33,51113} 100 = 75,2 \quad \frac{52,24446}{33,51113} 100 = 155,9 \quad \frac{23,08892}{33,51113} 100 = 68,9$$

	1º cuatrimestre	2º cuatrimestre	3º cuatrimestre	media anual
2002	26	52	22	33,3333
2003	26	53	23	34
2004	25	53	23	33,6667
2005	25	52	23	33,3333
2006	24	51	24	33
media por estación	25,2	52,2	23	media global corregida
media corregida	25,2	52,24446	23,08892	33,51113
I.V.E. (%)	75,2	155,9	68,9	

4. La tendencia de la serie de ventas cuatrimestrales de motocicletas en un país (en miles) y la variación estacional para cada cuatrimestre son:

$$\tau(t) = 12 + 6(t - 2010)$$

1º cuatrimestre	2º cuatrimestre	3º cuatrimestre
69,4	125,0	105,6

Estime las ventas del tercer cuatrimestre del año 2012.

Solución:

$\tau(2012) = 12 + 6(2012 - 2010) = 24$ es el valor de la tendencia en el punto central del año 2012, es decir en el punto central del 2º cuatrimestre de dicho año. Dado que la tendencia varía 6 unidades por cada año, variará 2 unidades por cada cuatrimestre. Así la tendencia en el 1º cuatrimestres de 2012 será $24 - 2 = 22$ y en el 3º cuatrimestre $24 + 2 = 26$.

La variación estacional está expresada en I.V.E. puesto que sus valores suman $300 = 3 \times 100$, por tanto el modelo asumido para la serie es el multiplicativo. Para estimar las ventas del tercer cuatrimestre del año 2012, multiplicaremos la estimación de la tendencia para dicho período (26) por el correspondiente I.V.E. expresado en tanto por uno

$$26 \times 1,056 = 27,456 \Rightarrow 27456 \text{ motocicletas}$$

5. Se han observado los beneficios trimestrales, en cientos de miles de euros, de una determinada empresa.

Año/Trimestre	1º	2º	3º	4º
2010	5	3	7	5
2011	5	2	6	6
2012	4	2	7	4

Ajustando una recta de mínimos cuadrados para la tendencia secular y usando el método de la razón a la tendencia para los índices de variación estacional, haga una predicción de los beneficios para el tercer trimestre de 2013.

Solución:

En primer lugar ajustamos la recta de tendencia sobre las medias anuales

$$\frac{5+3+7+5}{4} = 5 \quad \frac{5+2+6+6}{4} = 4,75 \quad \frac{4+2+7+4}{4} = 4,25$$

t_i	$x_i = t_i - 2009$	$y_i = \text{media anual}$	x_i^2	$x_i y_i$
2010	1	5	1	5
2011	2	4,75	4	9,5
2012	3	4,25	9	12,75
totales	6	14	14	27,25

$$n = 3 \quad \bar{x} = \frac{6}{3} = 2 \quad \bar{y} = \frac{14}{3} = 4,6667$$

$$S_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{14}{3} - 2^2 = 0,6667$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{27,25}{3} - (2 \times 4,6667) = -0,25$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Leftrightarrow y - 4,6667 = \frac{-0,25}{0,6667} (x - 2) \Leftrightarrow y = 5,4167 - 0,375x$$

Deshacemos el cambio de origen $x_i = t_i - 2009$ y expresamos y como la tendencia secular $\tau(t)$

$$y = 5,4167 - 0,375x \Leftrightarrow \tau(t) = 5,4167 - 0,375(t - 2009) \Leftrightarrow \tau(t) = 758,7917 - 0,375t$$

Estimamos la tendencia secular para cada trimestre de cada año, teniendo presente que la tendencia $\tau(t) = 758,7917 - 0,375t$ para 2010 se asigna al punto central del año (punto entre el 2º y 3º

trimestre), que la tendencia varía $\frac{b}{s} = \frac{-0,375}{4} = -0,09375$ cada trimestre que pasa y la mitad

$\left(\frac{-0,09375}{2} = -0,046875 \right)$ si sólo ha transcurrido medio trimestre.

Así: $\tau(2010) = 758,7917 - (0,375 \times 2010) = 5,0417 \rightarrow$ punto entre 2º y 3º trimestre de 2010

La tendencia centrada en el 3º trimestre de 2010 será: $5,0417 - 0,046875 = 4,994825$

A partir del anterior dato restaremos 0,09375 para obtener el valor de la tendencia en el siguiente trimestre (4º trimestre de 2010) y así sucesivamente hasta llegar al último trimestre de 2012.

Análogamente, para obtener el valor de la tendencia en el trimestre anterior (2º trimestre de 2010) le sumaremos 0,09375 a 4,994825 y así sucesivamente hasta llegar al primer trimestre de 2010. Los valores estimados de la tendencia se recogen en la siguiente tabla

$\tau(t)$	1º	2º	3º	4º
2010	5,182325	5,088575	4,994825	4,901075
2011	4,807325	4,713575	4,619825	4,526075
2012	4,432325	4,338575	4,244825	4,151075

Se elimina de las observaciones, $Y(t)$, el valor de la tendencia, $\tau(t)$, realizando los cocientes $\frac{Y(t)}{\tau(t)}$

$$\frac{5}{5,182325} = 0,9648 \quad \dots \quad \frac{4}{4,151075} = 0,9636$$

Calculamos la media por estación de los anteriores valores y la media de dichas medias (que es igual a la media global de todos los cocientes $\frac{Y(t)}{\tau(t)}$). Los índices de variación estacional (I.V.E.) se

obtienen como cociente de las medias por estación sobre la *media global* (el resultado se multiplica por 100 para expresarlo en tantos por ciento)

$$\begin{aligned} \frac{0,9648 + 1,0401 + 0,9025}{3} &= 0,9691 & \frac{0,5896 + 0,4243 + 0,461}{3} &= 0,4916 \\ \frac{1,4015 + 1,2988 + 1,6491}{3} &= 1,4498 & \frac{1,0202 + 1,3257 + 0,9636}{3} &= 1,1032 \end{aligned}$$

$$\frac{0,9691+0,4916+1,4498+1,1032}{4} = 1,003425$$

$$\frac{0,9691}{1,003425}100 = 96,58\% \quad \dots \quad \frac{1,1032}{1,003425}100 = 109,94\%$$

En la siguiente tabla se recogen todos los cálculos:

$Y(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2010	5	3	7	5	
2011	5	2	6	6	
2012	4	2	7	4	
$\tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2010	5,182325	5,088575	4,994825	4,901075	
2011	4,807325	4,713575	4,619825	4,526075	
2012	4,432325	4,338575	4,244825	4,151075	
$\frac{Y(t)}{\tau(t)}$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2010	0,9648	0,5896	1,4015	1,0202	
2011	1,0401	0,4243	1,2988	1,3257	
2012	0,9025	0,4610	1,6491	0,9636	media global
media por estación	0,9691	0,4916	1,4498	1,1032	1,003425
I.V.E.	96,58	48,99	144,49	109,94	

Para estimar los beneficios del tercer trimestre de 2013, estimamos en primer lugar la tendencia para dicho trimestre:

$$\tau(2013) = 5,4167 - 0,375(2013 - 2009) = 3,9167$$

La anterior estimación está referida al punto central del año 2013 (es decir punto en el que termina el segundo trimestre y comienza el tercero). Para obtener la estimación de la tendencia en el tercer trimestre de 2013 le sumamos al anterior valor lo que varía la tendencia en medio trimestre

$$3,9167 - 0,046875 = 3,8698$$

A dicho valor le aplicamos el I.V.E. del tercer trimestre (expresado en tantos por uno) y obtenemos la estimación del beneficio en dicho período

$$3,8698 \times 1,4449 = 5,591 \Rightarrow 5,591 \times 100000 = 559100\text{€}$$

6. Los siguientes datos expresan el número de toneladas producidas por una factoría. Halle los índices de variación estacional por el método de la razón a las medias móviles y desestacionalice la serie.

Año	1º trimestre	2º trimestre	3º trimestre	4º trimestre
2009	2	2,7	2	5
2010	3	3,3	3	6,1
2011	4	4,4	3,9	7,3
2012	5,3	5,7	4,9	7,8

Solución:

En primer lugar estimaremos la tendencia por el procedimiento de las medias móviles. Para ello comenzamos calculando las medias móviles de amplitud 4 que estarán asociadas al punto central de los 4 trimestres sobre los que se calcula:

$$2,925 = \frac{2+2,7+2+5}{4} \quad 3,175 = \frac{2,7+2+5+3}{4} \quad 3,325 = \frac{2+5+3+3,3}{4} \quad \dots$$

media móvil amplitud 4	1º trimestre	2º trimestre	3º trimestre	4º trimestre	1º trimestre (año siguiente)
2009	-	2,925	3,175	3,325	
2010	3,575	3,85	4,1	4,375	
2011	4,6	4,9	5,225	5,55	
2012	5,8	5,925	-	-	

A continuación procedemos a centrar las medias móviles en los mismos períodos que las observaciones de la serie, tomando medias móviles de amplitud 2 sobre las anteriores medias móviles de amplitud 4:

$$3,05 = \frac{2,925+3,175}{2} \quad 3,25 = \frac{3,175+3,325}{2} \quad 3,45 = \frac{3,325+3,575}{2} \quad \dots$$

medias móviles amplitud 2	1º trimestre	2º trimestre	3º trimestre	4º trimestre
2009	-	-	3,05	3,25
2010	3,45	3,7125	3,975	4,2375
2011	4,4875	4,75	5,0625	5,3875
2012	5,675	5,8625	-	-

Se elimina de las observaciones, $Y(t)$, el valor de la tendencia, $\tau(t)$, realizando los cocientes $\frac{Y(t)}{\tau(t)}$ en aquellos períodos donde tenemos estimaciones de la tendencia mediante medias móviles.

$$\frac{2}{3,05} = 0,6557 \quad \dots \quad \frac{5,7}{5,8625} = 0,9723$$

Calculamos la media por estación de los anteriores valores y la media de dichas medias (media global). Los índices de variación estacional (I.V.E.) se obtienen como cociente de las medias por estación sobre la *media global* (el resultado se multiplica por 100 para expresarlo en tantos por ciento)

$$\begin{aligned} \frac{0,8696+0,8914+0,9339}{3} &= 0,8983 & \frac{0,8889+0,9263+0,9723}{3} &= 0,9292 \\ \frac{0,6557+0,7547+0,7704}{3} &= 0,7269 & \frac{1,5385+1,4395+1,3550}{3} &= 1,4443 \end{aligned}$$

$$\frac{0,8983 + 0,9292 + 0,7269 + 1,4443}{4} = 0,999675$$

$$\frac{0,8983}{0,999675} 100 = 89,86\% \quad \dots \quad \frac{1,4443}{0,999675} 100 = 144,48\%$$

En la siguiente tabla se recogen todos los cálculos:

$Y(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2009	2	2,7	2	5	
2010	3	3,3	3	6,1	
2011	4	4,4	3,9	7,3	
2012	5,3	5,7	4,9	7,8	
$\tau(t)$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2009	-	-	3,05	3,25	
2010	3,45	3,7125	3,975	4,2375	
2011	4,4875	4,75	5,0625	5,3875	
2012	5,675	5,8625	-	-	
$\frac{Y(t)}{\tau(t)}$	1º trimestre	2º trimestre	3º trimestre	4º trimestre	
2009	-	-	0,6557	1,5385	
2010	0,8696	0,8889	0,7547	1,4395	
2011	0,8914	0,9263	0,7704	1,3550	
2012	0,9339	0,9723	-	-	media global
media por estación	0,8983	0,9292	0,7269	1,4443	0,999675
I.V.E. (%)	89,86	92,95	72,72	144,48	

Desestacionalizar una serie consiste en eliminar los altibajos observados en la misma debidos a factores estacionales. Para ello dividiremos las observaciones de cada estación por su I.V.E., expresado este último en tantos por uno:

$$\frac{2}{0,8986} = 2,2257 \quad \dots \quad \frac{7,8}{1,4448} = 5,3987$$

SERIE DESESTACIONALIZADA	1º trimestre	2º trimestre	3º trimestre	4º trimestre
2009	2,2257	2,9048	2,7503	3,4607
2010	3,3385	3,5503	4,1254	4,2220
2011	4,4514	4,7337	5,3630	5,0526
2012	5,8981	6,1323	6,7382	5,3987

7. Con los datos mensuales de una serie cronológica en el período 2007-2012 se ha estimado la tendencia:

$$\tau(t) = 29,92 + 25,958(t - 2007)$$

y los índices de variación estacional:

MES	I.V.E.	MES	I.V.E.
Enero	97,38	Julio	103,44
Febrero	97,49	Agosto	101,74
Marzo	96,89	Septiembre	100,03
Abril	98,9	Octubre	103,67
Mayo	98,75	Noviembre	102,35
Junio	97,91	Diciembre	101,64

- a) Haga una predicción para el mes de julio de 2013.
- b) Suponga que los datos de marzo y octubre han sido 126,18 y 145,64 respectivamente. Compárelos eliminando previamente la componente estacional.

Solución:

- a) $\tau(2013) = 29,92 + 25,958(2013 - 2007) = 185,668$ es la predicción de la tendencia para el punto central del año 2013 (fin de junio, principio de julio). La tendencia varía 25,958 cada año, por tanto $1,08158 = \frac{25,958}{24}$ cada medio mes. Según lo anterior, se estima un valor de la tendencia para el punto central de julio de 2013 de $185,668 + 1,08158 = 186,75$. Aplicando sobre la anterior estimación de la tendencia el I.V.E. de julio obtenemos la predicción para el mes de julio de 2013:

$$186,75 \times 1,0344 = 193,1742$$

- b) El valor desestacionalizado de marzo es:

$$\frac{126,18}{0,9689} = 130,23$$

El valor desestacionalizado de octubre es:

$$\frac{145,64}{1,0367} = 140,48$$

Aunque el mes de octubre no tuviera una componente estacional más favorable que la de marzo el valor de la serie en octubre hubiera sido mayor que en marzo ($140,48 > 130,23$).

8. En una provincia se ha analizado la serie cuatrimestral de las inversiones realizadas. Resultó estimada la tendencia por $\tau(t) = 3,8667 + 1,2(t - 2006)$ y los índices de variación estacional por

Cuatrimestre	Primero	Segundo	Tercero
I.V.E.	106,79	118,4	74,81

Haga una predicción de las inversiones en la provincia para cada uno de los cuatrimestres de 2013.

Solución:

Comenzamos estimando la tendencia en el punto central del año 2013, que coincide con el punto central de su segundo cuatrimestre

$$\tau(2013) = 3,8667 + 1,2(2013 - 2006) = 12,2667$$

La tendencia varía 1,2 cada año, $\frac{1,2}{3} = 0,4$ cada cuatrimestre. Por tanto la tendencia en el tercer cuatrimestre de 2013 será $12,2667 + 0,4 = 12,4667$ y en el primer cuatrimestre $12,2667 - 0,4 = 11,8667$.

Cuatrimestres de 2013	Primero	Segundo	Tercero
Valores de la tendencia	11,8667	12,2667	12,4667

Multiplicando los valores de la tendencia por los I.V.E. (en tantos por 1) se obtienen las estimaciones de las inversiones en cada cuatrimestre de 2013: $11,8667 \times 1,0679 = 12,67245 \dots$

Cuatrimestres de 2013	Primero	Segundo	Tercero
Estimaciones de las inversiones	12,67245	14,52377	9,32634

9. La siguiente tabla muestra las ventas trimestrales de automóviles en un concesionario.

Año/Trimestre	I	II	III	IV
2009	12	15	10	9
2010	14	18	13	10
2011	17	23	18	13
2012	21	27	21	15

Obtenga una predicción para las ventas del cuarto trimestre del año 2013 y desestacionalice la serie.

Utilice el método de las medias simples, modelo aditivo.

Solución:

Comenzamos ajustando la recta de tendencia sobre las medias anuales

$$\frac{12+15+10+9}{4} = 11,5 \quad \dots \quad \frac{21+27+21+15}{4} = 21$$

t_i	$x_i = t_i - 2008$	$y_i = \text{media anual}$	x_i^2	$x_i y_i$
2009	1	11,5	1	11,5
2010	2	13,75	4	27,5
2011	3	17,75	9	53,25
2012	4	21	16	84
totales	10	64	30	176,25

$$n = 4 \quad \bar{x} = \frac{10}{4} = 2,5 \quad \bar{y} = \frac{64}{4} = 16$$

$$S_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{30}{4} - 2,5^2 = 1,25$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{176,25}{4} - (2,5 \times 16) = 4,0625$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Leftrightarrow y - 16 = \frac{4,0625}{1,25} (x - 2,5) \Leftrightarrow y = 7,875 + 3,25x$$

Deshacemos el cambio de origen $x_i = t_i - 2008$ y expresamos y como la tendencia secular $\tau(t)$

$$y = 7,875 + 3,25x \Leftrightarrow \tau(t) = 7,875 + 3,25(t - 2008) \Leftrightarrow \tau(t) = -6518,125 + 3,25t$$

Seguidamente eliminamos la tendencia de los valores medios por estación, restando tantas veces

$$\frac{b}{s} = \frac{3,25}{4} = 0,8125 \text{ como estaciones del año han transcurrido desde el comienzo del mismo,}$$

obteniendo las medias corregidas:

$$20,75 - (0,8125) = 19,9375 \quad 15,5 - (2 \times 0,8125) = 13,875 \quad 11,75 - (3 \times 0,8125) = 9,3125$$

Calculamos la media de las medias corregidas (media global corregida):

$$14,78125 = \frac{16 + 19,9375 + 13,875 + 9,3125}{4}$$

Y comparando las medias corregidas con su promedio, por diferencia, obtenemos la variación estacional para el modelo aditivo.

$$16 - 14,78125 = 1,21875 \quad \dots \quad 9,3125 - 14,78125 = -5,46875$$

	1º trimestre	2º trimestre	3º trimestre	4º trimestre	media anual
2009	12	15	10	9	11,5
2010	14	18	13	10	13,75
2011	17	23	18	13	17,75
2012	21	27	21	15	21
media por estación	16	20,75	15,5	11,75	media global corregida
media corregida	16	19,9375	13,875	9,3125	14,78125
V.E.	1,21875	5,15625	-0,90625	-5,46875	

La estimación de la tendencia en el punto central del año 2013 es:

$$\tau(2013) = 7,875 + 3,25(2013 - 2008) = 24,125$$

La tendencia varía 3,25 cada año, $\frac{3,25}{4} = 0,8125$ cada trimestre, $\frac{0,8125}{2} = 0,40625$ cada medio

trimestre. Por tanto la tendencia estimada para el punto central del cuarto trimestre es:

$$24,125 + 0,8125 + 0,40625 = 25,34375$$

Valor que tenemos que corregir según la variación estacional de dicho trimestre para obtener la predicción pedida:

$$25,34375 - 5,46875 = 19,875$$

Para obtener la serie desestacionalizada, sencillamente se restan los valores de la variación estacional (V.E.) a los datos de la serie cronológica:

$$12 - 1,21875 = 10,78125 \quad 14 - 1,21875 = 12,78125 \quad \dots \quad 15 + 5,46875 = 20,46875$$

Serie desestacionalizada	I	II	III	IV
2009	10,78125	9,84375	10,90625	14,46875
2010	12,78125	12,84375	13,90625	15,46875
2011	15,78125	17,84375	18,90625	18,46875
2012	19,78125	21,84375	21,90625	20,46875

10. Una empresa de bebidas carbónicas ha vendido, en millones de litros, las siguientes cifras:

Estación / Año	2009	2010	2011
primavera	2,2	2,4	2,5
verano	3,5	3,6	3,6
otoño	4,3	4,5	4,8
invierno	2,1	2,2	2,5

Estime las ventas para el invierno de 2012, utilizando el método de las medias móviles y el modelo aditivo.

Solución:

Consideramos que el año comienza el primer día de la primavera y termina el último día del invierno.

En primer lugar estimaremos la tendencia por el procedimiento de las medias móviles. Para ello comenzamos calculando las medias móviles de amplitud 4 que estarán asociadas al punto central de las 4 estaciones sobre las que se calcula:

$$3,025 = \frac{2,2 + 3,5 + 4,3 + 2,1}{4} \quad 3,075 = \frac{3,5 + 4,3 + 2,1 + 2,4}{4} \quad 3,1 = \frac{4,3 + 2,1 + 2,4 + 3,6}{4} \quad \dots$$

media móvil amplitud 4	primavera	verano	otoño	invierno	primavera (año siguiente)
2009	-	3,025	3,075	3,1	
2010	3,15	3,175	3,2	3,2	
2012	3,275	3,35	-	-	

A continuación procedemos a centrar las medias móviles en los mismos períodos que las observaciones de la serie, tomando medias móviles de amplitud 2 sobre las anteriores medias móviles de amplitud 4:

$$3,05 = \frac{3,025 + 3,075}{2} \quad 3,0875 = \frac{3,075 + 3,1}{2} \quad 3,125 = \frac{3,1 + 3,15}{2} \quad \dots$$

medias móviles amplitud 2	primavera	verano	otoño	invierno
2009	-	-	3,05	3,0875
2010	3,125	3,1625	3,1875	3,2
2011	3,2375	3,3125	-	-

Se elimina de las observaciones, $Y(t)$, el valor de la tendencia, $\tau(t)$, realizando las diferencias $Y(t) - \tau(t)$ en aquellos períodos donde tenemos estimaciones de la tendencia mediante medias móviles.

$$4,3 - 3,05 = 1,25 \quad \dots \quad 3,6 - 3,3125 = 0,2875$$

Calculamos la media por estación de los anteriores valores y la media de dichas medias (media global). La variación estacional (V.E.) se obtiene como diferencia entre las medias por estación y la media global.

$$\frac{-0,725-0,7375}{2} = -0,73125 \quad \frac{0,4375+0,2875}{2} = 0,3625$$

$$\frac{1,25+1,3125}{2} = 1,28125 \quad \frac{-0,9875-1}{2} = -0,99375$$

$$\frac{-0,73125+0,3625+1,28125-0,99375}{4} = -0,0203125$$

$$-0,73125 - (-0,0203125) = \dots \quad -0,99375 - (-0,0203125) = -0,9734375$$

En la siguiente tabla se recogen todos los cálculos:

$Y(t)$	primavera	verano	otoño	invierno	
2009	2,2	3,5	4,3	2,1	
2010	2,4	3,6	4,5	2,2	
2011	2,5	3,6	4,8	2,5	
$\tau(t)$	primavera	verano	otoño	invierno	
2009	-	-	3,05	3,0875	
2010	3,125	3,1625	3,1875	3,2	
2011	3,2375	3,3125	-	-	
$Y(t) - \tau(t)$	primavera	verano	otoño	invierno	
2009	-	-	1,25	-0,9875	
2010	-0,725	0,4375	1,3125	-1	
2011	-0,7375	0,2875	-	-	media global
media por estación	-0,73125	0,3625	1,28125	-0,99375	-0,0203125
V.E.	-0,7109375	0,3828125	1,3015625	-0,9734375	

Para estimar las ventas del invierno de 2012 necesitamos una estimación de la tendencia en dicho momento y sumarle el valor de la variación estacional en invierno.

El método de obtención de la tendencia mediante medias móviles no permite estimar ésta para periodos futuros. Para estimar el valor de la tendencia en un futuro a partir de las medias móviles haría falta ajustar una recta sobre las medias móviles, esta recta no suele diferir mucho de la que se ajusta en los métodos de las medias simples y de la razón (diferencia) a la tendencia. Por este motivo, en la práctica, se procede a ajustar dicha recta para estimar la tendencia

t_i	$x_i = t_i - 2008$	$y_i = \text{media anual}$	x_i^2	$x_i y_i$
2009	1	3,025	1	3,025
2010	2	3,175	4	6,35
2011	3	3,35	9	10,05
totales	6	9,55	14	19,425

$$n = 3 \quad \bar{x} = \frac{6}{3} = 2 \quad \bar{y} = \frac{9,55}{3} = 3,1833$$

$$S_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{14}{3} - 2^2 = 0,6667$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{19,425}{3} - (2 \times 3,1833) = 0,1084$$

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \Leftrightarrow y - 3,1833 = \frac{0,1084}{0,6667} (x - 2) \Leftrightarrow y = 2,8581 + 0,1626x$$

Deshacemos el cambio de origen $x_i = t_i - 2008$ y expresamos y como la tendencia secular $\tau(t)$

$$y = 2,8581 + 0,1626x \Leftrightarrow \tau(t) = 2,8581 + 0,1626(t - 2008) \Leftrightarrow \tau(t) = -323,6427 + 0,1626t$$

La estimación de la tendencia en el punto central del año 2012 (en este caso, el punto que separa las estaciones verano y otoño) es:

$$\tau(2012) = 2,8581 + 0,1626(2012 - 2008) = 3,5085$$

La tendencia varía 0,1626 cada año, $\frac{0,1626}{4} = 0,04065$ cada estación, $\frac{0,04065}{2} = 0,020325$ cada media estación. Por tanto la tendencia estimada para el punto central del invierno de 2012 es:

$$3,5085 + 0,04065 + 0,020325 = 3,569475$$

Valor que tenemos que corregir según la variación estacional del invierno para obtener la predicción pedida:

$$3,569475 + (-0,9734375) = 2,596 \text{ millones de litros (2596000 litros)}$$

5. PROBABILIDAD.

5.1 Definición de probabilidad. Asignación de probabilidades.

La distinción entre *fenómenos determinísticos* y *aleatorios* conduce a plantearse el problema de la medida de la incertidumbre mostrada por estos últimos.

Conviene recordar algunas de las propiedades de las *frecuencias relativas* de *variables estadísticas* que van a proporcionar un excelente apoyo intuitivo para la **definición de la probabilidad**.

Consideremos una tabla estadística en la que se recogen los resultados del lanzamiento de un dado en 300 ocasiones:

x_i	n_i	f_i
1	45	0,15
2	57	0,19
3	51	0,17
4	48	0,16
5	54	0,18
6	45	0,15
total	$n=300$	1

1.- La frecuencia relativa de cualquier modalidad x_i es un número comprendido entre 0 y 1,

$$0 \leq f_i \leq 1.$$

2.- La frecuencia relativa de dos o más modalidades es la suma de las frecuencias relativas de cada una de las modalidades. Por ejemplo, la frecuencia relativa de obtener un número par es

$$f(\text{par}) = f(x_2 \cup x_4 \cup x_6) = f_2 + f_4 + f_6 = 0,19 + 0,16 + 0,15 = 0,50$$

3.- La frecuencia relativa de todas las modalidades (suma de todas las frecuencias relativas) es 1,

$$\sum_{i=1}^6 f_i = 1$$

Llamaremos **suceso** a cada uno de los posibles resultados de un *experimento* o *fenómeno aleatorio*. Utilizaremos letras mayúsculas para referirnos a ellos.

En el experimento de lanzar el dado, son ejemplos de sucesos: $A = \{\text{obtener par}\}$, $B = \{\text{obtener impar}\}$, $C = \{\text{obtener un 3}\}$, $D = \{\text{obtener un número mayor o igual que 4}\}$, $E = \{\text{obtener un número menor que 4}\}$... Utilizaremos estos sucesos como ejemplos para ilustrar las siguientes definiciones y operaciones sobre sucesos.

Al suceso formado por un único resultado se denomina **suceso elemental**. Para estos sucesos, además de la notación general con letras mayúsculas, utilizaremos la notación ω . Un ejemplo de suceso elemental sería $C = \{\text{obtener un 3}\}$, ω_3 según esta última notación (notación análoga a la utilizada en variables estadísticas, x_i).

Se llama **suceso seguro** al suceso que siempre ocurre, está formado por todos los sucesos elementales, se nota Ω . En el ejemplo del dado $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Al conjunto vacío, que se nota \emptyset , se denomina **suceso imposible**.

Operaciones y relaciones con sucesos:

El suceso A **implica** el suceso B si siempre que ocurre A ocurre B . Lo notaremos $A \subseteq B$. En el ejemplo del dado: C implica B , $C \subseteq B$, $\{3\} \subseteq \{\text{impares}\}$.

Se define la **unión de dos sucesos** A y B , $A \cup B$, como el suceso que ocurre cuando ocurre el suceso A o el suceso B . En el ejemplo del dado: $B = \{\text{impar}\}$, $D = \{\text{mayor o igual que 4}\}$, $B \cup D = \{1, 3, 4, 5, 6\}$.

La unión de sucesos cumple las **propiedades conmutativa, asociativa e idempotente**:

$$A \cup B = B \cup A$$

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$$A \cup A = A$$

Se define la **intersección de dos sucesos** A y B , $A \cap B$, como el suceso que ocurre cuando ocurre el suceso A y el suceso B . En el ejemplo del dado: $B = \{\text{impar}\}$, $D = \{\text{mayor o igual que 4}\}$, $B \cap D = \{5\}$.

La intersección de sucesos cumple las **propiedades conmutativa, asociativa e idempotente**:

$$A \cap B = B \cap A$$

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

$$A \cap A = A$$

Las operaciones de unión e intersección conjuntamente cumplen las **propiedades distributivas**:

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

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

y de **absorción**:

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

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

Se define la **diferencia de dos sucesos**, $A - B$, como el suceso formado por los sucesos elementales que son de A pero no de B . Para poder definir correctamente esta operación se ha de cumplir que $B \subseteq A$. En el ejemplo del dado: $B = \{\text{impar}\}$, $C = \{3\}$, $B - C = \{1, 5\}$.

El **suceso contrario (o complementario)** de A es el que ocurre cuando no ocurre A . Se nota como \overline{A} . $\overline{A} = \Omega - A$. En el ejemplo del dado: $A = \{\text{par}\}$, $\overline{A} = \{\text{impar}\}$, $D = \{\text{mayor o igual que } 4\}$, $\overline{D} = \{\text{menor que } 4\}$.

Esta operación tiene las siguientes **propiedades**:

$$\begin{aligned}\overline{\overline{A}} &= A \\ A \subseteq B &\Rightarrow \overline{B} \subseteq \overline{A} \\ \overline{\emptyset} &= \Omega \quad \overline{\Omega} = \emptyset \\ A \cap \overline{A} &= \emptyset \quad A \cup \overline{A} = \Omega \\ \overline{A_1 \cup \dots \cup A_n} &= \overline{A_1} \cap \dots \cap \overline{A_n} \\ \overline{A_1 \cap \dots \cap A_n} &= \overline{A_1} \cup \dots \cup \overline{A_n}\end{aligned}$$

Los sucesos A y B son **incompatibles** si no pueden ocurrir ambos simultáneamente, $A \cap B = \emptyset$. Un suceso y su complementario son siempre sucesos incompatibles. En el ejemplo del dado: $A = \{\text{par}\}$ y $C = \{3\}$ son incompatibles.

Representaremos los resultados de un experimento aleatorio mediante el par (Ω, \mathcal{A}) , donde Ω es el conjunto de todos los sucesos elementales de un experimento aleatorio y \mathcal{A} el conjunto de todos los sucesos (elementales y no elementales).

Definición de probabilidad.

Definimos una medida de probabilidad sobre (Ω, \mathcal{A}) como una aplicación:

$$\begin{aligned}(\Omega, \mathcal{A}) &\xrightarrow{P} \mathbb{R} \\ A \in \mathcal{A} &\rightarrow P(A) \in \mathbb{R}\end{aligned}$$

Verificando las siguientes **condiciones**:

$$1.- P(A) \geq 0 \quad \forall A \in \mathcal{A}$$

2.- Sean A_1, \dots, A_n sucesos incompatibles, $P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$

3.- $P(\Omega) = 1$

Obsérvese la analogía entre estas tres condiciones y las propiedades de las frecuencias relativas que destacábamos al comienzo de la lección.

Propiedades:

1.- $P(\bar{A}) = 1 - P(A)$

$$\text{Demostración: } 1 = P(\Omega) = P(A \cup \bar{A}) = P(A) + P(\bar{A}) \Rightarrow 1 - P(A) = P(\bar{A})$$

2.- Un caso particular de la propiedad anterior es $P(\emptyset) = 1 - P(\Omega) = 1 - 1 = 0$

3.- Si $B \subseteq A \Rightarrow P(A - B) = P(A) - P(B)$

$$\begin{aligned} \text{Demostración: } A &= (A - B) \cup B & (A - B) \cap B &= \emptyset & \Rightarrow & P(A) = P(A - B) + P(B) \\ &\Rightarrow P(A - B) = P(A) - P(B) \end{aligned}$$

4.- Si $B \subseteq A \Rightarrow P(B) \leq P(A)$

$$\text{Demostración: } \Rightarrow P(B) = P(A) - P(A - B) \Rightarrow P(B) \leq P(A)$$

5.- Un caso particular de la anterior propiedad nos conduce a que la probabilidad de todo suceso es menor o igual a 1. $A \subseteq \Omega \Rightarrow P(A) \leq 1$

6.- Probabilidad de la unión de sucesos compatibles: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$\text{Demostración: } A = (A \cap B) \cup (A \cap \bar{B}) \Rightarrow P(A) = P(A \cap B) + P(A \cap \bar{B})$$

$$B = (B \cap A) \cup (B \cap \bar{A}) \Rightarrow P(B) = P(B \cap A) + P(B \cap \bar{A})$$

$$P(A) + P(B) = 2P(A \cap B) + P(A \cap \bar{B}) + P(B \cap \bar{A})$$

$$A \cup B = (A \cap B) \cup (A \cap \bar{B}) \cup (B \cap \bar{A}) \Rightarrow P(A \cup B) = P(A \cap B) + P(A \cap \bar{B}) + P(B \cap \bar{A})$$

$$P(A) + P(B) = 2P(A \cap B) + P(A \cap \bar{B}) + P(B \cap \bar{A}) = P(A \cup B) + P(A \cap B)$$

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

7.- Consecuencia de la anterior propiedad es la subaditividad de P: $P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$

$$\text{Demostración: } P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$$

y por inducción se demuestra para $\bigcup_{i=1}^n A_i$.

Asignación de probabilidades.

Estas son las propiedades que tiene la medida de la probabilidad pero la teoría no especifica nada más sobre los valores numéricos concretos que toma la función $P(A)$.

Si el número de sucesos elementales, ω_i , es finito, es suficiente asignar probabilidades a los sucesos elementales, $P(\omega_i)$, y definir la probabilidad de un suceso A como

$$P(A) = \sum_{\omega_i \in A} P(\omega_i)$$

Si además de ser finito el conjunto de sucesos elementales, $\Omega = \{\omega_1, \dots, \omega_i, \dots, \omega_n\}$, suponemos que

los sucesos elementales son *equiprobables*, $P(\omega_i) = \frac{1}{n} \quad \forall \omega_i$, obtenemos la conocida **regla de**

Laplace

$$P(A) = \sum_{\omega_i \in A} P(\omega_i) = \sum_{\omega_i \in A} \frac{1}{n} = \frac{k}{n} = \frac{\text{casos favorables}}{\text{casos posibles}}$$

siendo k el número de sucesos elementales en A .

A esta **concepción de la probabilidad** se le denomina **clásica o de Laplace**.

Según esta concepción de la probabilidad, en el ejemplo del dado se supondría que todas las caras (sucesos elementales) tienen la misma probabilidad de aparición y esta es:

$$P(\omega_i) = \frac{1}{6} = 0,1\bar{6}$$

En la semejanza entre las propiedades de las frecuencias relativas y la definición de probabilidad se basa la **concepción frecuentista** de la probabilidad. Considera que la probabilidad de un suceso es el valor límite en torno al cual se estabiliza la frecuencia relativa de ese suceso cuando se observa indefinidamente el fenómeno aleatorio. En el ejemplo del dado, si continuamos observando los resultados cuando lo lanzamos muchas veces (más de las 300 iniciales), se observará que el valor de las frecuencias relativas ya no cambia apenas y se estabiliza en torno a un valor que tomaremos como medida de la probabilidad de aparición de dicho suceso.

x_i	n_i	f_i		x_i	n_i	f_i
1	45	0,15		1	456	0,152
2	57	0,19		2	561	0,187
3	51	0,17		3	516	0,172
4	48	0,16		4	483	0,161
5	54	0,18		5	534	0,178
6	45	0,15		6	450	0,150
total	$n=300$	1		total	$n=3000$	1

Cada una de estas concepciones tiene sus ventajas e inconvenientes. Por ejemplo, la concepción clásica no nos valdría para modelar el comportamiento de un dado, moneda, ... que estuviesen trucados para conseguir con mayor probabilidad determinados resultados o sencillamente mal contruidos. En contra de la concepción frecuentista se podría esgrimir el argumento de que multitud de fenómenos aleatorios no pueden repetirse de forma indefinida, especialmente fenómenos de tipo social y económico (elecciones, valores bursátiles, competiciones deportivas, ...). Esta y otras dificultades han llevado a considerar interpretaciones alternativas de la probabilidad como la **concepción subjetiva** en la que se considera un grado de creencia o confianza en la ocurrencia de un hecho (la probabilidad de que la selección española gane el próximo mundial de fútbol es de 4/5).

► EJEMPLO 5.1

Calcule las probabilidades de los siguientes sucesos en el lanzamiento de un dado según la concepción clásica de Laplace y según la concepción frecuentista (tomando para este último caso como valores estables de las frecuencias los observados en los 3000 lanzamientos). Expréselo de distintas formas donde se pueda.

$A = \{\text{obtener par}\}$, $B = \{\text{obtener impar}\}$, $C = \{\text{obtener un 3}\}$, $D = \{\text{obtener un número mayor o igual que 4}\}$, $E = \{\text{obtener un número menor que 4}\}$.

Solución:

Concepción clásica:

$$P(A) = \sum_{\omega_i \in A} P(\omega_i) = \sum_{\omega_i \in A} \frac{1}{6} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{3}{6} = \frac{1}{2} = 0,5$$

$$P(B) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{3}{6} = \frac{1}{2} = P(\overline{A}) = 1 - P(A) = 1 - 0,5 = 0,5$$

$$P(C) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{1}{6} = 0,1\bar{6}$$

$$P(D) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{3}{6} = \frac{1}{2} = 0,5$$

$$P(E) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{3}{6} = \frac{1}{2} = P(\overline{D}) = 1 - P(D) = 1 - 0,5 = 0,5$$

Concepción frecuentista:

$$P(A) = \sum_{\omega_i \in A} P(\omega_i) = 0,187 + 0,161 + 0,150 = 0,498$$

$$P(B) = \sum_{\omega_i \in B} P(\omega_i) = 0,152 + 0,172 + 0,178 = 0,502 = P(\overline{A}) = 1 - P(A) = 1 - 0,498 = 0,502$$

$$P(C) = 0,172$$

$$P(D) = \sum_{\omega_i \in D} P(\omega_i) = 0,161 + 0,178 + 0,150 = 0,489$$

$$P(E) = \sum_{\omega_i \in E} P(\omega_i) = 0,152 + 0,187 + 0,172 = 0,511 = P(\overline{D}) = 1 - P(D) = 1 - 0,489 = 0,511$$

► EJEMPLO 5.2

En una ciudad se publican tres periódicos (A , B y C). Se sabe que un 60% de la población está suscrito al periódico A , un 40% al B , un 30% al C , un 20% a A y B , un 15% a A y C , un 25% a B y C , y un 10% a los tres periódicos. ¿Qué parte de la población está suscrita al menos a un periódico?

Solución:

En este ejemplo se hace una interpretación de la probabilidad según la concepción frecuentista.

Sean A , B y C los sucesos comprar el periódico A , B y C respectivamente.

La pregunta se puede interpretar como qué parte de la población está suscrita a A o a B o a C . Lo que en la notación de sucesos equivale a la unión de los tres sucesos.

Utilizaremos la fórmula de la probabilidad de la unión de dos sucesos compatibles,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, y la propiedad asociativa de la unión de sucesos.

$$\begin{aligned} P(A \cup B \cup C) &= P(A \cup (B \cup C)) = P(A) + P(B \cup C) - P(A \cap (B \cup C)) = \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap (B \cup C)) = \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P((A \cap B) \cup (A \cap C)) = \\ &= P(A) + P(B) + P(C) - P(B \cap C) - (P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)) = \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C) = \\ &= \frac{60}{100} + \frac{40}{100} + \frac{30}{100} - \frac{25}{100} - \frac{20}{100} - \frac{15}{100} + \frac{10}{100} = \frac{80}{100} \end{aligned}$$

El 80% de la población está suscrita al menos a un periódico.

5.2 Definición de probabilidad condicionada. Sucesos dependientes e independientes.

Probabilidad condicionada.

En determinadas situaciones es necesario calcular la probabilidad de sucesos dada la condición adicional de que cierto suceso ha ocurrido. A este tipo de **probabilidad** se denomina **condicionada**.

Se define la probabilidad del suceso B condicionado a A , $P(B/A)$, como la probabilidad de que ocurra B supuesto que ha ocurrido A . Dicha *probabilidad condicionada* del suceso B al suceso A es igual a:

$$P(B/A) = \frac{P(B \cap A)}{P(A)} \quad P(A) > 0$$

Como en la probabilidad no condicionada, aquí se sigue tomando como modelo las frecuencias relativas. Supongamos el siguiente ejemplo sobre datos de los estudios completados por una población de 1000 habitantes

	Enseñanza obligatoria	Bachillerato	Universitarios	TOTAL
hombres	250	160	40	450
mujeres	350	140	60	550
TOTAL	600	300	100	1000

Entre los hombres (H sería el suceso conocido o que ha ocurrido) la frecuencia relativa de los que completaron los estudios universitarios (U sería el suceso sobre el que se quiere calcular la probabilidad, en este caso la frecuencia) es:

$$f(U/H) = \frac{40}{450} = \frac{\frac{40}{1000}}{\frac{450}{1000}} = \frac{f(U \cap H)}{f(H)}$$

Mientras que la frecuencia relativa de los que completaron los estudios universitarios (sin imponer ninguna condición sobre el sexo) es:

$$f(U) = \frac{100}{1000}$$

La definición de **probabilidad condicionada**, fijado A , verifica las **condiciones de una medida de probabilidad**:

1.- $P(B/A) \geq 0 \quad \forall B \in \mathcal{F}$

2.- Sean B_1, \dots, B_n sucesos incompatibles, $P\left(\frac{\bigcup_{i=1}^n B_i}{A}\right) = \sum_{i=1}^n P\left(\frac{B_i}{A}\right)$

3.- $P\left(\frac{\Omega}{A}\right) = 1$

Demostración:

1.- $P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)}$ $P(A) > 0$ y $P(B \cap A) \geq 0 \Rightarrow \frac{P(B \cap A)}{P(A)} \geq 0$

2.- $P\left(\frac{\bigcup_{i=1}^n B_i}{A}\right) = \frac{P\left(\left(\bigcup_{i=1}^n B_i\right) \cap A\right)}{P(A)} = \frac{P\left(\bigcup_{i=1}^n (B_i \cap A)\right)}{P(A)} = \frac{\sum_{i=1}^n P(B_i \cap A)}{P(A)} = \sum_{i=1}^n \frac{P(B_i \cap A)}{P(A)} = \sum_{i=1}^n P\left(\frac{B_i}{A}\right)$

3.- $P\left(\frac{\Omega}{A}\right) = \frac{P(\Omega \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1$

Por tanto las propiedades de la probabilidad (no condicionada) también se cumplen para la probabilidad condicionada.

De forma análoga a como se ha definido $P\left(\frac{B}{A}\right)$ se define $P\left(\frac{A}{B}\right)$ como:

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} \quad P(B) > 0$$

De las definiciones de $P\left(\frac{B}{A}\right)$ y $P\left(\frac{A}{B}\right)$ se deduce la **fórmula de la probabilidad compuesta o fórmula del producto de probabilidades:**

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

Para tres sucesos el anterior resultado se puede enunciar de 3!=6 formas alternativas, una de ellas es:

$$P(A \cap B \cap C) = P(A)P\left(\frac{B}{A}\right)P\left(\frac{C}{A \cap B}\right)$$

y para n sucesos una de sus posibles formas es:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P\left(\frac{A_2}{A_1}\right)P\left(\frac{A_3}{A_1 \cap A_2}\right) \dots P\left(\frac{A_n}{A_1 \cap A_2 \cap \dots \cap A_{n-1}}\right)$$

Sucesos dependientes e independientes.

En general $P(B/A) \neq P(B)$ y decimos en tal caso que B **depende** de A . Si $P(B/A) = P(B)$ diremos que B es **independiente** de A , es decir, que haya ocurrido A no modifica la probabilidad de B .

En el caso de **independencia**:

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

Y como consecuencia se sigue que también:

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

luego si B es independiente de A , entonces A es independiente de B .

Según todo lo anterior, se toma como definición de que los sucesos A y B son **independientes** la siguiente expresión:

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

La noción de independencia es fundamental en la teoría de la probabilidad. La mayoría de los resultados en probabilidad se obtienen bajo la hipótesis de independencia.

► EJEMPLO 5.3

En un dado con igual probabilidad de aparición para todas sus caras comprobar que los siguientes sucesos, dos a dos, son dependientes.

$A = \{\text{obtener par}\}$, $B = \{\text{obtener impar}\}$, $C = \{\text{obtener un 3}\}$, $D = \{\text{obtener un número mayor o igual que 4}\}$, $E = \{\text{obtener un número menor que 4}\}$.

$$P(A \cap B) = 0 \neq P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(A \cap C) = 0 \neq P(A)P(C) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

$$P(A \cap D) = \frac{2}{6} \neq P(A)P(D) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(A \cap E) = \frac{1}{6} \neq P(A)P(E) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(B \cap C) = \frac{1}{6} \neq P(B)P(C) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

$$P(B \cap D) = \frac{1}{6} \neq P(B)P(D) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(B \cap E) = \frac{2}{6} \neq P(B)P(E) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(C \cap D) = 0 \neq P(C)P(D) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

$$P(C \cap E) = \frac{1}{6} \neq P(C)P(E) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

$$P(D \cap E) = 0 \neq P(D)P(E) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

► EJEMPLO 5.4

En el bar de la Facultad de CC. EE. y EE. el 90% de los clientes son estudiantes. Se sabe que el 45% de los clientes toman café y que el 30% de los estudiantes toman café.

- ¿Cuál es la probabilidad de que un cliente elegido al azar sea estudiante y tome café?
- Se elige al azar un cliente que toma café, ¿cuál es la probabilidad de que sea estudiante?
- ¿Son independientes los sucesos *tomar café* y *ser estudiante*?

Solución:

Definimos los sucesos: $E = \{\text{el cliente es estudiante}\}$, $C = \{\text{el cliente toma café}\}$

- Utilizamos la fórmula de la probabilidad compuesta:

$$P(C \cap E) = P(E)P(C/E) = \frac{90}{100} \frac{30}{100} = 0,27$$

- Según la definición de probabilidad condicionada:

$$P(E/C) = \frac{P(C \cap E)}{P(C)} = \frac{0,27}{0,45} = 0,6$$

- $P(C \cap E) = 0,27 \neq P(C)P(E) = 0,45 \times 0,90 = 0,405$, por lo tanto no son independientes.



5.3 Fórmula de la probabilidad total. Fórmula de Bayes.

Sean los sucesos A_1, \dots, A_n una *partición* de Ω , es decir

$$A_i \cap A_j = \emptyset \quad \forall i \neq j \quad ; \quad \bigcup_{i=1}^n A_i = \Omega$$

en este contexto, la **fórmula de la probabilidad total** afirma que:

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

Demostración:

$$P(B) = P(B \cap \Omega) = P\left(B \cap \left(\bigcup_{i=1}^n A_i\right)\right) = P\left(\bigcup_{i=1}^n (B \cap A_i)\right) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B/A_i)P(A_i)$$

Las probabilidades condicionadas de un suceso cualquiera B sobre cada uno de los elementos de la partición, $P(B/A_i)$, son normalmente conocidas. No es así con las probabilidades $P(A_i/B)$.

Veamos cómo resolver este problema:

por la fórmula de la probabilidad compuesta $P(A_k \cap B) = P(B)P\left(\frac{A_k}{B}\right) = P(A_k)P\left(\frac{B}{A_k}\right)$

y según la fórmula de la probabilidad total $P(B) = \sum_{i=1}^n P\left(\frac{B}{A_i}\right)P(A_i)$

de ambas expresiones se sigue:

$$P\left(\frac{A_k}{B}\right) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P\left(\frac{B}{A_k}\right)}{\sum_{i=1}^n P\left(\frac{B}{A_i}\right)P(A_i)}$$

Esta última expresión se conoce como **fórmula de Bayes**.

► EJEMPLO 5.5

La producción de una factoría se realiza en cuatro máquinas, M_1, M_2, M_3 y M_4 . Diariamente la producción de cada una de las máquinas es la siguiente:

M_1	M_2	M_3	M_4	TOTAL
600	500	350	250	1700

Además sabemos que los porcentajes de piezas defectuosas producidas por cada una de las máquinas son:

M_1	M_2	M_3	M_4
4%	3,5%	4,6%	2%

- Si las piezas se almacenan conjuntamente, ¿cuál es la probabilidad de que al seleccionar una pieza al azar ésta sea defectuosa?
- Se ha seleccionado una pieza defectuosa, ¿cuál es la probabilidad de que haya sido producida en la máquina M_2 ?

Solución:

Sean M_1, M_2, M_3 y M_4 los sucesos *la pieza ha sido producida en M_1, M_2, M_3 y M_4 respectivamente*.

Sea D el suceso *la pieza es defectuosa*.

Los sucesos M_1, M_2, M_3 y M_4 constituyen una partición de la producción total de la factoría.

- Según la fórmula de la probabilidad total

$$P(D) = \sum_{i=1}^4 P\left(\frac{D}{M_i}\right)P(M_i) = \left(\frac{4}{100} \frac{600}{1700}\right) + \left(\frac{3,5}{100} \frac{500}{1700}\right) + \left(\frac{4,6}{100} \frac{350}{1700}\right) + \left(\frac{2}{100} \frac{250}{1700}\right) = \frac{6260}{170000} = 0,03682353$$

b) Según la fórmula de Bayes

$$P\left(\frac{M_2}{D}\right) = \frac{P\left(\frac{D}{M_2}\right)P(M_2)}{\sum_{i=1}^4 P\left(\frac{D}{M_i}\right)P(M_i)} = \frac{\frac{3,5}{100} \frac{500}{1700}}{\frac{6260}{170000}} = \frac{1750}{6260} = 0,27955$$

► EJEMPLO 5.6

En una empresa el 8% de los hombres y el 4,3% de las mujeres ganan más de 25000€ al año. Se sabe que el porcentaje de mujeres en la empresa es del 47%. Se selecciona al azar un empleado que gana menos de 25000€, ¿cuál es la probabilidad de que sea mujer?

Solución:

Definimos los sucesos, $H = \{\text{el empleado es hombre}\}$, $M = \{\text{el empleado es mujer}\}$ y $S = \{\text{el empleado supera los 25000€}\}$. Consecuentemente $\bar{S} = \{\text{el empleado gana menos de 25000€}\}$.

Sabemos:

$$P\left(\frac{S}{H}\right) = \frac{8}{100} \quad P\left(\frac{S}{M}\right) = \frac{4,3}{100} \quad P(M) = \frac{47}{100}$$

Por tanto:

$$P\left(\frac{\bar{S}}{H}\right) = 1 - P\left(\frac{S}{H}\right) = 1 - \frac{8}{100} = \frac{92}{100} \quad P\left(\frac{\bar{S}}{M}\right) = 1 - P\left(\frac{S}{M}\right) = 1 - \frac{4,3}{100} = \frac{95,7}{100}$$

$$P(H) = P(\bar{M}) = 1 - P(M) = 1 - \frac{47}{100} = \frac{53}{100}$$

Teniendo en cuenta que los sucesos H y M son una partición de los empleados de la empresa y aplicando la fórmula de Bayes, la probabilidad pedida es:

$$P\left(\frac{M}{\bar{S}}\right) = \frac{P\left(\frac{\bar{S}}{M}\right)P(M)}{P\left(\frac{\bar{S}}{M}\right)P(M) + P\left(\frac{\bar{S}}{H}\right)P(H)} = \frac{\frac{95,7}{100} \frac{47}{100}}{\left(\frac{95,7}{100} \frac{47}{100}\right) + \left(\frac{92}{100} \frac{53}{100}\right)} = \frac{4497,9}{9373,9} = 0,4798323$$

5.4 Ejercicios resueltos.

1. Una empresa compra bombillas a dos proveedores (A y B). Al primero le compra el 30% de las bombillas, siendo defectuosas el 3% de las mismas. También son defectuosas el 2% de las bombillas compradas al segundo proveedor.

- a) Calcule la probabilidad de que una bombilla comprada por la empresa no sea defectuosa.
- b) En un control se ha seleccionado una bombilla defectuosa. Calcule la probabilidad de que se haya comprado al segundo proveedor.

Solución:

A =bombilla comprada al proveedor A B =bombilla comprada al proveedor B

D =bombilla defectuosa

$$P(A) = 0,30 \quad P(D/A) = 0,03 \Rightarrow P(\bar{D}/A) = 1 - 0,03 = 0,97$$

$$P(B) = 1 - 0,30 = 0,70 \quad P(D/B) = 0,02 \Rightarrow P(\bar{D}/B) = 1 - 0,02 = 0,98$$

$$a) \quad P(\bar{D}) = P(\bar{D}/A)P(A) + P(\bar{D}/B)P(B) = (0,97 \times 0,30) + (0,98 \times 0,70) = 0,977 \Rightarrow 97,7\%$$

$$b) \quad P(B/D) = \frac{P(D/B)P(B)}{P(D)} = \frac{0,02 \times 0,70}{1 - 0,977} = \frac{0,014}{0,023} = 0,6087$$

2. El servicio municipal de recaudación de la tasa de basura en una ciudad establece dicha tasa en función de la categoría de la calle. En concreto, en este municipio hay 3 categorías: un 25% de las calles son de categoría I y un 40% son de categoría II. Además, el 4% de los propietarios de viviendas en calles de categoría I deben al menos un recibo. El 98% de los propietarios en calles de categoría II están al día en sus recibos de basura y este porcentaje es del 97% en las viviendas en calles de categoría III.

- a)Cuál es la probabilidad de que una vivienda esté al día en el pago de sus recibos.
- b) Se selecciona al azar una vivienda y está al día en el pago de sus recibos, ¿cuál será la categoría de la calle en la que se encuentra con mayor probabilidad?

Solución:

Con los datos del problema y según la siguiente notación obtenemos las probabilidades condicionadas y no condicionadas que aparecen más abajo:

I = calle de categoría I

II = calle de categoría II

III = calle de categoría III

D = el propietario debe al menos un recibo

\bar{D} = el propietario no debe ningún recibo, está al día en el pago de sus recibos

$$P(I) = 0,25 \quad P(II) = 0,40 \quad P(III) = 1 - 0,25 - 0,40 = 0,35$$

$$P(D/I) = 0,04 \Rightarrow P(\bar{D}/I) = 1 - 0,04 = 0,96$$

$$P(\overline{D}/II) = 0,98 \Rightarrow P(D/II) = 1 - 0,98 = 0,02$$

$$P(\overline{D}/III) = 0,97 \Rightarrow P(D/III) = 1 - 0,97 = 0,03$$

a)

$$P(\overline{D}) = P(\overline{D}/I)P(I) + P(\overline{D}/II)P(II) + P(\overline{D}/III)P(III) = \\ = (0,96 \times 0,25) + (0,98 \times 0,40) + (0,97 \times 0,35) = 0,9715$$

b)

$$P(I/D) = \frac{P(\overline{D}/I)P(I)}{P(\overline{D})} = \frac{0,96 \times 0,25}{0,9715} = 0,247$$

$$P(II/D) = \frac{P(\overline{D}/II)P(II)}{P(\overline{D})} = \frac{0,98 \times 0,40}{0,9715} = 0,4035$$

$$P(III/D) = \frac{P(\overline{D}/III)P(III)}{P(\overline{D})} = \frac{0,97 \times 0,35}{0,9715} = 0,349$$

Lo más probable es que pertenezca a una calle de categoría *II* (la mayor de las tres probabilidades es 0,4035).

3. Un inversionista está pensando comprar un número grande de acciones de una compañía. Se sabe que la cotización de las acciones se relaciona con el PNB. Si el PNB aumenta, la probabilidad de que suba el valor de las acciones es 0,8. Si el PNB no cambia, la probabilidad de que suban las acciones es de 0,2. Si el PNB disminuye, la probabilidad de subir las acciones es sólo de 0,1.

Si para los siguientes seis meses se asignan las probabilidades 0,4 , 0,3 y 0,3 a los eventos: el PNB aumenta, no cambia y disminuye, respectivamente. Determine la probabilidad de que las acciones suban de valor en los próximos seis meses.

Solución:

S=subir las acciones. *A*=aumentar el PNB. *N*=no cambiar el PNB. *D*=disminuir el PNB.

$$P(S/A) = 0,8 \quad P(S/N) = 0,2 \quad P(S/D) = 0,1$$

$$P(A) = 0,4 \quad P(N) = 0,3 \quad P(D) = 0,3$$

$$P(S) = P(S/A)P(A) + P(S/N)P(N) + P(S/D)P(D) = 0,32 + 0,06 + 0,03 = 0,41$$

4. Un avión realiza diariamente el mismo servicio. Estadísticamente se ha comprobado que la probabilidad de accidente en un día sin niebla es de 0,002 y en día con niebla 0,01. Cierta día de un mes, en el que hubo 18 días sin niebla y 12 con niebla, se produjo un accidente. Calcule la probabilidad de que el accidente haya ocurrido:

a) En día sin niebla.

b) En día con niebla.

Solución:

A =accidente del avión. N =día con niebla. \bar{N} =día sin niebla.

$$P\left(\frac{A}{\bar{N}}\right) = 0,002 \quad P\left(\frac{A}{N}\right) = 0,01$$

$$P(\bar{N}) = \frac{18}{30} = 0,6 \quad P(N) = \frac{12}{30} = 0,4$$

$$a) \quad P\left(\frac{\bar{N}}{A}\right) = \frac{P\left(\frac{A}{\bar{N}}\right)P(\bar{N})}{P\left(\frac{A}{\bar{N}}\right)P(\bar{N}) + P\left(\frac{A}{N}\right)P(N)} = \frac{0,002 \times 0,6}{(0,002 \times 0,6) + (0,01 \times 0,4)} = 0,23077$$

$$b) \quad P\left(\frac{N}{A}\right) = 1 - P\left(\frac{\bar{N}}{A}\right) = 1 - 0,2377 = 0,76923$$

5. En una caja hay 10 piezas de la fábrica A, 15 de la fábrica B y 25 de la C. La fábrica A produce un 80% de piezas excelentes, siendo excelentes el 90% de las piezas de la fábrica B y sólo el 70% de la fábrica C.

a) Calcule la probabilidad de que si extraemos una pieza al azar, ésta resulte de calidad excelente.

b) Se extrae una pieza al azar y resulta que no es excelente, calcule la probabilidad de que proceda de la fábrica B.

Solución:

E =pieza excelente. A =pieza de A. B =pieza de B. C =pieza de C.

$$P(A) = \frac{10}{50} = 0,2 \quad P(B) = \frac{15}{50} = 0,3 \quad P(C) = \frac{25}{50} = 0,5$$

$$P\left(\frac{E}{A}\right) = 0,80 \quad P\left(\frac{E}{B}\right) = 0,90 \quad P\left(\frac{E}{C}\right) = 0,70$$

a)

$$P(E) = P\left(\frac{E}{A}\right)P(A) + P\left(\frac{E}{B}\right)P(B) + P\left(\frac{E}{C}\right)P(C) = \\ = (0,8 \times 0,2) + (0,9 \times 0,3) + (0,7 \times 0,5) = 0,78$$

$$b) P\left(\frac{B}{\bar{E}}\right) = \frac{P\left(\frac{\bar{E}}{B}\right)P(B)}{P(\bar{E})} = \frac{P\left(\frac{\bar{E}}{B}\right)P(B)}{1-P(E)} = \frac{(1-0,9)0,3}{1-0,78} = 0,13636$$

6. En segundo curso de GECO en la Facultad de Económicas hay cuatro grupos. El grupo A tiene 50 alumnos matriculados de la provincia de Granada y otros 50 de fuera, en el grupo B hay 90 matriculados, de los que 40 son de la provincia de Granada, en el grupo C hay 80 matriculados, siendo 30 de fuera de la provincia y el grupo D tiene 30 matriculados de la provincia de Granada y 40 de fuera.

Se escoge al azar un alumno de dicho curso de GECO:

- Calcule la probabilidad de que el alumno sea de la provincia de Granada.
- El alumno responde que no es de la provincia de Granada, calcule la probabilidad de que no sea del grupo B.

Solución:

G =alumno de la provincia de Granada.

A =alumno del grupo A B =alumno del grupo B C =alumno del grupo C D =alumno del grupo D

$$P(A) = \frac{100}{100+90+80+70} = \frac{100}{340} \quad P(B) = \frac{90}{340} \quad P(C) = \frac{80}{340} \quad P(D) = \frac{70}{340}$$

$$P\left(\frac{G}{A}\right) = \frac{50}{100} \quad P\left(\frac{G}{B}\right) = \frac{40}{90} \quad P\left(\frac{G}{C}\right) = \frac{50}{80} \quad P\left(\frac{G}{D}\right) = \frac{30}{70}$$

$$\begin{aligned} a) P(G) &= P\left(\frac{G}{A}\right)P(A) + P\left(\frac{G}{B}\right)P(B) + P\left(\frac{G}{C}\right)P(C) + P\left(\frac{G}{D}\right)P(D) = \\ &= \left(\frac{50}{100} \times \frac{100}{340}\right) + \left(\frac{40}{90} \times \frac{90}{340}\right) + \left(\frac{50}{80} \times \frac{80}{340}\right) + \left(\frac{30}{70} \times \frac{70}{340}\right) = \frac{50+40+50+30}{340} = \frac{170}{340} = \frac{1}{2} = 0,5 \end{aligned}$$

$$b) P\left(\frac{\bar{B}}{\bar{G}}\right) = 1 - P\left(\frac{B}{\bar{G}}\right) = 1 - 0,2941 = 0,7059$$

$$P\left(\frac{B}{\bar{G}}\right) = \frac{P\left(\frac{\bar{G}}{B}\right)P(B)}{P(\bar{G})} = \frac{\left(1 - P\left(\frac{G}{B}\right)\right)P(B)}{1 - P(G)} = \frac{\left(1 - \frac{40}{90}\right)\frac{90}{340}}{1 - \frac{1}{2}} = \frac{50}{170} = 0,2941$$

7. En una comunidad autonómica, el 18% de los hombres y el 15% de las mujeres presentaron una declaración sobre el IRPF con rentas superiores a 35000€ al año. El 45% de todas las declaraciones recibidas corresponden a mujeres. Se pide:

- ¿Qué porcentaje de declaraciones no han superado los 35000€?

- b) Se selecciona al azar una declaración y resulta superior a los 35000€, determine la probabilidad de que corresponda a una mujer.

Solución:

S =declaración IRPF con renta superior a 35000€. H =hombre. M =mujer.

$$P(S/H) = 0,18 \quad P(S/M) = 0,15$$

$$P(H) = 1 - P(M) = 0,55 \quad P(M) = 0,45$$

a) $P(\bar{S}) = 1 - P(S) = 1 - 0,1665 = 0,8335 \Rightarrow 83,35\%$

$$P(S) = P(S/H)P(H) + P(S/M)P(M) = (0,18 \times 0,55) + (0,15 \times 0,45) = 0,1665$$

b) $P(M/S) = \frac{P(S/M)P(M)}{P(S)} = \frac{0,15 \times 0,45}{0,1665} = 0,4054$

8. La producción de una factoría se realiza en cuatro máquinas: A, B, C y D. Diariamente la máquina A produce 150 piezas, B produce 250, C produce 275 y D produce 325. Las probabilidades de que una pieza sea defectuosa son: 0,05 si la produce A, 0,04 si la produce B, 0,03 si la produce C y 0,06 si la produce D.

- a) Se ha elegido una pieza al azar y es defectuosa. Calcule la probabilidad de que haya sido producida por la máquina B.
- b) Calcule la probabilidad de que una pieza elegida al azar no sea defectuosa.

Solución:

A =pieza producida en A. B =pieza producida en B. C =pieza producida en C.

D =pieza producida en D. F =pieza defectuosa.

$$P(A) = \frac{150}{150 + 250 + 275 + 325} = \frac{150}{1000} \quad P(B) = \frac{250}{1000} \quad P(C) = \frac{275}{1000} \quad P(D) = \frac{325}{1000}$$

$$P(F/A) = 0,05 \quad P(F/B) = 0,04 \quad P(F/C) = 0,03 \quad P(F/D) = 0,06$$

a)
$$P(B/F) = \frac{P(F/B)P(B)}{P(F)} = \frac{P(F/B)P(B)}{P(F/A)P(A) + P(F/B)P(B) + P(F/C)P(C) + P(F/D)P(D)} =$$

$$= \frac{0,04 \times 0,25}{(0,05 \times 0,15) + (0,04 \times 0,25) + (0,03 \times 0,275) + (0,06 \times 0,325)} = \frac{0,01}{0,04525} = 0,22099$$

b) $P(\bar{F}) = 1 - P(F) = 1 - 0,04525 = 0,95475$

9. Una empresa de electrónica recibe suministros de tres mayoristas distintos. El 25% de los paquetes se los envía el mayorista A, el 40% el mayorista B y el resto el mayorista C. Cinco de cada mil envíos hechos por el mayorista A salen defectuosos, así como dos de cada mil envíos del mayorista B y uno de cada mil del mayorista C. Se ha recibido un envío defectuoso, calcule la probabilidad de que haya sido remitido por el mayorista A.

Solución:

A =suministro del mayorista A. B = suministro del mayorista B. C = suministro del mayorista C.
 D = suministro defectuoso.

$$P(A) = 0,25 \quad P(B) = 0,40 \quad P(C) = 1 - 0,25 - 0,40 = 0,35$$

$$P(D/A) = \frac{5}{1000} = 0,005 \quad P(D/B) = \frac{2}{1000} = 0,002 \quad P(D/C) = \frac{1}{1000} = 0,001$$

$$P(A/D) = \frac{P(D/A)P(A)}{P(D/A)P(A) + P(D/B)P(B) + P(D/C)P(C)} =$$

$$= \frac{0,005 \times 0,25}{(0,005 \times 0,25) + (0,002 \times 0,40) + (0,001 \times 0,35)} = \frac{0,00125}{0,0024} = 0,52083$$

10. Cada uno de los envíos que una empresa realiza a sus clientes está formado por 100 piezas. De ellas, 60 son fabricadas en la factoría A y las restantes proceden de diversas factorías. El 3% de las piezas fabricadas en la factoría A son defectuosas, también son defectuosas el 5% de las piezas que no han sido fabricadas en la factoría A. En uno de los envíos se elige una pieza al azar y resulta ser defectuosa, calcule la probabilidad de que no haya sido fabricada en la factoría A.

Solución:

A =pieza de la factoría A. \bar{A} =pieza de otra factoría. D =pieza defectuosa.

$$P(A) = \frac{60}{100} = 0,6 \quad P(\bar{A}) = \frac{40}{100} = 0,4$$

$$P(D/A) = \frac{3}{100} = 0,03 \quad P(D/\bar{A}) = \frac{5}{100} = 0,05$$

$$P(\bar{A}/D) = \frac{P(D/\bar{A})P(\bar{A})}{P(D/A)P(A) + P(D/\bar{A})P(\bar{A})} = \frac{0,05 \times 0,4}{(0,03 \times 0,6) + (0,05 \times 0,4)} = \frac{0,02}{0,038} = 0,5263$$

11. Una compañía de autobuses atiende tres líneas periféricas de una ciudad. El 60% de los autobuses cubren el servicio de la primera línea, el 30% la segunda y el resto de autobuses la tercera. Se sabe

que la probabilidad de que un autobús se averíe durante un mes es del 2% en la primera línea, del 4% en la segunda y del 1% en la tercera.

- Determine la probabilidad de que en un mes se averíe un autobús.
- Sabiendo que un autobús ha sufrido una avería, calcule la probabilidad de que no sea de la primera línea.
- Se elige al azar un autobús y se comprueba que durante el mes en estudio no ha sufrido avería, calcule la probabilidad de que el autobús elegido cubra el servicio en la segunda línea.

Solución:

I =autobús cubre servicio de la primera línea.

II = autobús cubre servicio de la segunda línea.

III = autobús cubre servicio de la tercera línea.

A =autobús sufre avería en un mes.

$$P(I) = 0,60 \quad P(II) = 0,30 \quad P(III) = 1 - 0,60 - 0,30 = 0,10$$

$$P\left(\frac{A}{I}\right) = 0,02 \quad P\left(\frac{A}{II}\right) = 0,04 \quad P\left(\frac{A}{III}\right) = 0,01$$

$$\begin{aligned} \text{a) } P(A) &= P\left(\frac{A}{I}\right)P(I) + P\left(\frac{A}{II}\right)P(II) + P\left(\frac{A}{III}\right)P(III) = \\ &= (0,02 \times 0,6) + (0,04 \times 0,3) + (0,01 \times 0,1) = 0,025 \end{aligned}$$

$$\text{b) } P\left(\frac{\bar{A}}{A}\right) = 1 - P\left(\frac{A}{A}\right) = 1 - 0,48 = 0,52$$

$$P\left(\frac{II}{A}\right) = \frac{P\left(\frac{A}{II}\right)P(II)}{P(A)} = \frac{0,04 \times 0,3}{0,025} = 0,48$$

$$\text{c) } P\left(\frac{II}{\bar{A}}\right) = \frac{P\left(\frac{\bar{A}}{II}\right)P(II)}{P(\bar{A})} = \frac{(1 - P\left(\frac{A}{II}\right))P(II)}{1 - P(A)} = \frac{(1 - 0,04) \times 0,3}{(1 - 0,025)} = \frac{0,288}{0,975} = 0,29358$$

12. La probabilidad de que ocurra un suceso A es 0,6. La probabilidad de que ocurra otro suceso B es p . La probabilidad de la unión de ambos sucesos es 0,8.

- ¿Pueden ser los sucesos A y B independientes? En caso afirmativo, calcule el valor de p para que los sucesos A y B sean independientes.
- ¿Pueden ser los sucesos A y B incompatibles? En caso afirmativo, calcule el valor p para que los sucesos A y B sean incompatibles.
- ¿Pueden ser los sucesos A y B complementarios? En caso afirmativo, calcule el valor p para que los sucesos A y B sean complementarios.

Solución:

$$P(A) = 0,6 \quad P(B) = p \quad P(A \cup B) = 0,8$$

a) Para que A y B sean independientes se tiene que cumplir: $P(A \cap B) = P(A)P(B) = 0,6 \times p$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,6 + p - (0,6 \times p) = 0,8$$

$$0,6 + p - (0,6 \times p) = 0,8 \Leftrightarrow 0,6 + (0,4 \times p) = 0,8 \Leftrightarrow (0,4 \times p) = 0,2 \Leftrightarrow p = \frac{0,2}{0,4} = 0,5$$

b) Para que A y B sean incompatibles se tiene que cumplir: $P(A \cap B) = 0$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,6 + p = 0,8 \Leftrightarrow p = 0,2$$

c) Si A es el complementario de B se tiene que cumplir: $P(A) = 1 - P(B)$

$$0,6 = 1 - p \Leftrightarrow p = 0,4$$

13. Con base en varios estudios, una compañía ha clasificado, de acuerdo con la probabilidad de descubrir petróleo, las formaciones geológicas en tres tipos. La compañía pretende perforar un pozo en un determinado sitio al que se le asignan las probabilidades 0,35 , 0,4 y 0,25 para los tres tipos de formaciones geológicas, respectivamente. De acuerdo con la experiencia se sabe que el petróleo se encuentra en el 40% de las formaciones tipo I, en el 20% de las formaciones tipo II y en el 30% de las formaciones tipo III.

a) Calcule la probabilidad de que al realizar una perforación no se encuentre petróleo.

b) Si se realiza la perforación y no se encuentra petróleo, determine la probabilidad de que en el sitio haya una formación de tipo II.

Solución:

I =formación geológica tipo I.

II = formación geológica tipo II.

III =formación geológica tipo III.

E =encuentre petróleo.

$$P(I) = 0,35 \quad P(II) = 0,40 \quad P(III) = 0,25$$

$$P\left(\frac{E}{I}\right) = 0,40 \quad P\left(\frac{E}{II}\right) = 0,20 \quad P\left(\frac{E}{III}\right) = 0,30$$

$$a) \quad P(\bar{E}) = 1 - P(E) = 1 - 0,295 = 0,705$$

$$\begin{aligned} P(E) &= P\left(\frac{E}{I}\right)P(I) + P\left(\frac{E}{II}\right)P(II) + P\left(\frac{E}{III}\right)P(III) = \\ &= (0,40 \times 0,35) + (0,20 \times 0,40) + (0,30 \times 0,25) = 0,295 \end{aligned}$$

$$\text{b) } P\left(\frac{H}{\bar{E}}\right) = \frac{P\left(\frac{\bar{E}}{H}\right)P(H)}{P(\bar{E})} = \frac{(1-P\left(\frac{E}{H}\right))P(H)}{0,705} = \frac{(1-0,20) \times 0,40}{0,705} = 0,4539$$

6. VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD.

6.1 Concepto de variable aleatoria y distribución de probabilidad.

Sea Ω el conjunto de sucesos elementales, ω_i , de un fenómeno aleatorio. Se define una **variable aleatoria** X como una función sobre Ω en \mathbb{R}

$$\begin{aligned}\Omega &\xrightarrow{X} \mathbb{R} \\ \omega_i &\rightarrow X(\omega_i) = x_i \in \mathbb{R}\end{aligned}$$

Es decir, mediante una variable aleatoria transformamos los resultados de un fenómeno aleatorio en valores numéricos que miden aquella característica del fenómeno que nos interesa destacar (*en el siguiente ejemplo la característica es el número de cruces en tres lanzamientos de una moneda*).

Las probabilidades de que la variable aleatoria X tome los distintos valores dependen de las probabilidades, $P(\omega_i)$, asociadas al fenómeno aleatorio sobre el que se ha definido.

Se denomina **distribución de probabilidad** de la variable aleatoria X a las probabilidades asociadas a los distintos valores que puede tomar la variable aleatoria, $p_i = P[X = x_i]$.

Toda distribución de probabilidad ha de cumplir:

- $p_i \geq 0 \quad \forall i$
- $\sum_{\forall i} p_i = 1$

(Esta definición es válida para variables que toman un conjunto discreto de valores, como el siguiente ejemplo. Sin embargo, cuando la variable aleatoria puede tomar un conjunto continuo de valores, la distribución de probabilidad no está caracterizada por probabilidades de valores aislados como veremos en el epígrafe 6.2).

► EJEMPLO 6.1

Represente la variable aleatoria X =número de cruces en tres lanzamientos de una moneda y obtenga la distribución de probabilidad asociada.

Solución:

C=cara, +=cruz

Ω	\xrightarrow{X}	\mathbb{R}
CCC	\rightarrow	$X(CCC) = 0$
+CC	\rightarrow	$X(+CC) = 1$
C+C	\rightarrow	1
CC+	\rightarrow	1
++C	\rightarrow	2
+C+	\rightarrow	2
C++	\rightarrow	2
+++	\rightarrow	3

Suponemos que la probabilidad de cara y cruz es la misma, $\frac{1}{2}$. Por tanto la probabilidad de cada

uno de los sucesos elementales de Ω es $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$ dado que son independientes los resultados de los tres lanzamientos de la moneda

$$P(CCC) = P(C \cap C \cap C) = P(C)P(C)P(C) = \frac{1}{2} \frac{1}{2} \frac{1}{2} = \frac{1}{8} \quad \dots$$

$$P(+++) = P(+)P(+)P(+) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

$$p_0 = P[X = 0] = P[(CCC)] = \frac{1}{8}$$

$$p_1 = P[X = 1] = P[(+CC) \cup (C+C) \cup (CC+)] = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

$$p_2 = P[X = 2] = P[(++C) \cup (+C+) \cup (C++)] = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

$$p_3 = P[X = 3] = P[(+++)] = \frac{1}{8}$$

La **distribución de probabilidad** se representa en una tabla como la que sigue

x_i	p_i
0	1/8
1	3/8
2	3/8
3	1/8

Como puede comprobarse $\sum_{i=0}^3 p_i = 1$.



6.2 Función de distribución. Variables aleatorias discretas y variables aleatorias continuas.

Se define la **función de distribución** de una variable aleatoria X como:

$$F(x) = P[X \leq x] = P[\omega \in \Omega / X(\omega) \leq x] \quad \forall x \in \mathbb{R}$$

Propiedades:

- Si $x_i < x_j \Rightarrow F(x_i) \leq F(x_j)$ (la función es no decreciente)
- $F(-\infty) = 0 \quad F(+\infty) = 1$
- $P[X > x] = 1 - F(x)$
- $P[x_1 < X \leq x_2] = F(x_2) - F(x_1)$

► EJEMPLO 6.2

Calcule la función de distribución asociada a la variable aleatoria X del ejemplo 6.1. Represente su gráfica.

Solución:

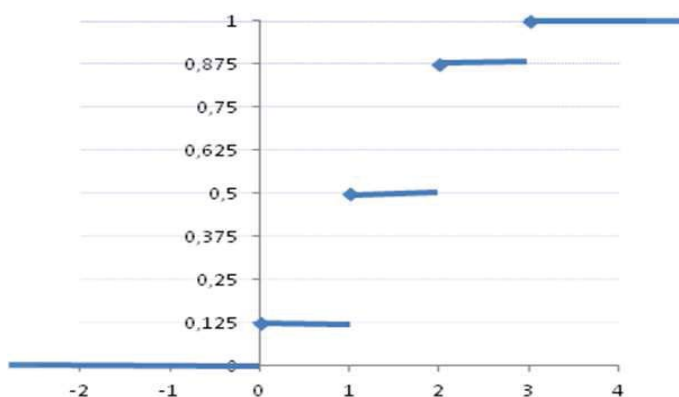
La distribución de probabilidad del ejemplo 6.1 está dada por:

x_i	p_i
0	1/8
1	3/8
2	3/8
3	1/8

Su función de distribución, por tanto, será:

$$F(x) = P[X \leq x] = \sum_{x_i \leq x} p_i = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{8} & \text{si } 0 \leq x < 1 \\ \frac{4}{8} & \text{si } 1 \leq x < 2 \\ \frac{7}{8} & \text{si } 2 \leq x < 3 \\ 1 & \text{si } 3 \leq x \end{cases}$$

y su representación gráfica:



Variables aleatorias discretas y variables aleatorias continuas.

Dependiendo del tipo de variable aleatoria, la función de distribución tendrá unas características diferentes.

Si la variable aleatoria toma un conjunto de valores aislados (como el ejemplo 6.1 donde la variable aleatoria sólo toma los valores 0, 1, 2 y 3) diremos que es una **variable aleatoria discreta**.

Si la variable aleatoria toma un conjunto continuo de valores (por ejemplo la duración de una bombilla, que puede tomar cualquier valor a partir de 0) diremos que es una **variable aleatoria continua**.

Como puede verse en el ejemplo 6.2, la **función de distribución** de una **variable aleatoria discreta** es una **función escalonada**, con saltos o discontinuidades de tamaño p_i en los valores x_i que toma la variable aleatoria. En este tipo de variables la función de distribución es igual a

$$F(x) = P[X \leq x] = \sum_{x_i \leq x} p_i$$

La **función de distribución** de una **variable aleatoria continua** es una **función continua**. La derivada de la función de distribución de una variable aleatoria continua se denomina **función de densidad**

$$F'(x) = f(x)$$

Como consecuencia de las propiedades de la función de distribución, toda función de densidad cumple que:

- $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x) dx = 1$

La probabilidad de que la variable aleatoria X tome valores en un determinado conjunto es igual a la integral definida de la función de densidad sobre dicho conjunto.

En particular:

$$P[X = a] = \int_a^a f(x) dx = 0$$

$$P[a < X < b] = \int_a^b f(x) dx$$

$$P[a < X] = \int_a^{+\infty} f(x) dx$$

$$P[X < a] = \int_{-\infty}^a f(x) dx$$

y por tanto la función de distribución se puede obtener a partir de la función de densidad:

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t) dt$$

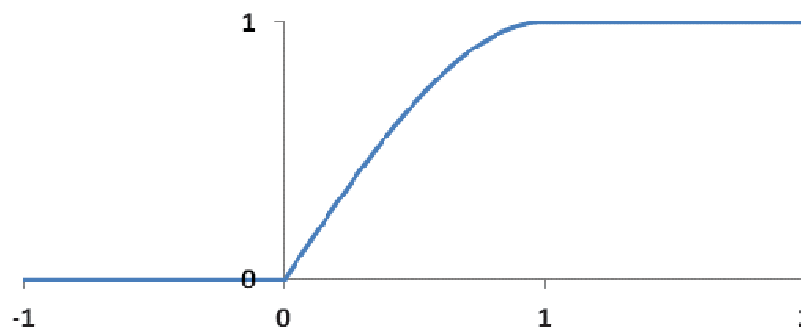
Como puede verse, en variables aleatorias continuas la probabilidad de un valor aislado es cero por lo que no tiene sentido definir para ellas el concepto de distribución de probabilidad tal y como apareció en el epígrafe 6.1. En **variables aleatorias continuas** su **distribución de probabilidad** está definida por su **función de densidad**.

► EJEMPLO 6.3

Sea la función de distribución de la variable aleatoria X :

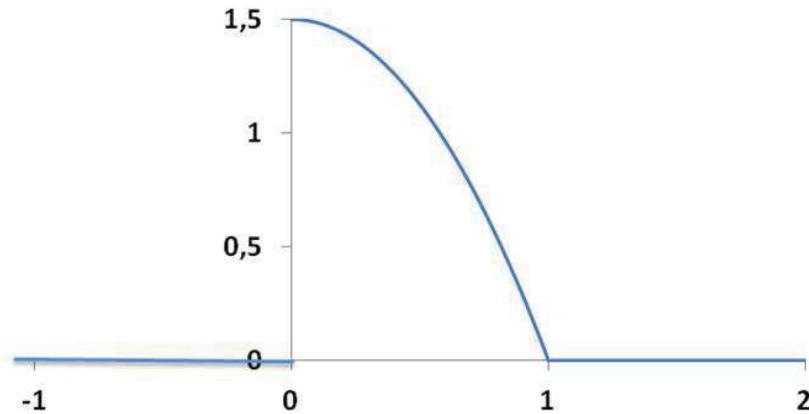
$$F(x) = \begin{cases} 0 & , x < 0 \\ \frac{3x - x^3}{2} & , 0 \leq x \leq 1 \\ 1 & , 1 < x \end{cases}$$

Como puede comprobarse es una función continua cuya gráfica es:



Derivando la función de distribución obtenemos la función de densidad de la variable aleatoria X

$$f(x) = \begin{cases} 0 & , x < 0 \\ \frac{3-3x^2}{2} = \frac{3}{2}(1-x^2) & , 0 \leq x \leq 1 \\ 0 & , 1 < x \end{cases}$$



Comprobemos que $f(x)$ cumple las dos propiedades que caracterizan a toda función de densidad:

- $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x) dx = 1$

La primera es inmediata a partir de su expresión, dado que $x \leq 1 \Rightarrow 0 \leq (1-x^2)$.

Y para comprobar la segunda, se calcula la integral:

$$\int_{-\infty}^{+\infty} f(x) dx = \int_{-\infty}^0 0 dx + \int_0^1 \frac{3}{2}(1-x^2) dx + \int_1^{+\infty} 0 dx = 0 + \frac{3}{2} \left[x - \frac{x^3}{3} \right]_0^1 + 0 = \frac{3}{2} \left[1 - \frac{1}{3} \right] = \frac{3}{2} \cdot \frac{2}{3} = 1$$

Calculemos la función de distribución a partir de la función de densidad

$$\text{Si } x < 0 \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt = 0$$

Si $0 \leq x \leq 1$

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t) dt = \int_{-\infty}^0 0 dt + \int_0^x \frac{3}{2}(1-t^2) dt = 0 + \frac{3}{2} \left[t - \frac{t^3}{3} \right]_0^x = \frac{3}{2} \left(x - \frac{x^3}{3} \right) = \frac{3x - x^3}{2}$$

Si $1 < x$

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t) dt = \int_{-\infty}^0 0 dt + \int_0^1 \frac{3}{2}(1-t^2) dt + \int_1^x 0 dt = 0 + \frac{3}{2} \left[t - \frac{t^3}{3} \right]_0^1 + 0 = \frac{3}{2} \left(1 - \frac{1}{3} \right) = \frac{3}{2} \cdot \frac{2}{3} = 1$$

Calculemos las probabilidades de que $\frac{1}{3} < x < \frac{1}{2}$ y de que $\frac{1}{2} < x$.

Lo haremos primero con la función de densidad y posteriormente con la función de distribución.

$$P\left[\frac{1}{3} < X < \frac{1}{2}\right] = \int_{\frac{1}{3}}^{\frac{1}{2}} \frac{3}{2}(1-t^2) dt = \frac{3}{2} \left[t - \frac{t^3}{3} \right]_{\frac{1}{3}}^{\frac{1}{2}} = \frac{3}{2} \left[\left(\frac{1}{2} - \frac{1}{24} \right) - \left(\frac{1}{3} - \frac{1}{81} \right) \right] = \frac{3}{2} \left[\frac{11}{24} - \frac{26}{81} \right] = 0,206$$

$$P\left[\frac{1}{2} < X\right] = \int_{\frac{1}{2}}^1 \frac{3}{2}(1-t^2) dt = \frac{3}{2} \left[t - \frac{t^3}{3} \right]_{\frac{1}{2}}^1 = \frac{3}{2} \left[\left(1 - \frac{1}{3} \right) - \left(\frac{1}{2} - \frac{1}{24} \right) \right] = \frac{5}{16} = 0,3125$$

$$P\left[\frac{1}{3} < X < \frac{1}{2}\right] = F\left(\frac{1}{2}\right) - F\left(\frac{1}{3}\right) = \left(\frac{\frac{3}{2} - \frac{1}{8}}{2} \right) - \left(\frac{1 - \frac{1}{27}}{2} \right) = \frac{11}{16} - \frac{26}{54} = 0,206$$

$$P\left[\frac{1}{2} < X\right] = 1 - P\left[X \leq \frac{1}{2}\right] = 1 - F\left(\frac{1}{2}\right) = 1 - \left(\frac{\frac{3}{2} - \frac{1}{8}}{2} \right) = 1 - \frac{11}{16} = \frac{5}{16} = 0,3125$$

6.3 Valor esperado de una variable aleatoria. Momentos.

Se define el **valor esperado, esperanza matemática o media** de una **variable aleatoria discreta** X como:

$$E[X] = \sum_{i=1}^{\infty} x_i p_i$$

(obsérvese la analogía con la media aritmética en Estadística Descriptiva)

Esta definición puede generalizarse para cualquier potencia, de orden r , de la variable aleatoria. Son los **momentos no centrados o respecto al origen**:

$$\alpha_r = E[X^r] = \sum_{i=1}^{\infty} x_i^r p_i$$

La **media** es el momento no centrado de orden 1, adopta las siguientes notaciones en distintos contextos:

$$E[X] = \alpha_1 = \alpha = \mu$$

Se definen los **momentos centrados, respecto a la esperanza o respecto a la media** como:

$$\mu_r = E\left[(X - E[X])^r\right] = \sum_{i=1}^{\infty} (x_i - E[X])^r p_i$$

La **varianza** es el momento centrado de orden 2, μ_2 , se suele notar como σ^2 . Su raíz cuadrada es la **desviación típica** que se nota como σ .

Las mismas relaciones que se encontraron entre los momentos en Estadística Descriptiva siguen siendo válidas, destacando por su interés las siguientes:

$$\sigma^2 = \mu_2 = \alpha_2 - \alpha_1^2$$

$$\mu_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3$$

$$\mu_4 = \alpha_4 - 4\alpha_1\alpha_3 + 6\alpha_1^2\alpha_2 - 3\alpha_1^4$$

De forma análoga, se definen los **momentos no centrados y centrados** para una **variable aleatoria continua** X como:

$$\alpha_r = E[X^r] = \int_{-\infty}^{+\infty} x^r f(x) dx$$

$$\mu_r = E[(X - E[X])^r] = \int_{-\infty}^{+\infty} (x - E[X])^r f(x) dx$$

Propiedades de la media y de la varianza.

- $E[aX + b] = aE[X] + b$

Demostración: (se hace para variables continuas, de forma análoga se haría para variables discretas)

$$E[aX + b] = \int_{-\infty}^{+\infty} (ax + b) f(x) dx = \int_{-\infty}^{+\infty} ax f(x) dx + \int_{-\infty}^{+\infty} b f(x) dx = a \int_{-\infty}^{+\infty} x f(x) dx + b \int_{-\infty}^{+\infty} f(x) dx = aE[X] + b$$

- $\sigma^2[aX + b] = a^2 \sigma^2[X]$

En particular ($a=0$), la varianza de una constante es cero, $\sigma^2[b] = 0$.

Las propiedades anteriores muestran como le afecta al valor de la media y de la varianza un cambio de origen y escala sobre la variable aleatoria.

- **Desigualdad de Tchebycheff:**

$$P[|X - E[X]| < k\sigma] = P[E[X] - k\sigma < X < E[X] + k\sigma] \geq 1 - \frac{1}{k^2}$$

► EJEMPLO 6.4

Calcule la media y la varianza de las distribuciones de probabilidad de los ejemplos 6.2 y 6.3

Solución:

Distribución de probabilidad del ejemplo 6.2:

x_i	p_i	$x_i p_i$	$x_i^2 p_i$
0	1/8	0	0
1	3/8	0,375	0,375
2	3/8	0,750	1,500
3	1/8	0,375	1,125
total		1,500	3

$$E[X] = \sum_{i=0}^3 x_i p_i = 1,5$$

$$\sigma^2 = \alpha_2 - \alpha_1^2 = \sum_{i=0}^3 x_i^2 p_i - \left(\sum_{i=0}^3 x_i p_i \right)^2 = 3 - 1,5^2 = 0,75$$

Distribución de probabilidad del ejemplo 6.3:

$$f(x) = \begin{cases} 0 & , x < 0 \\ \frac{3-3x^2}{2} = \frac{3}{2}(1-x^2) & , 0 \leq x \leq 1 \\ 0 & , 1 < x \end{cases}$$

$$\alpha_1 = E[X] = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^1 x \frac{3}{2} (1-x^2) dx = \frac{3}{2} \int_0^1 (x - x^3) dx = \frac{3}{2} \left[\frac{x^2}{2} - \frac{x^4}{4} \right]_0^1 = \frac{3}{2} \left[\frac{1}{2} - \frac{1}{4} \right] = \frac{3}{8} = 0,375$$

$$\sigma^2 = \alpha_2 - \alpha_1^2 = \frac{1}{5} - \left(\frac{3}{8} \right)^2 = 0,059375$$

$$\text{Donde } \alpha_2 = E[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \frac{3}{2} \int_0^1 (x^2 - x^4) dx = \frac{3}{2} \left[\frac{x^3}{3} - \frac{x^5}{5} \right]_0^1 = \frac{3}{2} \left[\frac{1}{3} - \frac{1}{5} \right] = \frac{1}{5} = 0,2 \quad \blacktriangleleft$$

6.4 Otras medidas de posición, dispersión y forma.

Moda.

Se define la moda, **Mo**, como aquel valor de la variable aleatoria donde la función de densidad o distribución de probabilidad alcanza su máximo (*según se trate de una variable aleatoria continua o discreta*)

$$f(Mo) \geq f(x) \quad \forall x \in \mathbb{R}$$

$$P[X = Mo] \geq p_i \quad \forall i$$

La variable aleatoria puede tener una o varias modas según se alcance el máximo en uno o más puntos.

Mediana, cuartiles y percentiles.

Se define la mediana, **Me**, como aquel valor de la variable aleatoria tal que

$$F(Me) = P[X \leq Me] = \frac{1}{2}$$

De forma análoga, se definen los cuartiles Q_i y los percentiles P_i como aquellos valores de la variable aleatoria que verifican:

$$F(Q_i) = P[X \leq Q_i] = \frac{i}{4}$$

$$F(P_i) = P[X \leq P_i] = \frac{i}{100}$$

(Estas definiciones valen para variables aleatorias continuas con función de distribución continua. En variables aleatorias discretas con función de distribución escalonada, aunque el concepto es el mismo, la forma de calcularlas varía).

Coefficiente de variación.

Se define el coeficiente de variación, **CV**, como

$$CV = \frac{\sigma}{E[X]}$$

Es la medida de dispersión relativa más utilizada. Nos permite comparar la dispersión de distintas variables aleatorias.

Coefficiente de asimetría.

A partir de los momentos se define el coeficiente de asimetría de Fisher

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

Si la distribución de probabilidad es simétrica $\gamma_1 = 0$, si es asimétrica a la izquierda $\gamma_1 < 0$ y si lo es a la derecha $\gamma_1 > 0$.

Coefficiente de apuntamiento o curtosis.

A partir de los momentos se define el coeficiente de apuntamiento o curtosis de Fisher

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Si la distribución de probabilidad tiene el mismo apuntamiento que la *campana de Gauss o curva normal* (mesocúrtica) $\gamma_2 = 0$, si es menos apuntada (platicúrtica) $\gamma_2 < 0$ y si es más apuntada (leptocúrtica) $\gamma_2 > 0$.

► EJEMPLO 6.5

Calcule la moda, mediana, percentil 75 y coeficiente de variación de las distribuciones de probabilidad de los ejemplos 6.2 y 6.3

Solución:

Distribución de probabilidad del ejemplo 6.2:

x_i	p_i	p_i acumulados
0	1/8	0,125
1	3/8	0,500
2	3/8	0,875
3	1/8	1

Como puede observarse se trata de una distribución de probabilidad bimodal, la distribución de probabilidad $p_i = P[X = x_i]$ alcanza el máximo en dos puntos

$$Mo=1 \quad y \quad Mo=2$$

Para el cálculo de la mediana y percentil 75 (tercer cuartil) procedemos de forma análoga a variables estadísticas en Estadística Descriptiva.

En el caso de la mediana, buscamos donde se alcanza o supera por primera vez la mitad de la probabilidad total $\left(\frac{1}{2}=0,5\right)$, para $x_i=1$ la probabilidad acumulada es igual a 0,5 por lo tanto se toma como mediana

$$Me = \frac{x_i + x_{i-1}}{2} = \frac{1+2}{2} = 1,5$$

En el caso del percentil 75, buscamos donde se alcanza o supera por primera vez el 75% de la probabilidad total (0,75), para $x_i=2$ la probabilidad acumulada es $0,875 > 0,75$ por lo tanto

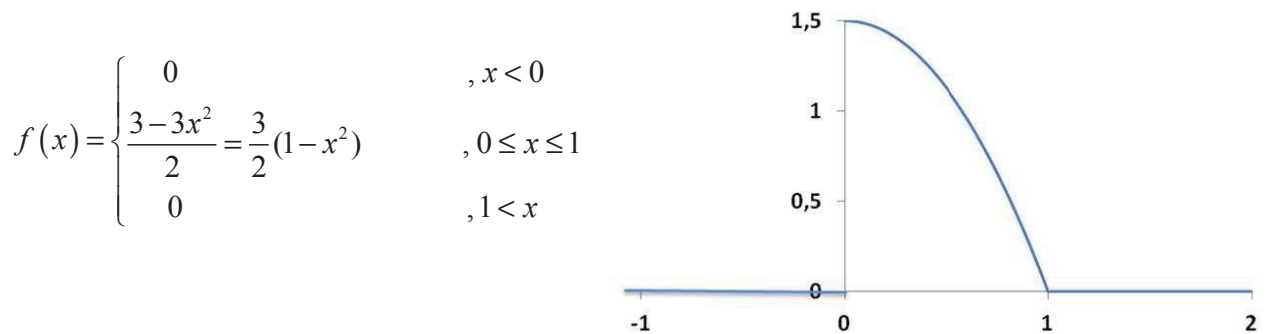
$$P_{75} = 2$$

Para calcular el coeficiente de variación utilizamos los valores de la media y varianza calculados en el ejemplo 6.4

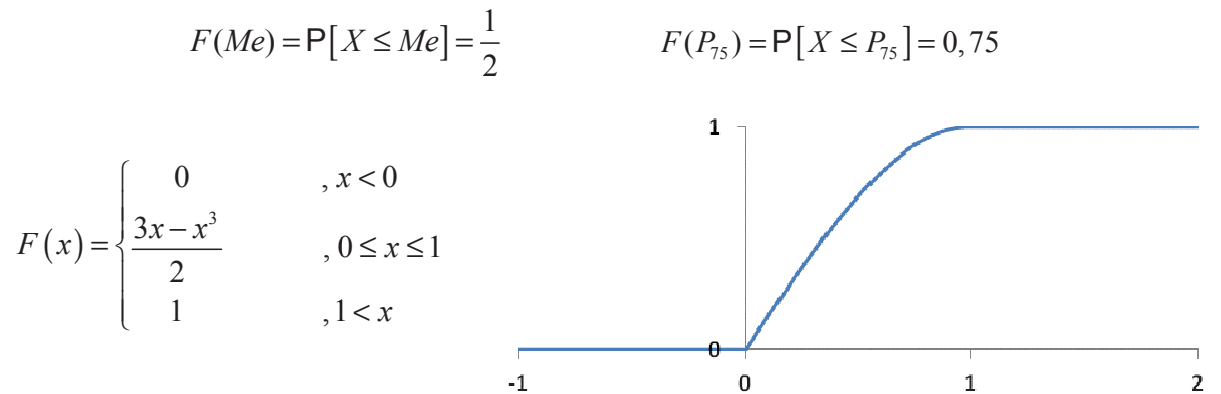
$$CV = \frac{\sigma}{E[X]} = \frac{\sqrt{0,75}}{1,5} = 0,577$$

Distribución de probabilidad del ejemplo 6.3:

Para hallar la moda en distribuciones de probabilidad continuas hay que calcular el punto o puntos donde la función de densidad alcanza su máximo.



Como puede observarse en la gráfica de la función de densidad, ésta alcanza su máximo en $Mo=0$. Para obtener la mediana y el percentil 75 utilizaremos la función de distribución



$$\frac{3x-x^3}{2} = \frac{1}{2} \Rightarrow 0 = x^3 - 3x + 1 \Rightarrow x \cong 0,3473$$

$$\frac{3x-x^3}{2} = 0,75 = \frac{3}{4} \Rightarrow 0 = 2x^3 - 6x + 3 \Rightarrow x \cong 0,557875$$

(Las anteriores ecuaciones no son fáciles de resolver a mano, se han resuelto con la ayuda del ordenador).

Para calcular el coeficiente de variación utilizamos los valores de la media y varianza calculados en el ejemplo 6.4

$$CV = \frac{\sigma}{E[X]} = \frac{\sqrt{0,059375}}{0,375} = \frac{0,24367}{0,375} = 0,6498$$

6.5 Variables aleatorias bidimensionales. Independencia de variables aleatorias.

Sea Ω el conjunto de sucesos elementales, ω_i , de un fenómeno aleatorio. Se define una **variable aleatoria bidimensional** (X, Y) como una función sobre Ω en $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$.

$$\begin{aligned}\Omega &\xrightarrow{(X, Y)} \mathbb{R} \times \mathbb{R} \\ \omega_i &\rightarrow (X(\omega_i), Y(\omega_i))\end{aligned}$$

Es decir, mediante una variable aleatoria bidimensional transformamos los resultados de un fenómeno aleatorio en dos valores numéricos que miden aquellas dos características del fenómeno que nos interesan destacar. Una variable aleatoria bidimensional es un par de variables aleatorias unidimensionales definidas sobre el mismo fenómeno aleatorio que determina a Ω .

Las probabilidades de que la variable aleatoria bidimensional (X, Y) tome los distintos valores dependen de las probabilidades, $P(\omega_i)$, asociadas al fenómeno aleatorio sobre el que se han definido.

Se denomina **distribución de probabilidad bidimensional** de las variables aleatorias (X, Y) a las probabilidades asociadas a los distintos valores que pueden tomar conjuntamente dichas variables aleatorias, $p_{ij} = P[(X, Y) = (x_i, y_j)] = P[X = x_i, Y = y_j]$.

Toda distribución de probabilidad bidimensional cumple (como en el caso unidimensional):

- $p_{ij} \geq 0 \quad \forall i, j$
- $\sum_i \sum_j p_{ij} = 1$

(Esta definición es válida cuando ambas variables son discretas, sin embargo cuando las variables aleatorias son continuas la distribución de probabilidad bidimensional está caracterizada por la función de densidad bidimensional como se verá más adelante).

► EJEMPLO 6.6

Veamos un sencillo ejemplo de variable aleatoria bidimensional discreta.

X =número de cruces en tres lanzamientos de una moneda.

Y =número de caras en los dos primeros lanzamientos de la moneda.

C=cara, +=cruz. La variable aleatoria bidimensional (X, Y) está definida por la siguiente aplicación:

$$\begin{array}{ll} \Omega & \xrightarrow{(X, Y)} \mathbb{R} \times \mathbb{R} \\ \text{CCC} & \rightarrow (X(\text{CCC}), Y(\text{CCC})) = (0, 2) \\ +\text{CC} & \rightarrow (X(+\text{CC}), Y(+\text{CC})) = (1, 1) \\ \text{C}+\text{C} & \rightarrow (1, 1) \\ \text{CC}+ & \rightarrow (1, 2) \\ ++\text{C} & \rightarrow (2, 0) \\ +\text{C}+ & \rightarrow (2, 1) \\ \text{C}++ & \rightarrow (2, 1) \\ +++ & \rightarrow (3, 0) \end{array}$$

Suponemos que la probabilidad de cara y cruz es la misma, $\frac{1}{2}$. Dado que son independientes los resultados de los tres lanzamientos de la moneda, la probabilidad de cada uno de los sucesos elementales de Ω es $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$.

$$P(\text{CCC}) = P(C \cap C \cap C) = P(C)P(C)P(C) = \frac{1}{2} \frac{1}{2} \frac{1}{2} = \frac{1}{8}$$

...

$$P(+++) = P(+)P(+)P(+) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

$$p_{02} = P[X = 0, Y = 2] = P[(\text{CCC})] = \frac{1}{8}$$

$$p_{11} = P[X = 1, Y = 1] = P[(+\text{CC}) \cup (\text{C}+\text{C})] = \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4}$$

$$p_{12} = P[X = 1, Y = 2] = P[(\text{CC}+)] = \frac{1}{8}$$

$$p_{20} = P[X = 2, Y = 0] = P[(++\text{C})] = \frac{1}{8}$$

$$p_{21} = P[X = 2, Y = 1] = P[(\text{C}++) \cup (+\text{C}+)] = \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4}$$

$$p_{30} = P[X = 3, Y = 0] = P[(+++)] = \frac{1}{8}$$

Las anteriores probabilidades se suelen recoger en una tabla de doble entrada (como *las variables estadísticas bidimensionales en Estadística Descriptiva*)

$X \backslash Y$	0	1	2
0	0	0	$\frac{1}{8}$
1	0	$\frac{2}{8}$	$\frac{1}{8}$
2	$\frac{1}{8}$	$\frac{2}{8}$	0
3	$\frac{1}{8}$	0	0

Como puede comprobarse $\sum_{i=0}^3 \sum_{j=0}^2 p_{ij} = 1$.



Función de distribución bidimensional. Función de densidad bidimensional.

Dada una variable aleatoria bidimensional (X, Y) , se define la **función de distribución** (conjunta) de (X, Y) como:

$$F(x, y) = P[X \leq x, Y \leq y] = P[\omega \in \Omega / X(\omega) \leq x, Y(\omega) \leq y] \quad \forall (x, y) \in \mathbb{R}^2$$

Si las variables aleatorias son discretas la función de distribución es escalonada (*discontinua*). Si las variables aleatorias son continuas la siguiente derivada de la función de distribución conjunta es la **función de densidad** (conjunta) de (X, Y) :

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y)$$

Toda función de densidad bidimensional cumple las siguientes propiedades que la caracterizan:

- $f(x, y) \geq 0 \quad \forall (x, y) \in \mathbb{R}^2$
- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$

La probabilidad de que la variable aleatoria bidimensional (X, Y) tome valores en un determinado conjunto es igual a la integral definida de la función de densidad bidimensional sobre dicho conjunto.

En particular:

$$F(x, y) = P[X \leq x, Y \leq y] = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$$

► EJEMPLO 6.7

Sea la variable aleatoria bidimensional de tipo continuo con función de distribución conjunta

$$F(x, y) = \begin{cases} \frac{3xy^2 - x^3y^2}{2} & , 0 \leq x \leq 1 \quad 0 \leq y \leq 1 \\ 0 & , \text{ en otro caso} \end{cases}$$

Calcule la función de densidad conjunta y compruebe que cumple las dos propiedades que caracterizan a toda función de densidad.

Solución:

Derivamos primero respecto de x

$$\frac{\partial F(x, y)}{\partial x} = \begin{cases} \frac{3y^2 - 3x^2y^2}{2} & , 0 \leq x \leq 1 \quad 0 \leq y \leq 1 \\ 0 & , \text{ en otro caso} \end{cases}$$

y a continuación derivamos de nuevo respecto de y

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \begin{cases} \frac{6y - 6x^2y}{2} = 3(1 - x^2)y & , 0 \leq x \leq 1 \quad 0 \leq y \leq 1 \\ 0 & , \text{ en otro caso} \end{cases}$$

- $f(x, y) \geq 0 \quad \forall (x, y) \in \mathbb{R}^2$

$$\left. \begin{array}{l} 0 \leq x \leq 1 \Rightarrow 0 \leq (1 - x^2) \\ 0 \leq y \leq 1 \Rightarrow 0 \leq y \end{array} \right\} \Rightarrow 0 \leq (1 - x^2)y$$

- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$

$$\int_0^1 \int_0^1 3y(1 - x^2) dx dy = \int_0^1 3y \left[x - \frac{x^3}{3} \right]_0^1 dy = \int_0^1 3y \left[1 - \frac{1}{3} \right] dy = 2 \int_0^1 y dy = 2 \left[\frac{y^2}{2} \right]_0^1 = 2 \left[\frac{1}{2} \right] = 1$$



Distribuciones de probabilidad marginales.

Para variables aleatorias discretas, se define la **distribución de probabilidad marginal** de X como

$$p_{i\bullet} = P[X = x_i] = P[X = x_i, Y = \text{cualquier valor}] = \sum_{\forall j} p_{ij}$$

Análogamente para la variable aleatoria Y

$$p_{\bullet j} = P[Y = y_j] = P[X = \text{cualquier valor}, Y = y_j] = \sum_{\forall i} p_{ij}$$

Para variables aleatorias continuas, se define la **función de densidad marginal** de X como

$$f_1(x) = f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

Análogamente para la variable aleatoria Y

$$f_2(y) = f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

► EJEMPLO 6.8

Calcule las distribuciones de probabilidad marginales para las distribuciones de probabilidad bidimensionales de los ejemplos 6.6 y 6.7.

Solución:

Distribución de probabilidad bidimensional del ejemplo 6.6:

Sumando en los márgenes de la tabla de la distribución de probabilidad bidimensional (*de forma análoga a como se hizo en Estadística Descriptiva*) se obtienen la distribución de probabilidad marginal de X ($p_{i\bullet}$) y la distribución de probabilidad marginal de Y ($p_{\bullet j}$).

Como puede comprobarse, la distribución de probabilidad marginal de X coincide con la distribución de probabilidad unidimensional de X =*número de cruces en tres lanzamientos de una moneda* que se obtuvo en el ejemplo 6.1

$X \backslash Y$	0	1	2	$p_{i\bullet}$
0	0	0	$\frac{1}{8}$	$\frac{1}{8}$
1	0	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{3}{8}$
2	$\frac{1}{8}$	$\frac{2}{8}$	0	$\frac{3}{8}$
3	$\frac{1}{8}$	0	0	$\frac{1}{8}$
$p_{\bullet j}$	$\frac{2}{8}$	$\frac{4}{8}$	$\frac{2}{8}$	1

Distribución de probabilidad bidimensional del ejemplo 6.7:

En variables aleatorias continuas la distribución de probabilidad está definida por su función de densidad. En este ejemplo por

$$f(x, y) = \begin{cases} \frac{6y - 6x^2y}{2} = 3(1 - x^2)y & , 0 \leq x \leq 1 \quad 0 \leq y \leq 1 \\ 0 & , \text{ en otro caso} \end{cases}$$

La distribución de probabilidad marginal de X está dada por su función de densidad marginal

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_0^1 3(1-x^2)y dy = 3(1-x^2) \left[\frac{y^2}{2} \right]_0^1 = \frac{3}{2}(1-x^2) \quad \text{si } 0 \leq x \leq 1$$

$$f_1(x) = 0 \quad \text{en otro caso}$$

Análogamente para la variable aleatoria Y

$$f_2(y) = \int_{-\infty}^{+\infty} f(x, y) dx = \int_0^1 3(1-x^2)y dx = 3y \left[x - \frac{x^3}{3} \right]_0^1 = 3y \left[1 - \frac{1}{3} \right] = 2y \quad \text{si } 0 \leq y \leq 1$$

$$f_2(y) = 0 \quad \text{en otro caso}$$



Independencia de variables aleatorias.

Recordemos que dos sucesos A y B son independientes si

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

En el caso discreto, diremos que las variables X e Y son **independientes** si

$$P[X = x_i, Y = y_j] = P[X = x_i]P[Y = y_j] \quad \forall x_i, y_j$$

o lo que es lo mismo

$$p_{ij} = p_{i \bullet} p_{\bullet j} \quad \forall i, j$$

(obsérvese la analogía con la independencia en Estadística Descriptiva donde $f_{ij} = f_{i \bullet} f_{\bullet j} \quad \forall i, j$)

En el caso continuo, diremos que las variables X e Y son **independientes** si

$$f(x, y) = f_1(x) f_2(y) \quad \forall x, y$$

► EJEMPLO 6.9

Estudiar la independencia de las variables aleatorias de los ejemplos 6.6 y 6.7

Solución:

Distribución de probabilidad bidimensional del ejemplo 6.6:

$X \backslash Y$	0	1	2	$p_{i\bullet}$
0	0	0	$\frac{1}{8}$	$\frac{1}{8}$
1	0	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{3}{8}$
2	$\frac{1}{8}$	$\frac{2}{8}$	0	$\frac{3}{8}$
3	$\frac{1}{8}$	0	0	$\frac{1}{8}$
$p_{\bullet j}$	$\frac{2}{8}$	$\frac{4}{8}$	$\frac{2}{8}$	1

Como puede comprobarse fácilmente, $p_{00} = 0 \neq p_{0\bullet} p_{\bullet 0} = \frac{1}{8} \frac{2}{8} = \frac{1}{32}$, luego no se cumple la condición de independencia, $p_{ij} = p_{i\bullet} p_{\bullet j} \quad \forall i, j$. En ese caso diremos que las variables aleatorias X e Y son **dependientes**.

Distribución de probabilidad bidimensional del ejemplo 6.7:

$$f(x, y) = \begin{cases} \frac{6y - 6x^2 y}{2} = 3(1 - x^2)y & , 0 \leq x \leq 1 \quad 0 \leq y \leq 1 \\ 0 & , \text{ en otro caso} \end{cases}$$

$$f_1(x) = \begin{cases} \frac{3}{2}(1 - x^2) & , 0 \leq x \leq 1 \\ 0 & , \text{ en otro caso} \end{cases} \quad f_2(y) = \begin{cases} 2y & , 0 \leq y \leq 1 \\ 0 & , \text{ en otro caso} \end{cases}$$

Es inmediato comprobar que $f(x, y) = f_1(x) f_2(y) \quad \forall x, y$. Luego X e Y son **independientes**. ◀

Principales momentos en variables aleatorias bidimensionales.

Definición y propiedades.

Media (marginal) de X :

Variables discretas: $E[X] = \sum_{i=1}^{\infty} x_i p_{i\bullet}$

Variables continuas: $E[X] = \int_{-\infty}^{+\infty} x f_1(x) dx$

Media (marginal) de Y:

Variables discretas:
$$E[Y] = \sum_{j=1}^{\infty} y_j p_{\bullet j}$$

Variables continuas:
$$E[Y] = \int_{-\infty}^{+\infty} y f_2(y) dy$$

Propiedades:

$$E[X + Y] = E[X] + E[Y]$$

Esta propiedad es igualmente válida si la suma se refiere a más de dos variables aleatorias.

Varianza (marginal) de X:

Variables discretas:
$$Var[X] = \sigma^2[X] = \sigma_x^2 = \sum_{i=1}^{\infty} (x_i - E[X])^2 p_{i\bullet} = E[X^2] - E[X]^2$$

Variables continuas:
$$Var[X] = \sigma^2[X] = \sigma_x^2 = \int_{-\infty}^{+\infty} (x - E[X])^2 f_1(x) dx = E[X^2] - E[X]^2$$

Varianza (marginal) de Y:

Variables discretas:
$$Var[Y] = \sigma^2[Y] = \sigma_y^2 = \sum_{j=1}^{\infty} (y_j - E[Y])^2 p_{\bullet j} = E[Y^2] - E[Y]^2$$

Variables continuas:
$$Var[Y] = \sigma^2[Y] = \sigma_y^2 = \int_{-\infty}^{+\infty} (y - E[Y])^2 f_2(y) dy = E[Y^2] - E[Y]^2$$

Propiedades:

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

$$Var[X - Y] = Var[X] + Var[Y] - 2Cov[X, Y]$$

Covarianza (momento centrado de órdenes 1,1):

Variables discretas:

$$Cov[X, Y] = \sigma[X, Y] = \sigma_{xy} = E[(X - E[X])(Y - E[Y])] = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (x_i - E[X])(y_j - E[Y]) p_{ij}$$

Variables continuas:

$$Cov[X, Y] = \sigma[X, Y] = \sigma_{xy} = E[(X - E[X])(Y - E[Y])] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E[X])(y - E[Y]) f(x, y) dx dy$$

Propiedades:

$$Cov[X, Y] = E[XY] - E[X]E[Y].$$

Si X e Y son **independientes** $E[XY] = E[X]E[Y]$

Demostración: (se hace para variables continuas, de forma análoga se haría para variables discretas)

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_1(x) f_2(y) dx dy = \int_{-\infty}^{\infty} y f_2(y) \int_{-\infty}^{\infty} x f_1(x) dx dy = \\ &= \int_{-\infty}^{\infty} y f_2(y) E[X] dy = E[X] \int_{-\infty}^{\infty} y f_2(y) dy = E[X] E[Y] \end{aligned}$$

Por tanto en caso de **independencia**: $Cov[X, Y] = 0$ y $Var[X \pm Y] = Var[X] + Var[Y]$.

6.6 Ejercicios resueltos.

1. A partir de 56 millones de llamadas de los abonados de la compañía *Noteoigo* en 2012, se ha construido la siguiente variable aleatoria X para modelizar la duración en minutos de las mismas:

$$f(x) = \begin{cases} k(4+x)(6-x) & 0 < x < 6 \\ 0 & \text{en otro caso} \end{cases}$$

Obtenga:

- a) k .
- b) Duración media de las llamadas.
- c) Probabilidad de que una llamada supere los 2 minutos de duración.

Solución:

$$\begin{aligned} \text{a) } 1 &= \int_{-\infty}^{\infty} f(x) dx = \int_0^6 k(4+x)(6-x) dx = k \int_0^6 (24 + 2x - x^2) dx = k \left[24x + x^2 - \frac{x^3}{3} \right]_0^6 = \\ &= k[144 + 36 - 72] = k108 \quad \Leftrightarrow \quad k = \frac{1}{108} \end{aligned}$$

$$\begin{aligned} \text{b) } E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^6 x \frac{1}{108} (4+x)(6-x) dx = \frac{1}{108} \int_0^6 (24x + 2x^2 - x^3) dx = \frac{1}{108} \left[24 \frac{x^2}{2} + 2 \frac{x^3}{3} - \frac{x^4}{4} \right]_0^6 = \\ &= \frac{1}{108} [432 + 144 - 324] = \frac{252}{108} = \frac{7}{3} = 2,3 \end{aligned}$$

$$\text{c) } P[X > 2] = \int_2^{\infty} f(x) dx = \int_2^6 \frac{1}{108} (4+x)(6-x) dx = \frac{1}{108} \int_2^6 (24 + 2x - x^2) dx = \frac{1}{108} \left[24x + x^2 - \frac{x^3}{3} \right]_2^6 =$$

$$= \frac{1}{108} \left[(108) - \left(48 + 4 - \frac{8}{3} \right) \right] = \frac{1}{108} \left[108 - \frac{148}{3} \right] = \frac{1}{108} \left[\frac{176}{3} \right] = 0,5432$$

2. Sea X el número de llantas defectuosas en un lote de cuatro llantas modeladas al mismo tiempo. La función de distribución de probabilidad viene dada por

x_i	0	1	2	3	4
$F(x_i)$	0,8	0,9	0,95	0,98	1

- a) Obtenga el valor esperado y la varianza de X .
- b) Cada una de las llantas defectuosas representa una pérdida de 120€, estime el coste esperado y la desviación típica de las pérdidas.

Solución:

A partir de la función de distribución $F(x_i) = P[X \leq x_i]$ obtenemos la distribución de probabilidad

$$p_i = P[X = x_i] = F(x_i) - F(x_{i-1})$$

y a partir de ésta responderemos ambos apartados.

x_i	p_i	$x_i p_i$	$x_i^2 p_i$
0	0,8	0	0
1	0,1	0,1	0,1
2	0,05	0,1	0,2
3	0,03	0,09	0,27
4	0,02	0,08	0,32
	1	0,37	0,89

a) $E[X] = \sum_{i=1}^{\infty} x_i p_i = 0,37$

$$\sigma^2[X] = E[X^2] - E^2[X] = \left(\sum_{i=1}^{\infty} x_i^2 p_i \right) - 0,37^2 = 0,89 - 0,1369 = 0,7531$$

b) $Y = \text{pérdidas} = 120X$

$$E[Y] = 120 E[X] = 120 \times 0,37 = 44,4 \quad \sigma^2[Y] = 120^2 \sigma^2[X] = 14400 \times 0,7531 = 10844,64$$

$$\sigma[Y] = \sqrt{10844,64} = 104,1376$$

3. De una variable aleatoria de tipo discreto se sabe que:

$$P[X = -2] = \frac{1}{3} \quad P[X = 1] = \frac{1}{6} \quad P[X = 3] = \frac{1}{2}$$

Se pide:

- a) El valor esperado de la variable $Y = 2X + 5$.
- b) La varianza de la variable $Y = 2X + 5$.

Solución:

$$\text{a) } E[X] = \sum_i x_i P[X = x_i] = \left(-2 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{2}\right) = \frac{-2}{3} + \frac{1}{6} + \frac{3}{2} = \frac{-4+1+9}{6} = 1$$

$$E[Y] = 2E[X] + 5 = (2 \times 1) + 5 = 7$$

$$\text{b) } E[X^2] = \sum_i x_i^2 P[X = x_i] = \left(4 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{6}\right) + \left(9 \times \frac{1}{2}\right) = \frac{8+1+27}{6} = \frac{36}{6} = 6$$

$$\sigma^2[X] = E[X^2] - E^2[X] = 6 - 1^2 = 5$$

$$\sigma^2[Y] = 2^2 \sigma^2[X] = 4 \times 5 = 20$$

4. Dada la siguiente función de distribución:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8} & 0 \leq x < 1 \\ \frac{1}{2} & 1 \leq x < 2 \\ \frac{7}{8} & 2 \leq x < 3 \\ 1 & 3 \leq x \end{cases}$$

- Determine la función de probabilidad de la variable X .
- Calcule las probabilidades $P[0 \leq X \leq 1]$ $P[1 < X \leq 2]$ $P[2 < X < 4]$.
- Calcule la varianza.
- Calcule la mediana.
- Calcule la moda.

Solución:

- Es una función de distribución escalonada que corresponde a una variable aleatoria discreta donde $P[X = x_i]$ es igual a los saltos de dicha función en cada escalón.

x_i	$F(x_i)$	$p_i = P[X = x_i]$	$x_i p_i$	$x_i^2 p_i$
0	$\frac{1}{8}$	$\frac{1}{8}$	0	0
1	$\frac{1}{2} = \frac{4}{8}$	$\frac{4}{8} - \frac{1}{8} = \frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$
2	$\frac{7}{8}$	$\frac{7}{8} - \frac{4}{8} = \frac{3}{8}$	$\frac{6}{8}$	$\frac{12}{8}$
3	1	$1 - \frac{7}{8} = \frac{1}{8}$	$\frac{3}{8}$	$\frac{9}{8}$
			$\frac{12}{8}$	$\frac{24}{8}$

b) $P[0 \leq X \leq 1] = P[X = 0] + P[X = 1] = \frac{4}{8} = \frac{1}{2} = 0,5$

$$P[1 < X \leq 2] = P[X = 2] = \frac{3}{8} = 0,375$$

$$P[2 < X < 4] = P[X = 3] = \frac{1}{8} = 0,125$$

c) $E[X] = \sum_i x_i p_i = \frac{12}{8} = \frac{3}{2} = 1,5$

$$E[X^2] = \sum_i x_i^2 p_i = \frac{24}{8} = 3$$

$$\sigma^2[X] = E[X^2] - E^2[X] = 3 - \left(\frac{3}{2}\right)^2 = 3 - \frac{9}{4} = \frac{3}{4} = 0,75$$

d) Buscamos $\frac{1}{2}$ en las probabilidades acumuladas $F(x_i)$. Encontramos que $F(1) = \frac{1}{2}$. Por tanto

$$Me = \frac{1+2}{2} = 1,5.$$

e) Buscamos el máximo valor de p_i . Encontramos que hay dos modas: $Mo = 1$, $Mo = 2$.

5. Sea X una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} k(5-2x) & 0 < x < 2 \\ 0 & \text{resto} \end{cases}$$

Calcule:

a) k .

b) $P[1 < X < 4]$.

c) El valor M tal que $P[X < M] = \frac{1}{2}$.

d) La media.

e) La varianza.

Solución:

$$\text{a) } 1 = \int_{-\infty}^{\infty} f(x) dx = k \int_0^2 (5-2x) dx = k \left[5x - x^2 \right]_0^2 = k6 \Rightarrow k = \frac{1}{6} = 0,1667$$

$$\text{b) } P[1 < X < 4] = \int_1^4 f(x) dx = \int_1^2 \frac{1}{6} (5-2x) dx = \frac{1}{6} \left[5x - x^2 \right]_1^2 = \frac{1}{6} (6-4) = \frac{2}{6} = \frac{1}{3} = 0,3333$$

$$\text{c) } \frac{1}{2} = P[X < M] = \int_{-\infty}^M f(x) dx = \int_0^M \frac{1}{6} (5-2x) dx = \frac{1}{6} \left[5x - x^2 \right]_0^M = \frac{1}{6} (5M - M^2)$$

$$\frac{1}{2} = \frac{1}{6} (5M - M^2) \Rightarrow M^2 - 5M + 3 = 0 \Rightarrow M = \frac{5 \pm \sqrt{25-12}}{2} = \frac{5 \pm 3,606}{2}$$

$$M = 0,697 \quad \text{o} \quad M = 4,303, \quad \text{pero } M = 4,303 \quad \text{no vale pues } 2 < M \Rightarrow P[X < M] = 1$$

$$\begin{aligned} \text{d) } E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^2 x \frac{1}{6} (5-2x) dx = \frac{1}{6} \int_0^2 (5x - 2x^2) dx = \frac{1}{6} \left[5 \frac{x^2}{2} - 2 \frac{x^3}{3} \right]_0^2 = \\ &= \frac{1}{6} \left[10 - \frac{16}{3} \right] = \frac{1}{6} \left[\frac{14}{3} \right] = \frac{14}{18} = \frac{7}{9} = 0,7778 \end{aligned}$$

$$\text{e) } Var[X] = E[X^2] - E^2[X] = \frac{8}{9} - \left(\frac{7}{9} \right)^2 = \frac{23}{81} = 0,284$$

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^2 x^2 \frac{1}{6} (5-2x) dx = \frac{1}{6} \int_0^2 (5x^2 - 2x^3) dx = \frac{1}{6} \left[5 \frac{x^3}{3} - 2 \frac{x^4}{4} \right]_0^2 = \\ &= \frac{1}{6} \left[\frac{40}{3} - 8 \right] = \frac{16}{18} = \frac{8}{9} = 0,8889 \end{aligned}$$

6. Dada la siguiente función de densidad:

$$f(x) = \begin{cases} 0 & x \leq 1 \\ x-1 & 1 < x \leq 2 \\ k-x & 2 < x \leq 3 \\ 0 & 3 < x \end{cases}$$

Calcule:

- a) El valor de k .
- b) $P[0 < X < 2]$.
- c) El tercer cuartil.
- d) $E[X]$.

Solución:

$$\begin{aligned} \text{a) } 1 &= \int_{-\infty}^{+\infty} f(x) dx = \int_1^2 (x-1) dx + \int_2^3 (k-x) dx = \left[\frac{x^2}{2} - x \right]_1^2 + \left[kx - \frac{x^2}{2} \right]_2^3 = \\ &= \left[(2-2) - \left(\frac{1}{2} - 1 \right) \right] + \left[\left(3k - \frac{9}{2} \right) - (2k-2) \right] = \frac{1}{2} + k - \frac{5}{2} = 1 \Leftrightarrow k = 1 + 2 = 3 \end{aligned}$$

$$\text{b) } P[0 < X < 2] = \int_0^1 0 dx + \int_1^2 (x-1) dx = \left[\frac{x^2}{2} - x \right]_1^2 = (2-2) - \left(\frac{1}{2} - 1 \right) = \frac{1}{2}$$

c) Por el apartado b) $P[X < 2] = \frac{1}{2}$ luego el tercer cuartil, $Q_3 = x_{0,75}$, estará dentro del intervalo $(2,3]$.

$$0,75 = \frac{3}{4} = P[X < Q_3] = \int_1^2 (x-1) dx + \int_2^{Q_3} (3-x) dx = \frac{1}{2} + \left[3x - \frac{x^2}{2} \right]_2^{Q_3} = \frac{1}{2} + \left[\left(3Q_3 - \frac{Q_3^2}{2} \right) - (6-2) \right]$$

$$\frac{3}{4} - \frac{1}{2} = \frac{1}{4} = 3Q_3 - \frac{Q_3^2}{2} - 4 \Leftrightarrow 1 = 12Q_3 - 2Q_3^2 - 16 \Leftrightarrow 2Q_3^2 - 12Q_3 + 17 = 0$$

$$Q_3 = \frac{12 \pm \sqrt{144 - 136}}{4} = \frac{12 \pm \sqrt{8}}{4}$$

$$Q_3 = \frac{12 + \sqrt{8}}{4} = 3,707 > 3 \quad \text{esta solución no es válida pues } 1 < X < 3.$$

$$Q_3 = \frac{12 - \sqrt{8}}{4} = 2,293 \quad \text{es la solución correcta.}$$

$$\begin{aligned} \text{d) } E[X] &= \int_{-\infty}^{+\infty} x f(x) dx = \int_1^2 x(x-1) dx + \int_2^3 x(3-x) dx = \int_1^2 (x^2 - x) dx + \int_2^3 (3x - x^2) dx = \\ &= \left[\frac{x^3}{3} - \frac{x^2}{2} \right]_1^2 + \left[\frac{3x^2}{2} - \frac{x^3}{3} \right]_2^3 = \left[\left(\frac{8}{3} - \frac{4}{2} \right) - \left(\frac{1}{3} - \frac{1}{2} \right) \right] + \left[\left(\frac{27}{2} - \frac{27}{3} \right) - \left(\frac{12}{2} - \frac{8}{3} \right) \right] = \left[\frac{5}{6} \right] + \left[\frac{7}{6} \right] = 2 \end{aligned}$$

7. En el presupuesto familiar, la parte que se dedica a la compra de productos alimenticios sigue una distribución con función de densidad $f(x) = 6x(1-x)$ para $0 < x < 1$,

a) ¿Cuál es la probabilidad de que se gaste más de la mitad del presupuesto familiar en alimentación?

b) ¿Cuál será el porcentaje medio que las familias dedican a la compra de estos productos?

Solución:

X = proporción del presupuesto dedicada a productos alimenticios.

$$\text{a) } P[X > 0,5] = \int_{0,5}^1 f(x) dx = \int_{0,5}^1 6x(1-x) dx = 6 \int_{0,5}^1 (x - x^2) dx = 6 \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_{0,5}^1 =$$

$$= 6 \left[\left(\frac{1}{2} - \frac{1}{3} \right) - \left(\frac{0,5^2}{2} - \frac{0,5^3}{3} \right) \right] = 0,5$$

$$\begin{aligned} \text{b) } E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x 6x(1-x) dx = 6 \int_0^1 (x^2 - x^3) dx = 6 \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^1 = \\ &= 6 \left[\frac{1}{3} - \frac{1}{4} \right] = 6 \frac{1}{12} = 0,50 \Rightarrow 50\% \end{aligned}$$

8. La demanda mensual de un artículo, en miles de unidades, viene dada por:

$$f(x) = \begin{cases} 1-kx & 0 \leq x \leq 2 \\ 0 & \text{en el resto} \end{cases}$$

Calcule:

- La probabilidad de que en un mes la demanda sea inferior a 3500 unidades.
- La probabilidad de que en un mes se demande más de 635 y menos de 1870 unidades.
- El número medio de artículos que esperamos que se demanden el mes próximo.

Solución:

$$1 = \int_{-\infty}^{\infty} f(x) dx = \int_0^2 (1-kx) dx = \left[x - k \frac{x^2}{2} \right]_0^2 = 2 - 2k \Leftrightarrow -1 = -2k \Leftrightarrow \frac{1}{2} = k$$

$$\text{a) } P[X < 3,5] = \int_{-\infty}^{3,5} f(x) dx = \int_0^2 (1-kx) dx = 1$$

$$\begin{aligned} \text{b) } P[0,635 < X < 1,87] &= \int_{0,635}^{1,87} f(x) dx = \int_{0,635}^{1,87} \left(1 - \frac{x}{2} \right) dx = \left[x - \frac{x^2}{4} \right]_{0,635}^{1,87} = \\ &= \left[\left(1,87 - \frac{1,87^2}{4} \right) - \left(0,635 - \frac{0,635^2}{4} \right) \right] = 0,46158 \end{aligned}$$

$$\text{c) } E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_0^2 x \left(1 - \frac{x}{2} \right) dx = \int_0^2 \left(x - \frac{x^2}{2} \right) dx = \left[\frac{x^2}{2} - \frac{x^3}{6} \right]_0^2 = \frac{4}{2} - \frac{8}{6} = \frac{2}{3} = 0,6667$$

9. El número de unidades demandadas mensualmente de un determinado artículo sigue una ley de probabilidad definida por la función de densidad:

$$f(x) = \begin{cases} kx & 0 \leq x < 5 \\ \frac{10-x}{25} & 5 \leq x \leq 10 \\ 0 & \text{en el resto} \end{cases}$$

Donde x viene expresada en miles de unidades. Se pide:

- El valor de k .
- Las unidades que en un mes se esperan vender.
- La probabilidad de que en un mes se vendan más de 3000 y menos de 6000 unidades.
- En un determinado mes se disponen de 8000 unidades para la venta, ¿Cuál es la probabilidad de no poder atender todas las peticiones que de dicho artículo se produzcan?
- ¿Cuántas unidades, como mínimo, debemos disponer mensualmente para atender la demanda con una probabilidad mayor o igual al 90%?

Solución:

$$\begin{aligned} \text{a) } 1 &= \int_{-\infty}^{\infty} f(x) dx = \int_0^5 kx dx + \int_5^{10} \frac{10-x}{25} dx = k \left[\frac{x^2}{2} \right]_0^5 + \frac{1}{25} \left[10x - \frac{x^2}{2} \right]_5^{10} = \\ &= k \frac{25}{2} + \frac{1}{25} \left[\left(100 - \frac{100}{2} \right) - \left(50 - \frac{25}{2} \right) \right] = \frac{25k}{2} + \left(\frac{1}{25} \frac{25}{2} \right) = \frac{25k+1}{2} \\ 1 &= \frac{25k+1}{2} \quad \Leftrightarrow \quad 2-1 = 25k \quad \Leftrightarrow \quad \frac{1}{25} = k \end{aligned}$$

$$\begin{aligned} \text{b) } E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^5 x \frac{1}{25} x dx + \int_5^{10} x \frac{10-x}{25} dx = \frac{1}{25} \int_0^5 x^2 dx + \frac{1}{25} \int_5^{10} (10x - x^2) dx = \\ &= \frac{1}{25} \left[\frac{x^3}{3} \right]_0^5 + \frac{1}{25} \left[\frac{10x^2}{2} - \frac{x^3}{3} \right]_5^{10} = \frac{1}{25} \frac{125}{3} + \frac{1}{25} \left[\left(\frac{1000}{2} - \frac{1000}{3} \right) - \left(\frac{250}{2} - \frac{125}{3} \right) \right] = \\ &= \frac{5}{3} + \frac{1}{25} \left[\frac{1000}{6} - \frac{500}{6} \right] = \frac{5}{3} + \frac{10}{3} = 5 \end{aligned}$$

$$\begin{aligned} \text{c) } P[3 < X < 6] &= \int_3^6 f(x) dx = \int_3^5 \frac{1}{25} x dx + \int_5^6 \frac{10-x}{25} dx = \frac{1}{25} \left[\frac{x^2}{2} \right]_3^5 + \frac{1}{25} \left[10x - \frac{x^2}{2} \right]_5^6 = \\ &= \frac{1}{25} \left[\frac{25}{2} - \frac{9}{2} \right] + \frac{1}{25} \left[\left(60 - \frac{36}{2} \right) - \left(50 - \frac{25}{2} \right) \right] = \frac{1}{25} 8 + \frac{1}{25} \frac{9}{2} = \frac{1}{25} \left(\frac{16}{2} + \frac{9}{2} \right) = \frac{1}{2} = 0,5 \end{aligned}$$

$$\begin{aligned} \text{d) } P[X > 8] &= \int_8^{\infty} f(x) dx = \int_8^{10} \frac{10-x}{25} dx = \frac{1}{25} \left[10x - \frac{x^2}{2} \right]_8^{10} = \\ &= \frac{1}{25} \left[\left(100 - \frac{100}{2} \right) - \left(80 - \frac{64}{2} \right) \right] = \frac{1}{25} \left[\frac{100}{2} - \frac{96}{2} \right] = \frac{1}{25} 2 = 0,08 \end{aligned}$$

$$\begin{aligned} \text{e) } 0,90 &= P[X < m] = \int_{-\infty}^m f(x) dx = \int_0^5 \frac{1}{25} x dx + \int_5^m \frac{10-x}{25} dx = \frac{1}{25} \left[\frac{x^2}{2} \right]_0^5 + \frac{1}{25} \left[10x - \frac{x^2}{2} \right]_5^m = \\ &= \frac{1}{25} \frac{25}{2} + \frac{1}{25} \left[\left(10m - \frac{m^2}{2} \right) - \left(50 - \frac{25}{2} \right) \right] = \frac{1}{2} + \frac{1}{25} \left(10m - \frac{m^2}{2} \right) - \frac{1}{25} \frac{75}{2} = \end{aligned}$$

$$0,90 = \frac{1}{25} \left(10m - \frac{m^2}{2} \right) - 1 \Leftrightarrow 1,90 \times 25 \times 2 = 20m - m^2 \Leftrightarrow m^2 - 20m + 95 = 0$$

$$m = \frac{20 \pm \sqrt{400 - 380}}{2} = \frac{20 \pm \sqrt{20}}{2}$$

$$m = \frac{20 + \sqrt{20}}{2} = 12,236 > 10 \quad \text{esta solución no es válida pues } 0 < X < 10.$$

$$m = \frac{20 - \sqrt{20}}{2} = 7,7639 \quad \text{es la solución correcta.}$$

Con un mínimo de 7764 unidades podremos atender la demanda mensual con una probabilidad mayor al 90%.

10. La longitud en centímetros de un tornillo fabricado por una máquina se distribuye según una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} k(x-1)(4-x) & 1 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

El tornillo sólo es válido si su longitud está comprendida entre 2 y 3 cm.

- ¿Cuál es la probabilidad de que el tornillo sea útil?
- Calcule la longitud media de los tornillos fabricados por dicha máquina.

Solución:

$$1 = \int_{-\infty}^{\infty} f(x) dx = \int_1^4 k(x-1)(4-x) dx = k \int_1^4 (4x - x^2 - 4 + x) dx = k \int_1^4 (5x - x^2 - 4) dx =$$

$$= k \left[5 \frac{x^2}{2} - \frac{x^3}{3} - 4x \right]_1^4 = k \left[\left(5 \frac{16}{2} - \frac{64}{3} - 16 \right) - \left(5 \frac{1}{2} - \frac{1}{3} - 4 \right) \right] = k \left[\frac{16}{6} - \left(\frac{-11}{6} \right) \right] = k \frac{27}{6}$$

$$1 = k \frac{27}{6} \Leftrightarrow \frac{6}{27} = 0,2222 = k$$

$$\text{a) } P[2 < X < 3] = \int_2^3 f(x) dx = \frac{6}{27} \int_2^3 (5x - x^2 - 4) dx = \frac{6}{27} \left[5 \frac{x^2}{2} - \frac{x^3}{3} - 4x \right]_2^3 =$$

$$= \frac{6}{27} \left[\left(5 \frac{9}{2} - \frac{27}{3} - 12 \right) - \left(5 \frac{4}{2} - \frac{8}{3} - 8 \right) \right] = \frac{6}{27} \left[\frac{9}{6} - \left(\frac{-4}{6} \right) \right] = \frac{13}{27} = 0,48148$$

$$\text{b) } E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_1^4 x \frac{6}{27} (5x - x^2 - 4) dx = \frac{6}{27} \int_1^4 (5x^2 - x^3 - 4x) dx =$$

$$= \frac{6}{27} \left[5 \frac{x^3}{3} - \frac{x^4}{4} - 4 \frac{x^2}{2} \right]_1^4 = \frac{6}{27} \left[\left(5 \frac{64}{3} - 64 - 32 \right) - \left(5 \frac{1}{3} - \frac{1}{4} - 2 \right) \right] = \frac{6}{27} \left[\frac{32}{3} - \left(\frac{-7}{12} \right) \right] =$$

$$= \frac{6}{27} \frac{135}{12} = \frac{5}{2} = 2,5$$

11. La demanda diaria de un artículo sigue una ley de probabilidad dada por la función de densidad

$$f(x) = \begin{cases} 1 - \frac{1}{2}x & 0 \leq x \leq a \\ 0 & \text{en el resto} \end{cases}$$

Donde x viene expresada en miles de unidades.

- Determine el valor de a .
- Obtenga la función de distribución.
- Calcule la probabilidad de que el número de unidades demandadas en un día
 - no supere las 3000.
 - sea igual a 1500.
 - esté comprendido entre 635 y 1870.

Solución:

$$a) \quad 1 = \int_{-\infty}^{\infty} f(x) dx = \int_0^a \left(1 - \frac{1}{2}x\right) dx = \left[x - \frac{1}{2} \frac{x^2}{2}\right]_0^a = a - \frac{a^2}{4}$$

$$1 = a - \frac{a^2}{4} \quad \Leftrightarrow \quad a^2 - 4a + 4 = 0 \quad \Leftrightarrow \quad a = \frac{4 \pm \sqrt{16 - 16}}{2} = 2$$

$$b) \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du$$

(**Nota:** se utiliza la variable u para la función de densidad porque se usa la variable x como límite de integración, variable de la función de distribución)

Para hallar la función de distribución distinguimos si $x < 0$, $0 \leq x \leq 2$ o $2 < x$.

$$\bullet \quad x < 0, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^x 0 du = 0$$

$$\bullet \quad 0 \leq x \leq 2,$$

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^0 0 du + \int_0^x \left(1 - \frac{1}{2}u\right) du = 0 + \left[u - \frac{1}{2} \frac{u^2}{2}\right]_0^x = x - \frac{x^2}{4}$$

$$\bullet \quad 2 < x, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^0 0 du + \int_0^2 \left(1 - \frac{1}{2}u\right) du + \int_2^x 0 du = 0 + 1 + 0 = 1$$

$$c.1) \quad P[X \leq 3] = F(3) = 1$$

c.2) $P[X = 1,5] = 0$, la probabilidad de todo valor aislado en una distribución de probabilidad continua es igual a cero.

$$c.3) \quad P\left[0,635 \leq X \leq 1,87\right] = F(1,87) - F(0,635) = \left(1,87 - \frac{1,87^2}{4}\right) - \left(0,635 - \frac{0,635^2}{4}\right) = 0,46158$$

12. Si X es una variable aleatoria con función de densidad

$$f(x) = \begin{cases} k(3-x) & 0 \leq x < 3 \\ 0 & \text{en el resto} \end{cases}$$

Obtenga:

- El valor esperado.
- El coeficiente de variación de Pearson.
- La probabilidad de que la variable sea mayor que 2.
- El valor de la variable por encima del que se situarían el 20% de las observaciones.

Solución:

$$1 = \int_{-\infty}^{\infty} f(x) dx = \int_0^3 k(3-x) dx = k \left[3x - \frac{x^2}{2} \right]_0^3 = k \left[9 - \frac{9}{2} \right] = k \frac{9}{2} \quad \Leftrightarrow \quad k = \frac{2}{9}$$

$$\begin{aligned} a) \quad E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^3 x \frac{2}{9} (3-x) dx = \frac{2}{9} \int_0^3 (3x - x^2) dx = \frac{2}{9} \left[3 \frac{x^2}{2} - \frac{x^3}{3} \right]_0^3 = \\ &= \frac{2}{9} \left[\frac{27}{2} - \frac{27}{3} \right] = \frac{2}{9} \frac{27}{6} = 1 \end{aligned}$$

$$b) \quad CV[X] = \frac{\sigma[X]}{E[X]} = \frac{\sqrt{\frac{1}{2}}}{1} = \sqrt{\frac{1}{2}} = 0,7071$$

$$\sigma^2[X] = E[X^2] - E^2[X] = \frac{3}{2} - 1^2 = \frac{1}{2}$$

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^3 x^2 \frac{2}{9} (3-x) dx = \frac{2}{9} \int_0^3 (3x^2 - x^3) dx = \frac{2}{9} \left[3 \frac{x^3}{3} - \frac{x^4}{4} \right]_0^3 = \\ &= \frac{2}{9} \left[27 - \frac{81}{4} \right] = \frac{2}{9} \frac{27}{4} = \frac{3}{2} = 1,5 \end{aligned}$$

$$c) \quad P[X > 2] = \int_2^{\infty} f(x) dx = \int_2^3 \frac{2}{9} (3-x) dx = \frac{2}{9} \left[3x - \frac{x^2}{2} \right]_2^3 = \frac{2}{9} \left[\left(9 - \frac{9}{2} \right) - \left(6 - 2 \right) \right] = \frac{1}{9} = 0,1111$$

$$d) \quad 0,20 = P[X > a] = \int_a^{\infty} f(x) dx = \int_a^3 \frac{2}{9} (3-x) dx = \frac{2}{9} \left[3x - \frac{x^2}{2} \right]_a^3 = \frac{2}{9} \left[\left(9 - \frac{9}{2} \right) - \left(3a - \frac{a^2}{2} \right) \right] =$$

$$= \frac{2}{9} \left[\frac{9}{2} - 3a + \frac{a^2}{2} \right] \Leftrightarrow 0,20 \times \frac{9}{2} = \frac{9}{2} - 3a + \frac{a^2}{2} \Leftrightarrow 1,8 = 9 - 6a + a^2 \Leftrightarrow a^2 - 6a + 7,2 = 0$$

$$a = \frac{6 + \sqrt{36 - 28,8}}{2} = 4,34, \text{ no vale pues } X \text{ sólo toma valores entre 0 y 3.}$$

$$a = \frac{6 - \sqrt{36 - 28,8}}{2} = 1,66 \text{ es el valor de la variable superado por un 20\% de las observaciones.}$$

13. Una variable aleatoria de tipo continuo tiene por función de densidad:

$$f(x) = \begin{cases} kx & 0 \leq x \leq 2 \\ 0 & \text{en el resto} \end{cases}$$

Se pide:

- El valor de k .
- La función de distribución.
- La mediana.
- La probabilidad de que la variable sea mayor que 1.
- La media.
- La varianza.

Solución:

$$\text{a) } 1 = \int_{-\infty}^{\infty} f(x) dx = \int_0^2 kx dx = k \left[\frac{x^2}{2} \right]_0^2 = k2 \Leftrightarrow k = \frac{1}{2}$$

b) Para hallar la función de distribución distinguimos si $x < 0$, $0 \leq x \leq 2$ o $2 < x$.

$$x < 0, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^x 0 du = 0$$

$$0 \leq x \leq 2, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^0 0 du + \int_0^x \frac{1}{2} u du = 0 + \frac{1}{2} \left[\frac{u^2}{2} \right]_0^x = \frac{x^2}{4}$$

$$2 < x, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^0 0 du + \int_0^2 \frac{1}{2} u du + \int_2^x 0 du = 0 + 1 + 0 = 1$$

$$\text{c) } \frac{1}{2} = P[X \leq Me] = F(Me) = \frac{Me^2}{4} \Leftrightarrow 2 = Me^2 \Leftrightarrow Me = +\sqrt{2}$$

(Nota: la raíz cuadrada con signo negativo no es solución pues la variable sólo toma valores entre 0 y 2)

Este apartado se podría haber resuelto también con la función de densidad:

$$\frac{1}{2} = \int_{-\infty}^{Me} f(x) dx = \int_0^{Me} \frac{1}{2} x dx = \dots$$

$$d) \quad P[X > 1] = 1 - P[X \leq 1] = 1 - F(1) = 1 - \frac{1}{4} = \frac{3}{4} = 0,75$$

Este apartado se podría haber resuelto también con la función de densidad:

$$P[X > 1] = \int_1^{\infty} f(x) dx = \int_1^2 \frac{1}{2} x dx = \dots$$

$$e) \quad E[X] = \int_{-\infty}^{\infty} x f(x) dx = \frac{1}{2} \int_0^2 x^2 dx = \frac{1}{2} \left[\frac{x^3}{3} \right]_0^2 = \frac{1}{2} \frac{8}{3} = \frac{4}{3} = 1,3333$$

$$f) \quad E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{2} \int_0^2 x^3 dx = \frac{1}{2} \left[\frac{x^4}{4} \right]_0^2 = \frac{1}{2} 4 = 2$$

$$\sigma^2[X] = E[X^2] - E^2[X] = 2 - \left(\frac{4}{3} \right)^2 = 2 - \frac{16}{9} = \frac{2}{9} = 0,2222$$

14. El precio de un determinado artículo fluctúa, según sus características, pudiendo llegar a costar dos mil euros la unidad. Sea X la variable aleatoria que representa el precio por unidad, en miles de euros. Supongamos que la función de distribución de la variable X viene dada por

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{4x - x^2}{4} & 0 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

Se pide:

- La probabilidad de que una unidad cueste más de 1500 euros.
- El tanto por ciento de unidades que tienen su precio comprendido entre 500 y 1000 euros.
- ¿Qué precio tienen el 25% de las unidades más caras? (Nota: tercer cuartil)
- El precio medio.
- La varianza de los precios.

Solución:

$$a) \quad P[X > 1,5] = 1 - P[X \leq 1,5] = 1 - F(1,5) = 1 - \frac{(4 \times 1,5) - 1,5^2}{4} = 0,0625$$

$$b) \quad P[0,5 < X < 1] = P[0,5 < X \leq 1] = P[X \leq 1] - P[X \leq 0,5] = F(1) - F(0,5) = \\ = \frac{4-1}{4} - \frac{(4 \times 0,5) - 0,5^2}{4} = \frac{3}{4} - \frac{1,75}{4} = \frac{1,25}{4} = 0,3125$$

$$c) \quad 0,75 = P[X \leq Q_3] = F(Q_3) = \frac{(4 \times Q_3) - Q_3^2}{4} \Leftrightarrow 4 \times 0,75 = 4Q_3 - Q_3^2 \Leftrightarrow 0 = Q_3 - 4Q_3 + 3$$

$$Q_3 = \frac{4 \pm \sqrt{16-12}}{2} = \frac{4 \pm 2}{2} = 3 \text{ y } 1 \quad P[X \leq 3] = F(3) = 1 \neq 0,75$$

Luego la solución válida es 1 (Las unidades más caras tienen un precio por encima de 1000€).

d) En primer lugar obtenemos la función de densidad derivando la función de distribución:

$$f(x) = F'(x) = \begin{cases} \frac{4-2x}{4} = \frac{2-x}{2} & 0 \leq x < 2 \\ 0 & x < 0 \text{ o } 2 \leq x \text{ (en otro caso)} \end{cases}$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_0^2 x \frac{2-x}{2} dx = \frac{1}{2} \int_0^2 (2x - x^2) dx = \frac{1}{2} \left[x^2 - \frac{x^3}{3} \right]_0^2 = \frac{1}{2} \left[4 - \frac{8}{3} \right] = \frac{1}{2} \frac{4}{3} = \frac{2}{3} = 0,6667$$

$$e) \quad E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^2 x^2 \frac{2-x}{2} dx = \frac{1}{2} \int_0^2 (2x^2 - x^3) dx = \frac{1}{2} \left[\frac{2x^3}{3} - \frac{x^4}{4} \right]_0^2 =$$

$$= \frac{1}{2} \left[\frac{16}{3} - \frac{16}{4} \right] = \frac{1}{2} \frac{16}{12} = \frac{2}{3} = 0,6667$$

$$\sigma^2[X] = E[X^2] - E^2[X] = \frac{2}{3} - \left(\frac{2}{3} \right)^2 = \frac{2}{3} - \frac{4}{9} = \frac{2}{9} = 0,2222$$

15. Sea X una variable aleatoria con función de distribución:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^3}{17} + \frac{9}{34}x & 0 \leq x \leq 2 \\ 1 & 2 < x \end{cases}$$

a) Calcule la función de densidad.

b) Calcule la esperanza de X .

c) Calcule la moda.

Solución:

$$a) \quad f(x) = F'(x) = \begin{cases} \frac{3x^2}{17} + \frac{9}{34} = \frac{6x^2+9}{34} & 0 \leq x \leq 2 \\ 0 & x < 0 \text{ o } 2 < x \text{ (en otro caso)} \end{cases}$$

$$b) \quad E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_0^2 x \frac{6x^2+9}{34} dx = \frac{1}{34} \int_0^2 (6x^3 + 9x) dx = \frac{1}{34} \left[6 \frac{x^4}{4} + 9 \frac{x^2}{2} \right]_0^2 =$$

$$= \frac{1}{34} [24 + 18] = \frac{42}{34} = \frac{21}{17} = 1,2353$$

- c) Para la moda buscamos el máximo valor que toma la función de densidad. Éste se encuentra entre los máximos relativos dentro del intervalo (0,2) y los valores de la función de densidad en los extremos de dicho intervalo.

Máximos relativos:

$$f'(x) = 0 = \frac{12x}{34} \Leftrightarrow x = 0 \notin (0,2) \Rightarrow \text{no hay máximos relativos en el intervalo } (0,2).$$

Valores de la función de densidad en los extremos del intervalo (0,2):

$$f(0) = \frac{9}{34} \quad f(2) = \frac{24+9}{34} = \frac{33}{34}$$

Por tanto, $Mo=2$.

16. Dada una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} \frac{x+1}{2} & -1 \leq x < 1 \\ 0 & \text{en el resto} \end{cases}$$

- Obtenga la función de distribución.
- Calcule la mediana.
- Calcule la probabilidad de que X se encuentre entre -0,5 y 0,5.
- Obtenga la media.

Solución:

- a) Para hallar la función de distribución distinguimos si $x < -1$, $-1 \leq x < 1$ o $1 \leq x$.

$$x < -1, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^x 0 du = 0$$

$$\begin{aligned} -1 \leq x < 1, \quad F(x) &= P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^{-1} 0 du + \int_{-1}^x \frac{u+1}{2} du = \frac{1}{2} \int_{-1}^x (u+1) du = \\ &= \frac{1}{2} \left[\frac{u^2}{2} + u \right]_{-1}^x = \frac{1}{2} \left[\left(\frac{x^2}{2} + x \right) - \left(\frac{1}{2} - 1 \right) \right] = \frac{x^2 + 2x + 1}{4} \end{aligned}$$

$$1 \leq x, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^{-1} 0 du + \int_{-1}^1 \frac{u+1}{2} du + \int_1^x 0 du = 0 + 1 + 0 = 1$$

$$\text{b) } \frac{1}{2} = P[X \leq Me] = F(Me) = \frac{Me^2 + 2Me + 1}{4} \Leftrightarrow 0 = Me^2 + 2Me - 1 \Leftrightarrow$$

$$Me = \frac{-2 \pm \sqrt{4+4}}{2} = -4,8284 \quad \text{y} \quad 0,8284$$

$-4,8284 \notin [-1,1)$ luego $Me=0,8284$.

$$\begin{aligned} \text{c) } P[-0,5 < X < 0,5] &= P[-0,5 < X \leq 0,5] = F(0,5) - F(-0,5) = \\ &= \frac{0,5^2 + (2 \times 0,5) + 1}{4} - \frac{(-0,5)^2 + (2 \times (-0,5)) + 1}{4} = \frac{2,25}{4} - \frac{0,25}{4} = \frac{2}{4} = \frac{1}{2} = 0,5 \end{aligned}$$

$$\begin{aligned} \text{d) } E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-1}^1 x \frac{x+1}{2} dx = \frac{1}{2} \int_{-1}^1 (x^2 + x) dx = \frac{1}{2} \left[\frac{x^3}{3} + \frac{x^2}{2} \right]_{-1}^1 = \\ &= \frac{1}{2} \left[\left(\frac{1}{3} + \frac{1}{2} \right) - \left(-\frac{1}{3} + \frac{1}{2} \right) \right] = \frac{1}{2} \frac{2}{3} = \frac{1}{3} = 0,3333 \end{aligned}$$

17. La variable aleatoria X expresa los miles de artículos que mensualmente se solicitan a un proveedor. Su función de densidad es:

$$f(x) = \begin{cases} kx(x+1) & 0 \leq x < 2 \\ 0 & \text{en el resto} \end{cases}$$

Se pide:

- La constante k .
- El número de artículos que el proveedor espera que le soliciten a lo largo de un año.
- La función de distribución.
- En un determinado mes, el proveedor sólo dispone de 1200 artículos, ¿cuál es la probabilidad de poder atender todas las peticiones que se produzcan en dicho mes?

Solución:

$$\begin{aligned} \text{a) } 1 &= \int_{-\infty}^{\infty} f(x) dx = \int_0^2 kx(x+1) dx = k \int_0^2 (x^2 + x) dx = k \left[\frac{x^3}{3} + \frac{x^2}{2} \right]_0^2 = k \left[\frac{8}{3} + 2 \right] = k \frac{14}{3} \\ k &= \frac{3}{14} \end{aligned}$$

b) X =miles de artículos solicitados mensualmente.

Y =miles de artículos solicitados anualmente= $12X$.

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^2 x \frac{3}{14} x(x+1) dx = \frac{3}{14} \int_0^2 (x^3 + x^2) dx = \frac{3}{14} \left[\frac{x^4}{4} + \frac{x^3}{3} \right]_0^2 = \\ &= \frac{3}{14} \left[4 + \frac{8}{3} \right] = \frac{3}{14} \frac{20}{3} = \frac{10}{7} = 1,429 \end{aligned}$$

$$E[Y] = 12E[X] = 12 \times \frac{10}{7} = \frac{120}{7} = 17,143$$

El proveedor espera que le soliciten 17143 artículos a lo largo del año.

c) Para hallar la función de distribución distinguimos si $x < 0$, $0 \leq x < 2$ o $2 \leq x$.

$$x < 0, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^x 0 du = 0$$

$$\begin{aligned} 0 \leq x < 2, \quad F(x) &= P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^0 0 du + \int_0^x \frac{3}{14} u(u+1) du = \int_0^x \frac{3}{14} (u^2 + u) du = \\ &= \frac{3}{14} \left[\frac{u^3}{3} + \frac{u^2}{2} \right]_0^x = \frac{3}{14} \left[\frac{x^3}{3} + \frac{x^2}{2} \right] = \frac{3}{14} \frac{2x^3 + 3x^2}{6} = \frac{2x^3 + 3x^2}{28} \end{aligned}$$

$$2 \leq x, \quad F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du = \int_{-\infty}^0 0 du + \int_0^2 \frac{3}{14} u(u+1) du + \int_2^x 0 du = 0 + 1 + 0 = 1$$

$$d) \quad P[X \leq 1, 2] = F(1, 2) = \frac{(2 \times 1, 2^3) + (3 \times 1, 2^2)}{28} = 0,2777$$

También se podría haber resuelto con la función de densidad:

$$P[X \leq 1, 2] = \int_{-\infty}^{1,2} f(x) dx = \int_0^{1,2} \frac{3}{14} x(x+1) dx = \dots$$

18. La cantidad de pan, en cientos de unidades, vendida diariamente en la panadería de un pueblo puede representarse mediante una variable aleatoria con la siguiente función de densidad:

$$f(x) = \begin{cases} kx & 0 \leq x < 4 \\ k(7-x) & 4 \leq x < 7 \\ 0 & \text{en el resto} \end{cases}$$

a) Calcule k .

b) Calcule la mediana.

c) Calcule la media.

Solución:

$$\begin{aligned} a) \quad 1 &= \int_{-\infty}^{\infty} f(x) dx = \int_0^4 kx dx + \int_4^7 k(7-x) dx = k \left[\frac{x^2}{2} \right]_0^4 + k \left[7x - \frac{x^2}{2} \right]_4^7 = \\ &= k[8] + k \left[\left(49 - \frac{49}{2} \right) - \left(28 - 8 \right) \right] = k \left[8 + \frac{49}{2} - 20 \right] = k \frac{25}{2} \Leftrightarrow k = \frac{2}{25} \end{aligned}$$

b) Veamos en primer lugar si $Me \in [0, 4)$ o $Me \in [4, 7)$

$$P[X \leq 4] = \int_{-\infty}^4 f(x) dx = \int_0^4 \frac{2}{25} x dx = \frac{2}{25} \left[\frac{x^2}{2} \right]_0^4 = \frac{2}{25} 8 = 0,64 > \frac{1}{2}$$

Luego Me es menor que 4.

$$\frac{1}{2} = P[X \leq Me] = \int_0^{Me} \frac{2}{25} x dx = \frac{2}{25} \left[\frac{x^2}{2} \right]_0^{Me} = \frac{Me^2}{25} \Leftrightarrow \frac{25}{2} = Me^2 \Leftrightarrow Me = +\sqrt{\frac{25}{2}} = 3,5355$$

La solución negativa no vale puesto que la variable aleatoria X no toma valores negativos.

$$\begin{aligned} \text{c) } E[X] &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^4 x \frac{2}{25} x dx + \int_4^7 x \frac{2}{25} (7-x) dx = \frac{2}{25} \left(\int_0^4 x^2 dx + \int_4^7 (7x - x^2) dx \right) = \\ &= \frac{2}{25} \left(\left[\frac{x^3}{3} \right]_0^4 + \left[7 \frac{x^2}{2} - \frac{x^3}{3} \right]_4^7 \right) = \frac{2}{25} \left(\left[\frac{64}{3} \right] + \left[\left(\frac{343}{2} - \frac{343}{3} \right) - \left(56 - \frac{64}{3} \right) \right] \right) = \frac{2}{25} \frac{263}{6} = \frac{263}{75} = 3,5067 \end{aligned}$$

19. Se sabe que la función de distribución de la variable aleatoria $X = \text{horas de duración de un tipo de bombillas}$ es:

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-\frac{x}{100}} & \text{si } x \geq 0 \end{cases}$$

Obtenga:

- Función de densidad y compruebe sus propiedades.
- Probabilidad de que una bombilla dure entre 100 y 200 horas.
- Probabilidad de que una bombilla dure más de 200 horas.
- Media.
- Varianza.
- Moda.
- Mediana.
- Percentil 75.
- Coefficiente de variación.

Solución:

- Derivando la función de distribución obtenemos la función de densidad:

$$F'(x) = f(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{100} e^{-\frac{x}{100}} & \text{si } x \geq 0 \end{cases}$$

Comprobemos que $f(x)$ cumple las dos propiedades que caracterizan a toda función de densidad:

- $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x) dx = 1$

La primera es inmediata a partir de su expresión y para comprobar la segunda se calcula la integral:

$$\int_{-\infty}^{+\infty} f(x) dx = \int_{-\infty}^0 0 dx + \int_0^{+\infty} \frac{1}{100} e^{-\frac{x}{100}} dx = 0 + \left[-e^{-\frac{x}{100}} \right]_0^{+\infty} = -e^{-\infty} + e^0 = -0 + 1 = 1$$

b) Haremos los apartados *b* y *c* con la función de densidad y con la función de distribución:

$$P[100 < X < 200] = \int_{100}^{200} \frac{1}{100} e^{-\frac{x}{100}} dx = \left[-e^{-\frac{x}{100}} \right]_{100}^{200} = -e^{-2} + e^{-1} = -0,135 + 0,368 = 0,233$$

$$P[100 < X < 200] = F(200) - F(100) = \left(1 - e^{-\frac{200}{100}} \right) - \left(1 - e^{-\frac{100}{100}} \right) = -e^{-2} + e^{-1} = 0,233$$

$$c) P[200 < X] = \int_{200}^{+\infty} \frac{1}{100} e^{-\frac{x}{100}} dx = \left[-e^{-\frac{x}{100}} \right]_{200}^{+\infty} = -e^{-\infty} + e^{-2} = -0 + 0,135 = 0,135$$

$$P[200 < X] = 1 - P[X \leq 200] = 1 - F(200) = 1 - \left(1 - e^{-\frac{200}{100}} \right) = e^{-2} = 0,135$$

$$d) E[X] = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^{+\infty} x \frac{1}{100} e^{-\frac{x}{100}} dx = \left\{ \begin{array}{ll} u = x & du = dx \\ dv = \frac{1}{100} e^{-\frac{x}{100}} dx & v = -e^{-\frac{x}{100}} \end{array} \right\} =$$

$$= \left[-xe^{-\frac{x}{100}} \right]_0^{+\infty} - \int_0^{+\infty} -e^{-\frac{x}{100}} dx = \left[-xe^{-\frac{x}{100}} \right]_0^{+\infty} - \left[100e^{-\frac{x}{100}} \right]_0^{+\infty} =$$

$$= [0 - (-0 \times e^0)] - [(100 \times e^{-\infty}) - (100 \times e^0)] =$$

$$= [0 - (-0 \times 1)] - [(100 \times 0) - (100 \times 1)] = 100$$

Hemos utilizado que:

$$e^{+\infty} = +\infty, \quad e^{-\infty} = \frac{1}{e^{\infty}} = 0, \quad e^0 = 1, \quad \lim_{x \rightarrow +\infty} -xe^{-\frac{x}{100}} = \lim_{x \rightarrow +\infty} \frac{-x}{e^{\frac{x}{100}}} = \frac{-\infty}{\infty} = \lim_{x \rightarrow +\infty} \frac{-1}{\frac{1}{100} e^{\frac{x}{100}}} = \frac{-1}{\infty} = 0$$

$$e) \sigma^2[X] = E[X^2] - E^2[X] = 20000 - 10000 = 10000$$

$$E[X^2] = \int_0^{+\infty} x^2 \frac{1}{100} e^{-\frac{x}{100}} dx = \left\{ \begin{array}{ll} u = x^2 & du = 2x dx \\ dv = \frac{1}{100} e^{-\frac{x}{100}} dx & v = -e^{-\frac{x}{100}} \end{array} \right\} =$$

$$= \left[-x^2 e^{-\frac{x}{100}} \right]_0^{+\infty} - \int_0^{+\infty} -2x e^{-\frac{x}{100}} dx = [0 - (-0 \times e^0)] + (2 \times 100) \int_0^{+\infty} x \frac{1}{100} e^{-\frac{x}{100}} dx =$$

$$= 2 \times 100 \times 100 = 20000$$

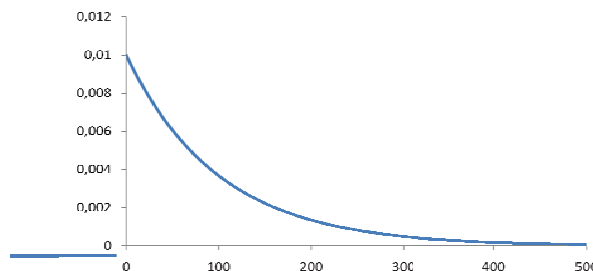
Hemos utilizado que:

$$e^{+\infty} = +\infty, \quad e^{-\infty} = \frac{1}{e^{\infty}} = 0, \quad e^0 = 1$$

$$\lim_{x \rightarrow +\infty} -x^2 e^{-\frac{x}{100}} = \lim_{x \rightarrow +\infty} \frac{-x^2}{e^{\frac{x}{100}}} = \frac{-\infty}{\infty} = \lim_{x \rightarrow +\infty} \frac{-2x}{\frac{1}{100} e^{\frac{x}{100}}} = \frac{-\infty}{\infty} = \lim_{x \rightarrow +\infty} \frac{-2}{\left(\frac{1}{100}\right)^2 e^{\frac{x}{100}}} = \frac{-2}{\infty} = 0$$

- f) Para hallar la moda en distribuciones de probabilidad continuas hay que calcular el punto o puntos donde la función de densidad alcanza su máximo.

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{100} e^{-\frac{x}{100}} & \text{si } x \geq 0 \end{cases}$$



Como puede observarse en la gráfica de la función de densidad, ésta alcanza su máximo en $Mo=0$.

Si no conociéramos la representación gráfica de la función de densidad, habría que calcular los máximos relativos de dicha función y compararlos con el valor de la función de densidad en los extremos del dominio donde está definida, $[0, \infty)$, es decir en $x=0$ y $x=\infty$.

Derivando la función de densidad en $x > 0$,

$$f'(x) = \frac{-1}{100^2} e^{-\frac{x}{100}} \quad \text{si } x > 0$$

vemos que su derivada no se anula nada más que en $x=\infty$, luego no tiene máximos relativos. Además, su signo es negativo para todos los valores $x > 0$, es decir, es siempre decreciente, por tanto alcanza su máximo en $x=0$.

- g) Para obtener la mediana y el percentil 75 utilizaremos la función de distribución

$$F(Me) = P[X \leq Me] = \frac{1}{2}$$

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-\frac{x}{100}} & \text{si } x \geq 0 \end{cases}$$

$$1 - e^{-\frac{x}{100}} = \frac{1}{2} \Leftrightarrow \frac{1}{2} = e^{-\frac{x}{100}} \Leftrightarrow \ln\left(\frac{1}{2}\right) = -\frac{x}{100} \Leftrightarrow$$

$$\Leftrightarrow -0,693 = -\frac{x}{100} \Leftrightarrow x = 69,3 \quad Me = 69,3$$

- h) $F(P_{75}) = P[X \leq P_{75}] = 0,75$

$$1 - e^{-\frac{x}{100}} = 0,75 \Leftrightarrow 0,25 = e^{-\frac{x}{100}} \Leftrightarrow \ln(0,25) = -\frac{x}{100} \Leftrightarrow$$

$$\Leftrightarrow -1,386 = -\frac{x}{100} \Leftrightarrow P_{75} = 138,6$$

- i) Para calcular el coeficiente de variación utilizamos los valores de la media y varianza calculados anteriormente.

$$CV = \frac{\sigma}{E[X]} = \frac{\sqrt{10000}}{100} = 1$$

20. Una variable aleatoria bidimensional tiene la siguiente función de densidad conjunta:

$$f(x, y) = \begin{cases} k \left(x^2 + \frac{xy}{3} \right) & 0 \leq x \leq 1 \quad 0 \leq y \leq 2 \\ 0 & \text{en el resto} \end{cases}$$

Se pide:

- El valor esperado de X .
- El valor esperado de Y .
- La varianza de X .
- ¿Son independientes X e Y ?
- La covarianza.

Solución:

En primer lugar vamos a calcular el valor de la constante k y las funciones de distribución marginales de X e Y .

$$\begin{aligned} \bullet \quad 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = \int_{y=0}^{y=2} \int_{x=0}^{x=1} k \left(x^2 + \frac{xy}{3} \right) dx dy = \int_{y=0}^{y=2} \left[\int_{x=0}^{x=1} k \left(x^2 + \frac{xy}{3} \right) dx \right] dy = \\ &= \int_{y=0}^{y=2} k \left(\frac{2+y}{6} \right) dy = \frac{k}{6} \int_{y=0}^{y=2} (2+y) dy = \frac{k}{6} \left[2y + \frac{y^2}{2} \right]_0^2 = \frac{k}{6} (4+2) = k \quad \Leftrightarrow \quad 1 = k \end{aligned}$$

$$\text{Donde} \quad \int_{x=0}^{x=1} k \left(x^2 + \frac{xy}{3} \right) dx = k \left[\frac{x^3}{3} + \frac{y}{3} \frac{x^2}{2} \right]_0^1 = k \left[\frac{1}{3} + \frac{y}{6} \right] = k \left(\frac{2+y}{6} \right)$$

- Para hallar la función de densidad marginal de la variable X distinguimos si $0 \leq x \leq 1$ o no.

$$\text{Si } 0 \leq x \leq 1, \quad f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{y=0}^{y=2} \left(x^2 + \frac{xy}{3} \right) dy = \left[x^2 y + \frac{x}{3} \frac{y^2}{2} \right]_0^2 = 2x^2 + \frac{2}{3}x$$

$$\text{Si } x < 0 \text{ o } 1 < x, \quad f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-\infty}^{\infty} 0 dy = 0$$

- Para hallar la función de densidad marginal de la variable Y distinguimos si $0 \leq y \leq 2$ o no.

$$\text{Si } 0 \leq y \leq 2, \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{x=0}^{x=1} \left(x^2 + \frac{xy}{3} \right) dx = \left[\frac{x^3}{3} + \frac{y}{3} \frac{x^2}{2} \right]_0^1 = \frac{1}{3} + \frac{y}{6} = \frac{2+y}{6}$$

$$\text{Si } y < 0 \text{ o } 2 < y, \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{-\infty}^{\infty} 0 dx = 0$$

$$\begin{aligned} \text{a) } E[X] &= \int_{-\infty}^{\infty} x f_1(x) dx = \int_0^1 x \left(2x^2 + \frac{2}{3}x \right) dx = \int_0^1 \left(2x^3 + \frac{2}{3}x^2 \right) dx = \left[2 \frac{x^4}{4} + \frac{2}{3} \frac{x^3}{3} \right]_0^1 = \frac{2}{4} + \frac{2}{9} = \\ &= \frac{26}{36} = \frac{13}{18} = 0,7222 \end{aligned}$$

$$\begin{aligned} \text{b) } E[Y] &= \int_{-\infty}^{\infty} y f_2(y) dy = \int_0^2 y \left(\frac{2+y}{6} \right) dy = \frac{1}{6} \int_0^2 (2y + y^2) dy = \frac{1}{6} \left[y^2 + \frac{y^3}{3} \right]_0^2 = \frac{1}{6} \left[4 + \frac{8}{3} \right] = \\ &= \frac{20}{18} = \frac{10}{9} = 1,1111 \end{aligned}$$

$$\text{c) } \sigma^2[X] = E[X^2] - (E[X])^2 = E[X^2] - E^2[X]$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_1(x) dx = \int_0^1 x^2 \left(2x^2 + \frac{2}{3}x \right) dx = \int_0^1 \left(2x^4 + \frac{2}{3}x^3 \right) dx = \left[2 \frac{x^5}{5} + \frac{2}{3} \frac{x^4}{4} \right]_0^1 = \frac{2}{5} + \frac{2}{12} = \frac{17}{30}$$

$$\sigma^2[X] = \frac{17}{30} - \left(\frac{13}{18} \right)^2 = 0,04506$$

- d) Para comprobar la independencia de ambas variables, observamos si $f_1(x)f_2(y) = f(x, y)$ para todo valor de x e y .

En particular si $0 \leq x \leq 1$, $0 \leq y \leq 2$

$$f_1(x)f_2(y) = \left(2x^2 + \frac{2}{3}x \right) \left(\frac{2+y}{6} \right) \neq \left(x^2 + \frac{xy}{3} \right) = f(x, y) \quad \text{luego no son independientes.}$$

- e) Si las variables fuesen independientes la covarianza es cero (no sería necesario ningún cálculo).

$$\text{Cov}[X, Y] = E[XY] - (E[X]E[Y]) = \frac{43}{54} - \left(\frac{13}{18} \frac{10}{9} \right) = -0,006173$$

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy = \int_{y=0}^{y=2} \int_{x=0}^{x=1} xy \left(x^2 + \frac{xy}{3} \right) dx dy = \int_{y=0}^{y=2} \left[\int_{x=0}^{x=1} \left(x^3 y + \frac{x^2 y^2}{3} \right) dx \right] dy = \\ &= \int_{y=0}^{y=2} \left(\frac{y}{4} + \frac{y^2}{9} \right) dy = \left[\frac{1}{4} \frac{y^2}{2} + \frac{1}{9} \frac{y^3}{3} \right]_0^2 = \frac{1}{2} + \frac{8}{27} = \frac{43}{54} = 0,7963 \end{aligned}$$

$$\text{Donde } \int_{x=0}^{x=1} \left(x^3 y + \frac{x^2 y^2}{3} \right) dx = \left[y \frac{x^4}{4} + \frac{y^2}{3} \frac{x^3}{3} \right]_0^1 = \frac{y}{4} + \frac{y^2}{9}$$

21. Las variables aleatorias X e Y tienen la siguiente función de densidad conjunta:

$$f(x, y) = \begin{cases} \frac{x+y}{3} & 0 \leq x \leq 1 \quad 0 < y < 2 \\ 0 & \text{en el resto} \end{cases}$$

a) ¿Son independientes las variables?

b) Calcule la varianza de X .

Solución:

a) Para hallar la función de densidad marginal de la variable X distinguimos si $0 \leq x \leq 1$ o no.

$$\text{Si } 0 \leq x \leq 1, \quad f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{y=0}^{y=2} \frac{1}{3}(x+y) dy = \frac{1}{3} \left[xy + \frac{y^2}{2} \right]_0^2 = \frac{1}{3} [2x + 2] = \frac{2}{3}(x+1)$$

$$\text{Si } x < 0 \text{ o } 1 < x, \quad f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-\infty}^{\infty} 0 dy = 0$$

Para hallar la función de densidad marginal de la variable Y distinguimos si $0 \leq y \leq 2$ o no.

$$\text{Si } 0 \leq y \leq 2, \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{x=0}^{x=1} \frac{1}{3}(x+y) dx = \frac{1}{3} \left[\frac{x^2}{2} + yx \right]_0^1 = \frac{1}{3} \left(\frac{1}{2} + y \right)$$

$$\text{Si } y < 0 \text{ o } 2 < y, \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{-\infty}^{\infty} 0 dx = 0$$

Para comprobar la independencia de ambas variables, observamos si $f_1(x)f_2(y) = f(x, y)$ para todo valor de x e y .

En particular si $0 \leq x \leq 1, 0 \leq y \leq 2$

$$f_1(x)f_2(y) = \frac{2}{3}(x+1) \frac{1}{3} \left(\frac{1}{2} + y \right) \neq \frac{x+y}{3} = f(x, y) \quad \text{luego no son independientes.}$$

$$\text{b) } E[X] = \int_{-\infty}^{\infty} x f_1(x) dx = \int_0^1 x \frac{2}{3}(x+1) dx = \frac{2}{3} \int_0^1 (x^2 + x) dx = \frac{2}{3} \left[\frac{x^3}{3} + \frac{x^2}{2} \right]_0^1 =$$

$$= \frac{2}{3} \left[\frac{1}{3} + \frac{1}{2} \right] = \frac{2}{3} \frac{5}{6} = \frac{5}{9} = 0,5556$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_1(x) dx = \int_0^1 x^2 \frac{2}{3}(x+1) dx = \frac{2}{3} \int_0^1 (x^3 + x^2) dx = \frac{2}{3} \left[\frac{x^4}{4} + \frac{x^3}{3} \right]_0^1 =$$

$$= \frac{2}{3} \left[\frac{1}{4} + \frac{1}{3} \right] = \frac{2}{3} \frac{7}{12} = \frac{7}{18} = 0,3889$$

$$\sigma^2[X] = E[X^2] - E^2[X] = \frac{7}{18} - \left(\frac{5}{9} \right)^2 = \frac{7}{18} - \frac{25}{81} = 0,080247$$

22. Dada la variable aleatoria bidimensional (X, Y) con función de densidad conjunta

$$f(x, y) = \begin{cases} kxy^2 & 0 \leq x < 4 \quad 0 < y < 2 \\ 0 & \text{en el resto} \end{cases}$$

- a) Calcule k .
b) ¿Son independientes las variables?

Solución:

a)

$$1 = \int_{y=0}^{y=2} \int_{x=0}^{x=4} kxy^2 dx dy = \int_{y=0}^{y=2} ky^2 \left[\frac{x^2}{2} \right]_0^4 dy = \int_{y=0}^{y=2} ky^2 [8-0] dy = 8k \left[\frac{y^3}{3} \right]_0^2 = 8k \left[\frac{8}{3} - 0 \right] = \frac{64}{3} k$$

$$k = \frac{3}{64}$$

b)

$$f_1(x) = \int_0^2 \frac{3}{64} xy^2 dy = \frac{3}{64} x \left[\frac{y^3}{3} \right]_0^2 = \frac{3}{64} x \left[\frac{8}{3} - 0 \right] = \frac{1}{8} x \quad 0 \leq x < 4$$

$$f_1(x) = 0 \quad \text{en otro caso}$$

$$f_2(y) = \int_0^4 \frac{3}{64} xy^2 dx = \frac{3}{64} y^2 \left[\frac{x^2}{2} \right]_0^4 = \frac{3}{64} y^2 [8-0] = \frac{3}{8} y^2 \quad 0 < y < 2$$

$$f_2(y) = 0 \quad \text{en otro caso}$$

$$f_1(x)f_2(y) = \frac{3}{64} xy^2 = f(x, y) \quad \text{si } 0 \leq x < 4 \quad 0 < y < 2$$

$$f_1(x)f_2(y) = 0 = f(x, y) \quad \text{en otro caso}$$

$$f_1(x)f_2(y) = f(x, y) \quad \forall x, y \Rightarrow \text{Las variables son independientes.}$$

7. DISTRIBUCIONES DISCRETAS DE PROBABILIDAD.

7.1 Distribución Uniforme discreta.

Una variable aleatoria discreta X sigue una **distribución Uniforme** en n puntos x_1, \dots, x_n , si su distribución de probabilidad es:

$$p_k = P[X = x_k] = \frac{1}{n} \quad k = 1, \dots, n$$

(Esta es la distribución que se asume en la concepción de la probabilidad clásica o de Laplace)

La variable aleatoria que representa el valor obtenido al lanzar un dado no cargado sigue una distribución Uniforme discreta con probabilidad $\frac{1}{6}$ para cada uno de los seis posibles valores.

La media y varianza para estas variables aleatorias se calculan como sigue:

Media:

$$E[X] = \sum_{k=1}^n x_k p_k = \sum_{k=1}^n x_k \frac{1}{n} = \frac{\sum_{k=1}^n x_k}{n}$$

Varianza:

$$E[X^2] = \sum_{k=1}^n x_k^2 p_k = \sum_{k=1}^n x_k^2 \frac{1}{n} = \frac{\sum_{k=1}^n x_k^2}{n} \quad \text{Var}[X] = \frac{\sum_{k=1}^n x_k^2}{n} - \left(\frac{\sum_{k=1}^n x_k}{n} \right)^2$$

7.2 Distribución Binomial.

Distribución Binomial de parámetros 1 y p (distribución de Bernoulli).

Consideramos un experimento que puede presentar dos resultados: $A=\text{éxito}$ y $\bar{A}=\text{fracaso}$ con probabilidades $P(A) = p$, $P(\bar{A}) = q$, donde evidentemente $p + q = 1$.

A dicho experimento le asociamos la variable aleatoria:

$$X = \begin{cases} 1 & \text{si ocurre } A \\ 0 & \text{si ocurre } \bar{A} \end{cases}$$

Su distribución de probabilidad será:

$$\begin{aligned}P[X = 1] &= p \\P[X = 0] &= q = 1 - p\end{aligned}$$

La anterior distribución de probabilidad se conoce como **distribución Binomial de parámetros 1 y p** . Se nota como $X \sim \mathcal{B}(1, p)$, 1 es el número de veces que se observa o realiza el experimento y p es la probabilidad de obtener *éxito*.

Media:

$$E[X] = \sum_{k=0}^1 x_k p_k = (0 \times q) + (1 \times p) = p$$

Varianza:

$$E[X^2] = \sum_{k=0}^1 x_k^2 p_k = (0^2 \times q) + (1^2 \times p) = p$$

$$\sigma^2[X] = E[X^2] - E^2[X] = p - p^2 = p(1 - p) = pq$$

Distribución Binomial de parámetros n y p .

Sea la variable aleatoria X la suma de n variables aleatorias independientes e idénticamente distribuidas (i.i.d.) según una distribución de probabilidad $\mathcal{B}(1, p)$, es decir $X_i \sim \mathcal{B}(1, p) \quad i = 1, \dots, n$

$$X = X_1 + X_2 + \dots + X_n$$

La variable aleatoria X puede tomar todos los valores enteros comprendidos entre 0 y n , ambos inclusive, y representa el **número de éxitos que se obtienen en n pruebas o experimentos idénticos e independientes**.

A la distribución de esta variable aleatoria se le llama **distribución Binomial de parámetros n y p** .

Se nota $X \sim \mathcal{B}(n, p)$, n es el número de pruebas idénticas e independientes que se realizan u observan y p la probabilidad de obtener *éxito* en cada prueba.

Vamos a calcular la distribución de probabilidad asociada a esta variable aleatoria X , es decir:

$$p_x = P[X = x] \quad x = 0, \dots, n$$

X toma el valor x si por ejemplo se observa en las n realizaciones del experimento una secuencia de *éxitos y fracasos* como la que sigue:

$$\underbrace{\underbrace{AA \dots A}_x \underbrace{\overline{A} \overline{A} \dots \overline{A}}_{n-x}}_n$$

Dado que las pruebas o realizaciones del experimento son independientes:

$$P\left(\underbrace{AA\dots A}_x \underbrace{\overline{AA}\dots\overline{A}}_{n-x}\right) = \underbrace{pp\dots p}_x \underbrace{qq\dots q}_{n-x} = p^x q^{n-x}$$

Además, cualquier otra reordenación de la anterior secuencia de resultados tiene también x éxitos.

Dado que el número total de distintas reordenaciones es $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, se sigue que

$$p_x = P[X = x] = \binom{n}{x} p^x q^{n-x} \quad x = 0, \dots, n$$

Se puede comprobar que los valores p_x verifican las dos condiciones que toda distribución de probabilidad ha de cumplir:

- $p_x \geq 0 \quad \forall x$
- $\sum_{x=0}^n p_x = 1$

Dado que p , q , n y x son no negativos p_x también lo será.

La suma de todas las probabilidades se corresponden con el desarrollo del binomio $(p+q)^n = 1^n = 1$

$$\sum_{x=0}^n p_x = \sum_{x=0}^n \binom{n}{x} p^x q^{n-x} = (p+q)^n$$

Media:

$$E[X] = E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n] = p + p + \dots + p = np$$

Varianza:

Dado que X_1, X_2, \dots, X_n son independientes la varianza de la suma de dichas variables es igual a la suma de las varianzas de cada variable

$$\sigma^2[X] = \sigma^2[X_1 + X_2 + \dots + X_n] = \sigma^2[X_1] + \sigma^2[X_2] + \dots + \sigma^2[X_n] = pq + pq + \dots + pq = npq$$

Propiedad de aditividad o reproductividad.

Sean X e Y dos variables aleatorias independientes con distribuciones de probabilidad

$X \sim \mathcal{B}(n, p)$ e $Y \sim \mathcal{B}(m, p)$. Entonces $X + Y \sim \mathcal{B}(n+m, p)$.

► EJEMPLO 7.1

En un almacén hay 6000 piezas de las cuales 900 son defectuosas. Se seleccionan aleatoriamente (con reemplazamiento) 10 piezas. Calcular la probabilidad de que todas estén en buen estado.

Solución:

$X = \text{n}^\circ \text{ de piezas defectuosas en las 10 seleccionadas} \sim \mathcal{B}(10, \frac{900}{6000})$

$$p = \frac{900}{6000} = 0,15 \quad P[X = 0] = \binom{10}{0} 0,15^0 \times 0,85^{10} = \frac{10!}{0!10!} 1 \times 0,1969 = 0,1969$$

O bien:

$Y = \text{n}^\circ \text{ de piezas en buen estado de las 10 seleccionadas} \sim \mathcal{B}(10, \frac{5100}{6000})$

$$p = \frac{5100}{6000} = 0,85 \quad P[Y = 10] = \binom{10}{10} 0,85^{10} \times 0,15^0 = \frac{10!}{10!0!} 0,1969 \times 1 = 0,1969 \quad \blacktriangleleft$$

► EJEMPLO 7.2

Un proceso de fabricación produce un 1% de piezas defectuosas. Las piezas se empaquetan en cajas de 20 unidades. ¿Cuál es la probabilidad de que si se seleccionan 5 cajas al azar, ninguna de ellas tenga piezas defectuosas?

Solución:

$X = \text{n}^\circ \text{ de piezas defectuosas en una caja de 20 unidades} \sim \mathcal{B}(20, \frac{1}{100})$

$$P[X = 0] = \binom{20}{0} 0,01^0 \times 0,99^{20} = \frac{20!}{0!20!} 1 \times 0,8179 = 0,8179$$

$Y = \text{n}^\circ \text{ de cajas sin piezas defectuosas} \sim \mathcal{B}(5, 0,8179)$

$$P[Y = 5] = \binom{5}{5} 0,8179^5 \times 0,1821^0 = \frac{5!}{5!0!} 0,366 \times 1 = 0,366 \quad \blacktriangleleft$$

7.3 Distribución de Poisson.

Una variable aleatoria X sigue una **distribución de Poisson de parámetro λ** si puede tomar todos los valores enteros no negativos (0, 1, 2, ...) con probabilidades:

$$p_x = P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

Se nota abreviadamente como $X \sim \mathcal{P}(\lambda)$.

Una de las aplicaciones comunes de la distribución de Poisson es calcular la probabilidad de un cierto número de eventos en un determinado período de tiempo (por ejemplo, el número de automóviles que se presenta a una estación de peaje de una autopista en el intervalo de un minuto).

Media:

$$E[X] = \lambda$$

Varianza:

$$\sigma^2[X] = \lambda$$

Propiedad de aditividad o reproductividad.

Sean X e Y dos variables aleatorias independientes con distribuciones de probabilidad $X \sim \mathcal{P}(\lambda_1)$ e $Y \sim \mathcal{P}(\lambda_2)$. Entonces $X + Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$.

► **EJEMPLO 7.3**

El número de fallos por hora en un determinado mecanismo sigue una distribución de Poisson de media 1,5.

- a) ¿Cuál es la probabilidad de que haya algún fallo durante una hora?
- b) ¿Cuál es la probabilidad de que haya algún fallo durante tres horas?

Solución:

a) $X_1 = \text{nº de fallos durante una hora} \sim \mathcal{P}(1,5)$

$$P[X_1 \geq 1] = 1 - P[X_1 < 1] = 1 - P[X_1 = 0] = 1 - \frac{e^{-1,5} 1,5^0}{0!} = 1 - 0,2231 = 0,7769$$

b) $X = \text{número de fallos durante tres horas} = X_1 + X_2 + X_3$

$$X_1 = \text{nº de fallos durante la primera hora} \sim \mathcal{P}(1,5)$$

$$X_2 = \text{nº de fallos durante la segunda hora} \sim \mathcal{P}(1,5)$$

$$X_3 = \text{nº de fallos durante la tercera hora} \sim \mathcal{P}(1,5)$$

$$X \sim \mathcal{P}(1,5 + 1,5 + 1,5) = \mathcal{P}(4,5)$$

$$P[X \geq 1] = 1 - P[X < 1] = 1 - P[X = 0] = 1 - \frac{e^{-4,5} 4,5^0}{0!} = 1 - 0,0111 = 0,9889$$



Distribución de Poisson como límite de la distribución Binomial.

Cuando en la distribución $\mathcal{B}(n, p)$ p es muy pequeño y n suficientemente grande, se puede aproximar por la distribución de Poisson $\mathcal{P}(np)$ para el cálculo de las probabilidades asociadas. Esta propiedad sirve para caracterizar a una gran cantidad de fenómenos que siguen una distribución de Poisson, son fenómenos con una pequeña probabilidad de éxito que son observados en un número elevado de ocasiones.

► EJEMPLO 7.4

La probabilidad de que aparezca una pieza defectuosa en un proceso es 0,0001. La producción de un año es de 36000 piezas. Calcular la probabilidad de que en la producción anual el número de piezas defectuosas sea por lo menos dos.

Solución:

X = número de piezas defectuosas en la producción anual

$$X \sim \mathcal{B}(36000, 0,0001) \rightarrow \mathcal{P}(36000 \times 0,0001) = \mathcal{P}(3,6)$$

$$P[X \geq 2] = 1 - P[X < 2] = 1 - P[X = 0] - P[X = 1] = 1 - \frac{e^{-3,6} 3,6^0}{0!} - \frac{e^{-3,6} 3,6^1}{1!} = 1 - 0,0273 - 0,0984 = 0,8743$$

Se podría haber resuelto también usando la distribución Binomial:

$$\begin{aligned} P[X \geq 2] &= 1 - P[X < 2] = 1 - P[X = 0] - P[X = 1] = \\ &= 1 - \binom{36000}{0} 0,0001^0 0,9999^{36000} - \binom{36000}{1} 0,0001^1 0,9999^{35999} = 0,874323663 \end{aligned}$$

Pero es más complicado e incluso a veces imposible realizar los cálculos debido a la magnitud de n y p . ◀

7.4 Distribución Hipergeométrica.

Se selecciona una muestra (subconjunto) de una población finita con N elementos donde los individuos están clasificados en dos categorías (*éxito y fracaso*). Se puede hacer básicamente de dos formas:

- *Muestreo con reemplazamiento*: los individuos se extraen, observan y devuelven a la población, de forma que la composición de la población es constante en cada extracción. Para calcular la probabilidad de que cierto número de individuos de la muestra presentan la característica *éxito* se utiliza la distribución *Binomial*.
- *Muestreo sin reemplazamiento*: los individuos extraídos se dejan fuera de la población con lo que la composición de la población va cambiando con cada extracción (es equivalente extraer de uno en uno o en conjunto). Para calcular la probabilidad de que cierto número de individuos de la muestra presentan la característica *éxito* se utiliza la distribución *Hipergeométrica* que vamos a estudiar a continuación.

Sea una población finita con N individuos de los cuales Np individuos presentan la característica *éxito* y $N(1-p) = Nq$ individuos no la presentan. La probabilidad de obtener *éxito* en la primera

extracción es $\frac{Np}{N} = p$ y la de no obtenerlo (u obtener *fracaso*) es q . En las sucesivas extracciones estas probabilidades de obtener *éxito* o *fracaso* van cambiando al modificarse la composición de la población.

La variable aleatoria X =*número de éxitos en una muestra sin reemplazamiento de tamaño n* sigue una **distribución de probabilidad Hipergeométrica**, abreviadamente $X \sim \mathcal{H}(N, n, p)$, cuyas probabilidades asociadas son:

$$p_x = P[X = x] = \frac{\binom{Np}{x} \binom{Nq}{n-x}}{\binom{N}{n}}$$

La variable aleatoria X puede tomar todos los valores enteros comprendidos entre $\max\{0, n-Nq\}$ y $\min\{n, Np\}$.

Cuando N es infinito (en la práctica un valor muy elevado) las distribuciones *Hipergeométrica* y *Binomial* coinciden.

Media:

$$E[X] = np$$

Varianza:

$$\sigma^2[X] = \frac{N-n}{N-1} npq$$

► EJEMPLO 7.5

Una urna contiene 15 bolas blancas y 5 bolas negras. Se extraen 5 bolas, ¿cuál es la probabilidad de que 2 sean negras?

- Se extraen las bolas una a una, se observa el color y se devuelven a la urna antes de la siguiente extracción.
- Se extraen las 5 bolas de una vez.

Solución:

X = número de bolas negras entre las 5 extraídas

$$\text{a) } n = 5 \quad p = \frac{5}{20} \quad X \sim \mathcal{B}\left(5, \frac{5}{20}\right) = \mathcal{B}(5, 0,25)$$

$$P[X = 2] = \binom{5}{2} 0,25^2 (1-0,25)^3 = 0,26367$$

$$\text{b) } N = 20 \quad n = 5 \quad p = \frac{5}{20} = 0,25 \quad X \sim \mathcal{H}(20, 5, 0,25)$$

$$Np = 20 \times 0,25 = 5 \quad Nq = 20 \times (1 - 0,25) = 15$$

$$P[X = 2] = \frac{\binom{5}{2} \binom{15}{3}}{\binom{20}{5}} = 0,29347$$

► EJEMPLO 7.6

Si jugamos una combinación de la lotería primitiva. ¿Cuál es la probabilidad de que acertemos 4 números o más?

Solución:

En el sorteo de la lotería primitiva las extracciones se hacen sin reemplazamiento.

X = número de bolas extraídas iguales a alguno de nuestros 6 números.

$$N = 49 \quad n = 6 \quad p = \frac{6}{49} \quad X \sim \mathcal{H}(49, 6, \frac{6}{49})$$

$$Np = 49 \frac{6}{49} = 6 \quad Nq = N(1 - p) = N - Np = 49 - 6 = 43$$

$$\begin{aligned} P[X \geq 4] &= P[X = 4] + P[X = 5] + P[X = 6] = \frac{\binom{6}{4} \binom{43}{2}}{\binom{49}{6}} + \frac{\binom{6}{5} \binom{43}{1}}{\binom{49}{6}} + \frac{\binom{6}{6} \binom{43}{0}}{\binom{49}{6}} = \\ &= 0,00096862 + 0,00001845 + 0,000000071 = 0,000987141 \end{aligned}$$

7.5 Distribución Geométrica.

Sea X la variable aleatoria que indica el *número de fracasos* antes de obtener el *primer éxito* en repeticiones idénticas e independientes de un experimento. La variable aleatoria X sigue una **distribución Geométrica de parámetro p** , abreviadamente $X \sim \mathcal{G}(p)$, donde p es la probabilidad de éxito. X puede tomar todos los valores enteros no negativos $(0, 1, 2, \dots)$ con probabilidades:

$$p_x = P[X = x] = (1 - p)^x p = q^x p \quad x = 0, 1, 2, \dots$$

Si llamamos A al suceso *éxito*:

$$p_x = P[X = x] = P\left[\underbrace{\overline{A} \dots \overline{A}}_x A\right] = \underbrace{P[\overline{A}] \dots P[\overline{A}]}_x P[A] = \underbrace{q \dots q}_x p = q^x p$$

Dado que p y q son positivos lo serán las probabilidades $p_x = P[X = x]$. Y la suma de todas ellas es 1:

$$\begin{aligned} P[X=0] + P[X=1] + P[X=2] + \dots + P[X=x] + \dots = \\ = p + pq + pq^2 + \dots + pq^x + \dots = p(1 + q + q^2 + \dots) = p \frac{1}{1-q} = p \frac{1}{p} = 1 \end{aligned}$$

Donde se ha utilizado que la suma de una progresión geométrica de razón $r < 1$ es igual a $\frac{a_1}{1-r}$ (a_1 = primer término de la progresión geométrica).

Media:

$$E[X] = \frac{1-p}{p} = \frac{q}{p}$$

Varianza:

$$\sigma^2[X] = \frac{1-p}{p^2} = \frac{q}{p^2}$$

► EJEMPLO 7.7

En una Facultad suelen aprobar el 70% de los alumnos. Hallar la probabilidad de que un alumno apruebe en alguna de las 6 convocatorias de examen que dispone.

Solución:

X = número de suspensos antes de aprobar sigue una distribución Geométrica de parámetro $p=0,70$.
 $p=0,70$ $q=1-0,70=0,30$

$$\begin{aligned} P[X \leq 5] &= P[X=0] + P[X=1] + P[X=2] + P[X=3] + P[X=4] + P[X=5] = \\ &= (0,3^0 \times 0,7) + (0,3 \times 0,7) + (0,3^2 \times 0,7) + (0,3^3 \times 0,7) + (0,3^4 \times 0,7) + (0,3^5 \times 0,7) = \\ &= 0,7 + 0,21 + 0,063 + 0,0189 + 0,00567 + 0,001701 = 0,999271 \end{aligned}$$



7.6 Ejercicios resueltos.

1. Una ciudad está dividida en dos barrios (A y B). El número de apagones que se producen en el barrio A sigue una distribución de Poisson con media dos apagones por mes. Siendo en el barrio B la media de un apagón por mes. Calcule la probabilidad de que en un mes:
- Haya algún apagón en el barrio A.
 - Hayan, a lo sumo, dos apagones en la ciudad.

Solución:

$X = \text{n}^\circ \text{ de apagones por mes en el barrio A} \sim \mathcal{P}(2)$

$Y = \text{n}^\circ \text{ de apagones por mes en el barrio B} \sim \mathcal{P}(1)$

$Z = X + Y = \text{n}^\circ \text{ de apagones por mes en la ciudad} \sim \mathcal{P}(3) = \mathcal{P}(2+1)$

- $P[X \geq 1] = 1 - P[X = 0] = 1 - \frac{e^{-2} 2^0}{0!} = 1 - e^{-2} = 1 - 0,1353 = 0,8647$
- $P[Z \leq 2] = P[Z = 0] + P[Z = 1] + P[Z = 2] = \frac{e^{-3} 3^0}{0!} + \frac{e^{-3} 3^1}{1!} + \frac{e^{-3} 3^2}{2!} =$
 $= 0,0498 + 0,1494 + 0,2240 = 0,4232$

2. Una compañía de seguros tiene contratadas anualmente 10000 pólizas contra incendios:
- Sabiendo que hay una probabilidad $1/5000$ de que se produzca un incendio en una vivienda, calcule la probabilidad de que tenga que pagar a alguno de sus asegurados.
 - ¿Cuántos incendios, por término medio, hay que atender en un año?
 - La compañía, en una campaña de publicidad, decide cobrar sólo la mitad a 1000 de sus asegurados. En un edificio en el que 20 de las viviendas tienen contratada una póliza con esta compañía, calcúlese la probabilidad de que alguno de los asegurados de este edificio se vea agraciado con el descuento. Calcúlese también la probabilidad de que todos los asegurados en ese edificio se vean agraciados por la reducción en el precio de la póliza.

Solución:

a) $X = \text{número de incendios} \sim \mathcal{B}\left(10000, \frac{1}{5000}\right) \approx \mathcal{P}\left(\frac{10000}{5000}\right) = \mathcal{P}(2)$

Cuando en una distribución Binomial n es muy grande y p muy pequeño, ésta se puede aproximar por una distribución de Poisson con $\lambda = np$.

Vamos a resolverlo con ambas distribuciones y de paso comprobaremos la aproximación de una por otra.

$$P[X \geq 1] = 1 - P[X = 0] = 1 - \binom{10000}{0} \left(\frac{1}{5000}\right)^0 \left(\frac{4999}{5000}\right)^{10000} = 1 - 0,135308 = 0,864692$$

$$P[X \geq 1] = 1 - P[X = 0] = 1 - \frac{e^{-2} 2^0}{0!} = 1 - 0,135335 = 0,864665$$

$$b) E[X] = np = 10000 \frac{1}{5000} = 2$$

$$c) p = \text{probabilidad de cobrar la mitad} = \frac{1000}{10000} = 0,1.$$

$Y = \text{número de asegurados que pagan la mitad en dicho edificio} \sim \mathcal{B}(20; 0,1).$

$$P[Y \geq 1] = 1 - P[Y = 0] = 1 - \binom{20}{0} 0,1^0 0,9^{20} = 1 - 0,121577 = 0,878423$$

$$P[Y = 20] = \binom{20}{20} 0,1^{20} 0,9^0 = 10^{-20} = \frac{1}{10^{20}} \approx 0$$

3. La producción de una factoría se realiza en dos máquinas A y B. La probabilidad de que la máquina A produzca x piezas defectuosas al día viene dada por la expresión

$$e^{-1} \frac{1^x}{x!}$$

La probabilidad de que la máquina B produzca y piezas defectuosas al día viene dada por la expresión

$$e^{-3} \frac{3^y}{y!}$$

Se pide la probabilidad de que:

- En un día no haya piezas defectuosas en la factoría.
- De cinco días, elegidos al azar, en dos de ellos no se produzcan piezas defectuosas en la factoría.

Solución:

$$a) X = \text{número de piezas defectuosas producidas por la máquina A} \sim \mathcal{P}(1)$$

$$Y = \text{número de piezas defectuosas producidas por la máquina B} \sim \mathcal{P}(3)$$

$$Z = X + Y = \text{nº de piezas defectuosas en la factoría (producidas por A o B)} \sim \mathcal{P}(1+3) = \mathcal{P}(4)$$

$$P[Z = 0] = \frac{e^{-4} 4^0}{4!} = e^{-4} = 0,018316$$

$$b) V = \text{número de días sin piezas defectuosas} \sim \mathcal{B}(5; 0,018316)$$

$$P[V=2] = \binom{5}{2} 0,018316^2 (1-0,018316)^{5-2} = \frac{5!}{2!3!} 0,018316^2 0,981684^3 = 0,003174$$

4. Un concesionario de automóviles vende a sus clientes vehículos de la misma marca. Sabiendo que la probabilidad de que este tipo de vehículos esté en servicio cinco años después es de 0,9 , determinar la probabilidad de que:

- De los cinco vehículos comprados por una empresa, al menos cuatro de ellos estén en servicio dentro de cinco años.
- De los cuarenta vehículos comprados por una empresa, al menos 38 de ellos estén en servicio dentro de cinco años.

Solución:

p =probabilidad de que un vehículo esté en servicio cinco años después=0,9

a) X =número de vehículos en servicio (de los cinco) $\sim \mathcal{B}(5; 0,9)$

$$\begin{aligned} P[X \geq 4] &= P[X=4] + P[X=5] = \binom{5}{4} 0,9^4 (1-0,9)^{5-4} + \binom{5}{5} 0,9^5 (1-0,9)^{5-5} = \\ &= 0,32805 + 0,59049 = 0,91854 \end{aligned}$$

b) Y =número de vehículos en servicio (de los cuarenta) $\sim \mathcal{B}(40; 0,9)$

$$\begin{aligned} P[Y \geq 38] &= P[Y=38] + P[Y=39] + P[Y=40] = \\ &= \binom{40}{38} 0,9^{38} 0,1^2 + \binom{40}{39} 0,9^{39} 0,1^1 + \binom{40}{40} 0,9^{40} 0,1^0 = \\ &= 0,142334 + 0,065693 + 0,014781 = 0,222808 \end{aligned}$$

5. La probabilidad de que un individuo presente una cierta característica económica es 1/20.

- Tomando una muestra de 8 individuos al azar, calcule la probabilidad de que 3 de ellos presenten la característica.
- Tomando una muestra de 120 individuos al azar, ¿cuál es la probabilidad de que a lo sumo 2 presenten la característica?

Solución:

a) X =número de individuos, entre los 8, con la característica económica $\sim \mathcal{B}\left(8, \frac{1}{20}\right)$

$$P[X=3] = \binom{8}{3} \left(\frac{1}{20}\right)^3 \left(\frac{19}{20}\right)^{8-3} = 0,005416$$

b) $Y = \text{número de individuos, entre los 120, con la característica económica} \sim \mathcal{B}\left(120, \frac{1}{20}\right)$

$$\begin{aligned} P[Y \leq 2] &= P[Y = 0] + P[Y = 1] + P[Y = 2] = \\ &= \binom{120}{0} \left(\frac{1}{20}\right)^0 \left(\frac{19}{20}\right)^{120} + \binom{120}{1} \left(\frac{1}{20}\right)^1 \left(\frac{19}{20}\right)^{119} + \binom{120}{2} \left(\frac{1}{20}\right)^2 \left(\frac{19}{20}\right)^{118} = \\ &= 0,002122 + 0,013405 + 0,041978 = 0,057505 \end{aligned}$$

6. Una entidad realiza anualmente tres mil préstamos hipotecarios. La probabilidad de impago es 0,002. Se pide:

- La probabilidad de que en un año se produzcan más de dos impagos.
- La probabilidad de que en tres años, en uno de ellos no se produzcan impagos.
- ¿Cuántos impagos se esperan en cinco años?

Solución:

a) $X = \text{número de impagos en un año} \sim \mathcal{B}(3000; 0,002) \approx \mathcal{P}(3000 \times 0,002) = \mathcal{P}(6)$

Cuando en una distribución Binomial n es muy grande y p muy pequeño, ésta se puede aproximar por una distribución de Poisson con $\lambda = np$.

Vamos a resolverlo con ambas distribuciones y de paso comprobaremos la aproximación de una por otra.

$$\begin{aligned} P[X > 2] &= 1 - P[X \leq 2] = 1 - P[X = 0] - P[X = 1] - P[X = 2] = \\ &= 1 - \binom{3000}{0} 0,002^0 0,998^{3000} - \binom{3000}{1} 0,002^1 0,998^{2999} - \binom{3000}{2} 0,002^2 0,998^{2998} = \\ &= 1 - 0,002464 - 0,014813 - 0,044513 = 0,93821 \end{aligned}$$

$$\begin{aligned} P[X > 2] &= 1 - P[X \leq 2] = 1 - P[X = 0] - P[X = 1] - P[X = 2] = \\ &= 1 - \frac{e^{-6} 6^0}{0!} - \frac{e^{-6} 6^1}{1!} - \frac{e^{-6} 6^2}{2!} = 1 - 0,002479 - 0,014873 - 0,044618 = 0,93803 \end{aligned}$$

b) $Y = \text{número de años sin impagos} \sim \mathcal{B}(3; 0,002464)$

$p = \text{probabilidad de un años sin impagos} = P[X = 0] = 0,002464$

$$P[Y = 1] = \binom{3}{1} 0,002464^1 (1 - 0,002464)^{3-1} = 0,007356$$

c) $X_i = \text{número de impagos en el año } i \sim \mathcal{B}(3000; 0,002) \approx \mathcal{P}(6)$

Por la propiedad de aditividad de la distribución Binomial y de la distribución de Poisson:

$$Z = X_1 + X_2 + X_3 + X_4 + X_5 \sim \mathcal{B}(3000; 0,002) + \mathcal{B}(3000; 0,002) + \mathcal{B}(3000; 0,002) + \mathcal{B}(3000; 0,002) + \mathcal{B}(3000; 0,002) = \\ = \mathcal{B}(5 \times 3000; 0,002) = \mathcal{B}(15000; 0,002)$$

O bien:

$$Z = X_1 + X_2 + X_3 + X_4 + X_5 \sim \mathcal{P}(6) + \mathcal{P}(6) + \mathcal{P}(6) + \mathcal{P}(6) + \mathcal{P}(6) = \mathcal{P}(5 \times 6) = \mathcal{P}(30)$$

$$\text{En ambos casos: } E[Z] = np = 15000 \times 0,002 = \lambda = 30.$$

$$\text{O sencillamente } Z = 5X \Rightarrow E[Z] = 5E[X] = 5 \times np = 5 \times 3000 \times 0,002 = 30.$$

7. Diariamente se empaquetan dos mil quinientas unidades de un determinado producto. La probabilidad de que una de las unidades sea defectuosa es 0,001. Se pide:

- La probabilidad de que en un día aparezcan más de dos artículos defectuosos.
- ¿Cuántos artículos defectuosos se espera que aparezcan en diez días de empaquetado?

Solución:

$$\text{a) } X = \text{número de unidades defectuosas diariamente} \sim \mathcal{B}(2500; 0,001) \approx \mathcal{P}(2,5)$$

Cuando en una distribución Binomial n es muy grande y p muy pequeño, ésta se puede aproximar por una distribución de Poisson con $\lambda = np$.

Vamos a resolverlo con ambas distribuciones y de paso comprobaremos la aproximación de una por otra.

$$\begin{aligned} P[X > 2] &= 1 - P[X \leq 2] = 1 - P[X = 0] - P[X = 1] - P[X = 2] = \\ &= 1 - \binom{2500}{0} 0,001^0 0,999^{2500} - \binom{2500}{1} 0,001^1 0,999^{2499} - \binom{2500}{2} 0,001^2 0,999^{2498} = \\ &= 1 - 0,081982 - 0,205161 - 0,256605 = 0,456252 \end{aligned}$$

$$\begin{aligned} P[X > 2] &= 1 - P[X \leq 2] = 1 - P[X = 0] - P[X = 1] - P[X = 2] = \\ &= 1 - \frac{e^{-2,5} 2,5^0}{0!} - \frac{e^{-2,5} 2,5^1}{1!} - \frac{e^{-2,5} 2,5^2}{2!} = 1 - 0,082085 - 0,205212 - 0,256516 = 0,456187 \end{aligned}$$

$$\text{b) } X_i = \text{número de unidades defectuosas en el día } i \sim \mathcal{B}(2500; 0,001) \approx \mathcal{P}(2,5)$$

$Y = \text{número de unidades defectuosas en 10 días}$

Por la propiedad de aditividad de la distribución Binomial y de la distribución de Poisson:

$$Y = X_1 + \dots + X_{10} \sim \mathcal{B}(2500; 0,001) + \dots + \mathcal{B}(2500; 0,001) = \mathcal{B}(10 \times 2500; 0,001) = \mathcal{B}(25000; 0,001)$$

O bien:

$$Y = X_1 + \dots + X_{10} \sim \mathcal{P}(2,5) + \dots + \mathcal{P}(2,5) = \mathcal{P}(10 \times 2,5) = \mathcal{P}(25)$$

En ambos casos: $E[Y] = np = 25000 \times 0,001 = \lambda = 25$.

O sencillamente $Y=10X \Rightarrow E[Y] = 10E[X] = 10 \times np = 10 \times 2500 \times 0,001 = 25$.

8. El número de accidentes por año en una obra sigue una distribución de probabilidad con media y varianza iguales a 6.

- Sin suponer ningún modelo para la anterior variable aleatoria, ¿tiene alguna idea sobre la probabilidad de que en un año el número de accidentes esté comprendido entre 2 y 10?
- Suponiendo que el número de accidentes sigue una distribución de Poisson, ¿cuál es la probabilidad del anterior número de accidentes? ¿Cuál es la probabilidad de que haya no más de 6 y no menos de 4 accidentes en un año?

Solución:

- Según la desigualdad de Tchebycheff:

$$P[|X - E[X]| < k\sigma] = P[E[X] - k\sigma < X < E[X] + k\sigma] \geq 1 - \frac{1}{k^2}$$

$$P[2 < X < 10] = P[|X - 6| < 4] \geq 1 - \frac{1}{k^2} = 0,625$$

$$4 = k\sigma = k\sqrt{6} \Leftrightarrow k = \frac{4}{\sqrt{6}} \Leftrightarrow k^2 = \frac{16}{6} \Leftrightarrow 1 - \frac{1}{k^2} = 1 - \frac{6}{16} = \frac{10}{16} = 0,625$$

Si consideramos que los valores 2 y 10 están incluidos, vale la misma desigualdad:

$$P[2 \leq X \leq 10] \geq P[2 < X < 10] \geq 0,625$$

- $X = \text{número de accidentes/año} \sim \mathcal{P}(6)$

$$\begin{aligned} P[2 \leq X \leq 10] &= P[X=2] + P[X=3] + P[X=4] + P[X=5] + P[X=6] + P[X=7] + \\ &+ P[X=8] + P[X=9] + P[X=10] = e^{-6} \frac{6^2}{2!} + e^{-6} \frac{6^3}{3!} + e^{-6} \frac{6^4}{4!} + e^{-6} \frac{6^5}{5!} + e^{-6} \frac{6^6}{6!} + e^{-6} \frac{6^7}{7!} + \\ &+ e^{-6} \frac{6^8}{8!} + e^{-6} \frac{6^9}{9!} + e^{-6} \frac{6^{10}}{10!} = 0,044618 + 0,089235 + 0,133853 + 0,160623 + 0,160623 + \\ &+ 0,137677 + 0,103258 + 0,068838 + 0,041303 = 0,940028 \end{aligned}$$

$$P[4 \leq X \leq 6] = P[X=4] + P[X=5] + P[X=6] = 0,133853 + 0,160623 + 0,160623 = 0,455099$$

9. Una empresa aseguradora ha estimado que la probabilidad de que un cliente realice una reclamación es del 2 por mil.

- Una cartera en concreto está compuesta por 9 asegurados. Calcular la probabilidad de que al menos tres asegurados demanden reclamaciones.

- b) Si en la provincia de Granada esta compañía dispone de 1500 asegurados, halle la probabilidad de que realicen reclamaciones más de 4 personas.

Solución:

a) $X = \text{número de reclamaciones en 9 asegurados} \sim \mathcal{B}\left(9, \frac{2}{1000}\right)$

$$\begin{aligned} P[X \geq 3] &= 1 - P[X = 0] - P[X = 1] - P[X = 2] = \\ &= 1 - \binom{9}{0} 0,002^0 0,998^9 - \binom{9}{1} 0,002^1 0,998^8 - \binom{9}{2} 0,002^2 0,998^7 = \\ &= 1 - 0,982143 - 0,017714 - 0,000142 = 0,000001 \end{aligned}$$

b) $Y = \text{número de reclamaciones en 1500 asegurados} \sim \mathcal{B}(1500; 0,002) \approx \mathcal{P}(3)$

Cuando en una distribución Binomial n es muy grande y p muy pequeño, ésta se puede aproximar por una distribución de Poisson con $\lambda = np$.

Vamos a resolverlo con ambas distribuciones y de paso comprobaremos la aproximación de una por otra.

$$\begin{aligned} P[X > 4] &= 1 - P[X \leq 4] = 1 - P[X = 0] - P[X = 1] - P[X = 2] - P[X = 3] - P[X = 4] = \\ &= 1 - \binom{1500}{0} 0,002^0 0,998^{1500} - \binom{1500}{1} 0,002^1 0,998^{1499} - \binom{1500}{2} 0,002^2 0,998^{1498} - \\ &\quad - \binom{1500}{3} 0,002^3 0,998^{1497} - \binom{1500}{4} 0,002^4 0,998^{1496} = \\ &= 1 - 0,049638 - 0,149212 - 0,224116 - 0,224266 - 0,1682 = 0,184568 \end{aligned}$$

$$\begin{aligned} P[X > 4] &= 1 - P[X \leq 4] = 1 - P[X = 0] - P[X = 1] - P[X = 2] - P[X = 3] - P[X = 4] = \\ &= 1 - \frac{e^{-3} 3^0}{0!} - \frac{e^{-3} 3^1}{1!} - \frac{e^{-3} 3^2}{2!} - \frac{e^{-3} 3^3}{3!} - \frac{e^{-3} 3^4}{4!} = \\ &= 1 - 0,049787 - 0,149361 - 0,224042 - 0,224042 - 0,168031 = 0,184737 \end{aligned}$$

10. En un almacén hay 5000 piezas de las cuales 400 son defectuosas. Se seleccionan aleatoriamente 10 piezas, en una ocasión con reemplazamiento y en otra sin reemplazamiento. En cada caso determine la probabilidad de que todas las piezas estén en buen estado.

Solución:

$P[\text{pieza defectuosa}] = 400/5000 = 0,08$

$X = \text{número de piezas defectuosas entre 10.}$

Con reemplazamiento:

$$X \sim \mathcal{B}(10; 0,08)$$

$$P[X = 0] = \binom{10}{0} 0,08^0 0,92^{10} = 0,92^{10} = 0,434388$$

Sin reemplazamiento:

$$X \sim \mathcal{H}(5000; 10; 0,08)$$

$$\begin{aligned} P[X = 0] &= \frac{\binom{400}{0} \binom{4600}{10}}{\binom{5000}{10}} = \frac{400!}{0!400!} \frac{4600!}{10!4590!} = \\ &= \frac{4600 \times 4599 \times 4598 \times 4597 \times 4596 \times 4595 \times 4594 \times 4593 \times 4592 \times 4591}{5000 \times 4999 \times 4998 \times 4997 \times 4996 \times 4995 \times 4994 \times 4993 \times 4992 \times 4991} = 0,434048 \end{aligned}$$

Como puede observarse, cuando el tamaño de la población es grande ($N=5000$), no hay mucha diferencia entre hacer la selección de la muestra con o sin reemplazamiento.

11. Una empresa recibe piezas de un proveedor en lotes de 2000 que se someten al siguiente control de calidad: se toman 20 piezas al azar y si hay más de una defectuosa se rechaza el lote, en caso contrario se acepta. La calidad garantizada por el proveedor es de un 0,8% de piezas defectuosas. Calcule la probabilidad de:

- a) Aceptar un lote que contenga un 2% de defectuosas.
- b) Rechazar un lote que debiera ser aceptado por tener sólo un 0,8% de defectuosas.

Solución:

X =número de piezas defectuosas entre las 20 seleccionadas.

- a) $P[\text{pieza defectuosa}] = 0,02$

$$X \sim \mathcal{B}(20; 0,02)$$

$$\begin{aligned} P[\text{aceptar el lote}] &= P[X \leq 1] = \binom{20}{0} 0,02^0 0,98^{20} + \binom{20}{1} 0,02^1 0,98^{19} = \\ &= 0,667608 + 0,272493 = 0,940101 \end{aligned}$$

- b) $P[\text{pieza defectuosa}] = 0,008$

$$X \sim \mathcal{B}(20; 0,008)$$

$$\begin{aligned} P[\text{rechazar el lote}] &= P[X > 1] = 1 - P[X \leq 1] = 1 - \binom{20}{0} 0,008^0 0,992^{20} - \binom{20}{1} 0,008^1 0,992^{19} = \\ &= 1 - 0,851596 - 0,137354 = 0,01105 \end{aligned}$$

12. Un tirador dispara contra una diana circular. La distancia, en centímetros, de los impactos al centro de la diana es una variable aleatoria con función de densidad

$$f(x) = \begin{cases} \frac{10-x}{50} & 0 \leq x \leq 10 \\ 0 & \text{en otro caso} \end{cases}$$

Obtiene premio cuando la distancia del impacto al centro no es superior a 2,5 cm. ¿Cuál es la probabilidad de a lo sumo 2 premios en cuatro disparos?

Solución:

$$P[\text{premio}] = P[0 < X \leq 2,5] = \int_0^{2,5} f(x) dx = \int_0^{2,5} \frac{10-x}{50} dx = \frac{1}{50} \left[10x - \frac{x^2}{2} \right]_0^{2,5} = \frac{1}{50} (25 - 3,125) = 0,4375$$

$Y = \text{número de premios en 4 disparos. } Y \sim \mathcal{B}(4; 0,4375)$

$$P[Y \leq 2] = \binom{4}{0} 0,4375^0 0,5625^{20} + \binom{4}{1} 0,4375^1 0,5625^{19} + \binom{4}{2} 0,4375^2 0,5625^{18} = \\ = 0,100113 + 0,311462 + 0,363373 = 0,774948$$

13. El número de personas que esperan ser atendidas en una oficina sigue una distribución de Poisson con media 5. ¿Qué número de sillas debe haber en la sala de espera para que todas puedan estar sentadas con un probabilidad del 95%?

Solución:

$X = \text{número de personas que esperan ser atendidas. } X \sim \mathcal{P}(5).$

Sea s el número de sillas en la sala de espera.

$$P[X \leq s] = 0,95$$

Utilizando las tablas de la distribución de Poisson o con Excel, obtenemos que:

$$P[X \leq 8] = 0,9319 \quad P[X \leq 9] = 0,9682$$

Para que estén sentados todos con una probabilidad mayor del 95%, el número de sillas debe ser 9 como mínimo.

14. Un proceso de fabricación produce por término medio 3 piezas defectuosas por cada 1000. Las piezas se empaquetan en cajas de 2500. ¿Cuál es la probabilidad de que entre 5 cajas, elegidas al azar, en dos de ellas no haya piezas defectuosas?

Solución:

$$P[\text{pieza defectuosa}] = 0,003$$

X =número de piezas defectuosas por caja. $X \sim \mathcal{B}(2500; 0,003) \cong \mathcal{P}(7,5)$.

Veamos que efectivamente para esos valores de n y p la distribución Binomial se aproxima bien mediante una distribución de Poisson con $\lambda = np = 7,5$.

$$P[X = 0] = \binom{2500}{0} 0,003^0 0,997^{2500} = 0,000547$$

$$P[X = 0] = \frac{e^{-7,5} 7,5^0}{0!} = 0,000553$$

Y =número de cajas sin piezas defectuosas. $Y \sim \mathcal{B}(5; 0,000547)$

$$q = 1 - 0,000547 = 0,999453$$

$$P[Y = 2] = \binom{5}{2} 0,000547^2 0,999453^3 = 0,000002987$$