Apuntes de Estadística

David Ruiz Muñoz

Ana María Sánchez Sánchez

ISBN: xxxxxxxxx

ÍNDICE

Capítulo I	HISTORIA DE LA ESTADISTICA

Como dijera Huntsberger: "La palabra estadística a menudo nos trae a la mente imágenes de números apilados en grandes arreglos y tablas, de volúmenes de cifras relativas a nacimientos, muertes, impuestos, poblaciones, ingresos, deudas, créditos y así sucesivamente. Huntsberger tiene razón pues al instante de escuchar esta palabra estas son las imágenes que llegan a nuestra cabeza.
La Estadística es mucho más que sólo números apilados y gráficas bonitas. Es una ciencia con tanta antigüedad como la escritura, y es por sí misma auxiliar de todas las demás ciencias. Los mercados, la medicina, la ingeniería, los gobiernos, etc. Se nombran entre los más destacados clientes de ésta.
La ausencia de ésta conllevaría a un caos generalizado, dejando a los administradores y ejecutivos sin información vital a la hora de tomar decisiones en tiempos de incertidumbre.
La Estadística que conocemos hoy en día debe gran parte de su realización a los trabajos matemáticos de aquellos hombres que desarrollaron la teoría de las probabilidades, con la cual se adhirió a la Estadística a las ciencias formales.
En este breve material se expone los conceptos, la historia, la división así como algunos errores básicos cometidos al momento de analizar datos Estadísticos.

Definición de Estadística

La Estadística es la ciencia cuyo objetivo es reunir una información cuantitativa concerniente a individuos, grupos, series de hechos, etc. y deducir de ello gracias al análisis de estos datos unos significados precisos o unas previsiones para el futuro.
La estadística, en general, es la ciencia que trata de la recopilación, organización presentación, análisis e interpretación de datos numéricos con e fin de realizar una toma de decisión más efectiva.
Otros autores tienen definiciones de la Estadística semejantes a las anteriores, y algunos otros no tan semejantes. Para Chacón esta se define como "la ciencia que tiene por objeto el estudio cuantitativo de los colectivos"; otros la definen como la expresión cuantitativa del conocimiento dispuesta en forma adecuada para el escrutinio y análisis. La más aceptada, sin embargo, es la de Minguez, que define la Estadística como "La ciencia que tiene por objeto aplicar las leyes de la cantidad a los hechos sociales para medir su intensidad, deducir las leyes que los rigen y hacer su predicción próxima".
Los estudiantes confunden comúnmente los demás términos asociados con las Estadísticas, una confusión que es conveniente aclarar debido a que esta palabra tiene tres significados: la palabra estadística, en primer término se usa para referirse a la información estadística; también se utiliza para referirse al conjunto de técnicas y métodos que se utilizan para analizar la información estadística; y el término estadístico, en singular y en masculino, se refiere a una medida derivada de una muestra.

Utilidad e Importancia

Los métodos estadísticos tradicionalmente se utilizan para propósitos descriptivos, para organizar y resumir datos numéricos. La estadística descriptiva, por ejemplo trata de la tabulación de datos, su presentación en forma gráfica o ilustrativa y el cálculo de medidas descriptivas.
Ahora bien, las técnicas estadísticas se aplican de manera amplia en mercadotecnia, contabilidad, control de calidad y en otras actividades; estudios de consumidores; análisis de resultados en deportes; administradores de instituciones; en la educación; organismos políticos; médicos; y por otras personas que intervienen en la toma de decisiones.

Historia de la Estadística

Los comienzos de la estadística pueden ser hallados en el antiguo Egipto, cuyos faraones lograron recopilar, hacia el año 3050 antes de Cristo, prolijos datos relativos a la población y la riqueza del país. De acuerdo al historiador griego Heródoto, dicho registro de riqueza y población se hizo con el objetivo de preparar la construcción de las pirámides. En el mismo Egipto, Ramsés II hizo un censo de las tierras con el objeto de verificar un nuevo reparto.

En el antiguo Israel la Biblia da referencias, en el libro de los Números, de los datos estadísticos obtenidos en dos recuentos de la población hebrea. El rey David por otra parte, ordenó a Joab, general del ejército hacer un censo de Israel con la finalidad de conocer el número de la población.

También los chinos efectuaron censos hace más de cuarenta siglos. Los griegos efectuaron censos periódicamente con fines tributarios, sociales (división de tierras) y militares (cálculo de recursos y hombres disponibles). La investigación histórica revela que se realizaron 69 censos para calcular los impuestos, determinar los derechos de voto y ponderar la potencia guerrera.

Pero fueron los romanos, maestros de la organización política, quienes mejor supieron emplear los recursos de la estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio.

Durante los mil años siguientes a la caída del imperio Romano se realizaron muy pocas operaciones Estadísticas, con la notable excepción de las relaciones de tierras pertenecientes a la Iglesia, compiladas por Pipino el Breve en el 758 y por Carlomagno en el 762 DC. Durante el siglo IX se realizaron en Francia algunos censos parciales de siervos. En Inglaterra, Guillermo el Conquistador recopiló el Domesday Book o libro del Gran Catastro para el año 1086, un documento de la propiedad, extensión y valor de las tierras de Inglaterra. Esa obra fue el primer compendio estadístico de Inglaterra.

Aunque Carlomagno, en Francia; y Guillermo el Conquistador, en Inglaterra, trataron de revivir la técnica romana, los métodos estadísticos permanecieron casi olvidados durante la Edad Media.

Durante los siglos XV, XVI, y XVII, hombres como Leonardo de Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, hicieron grandes operaciones al método científico, de tal forma que cuando se crearon los Estados Nacionales y surgió como fuerza el comercio internacional existía ya un método capaz de aplicarse a los datos económicos.

Para el año 1532 empezaron a registrarse en Inglaterra las defunciones debido al temor que Enrique VII tenía por la peste. Más o menos por la misma época, en Francia la ley exigió a los clérigos registrar los bautismos, fallecimientos y matrimonios. Durante un brote de peste que apareció a fines de la década de 1500, el gobierno inglés
comenzó a publicar estadística semanales de los decesos. Esa costumbre continuó muchos años, y en 1632 estos Bills of Mortality (Cuentas de Mortalidad) contenían los nacimientos y fallecimientos por sexo. En 1662, el capitán John Graunt usó documentos que abarcaban treinta años y efectuó predicciones sobre el número de personas que morirían de varias enfermedades y sobre las proporciones de nacimientos de varones y mujeres que cabría esperar. El trabajo de Graunt, condensado en su obra Natural and Political Observations...Made upon the Bills of Mortality (Observaciones Políticas y Naturales ... Hechas a partir de las Cuentas de Mortalidad), fue un esfuerzo innovador en el análisis estadístico.

Por el año 1540 el alemán Sebastián Muster realizó una compilación estadística de los recursos nacionales, comprensiva de datos sobre organización política, instrucciones sociales, comercio y poderío militar. Durante el siglo XVII aportó indicaciones más

concretas de métodos de observación y análisis cuantitativo y amplió los campos de la inferencia y la teoría Estadística.
Los eruditos del siglo XVII demostraron especial interés por la Estadística Demográfica como resultado de la especulación sobre si la población aumentaba, decrecía o permanecía estática.
En los tiempos modernos tales métodos fueron resucitados por algunos reyes que necesitaban conocer las riquezas monetarias y el potencial humano de sus respectivos países. El primer empleo de los datos estadísticos para fines ajenos a la política tuvo lugar en 1691 y estuvo a cargo de Gaspar Neumann, un profesor alemán que vivía en Breslau. Este investigador se propuso destruir la antigua creencia popular de que en los años terminados en siete moría más gente que en los restantes, y para lograrlo hurgó pacientemente en los archivos parroquiales de la ciudad. Después de revisar miles de partidas de defunción pudo demostrar que en tales años no fallecían más personas que en los demás. Los procedimientos de Neumann fueron conocidos por el astrónomo inglés Halley, descubridor del cometa que lleva su nombre, quien los aplicó al estudio de la vida humana. Sus cálculos sirvieron de base para las tablas de mortalidad que hoy utilizan todas las compañías de seguros.
Durante el siglo XVII y principios del XVIII, matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace desarrollaron la teoría de probabilidades. No obstante durante cierto tiempo, la teoría de las probabilidades limitó su aplicación a los juegos de azar y hasta el siglo XVIII no comenzó a aplicarse a los grandes problemas científicos.
Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística, que extrajo del término italiano statista (estadista). Creía, y con sobrada razón, que los datos de la nueva ciencia serían el aliado más eficaz del gobernante consciente. La raíz remota de la palabra se halla, por otra parte, en el término latino status, que significa estado o situación; Esta etimología aumenta el valor intrínseco de la palabra, por cuanto la estadística revela el sentido cuantitativo de las más variadas situaciones.
Jacques Quételect es quien aplica las Estadísticas a las ciencias sociales. Este interpretó la teoría de la probabilidad para su uso en las ciencias sociales y resolver la aplicación del principio de promedios y de la variabilidad a los fenómenos sociales. Quételect fue el primero en realizar la aplicación práctica de todo el método Estadístico, entonces conocido, a las diversas ramas de la ciencia.
Entretanto, en el período del 1800 al 1820 se desarrollaron dos conceptos matemáticos fundamentales para la teoría Estadística; la teoría de los errores de observación, aportada por Laplace y Gauss; y la teoría de los mínimos cuadrados desarrollada por Laplace, Gauss y Legendre. A finales del siglo XIX, Sir Francis Gaston ideó el método conocido por Correlación, que tenía por objeto medir la influencia relativa de los factores sobre las variables. De aquí partió el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica como J. Pease Norton, R. H. Hooker y G. Udny Yule, que efectuaron amplios estudios sobre la medida de las relaciones.
Los progresos más recientes en el campo de la Estadística se refieren al ulterior desarrollo del cálculo de probabilidades, particularmente en la rama denominada indeterminismo o relatividad, se ha demostrado que el determinismo fue reconocido en la Física como resultado de las investigaciones atómicas y que este principio se juzga aplicable tanto a las ciencias sociales como a las físicas.

Etapas de Desarrollo de la Estadística

La historia de la estadística está resumida en tres grandes etapas o fases.
1.- Primera Fase: Los Censos:
Desde el momento en que se constituye una autoridad política, la idea de inventariar de una forma más o menos regular la población y las riquezas existentes en el territorio está ligada a la conciencia de soberanía y a los primeros esfuerzos administrativos.

2.- Segunda Fase: De la Descripción de los Conjuntos a la Aritmética Política:
Las ideas mercantilistas extrañan una intensificación de este tipo de investigación. Colbert multiplica las encuestas sobre artículos manufacturados, el comercio y la población: los intendentes del Reino envían a París sus memorias. Vauban, más conocido por sus fortificaciones o su Dime Royale, que es la primera propuesta de un impuesto sobre los ingresos, se señala como el verdadero precursor de los sondeos. Más tarde, Bufón se preocupa de esos problemas antes de dedicarse a la historia natural.
La escuela inglesa proporciona un nuevo progreso al superar la fase puramente descriptiva. Sus tres principales representantes son Graunt, Petty y Halley. El penúltimo es autor de la famosa Aritmética Política.
Chaptal, ministro del interior francés, publica en 1801 el primer censo general de población, desarrolla los estudios industriales, de las producciones y los cambios, haciéndose sistemáticos durantes las dos terceras partes del siglo XIX.
3.- Tercera Fase: Estadística y Cálculo de Probabilidades:
El cálculo de probabilidades se incorpora rápidamente como un instrumento de análisis extremadamente poderoso para el estudio de los fenómenos económicos y sociales y en general para el estudio de fenómenos "cuyas causas son demasiados complejas para conocerlos totalmente y hacer posible su análisis".

División de la Estadística

La Estadística para su mejor estudio se ha dividido en dos grandes ramas: la Estadística Descriptiva y la Inferencial.

Estadística Descriptiva: consiste sobre todo en la presentación de datos en forma de tablas y gráficas. Esta comprende cualquier actividad relacionada con los datos y está diseñada para resumir o describir los mismos sin factores pertinentes adicionales; esto es, sin intentar inferir nada que vaya más allá de los datos, como tales.

Estadística Inferencial: se deriva de muestras, de observaciones hechas sólo acerca de una parte de un conjunto numeroso de elementos y esto implica que su análisis requiere de generalizaciones que van más allá de los datos. Como consecuencia, la característica más importante del reciente crecimiento de la estadística ha sido un cambio en el énfasis de los métodos que describen a métodos que sirven para hacer generalizaciones. La Estadística Inferencial investiga o analiza una población partiendo de una muestra tomada.

Método Estadístico

El conjunto de los métodos que se utilizan para medir las características de la información, para resumir los valores individuales, y para analizar los datos a fin de extraerles el máximo de información, es lo que se llama métodos estadísticos. Los métodos de análisis para la información cuantitativa se pueden dividir en los siguientes seis pasos:
1. Definición del problema.
2. Recopilación de la información existente.
3. Obtención de información original.
4. Clasificación.
5. Presentación.
6. Análisis.

Errores Estadísticos Comunes

Al momento de recopilar los datos que serán procesados se es susceptible de cometer errores así como durante los cómputos de los mismos. No obstante, hay otros errores que no tienen nada que ver con la digitación y que no son tan fácilmente identificables. Algunos de éstos errores son:
Sesgo: Es imposible ser completamente objetivo o no tener ideas preconcebidas antes de comenzar a estudiar un problema, y existen muchas maneras en que una perspectiva o

estado mental pueda influir en la recopilación y en el análisis de la información. En estos casos se dice que hay un sesgo cuando el individuo da mayor peso a los datos que apoyan su opinión que a aquellos que la contradicen. Un caso extremo de sesgo sería la situación donde primero se toma una decisión y después se utiliza el análisis estadístico para justificar la decisión ya tomada.

Datos no comparables: el establecer comparaciones es una de las partes más importantes del análisis estadístico, pero es extremadamente importante que tales comparaciones se hagan entre datos que sean comparables.

Proyección descuidada de tendencias: la proyección simplista de tendencias pasadas hacia el futuro es uno de los errores que más ha desacreditado el uso del análisis estadístico.

Muestreo Incorrecto: en la mayoría de los estudios sucede que el volumen de información disponible es tan inmenso que se hace necesario estudiar muestras, para derivar conclusiones acerca de la población a que pertenece la muestra. Si la muestra se selecciona correctamente, tendrá básicamente las mismas propiedades que la población de la cual fue extraída; pero si el muestreo se realiza incorrectamente, entonces puede suceder que los resultados no signifiquen nada

Capítulo II	CARACTERISTICAS DE UNA DISTRIBUCIÓN DE FRECUENCIAS

2.1. Introducción

La fase previa de cualquier estudio estadístico se basa en la recogida y ordenación de datos; esto se realiza con la ayuda de los resúmenes numéricos y gráficos visto en los temas anteriores.

2.2. Medidas de posición

Son aquellas medidas que nos ayudan a saber *donde* están los datos pero sin indicar como se distribuyen.

2.2.1. Medidas de posición central

a) Media aritmética ($\overline{X}$)

La media aritmética o simplemente media, que denotaremos por $\overline{X}$, es el número obtenido al dividir la suma de todos los valores de la variable entre el numero total de observaciones, y se define por la siguiente expresión:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N}$$

Ejemplo:
Si tenemos la siguiente distribución, se pide hallar la media aritmética, de los siguientes datos expresados en kg.

xi	ni	xi ni
54	2	108
59	3	177
63	4	252
64	1	64
	N=10	601

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{601}{10} = \underline{\underline{60{,}1}} \text{ kg}$$

Si los datos están agrupados en intervalos, la expresión de la media aritmética, es la misma, pero utilizando la marca de clase (Xi).

Ejemplo:

$(L_{i-1}, L_i]$	x_i	n_i	$x_i n_i$

[30 , 40]	35	3	105
(40 , 50]	45	2	90
(50 , 60]	55	5	275
		10	470

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{470}{10} = \underline{\underline{47}}$$

Propiedades:
1ª) Si sometemos a una variable estadística X, a un cambio de origen y escala Y = a + b X, la media aritmética de dicha variable X, varía en la misma proporción.

$$Y = a + bX \quad \Rightarrow \quad \overline{Y} = a + b\overline{X}$$

2ª) La suma de las desviaciones de los valores o datos de una variable X, respecto a su media aritmética es cero.

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

Ventajas e inconvenientes:
- La media aritmética viene expresada en las mismas unidades que la variable.
- En su cálculo intervienen todos los valores de la distribución.
- Es el centro de gravedad de toda la distribución, representando a todos los valores observados.
- Es única.
- Su principal inconveniente es que se ve afectada por los valores extremadamente grandes o pequeños de la distribución.

- Media aritmética ponderada

Es una media aritmética que se emplea en distribuciones de tipo unitario, en las que se introducen unos coeficientes de ponderación, denominados ω_i, que son valores positivos, que representan el número de veces que un valor de la variable es más importante que otro.

$$W = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

b) <u>Media geométrica</u>

Sea una distribución de frecuencias (x_i, n_i). La media geométrica, que denotaremos por G. se define como la raíz N-ésima del producto de los N valores de la distribución.

$G = \sqrt[N]{x_1^{n1} x_2^{n2} \cdots\cdots x_k^{nk}}$

Si los datos están agrupados en intervalos, la expresión de la media geométrica, es la misma, pero utilizando la marca de clase (Xi).

El empleo más frecuente de la media geométrica es el de promediar variables tales como porcentajes, tasas, números índices. etc., es decir, en los casos en los que se supone que la variable presenta variaciones acumulativas.

Ventajas e inconvenientes:
- En su cálculo intervienen todos los valores de la distribución.
- Los valores extremos tienen menor influencia que en la media aritmética.
- Es única.
- Su cálculo es más complicado que el de la media aritmética.

Además, cuando la variable toma al menos un $x_i = 0$ entonces G se anula, y si la variable toma valores negativos se pueden presentar una gama de casos particulares en los que tampoco queda determinada debido al problema de las raíces de índice par de números negativos.

c) Media armónica

La media armónica, que representaremos por H, se define como sigue:

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

Obsérvese que la inversa de la media armónica es la media aritmética de los inversos de los valores de la variable. No es aconsejable en distribuciones de variables con valores pequeños. Se suele utilizar para promediar variables tales como productividades, velocidades, tiempos, rendimientos, cambios, etc.

Ventajas e inconvenientes:
- En su cálculo intervienen todos los valores de la distribución.
- Su cálculo no tiene sentido cuando algún valor de la variable toma valor cero.
- Es única.

- Relación entre las medias:

$$\boxed{H \leq G \leq \overline{X}}$$

d) Mediana (Me)

Dada una distribución de frecuencias con los valores ordenados de menor a mayor, llamamos mediana y la representamos por Me, al valor de la variable, que deja a su izquierda el mismo número de frecuencias que a su derecha.

- Calculo de la mediana:

Variara según el tipo de dato:

a) Variables discretas no agrupadas:

1º) Se calcula $\frac{N}{2}$ y se construye la columna de las Ni (frecuencias acumuladas)

2º) Se observa cual es la primera Ni que supera o iguala a $\frac{N}{2}$, distinguiéndose dos casos:

- Si existe un valor de Xi tal que $N_{i-1} \prec \frac{N}{2} \prec N_i$, entonces se toma como $\boxed{Me = x_i}$
- Si existe un valor i tal que $N_i = \frac{N}{2}$, entonces la $\boxed{Me = \frac{x_i + x_{i+1}}{2}}$

Ejemplo: Sea la distribución

xi	ni	Ni
1	3	3
2	4	7
5	9	16
7	10	26
10	7	33
13	2	35
	n = 35	

lugar que ocupa $\frac{N}{2} = \frac{35}{2} = 17{,}5$

como se produce que $N_{i-1} < \frac{N}{2} < N_i \Rightarrow 16 < 17{,}7 < 26 \Rightarrow Me = x_i$,por lo tanto Me = 7

El otro caso lo podemos ver en la siguiente distribución:

xi	ni	Ni
1	3	3
2	4	7

5	9	16
7	10	26
10	6	32
	n= 32	

Lugar que ocupa = 32/2 = 16 ==> $Me = \frac{x_1 + x_{i+1}}{2} = \frac{5+7}{2} = 6$

Notar que en este caso se podría haber producido que hubiera una frecuencia absoluta acumulada superior a 16. En este caso se calcularía como en el ejemplo anterior.

b) Variables agrupadas por intervalos

En este caso hay que detectar en que intervalo está el valor mediano. Dicho intervalo se denomina " intervalo mediano ".

Cada intervalo I_i vendrá expresado según la notación I_i = (L_{i-1} , L_i]; observando la columna de las frecuencias acumuladas, buscaremos el primer intervalo cuya Ni sea mayor o igual que $\frac{N}{2}$, que será el intervalo modal; una vez identificado dicho intervalo, procederemos al cálculo del valor mediano, debiendo diferenciar dos casos:

1º) Si existe Ii tal que $N_{i-1} \prec \frac{N}{2} \prec N_i$, entonces el intervalo mediano es el (L_{i-1} , L_i] y la mediana es:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} c_i$$

2º) Análogamente si existe un Ii tal que $N_i = \frac{N}{2}$, la mediana es $\boxed{Me = L_i}$

Ejemplo:

(Li-1, Li]	ni	Ni
[20 , 25]	100	100
(25 , 30]	150	250
(30 , 35]	200	450
(35 , 40]	180	630
(40 , 45]	41	671
	N = 671	

671/2 = 335.5 ; Me estará en el intervalo (30 - 35]. Por tanto realizamos el cálculo:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 30 + \frac{33{,}5 - 250}{200} * 5 = \underline{\underline{32{,}138}}$$

Ventajas e inconvenientes :

- Es la medida más representativa en el caso de variables que solo admitan la escala ordinal.
- Es fácil de calcular.
- En la mediana solo influyen los valores centrales y es insensible a los valores extremos u "outliers ".
- En su determinación no intervienen todos los valores de la variable.

e) Moda

La moda es el valor de la variable que más veces se repite, y en consecuencia, en una distribución de frecuencias, es el valor de la variable que viene afectada por la máxima frecuencia de la distribución. En distribuciones no agrupadas en intervalos se observa la columna de las frecuencias absolutas, y el valor de la distribuci6n al que corresponde la mayor frecuencia será la moda. A veces aparecen distribuciones de variables con más de una moda (bimodales, trimodales, etc), e incluso una distribución de frecuencias que presente una moda absoluta y una relativa.

En el caso de estar la variable agrupada en intervalos de distinta amplitud, se define el intervalo modal, y se denota por (L_{i-1} , L_i], como aquel que posee mayor densidad de frecuencia (h_i); la densidad de frecuencia se define como : $\boxed{h_i = \frac{n_i}{a_i}}$

Una vez identificado el intervalo modal procederemos al cálculo de la moda, a través de la fórmula:

$$\boxed{Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} c_i}$$

En el caso de tener todos los intervalos la misma amplitud, el intervalo modal será el que posea una mayor frecuencia absoluta (n_i) y una vez identificado este, empleando la fórmula:

$$\boxed{Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} c_i}$$

Ventajas e inconvenientes:

- Su cálculo es sencillo.
- Es de fácil interpretación.

- Es la única medida de posición central que puede obtenerse en las variables de tipo cualitativo.
- En su determinación no intervienen todos lo valores de la distribución.

2.2.2. **Medidas de posición no central (Cuantiles)**

Los cuantiles son aquellos valores de la variable, que ordenados de menor a mayor, dividen a la distribución en partes, de tal manera que cada una de ellas contiene el mismo número de frecuencias.

Los cuantiles más conocidos son:

a) Cuartiles (Qi)

Son valores de la variable que dividen a la distribución en 4 partes, cada una de las cuales engloba el 25 % de las mismas. Se denotan de la siguiente forma: Q_1 es el primer cuartil que deja a su izquierda el 25 % de los datos; Q_2 es el segundo cuartil que deja a su izquierda el 50% de los datos, y Q_3 es el tercer cuartil que deja a su izquierda el 75% de los datos. (Q_2 = Me)

b) Deciles (Di)

Son los valores de la variable que dividen a la distribución en las partes iguales, cada una de las cuales engloba el 10 % de los datos. En total habrá 9 deciles. ($Q_2 = D_5$ = Me)

c) Centiles o Percentiles (Pi)

Son los valores que dividen a la distribución en 100 partes iguales, cada una de las cuales engloba el 1 % de las observaciones. En total habrá 99 percentiles. ($Q_2 = D_5$ = Me = P_{50})

- **Cálculo de los cuantiles en distribuciones no agrupadas en intervalos**

- Se calculan a través de la siguiente expresión: $\boxed{\frac{rN}{q}}$, siendo :

  r = el orden del cuantil correspondiente

  q = el número de intervalos con iguales frecuencias u observaciones (q = 4, 10, ó 100).

  N = número total de observaciones

- La anterior expresión nos indica que valor de la variable estudiada es el cuantil que nos piden, que se corresponderá con el primer valor cuya frecuencia acumulada sea mayor o igual a $\frac{rN}{q}$

Ejemplo: DISTRIBUCIONES NO AGRUPADAS: En la siguiente distribución

xi	ni
5	3
10	7
15	5
20	3
25	2
	N = 20

Ni
3
10
15
18
20

Calcular la mediana (Me); el primer y tercer cuartil (C1,C3); el 4º decil (D4) y el 90 percentil (P90)

Mediana (Me)

Lugar que ocupa la mediana ➔ lugar 20/2 = 10

Como es igual a un valor de la frecuencia absoluta acumulada, realizaremos es cálculo: $Me = \frac{x_i + x_{i+1}}{2} = \frac{10+15}{2} = \underline{\underline{12,5}}$

Primer cuartil (C1)

Lugar que ocupa en la distribución (¼). 20 = 20/4 = 5 Como Ni-1 < $\frac{rN}{q}$ < Ni , es decir 3 < 5 < 10 esto implicara que C1 = xi = 10

Tercer cuartil (C3)

Lugar que ocupa en la distribución (3/4).20 = 60/4 = 15, que coincide con un valor de la frecuencia absoluta acumulada, por tanto realizaremos el cálculo: $C_3 = \frac{x_i + x_{i+1}}{2} = \frac{15+20}{2} = \underline{\underline{17,5}}$

Cuarto decil (D4)

Lugar que ocupa en la distribución (4/10) . 20 = 80/10 = 8. Como Ni-1 < $\frac{rN}{q}$ < Ni ya que 3 < 8 < 10 por tanto D4 =10.

Nonagésimo percentil (P90)

Lugar que ocupa en la distribución (90/100). 20 = 1800/100 = 18. que coincide con un valor de la frecuencia absoluta acumulada, por tanto realizaremos el cálculo: $P_{90} = \frac{x_i + x_{i+1}}{2} = \frac{20+25}{2} = \underline{\underline{22,5}}$

- **Cálculo de los cuantiles en distribuciones agrupadas en intervalos**

- Este cálculo se resuelve de manera idéntica al de la mediana.
- El intervalo donde se encuentra el cuantil i-esimo, es el primero que una vez ordenados los datos de menor a mayor, tenga como frecuencia acumulada (Ni) un valor superior o igual a $\frac{rN}{q}$; una vez

identificado el intervalo Ii (Li-1 , Li], calcularemos el cuantil correspondiente, a través de la fórmula:

$$C_{r/q} = L_{i-1} + \frac{\frac{rN}{q} - N_{i-1}}{n_i} c_i$$

r=1,2,...,q-1.

Cuartil: q=4; Decil: q=10; Percentil: q=100

Ejemplo:
DISTRIBUCIONES AGRUPADAS: Hallar el primer cuartil, el cuarto decil y el 90 percentil de la siguiente distribución:

[Li-1 , Li)	ni	Ni
[0 , 100]	90	90
(100 , 200]	140	230
(200 , 300]	150	380
(300 , 800]	120	500
	N = 500	

- Primer cuartil (Q1)
- Lugar ocupa el intervalo del primer cuartil: (1/4). 500 = 500/4 = 125. Por tanto Q1 estará situado en el intervalo (100 – 200].Aplicando la expresión directamente, tendremos: $Q_1 = 100 + \frac{125-90}{140}100 = \underline{\underline{125}}$

- Cuarto decil (D4)
- Lugar que ocupa: (4/10) . 500 = 200 . Por tanto D4 estará situado en el intervalo (100 – 200]. Aplicando la expresión tendremos: $D_4 = 100 + \frac{200-90}{140}100 = \underline{\underline{178,57}}$
-
- Nonagésimo percentil (P 90)
- Lugar que ocupa: (90/100) . 500 = 450, por tanto P90 estará situado en el intervalo (300 – 800]. Aplicando la expresión tendremos: $P_{90} = 300 + \frac{450-380}{120}500 = 300 + \frac{70}{120}500 = \underline{\underline{591,67}}$

2.3. Momentos potenciales

Los momentos son medidas obtenidas a partir de todos los datos de una variable estadística y sus frecuencias absolutas. Estas medidas caracterizan a las distribuciones

de frecuencias de tal forma que si los momentos coinciden en dos distribuciones, diremos que son iguales.

2.3.1. Momentos respecto al origen

Se define el momento de orden h respecto al origen de una variable estadística a la expresión:

$$a_h = \frac{\sum_{i=1}^{n} x_i^{\,h} n_i}{N}$$

Particularidades:

Si h = 1, a1 es igual a la media aritmética.
Si h = 0, a0 es igual a uno (a0 = 1)

2.3.2. Momentos centrales o momentos con respecto a la media aritmética

$$m_h = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^h n_i}{N}$$

Particularidades:

- Si h = 1, entonces m1 = 0
- Si h = 2, entonces m2 = S^2

2.4. Medidas de dispersión

Las medidas de dispersión tratan de medir el grado de dispersión que tiene una variable estadística en torno a una medida de posición o tendencia central, indicándonos lo representativa que es la medida de posición. A mayor dispersión menor representatividad de la medida de posición y viceversa.

2.4.1 **Medidas de dispersión absoluta**

a) Recorrido (Re)

Se define como la diferencia entre el máximo y el mínimo valor de la variable:

$$R = \text{máx } x_i - min\ x_i$$

Ej: Sea X, las indemnizaciones recibidas por cuatro trabajadores de dos empresas A y B

A	100	120	350	370
B	225	230	240	245

Re (A) = 370 – 100= 270
Re (B) = 245 – 225= 20 ---→ Distribución menos dispersa

- Otros recorridos:
- intervalo intercuartílico $I = Q_3 - Q_1$
- intervalo interdecílico $I = (D_9 - D_1)$
- intervalo intercentílico $I = (P_{99} - P_1)$

b) Desviación absoluta media con respecto a la media (d_e)
Nos indica las desviaciones con respecto a la media con respecto a la media aritmética en valor absoluto.

$$d_e = \frac{\sum_{i=1}^{r} \left| x_i - \bar{x} \right| n_i}{N}$$

c) Varianza
La varianza mide la mayor o menor dispersión de los valores de la variable respecto a la media aritmética. Cuanto mayor sea la varianza mayor dispersión existirá y por tanto menor representatividad tendrá la media aritmética.
La varianza se expresa en las mismas unidades que la variable analizada, pero elevadas al cuadrado.

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

$$S^2 = \frac{\sum_{i=1}^{r} x_i^{\,2} n_i}{N} - \overline{x^2}$$

Propiedades:

1ª) La varianza siempre es mayor o igual que cero y menor que infinito $\left(S^2{}_x \geq 0\right)$

2ª) Si a una variable X la sometemos a un cambio de origen " a " y un cambio de escala " b ", la varianza de la nueva variable Y= a + bX, será:

$$\left(S_y{}^2 = b^2 S_x{}^2\right)$$

d) Desviación típica o estándar

Se define como la raíz cuadrada con signo positivo de la varianza.

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

2.4.2. **Medidas de dispersión relativa**

Nos permiten comparar la dispersión de distintas distribuciones.

a) Coeficiente de variación de Pearson (CVx)

Indica la relación existente entre la desviación típica de una muestra y su media.

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

Al dividir la desviación típica por la media se convierte en un valor excento de unidad de medida. Si comparamos la dispersión en varios conjuntos de observaciones tendrá menor dispersión aquella que tenga menor coeficiente de variación.

El principal inconveniente, es que al ser un coeficiente inversamente proporcional a la media aritmética, cuando está tome valores cercanos a cero, el coeficiente tenderá a infinito.

Ejemplo: Calcula la varianza, desviación típica y la dispersión relativa de esta distribución.

Sea x el número de habitaciones que tienen los 8 pisos que forman un bloque de vecinos

X	ni
2	2
3	2
5	1
6	3

	N= 8

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{2*2+3*2+5*1+6*3}{8} = 4.125$$ habitaciones

$$S^2 = \frac{\sum_{i=1}^{r} x_i{}^2 n_i}{N} - \overline{x^2} = \frac{2^2*2+3^2*2+5^2*1+6^2*3}{8} - (4.125)^2 = 2.86$$ (habitaciones)2

$$S_x = +\sqrt{S_x^2} = +\sqrt{2.86} = 1.69$$ habitaciones

$$CV = \frac{S}{\bar{x}} = \frac{1.69}{4.125} = 0.41$$

2.5. Medidas de forma

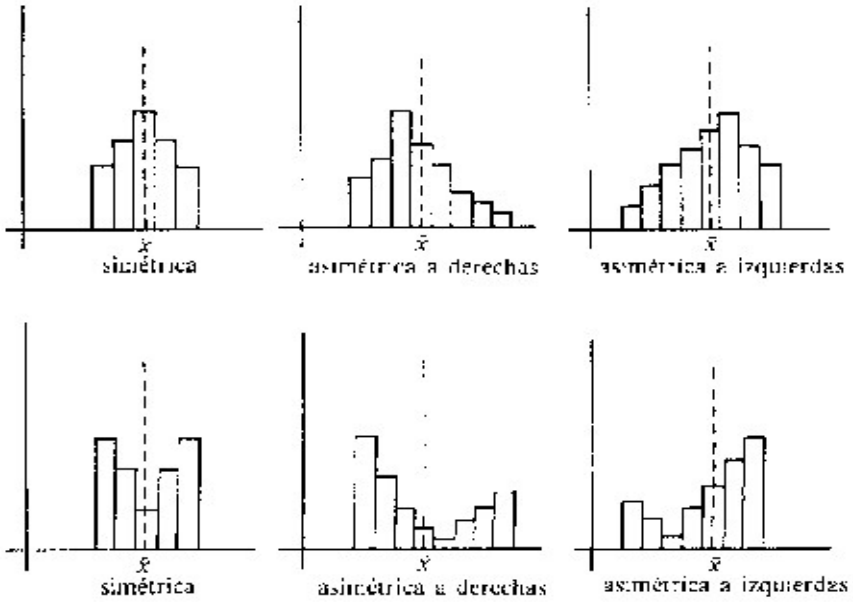
- Asimetría
- Curtosis o apuntamiento.

Hasta ahora, hemos estado analizando y estudiando la dispersión de una distribución, pero parece evidente que necesitamos conocer más sobre el comportamiento de una distribución. En esta parte, analizaremos las medidas de forma, en el sentido de histograma o representación de datos, es decir, que información nos aporta según la forma que tengan la disposición de datos.

Las medidas de forma de una distribución se pueden clasificar en dos grandes grupos o bloques: medidas de asimetría y medidas de curtosis.

2.5.1. **Medidas de asimetría o sesgo : Coeficiente de asimetría de Fisher.**

Cuando al trazar una vertical, en el diagrama de barras o histograma, de una variable, según sea esta discreta o continua, por el valor de la media, esta vertical, se transforma en eje de simetría, decimos que la distribución es simétrica. En caso contrario, dicha distribución será asimétrica o diremos que presenta asimetría.

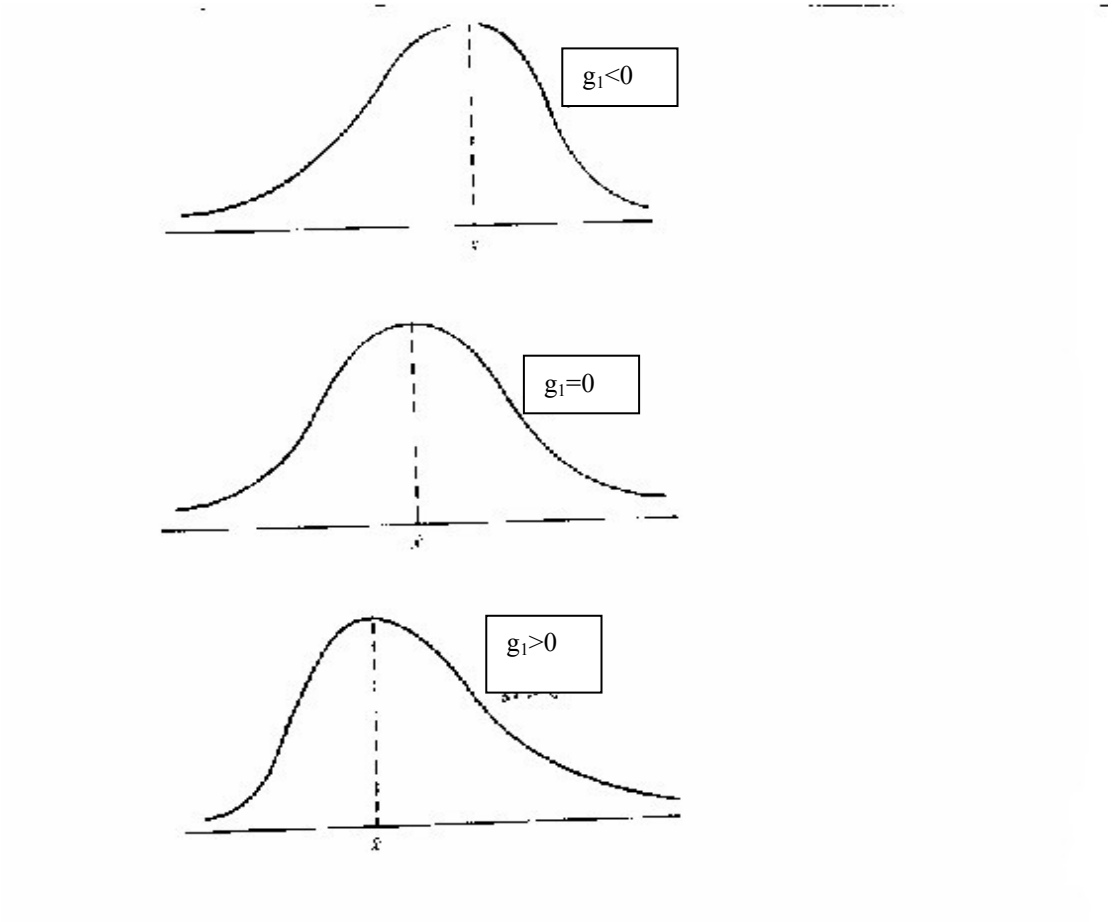


El coeficiente de asimetría más preciso es el de Fisher, que se define por:

$$g_1 = \frac{\sum_{i=1}^{r} \left(x_i - \bar{x}\right)^3 n_i}{\frac{N}{S^3}}$$

Según sea el valor de g1, diremos que la distribución es asimétrica a derechas o positiva, a izquierdas o negativa, o simétrica, o sea:

Si g1 > 0 ➔ la distribución será asimétrica positiva o a derechas (desplazada hacia la derecha).
Si g1 < 0 ➔ la distribución será asimétrica negativa o a izquierdas (desplazada hacia la izquierda).
Si g1 = 0 ➔ la distribución puede ser simétrica; si la distribución es simétrica, entonces si podremos afirmar que g1 = 0.



- Si existe simetría, entonces $g_1 = 0$, y $\overline{X} = Me$; si además la distribución es unimodal, también podemos afirmar que: $\overline{X} = Me = Mo$
- Si $g_1 > 0$, entonces : $\overline{X} > Me > Mo$
- Si $g_1 < 0$, entonces : $\overline{X} < Me < Mo$

2.5.2. **Medidas de apuntamiento o curtosis: coeficiente de curtosis de Fisher**

Con estas medidas nos estamos refiriendo al grado de apuntamiento que tiene una distribución; para determinarlo, emplearemos el coeficiente de curtosis de Fisher. (g2)

$$g2 = \frac{\dfrac{\sum_{i=1}^{r}\left(x_i - \bar{x}\right)^4 n_i}{N}}{S^4}$$

Si g2 > 3 la distribución será leptocúrtica o apuntada
Si g2 = 3 la distribución será mesocúrtica o normal
Si g2 < 3 la distribución será platicúrtica o menos apuntada que lo normal.

2.6. Medidas de concentración

Las medidas de concentración tratan de poner de relieve el mayor o menor grado de igualdad en el reparto del total de los valores de la variable, son por tanto indicadores del grado de distribución de la variable.

Para este fin, están concebidos los estudios sobre concentración.

Denominamos concentración a la mayor o menor equidad en el reparto de la suma total de los valores de la variable considerada (renta, salarios, etc.).

Las infinitas posibilidades que pueden adoptar los valores, se encuentran entre los dos extremos:

1.- Concentración máxima, cuando uno solo percibe el total y los demás nada, en este caso, nos encontraremos ante un reparto no equitativo:
x1 = x2 = x3 = …… = xn-1 = 0 y xn.
2.- Concentración mínima, cuando el conjunto total de valores de la variable esta repartido por igual, en este caso diremos que estamos ante un reparto equitativo
x1 = x2 = x3 = …… = xn-1 = xn

De las diferentes medidas de concentración que existen nos vamos a centrar en dos:
Indice de Gini, Coeficiente, por tanto será un valor numérico.
Curva de Lorenz, gráfico, por tanto será una representación en ejes coordenados.

Sea una distribución de rentas (xi, ni) de la que formaremos una tabla con las siguientes columnas:

1.- Los productos xi ni, que nos indicarán la renta total percibida por los ni rentistas de renta individual xi .

2.- Las frecuencias absolutas acumuladas Ni .

3.- Los totales acumulados ui que se calculan de la siguiente forma:

u1= x1 n1

u2 = x1 n1 + x2 n2

u3 = x1 n1 + x2 n2 + x3 n3

u4 = x1 n1 + x2 n2 + x3 n3 + x4 n4

un = x1 n1 + x2 n2 + x3 n3 + x4 n4 + ……. + xn nn

Por tanto podemos decir que $\boxed{u_n = \sum_{i=1}^{n} x_i n_i}$

4.- La columna total de frecuencias acumuladas relativas, que expresaremos en tanto por ciento y que representaremos como pi y que vendrá dada por la siguiente notación

$$\boxed{p_i = \frac{N_i}{n} 100}$$

5.- La renta total de todos los rentistas que será un y que dada en tanto por ciento, la cual representaremos como qi y que responderá a la siguiente notación:

$$\boxed{q_i = \frac{u_i}{u_n} 100}$$

Por tanto ya podemos confeccionar la tabla que será la siguiente:

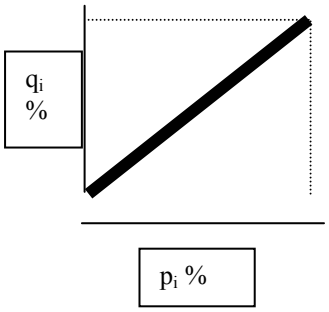
xi	ni	xi ni	Ni	ui	$p_i = \frac{N_i}{n} 100$	$q_i = \frac{u_i}{u_n} 100$	pi - qi
x1	n1	x1 n1	N1	u1	p1	q1	p1 - q1
x2	n2	x2 n2	N2	u2	p2	q2	p2 - q2
...	...	...	...	...	...	...	...
xn	nn	xn nn	Nn	un	pn	qn	pn - qn

Como podemos ver la última columna es la diferencia entre las dos penúltimas, esta diferencia seria 0 para la concentración mínima ya que pi = qi y por tanto su diferencia seria cero.

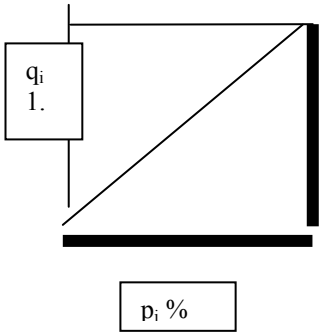
Si esto lo representamos gráficamente obtendremos la curva de concentración o curva de Lorenz .La manera de representarlo será, en el eje de las X, los valores pi en % y en el de las Y los valores de qi en %. Al ser un %, el gráfico siempre será un cuadrado, y la gráfica será una curva que se unirá al cuadrado, por los valores (0,0), y (100,100), y quedará siempre por debajo de la diagonal.

La manera de interpretarla será: cuanto más cerca se sitúe esta curva de la diagonal, menor concentración habrá, o más homogeneidad en la distribución. Cuanto más se acerque a los ejes, por la parte inferior del cuadrado, mayor concentración.

Los extremos son



Distribución de concentración mínima



Distribución de concentración

Analíticamente calcularemos el índice de Gini el cual responde a la siguiente ecuación

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

Este índice tomara los valores de $I_G = 0$ cuando pi = qi concentración mínima y de $I_G = 1$ cuando qi = 0

Esto lo veremos mejor con un ejemplo :

Li-1 - Li	marca xi	Frecuencia ni	Ni	xini	Σun	qi = (ui/un)* 100	pi=(Ni/n) *100	pi - qi
0 - 50	25	23	23	575	575	1,48	8,85	7,37
50 - 100	75	72	95	5400	5975	15,38	36,54	21,16
100 - 150	125	62	157	7750	13725	35,33	60,38	25,06
150 - 200	175	48	205	8400	22125	56,95	78,85	21,90
200 - 250	225	19	224	4275	26400	67,95	86,15	18,20
250 - 300	275	8	232	2200	28600	73,62	89,23	15,61
300 - 350	325	14	246	4550	33150	85,33	94,62	9,29
350 - 400	375	7	253	2625	35775	92,08	97,31	5,22
400 - 450	425	5	258	2125	37900	97,55	99,23	1,68
450 - 500	475	2	260	950	38850	100,00	100,00	0,00
		260		38850			651,15	125,48

Se pide Indice de concentración y Curva de Lorenz correspondiente

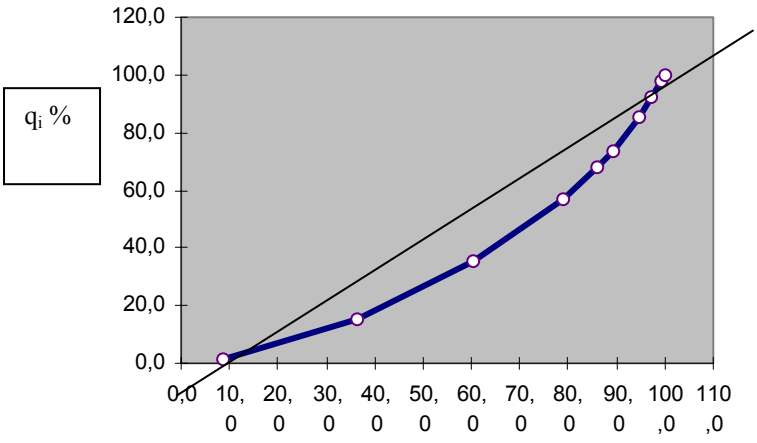
Indice de concentración de GINI

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{125{,}48}{651{,}15} = 0{,}193$$

, Observamos que hay poca concentración por encontrarse cerca del 0.

Curva de Lorenz

La curva la obtenemos cerca de la diagonal, que indica que hay poca concentración:


q_i %
120,0
100,0
80,0
60,0
40,0
20,0
0,0
0,0
10,00
20,00
30,00
40,00
50,00
60,00
70,00
80,00
90,00
100,0
110,0
p_i %

-

Capítulo III	DISTRIBUCIONES BIDIMENSIONALES

3.1. Introducción.

Estudiaremos dos características de un mismo elemento de la población (altura y peso, dos asignaturas, longitud y latitud).
De forma general, si se estudian sobre una misma población y se miden por las mismas unidades estadísticas una variable X y una variable Y, se obtienen series estadísticas de las variables X e Y.
Considerando simultáneamente las dos series, se suele decir que estamos ante una variable estadística bidimensional.

3.2. Tabulación de variables estadísticas bidimensionales.

Vamos a considerar 2 tipos de tabulaciones:
1º) Para variables cuantitativas, que reciben el nombre de tabla de correlación.
2º) Para variables cualitativas, que reciben el nombre de tabla de contingencia.

3.2.1.Tablas de correlación.

Sea una población estudiada simultaneamente según dos caracteres X e Y; que representaremos genéricamente como $\mathbf{(x_i; y_j ; n_{ij})}$, donde x_i; y_j, son dos valores cualesquiera y n_{ij} es la frecuencia absoluta conjunta del valor i-ésimo de X con el j-ésimo de Y.

Una forma de disponer estos resultados es la conocida como tabla de doble entrada o tabla de correlación, la cual podemos representar como sigue:

Y X	y_1	y_2	.....	y_j	.....	y_s	n_i .	f_i .
x_1	n11	n12	.....	n1j	.....	n1k	n_1 .	f_1 .
x_2	n21	n22	.....	n2j	.....	n2k	n_2 .	f_2 .
. . .	. .	. .	. .... .	. .	. .... .	. .	. .	. .
x_i	ni1	ni2	.....	nij	.....	nik	n_i .	f_i .
. . .	. . .	. . .	. .... .	. . .	. .... .	. . .	. . .	. . .
x_r	nh1	nh2	.....	nhj	.....	nhk	n_h .	f_h .
$n._j$	$n._1$	$n._2$	.....	$n._j$	.....	$n._k$	N	
$f._j$	$f._1$	$f._2$	.....	$f._j$	.....	$f._k$		1

En este caso, n_{11} nos indica el número de veces que aparece x_1 conjuntamente con y_1; n_{12}, nos indica la frecuencia conjunta de x_1 con y_2, etc.

- **Tipos de distribuciones**

Cuando se estudian conjuntamente dos variables, surgen tres tipo de distribuciones: Distribuciones conjuntas, distribuciones marginales y distribuciones condicionadas.

a) Distribución conjunta

- *La frecuencia absoluta conjunta*, viene determinada por el número de veces que aparece el par ordenado (x_i , y_j), y se representa por " **n_{ij} ".**

- *La frecuencia relativa conjunta*, del par (x_i , y_j) es el cociente entre la frecuencia absoluta conjunta y el número total de observaciones. Se trata de " **f_{ij} ".**

Se cumplen las siguientes relaciones entre las frecuencias de distribución conjunta:
1ª) La suma de las frecuencias absolutas conjuntas, extendida a todos los pares es igual al total de observaciones.

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = N$$

2ª) La suma de todas las frecuencias relativas conjuntas extendida a todos los pares es igual a la unidad.

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = 1$$

b) Distribuciones marginales

Cuando trabajamos con más de una variable y queremos calcular las distribuciones de frecuencias de cada una de manera independiente, nos encontramos con las distribuciones marginales.

Variable X

x_i	$n_i.$	$f_i.$
x_1	$n_1.$	$f_1.$
x_2	$n_2.$	$f_2.$
x_3	$n_3.$	$f_3.$
x_4	$n_4.$	$f_4.$
	N	1

Variable Y

y_j	$n.j$	$f.j$
y_1	$n.1$	$f.1$
y_2	$n.2$	$f.2$
y_3	$n.3$	$f.3$
y_4	$n.4$	$f.4$
	N	1

- *Frecuencia absoluta marginal: el valor* n_i. Representa el número de veces que aparece el valor x_i de X, sin tener en cuenta cual es el valor de la variable Y. A $n_{i.}$ se le denomina frecuencia absoluta marginal del valor x_i de X, de forma que:

-

$$n_{i.} = n_{i1+}n_{i2} + ... + n_{is}$$

De la misma manera, la frecuencia absoluta marginal del valor y_j de Y se denotará por $n_{.j}$

$$n_{.j} = n_{1j+}n_{2j} + ... + n_{rj}$$

- *Frecuencia relativa marginal*

La frecuencia relativa marginal de x_i de X, viene dada por:

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

La frecuencia relativa marginal de y_j de Y, viene dada por:

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

- Se cumplen las siguientes relaciones entre las frecuencias de distribución marginales:

1ª) La suma de frecuencias absolutas marginales de la variable X, es igual al número de observaciones que componen la muestra

2ª) La suma de las frecuencias relativas marginales de la variable X, es igual a 1.

3ª) Las dos propiedades anteriores se cumplen también para la variable Y.

c) Distribuciones condicionadas

Consideremos a los $n_{.j}$ individuos de la población que representan la modalidad y_j de la variable Y, y obsérvese la columna j-esima de la tabla. Sus $n_{.j}$ elementos constituyen una población, que es un subconjunto de la población total. Sobre este subconjunto se define la distribución de X condicionada por y_j, que se representa por X / y_j ;su frecuencia absoluta se representa por n_i / j , y su frecuencia relativa por $f_i / _j$, para i = 1, 2, 3, …., r siendo $f_i / _j = \frac{n_{ij}}{n_{.j}}$

El razonamiento es análogo cuando condicionamos la variable Y a un determinado valor de X, es decir Y / x_i

Ejemplo:

Sea X= salario en u.m.

Sea Y = antigüedad en la empresa (años)

X / Y	**1**	**3**	**5**	**7**	**9**	**11**	**ni.**	**fi.**
90	1	2	1	1	0	0	5	0,053
110	2	4	4	5	2	1	18	0,189
130	1	7	3	1	2	0	14	0,147
150	4	6	6	4	3	0	23	0,242
170	2	3	4	6	4	1	20	0,211
190	0	0	2	5	5	3	15	0,158
n.j	10	22	20	22	16	5	95	1
f.j	0,105	0,232	0,211	0,232	0,168	0,053		1

¿ Cuál es la distribución de la retribución, pero únicamente de los empleados con una antigüedad de 5 años?, es decir ¿ cual es la distribución condicionada de la variable X condicionada a que Y sea igual a 5?

X / Y	ni/ y=5	fi/ y=5
90	1	1/20
110	4	4/20
130	3	3/20
150	6	6/20
170	4	4/20
190	2	2/20
n.j	20	1

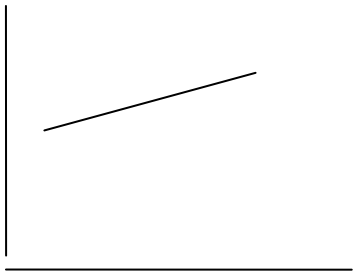
- Covarianza

La covarianza mide la forma en que varía conjuntamente dos variables X e Y
En el estudio conjunto de dos variables, lo que nos interesa principalmente es saber si existe algún tipo de relación entre ellas. Veremos ahora una medida descriptiva que sirve para medir o cuantificar esta relación:

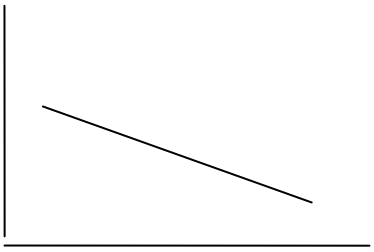
$$S_{xy} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{N}$$

Si $S_{xy} > 0$ hay dependencia directa (positiva), es decir las variaciones de las variables tienen el mismo sentido
Si $S_{xy} = 0$ las variables están incorreladas, es decir no hay relación lineal, pero podría existir otro tipo de relación.
Si $S_{xy} < 0$ hay dependencia inversa o negativa, es decir las variaciones de las variables tienen sentido opuesto.

Gráficamente, indicaría la Covarianza, que los datos, se ajustan a una recta, en los siguientes casos:



$S_{xy} > 0$



$S_{xy} < 0$

- Otra forma de calcular la Covarianza sería: $S_{xy} = m_{11} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j n_{ij}}{N} - \bar{x}\bar{y}$. Será la que utilizaremos en la práctica.
- La covarianza no es un parámetro acotado, y puede tomar cualquier valor real, por lo que su magnitud no es importante; lo significativo es el signo que adopte la misma.

Ejemplo:
Sea X el tiempo de vida de un insecto (años) e Y la longitud del mismo, ¿ podrías deducir si existe relación entre la "edad" del insecto y su tamaño.

X / Y	**2**	**3**	**4**	**ni.**
1	3	1	0	4
2	1	3	1	5
3	0	1	3	4
n.j	4	5	4	13

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_{i.}}{N} = \frac{1*4+2*5+3*4}{13} = 2 \text{ años}$$

$$\bar{y} = \frac{\sum_{j=1}^{s} y_j n_{.j}}{N} = \frac{2*4+3*5+4*4}{13} = 3 \text{ cms}$$

$$S_{xy} = \frac{1*2*3+1*3*1+1*4*0+2*2*1+2*3*3+2*4*1+3*2*0+3*3*1+3*4*3}{13} - 2*3 = 0.461$$

Al tener la covarianza entre ambas variables signo positivo, podemos deducir que existe una relación directa o positiva entre ambas variables, es decir, cuando aumenta la " edad " del insecto también aumenta su tamaño.

3.2.2.Tablas de contingencia
Cuando tenemos la información de 2 variables de tipo cualitativo o de una variable cualitativa y otra cuantitativa, se dispone de una tabla de contingencia. Nos limitaremos al caso de 2 variables. Es una tabla de doble entrada en la que en las filas se ubican las modalidades de una de las variables (atributos) y en las columnas las del otro; en las celdas resultantes del cruce de las filas y las columnas se incluye el número de elementos de la distribución que presentan ambas modalidades.
Si se tiene información de N elementos acerca de las variables A y B de tal forma que presentan " r " y " s " modalidades respectivamente, la tabla de contingencia sería de la forma:

B / A	**B_1**	**B_2**	**.....**	**B_j**	**.....**	**B_s**	**$n_{i.}$**	**$f_{i.}$**
A_1	**n11**	**n12**	**.....**	**n1j**	**.....**	**n1s**	**$n_{1.}$**	**$f_{1.}$**
A_2	**n21**	**n22**	**.....**	**n2j**	**.....**	**n2s**	**$n_{2.}$**	**$f_{2.}$**

. . .	. .	. .	. .... .	. .	. .... .	. .	. .	. .
A_i	**ni1**	**ni2**	**.....**	**nij**	**.....**	**nis**	**n_i .**	**fi .**
. . .	. . .	. . .	. .... .	. . .	. .... .	. . .	. . .	. . .
A_r	**nr1**	**nr2**	**.....**	**nrj**	**.....**	**nrs**	**n_r .**	**f_r .**
n. s	**n. 1**	**n. 2**	**.....**	**n. j**	**.....**	**n. s**	**N**	
f. s	**f. 1**	**f. 2**	**.....**	**f. j**	**.....**	**f. s**		**1**

tabla de contingencia r x s

n_{ij}= número de elementos de la distribución que presentan la modalidad i –ésima del atributo A y la modalidad j – esima del atributo B.

$n_{i.}$= n_{i1+} n_{i2+} ... + n_{is} --→ número de elementos de la distribución con la i – ésima modalidad del atributo A.

Como a las variables cualitativas no se les puede someter a operaciones de sumas, restas y divisiones, al venir expresadas en escalas nominales u ordinales no tiene sentido hablar de medias marginales, condicionadas, varianzas, etc; si podríamos calcular la moda en el caso de que se empleara una escala nominal y de la mediana si utilizamos escalas ordinales.

3.3. Dependencia e independencia

3.3.1.Independencia

Cuando no se da ningún tipo de relación entre 2 variables o atributos, diremos que son independientes.
Dos variables X e Y, son independientes entre si, cuando una de ellas no influye en la distribución de la otra condicionada por el valor que adopte la primera. Por el contrario existirá dependencia cuando los valores de una distribución condicionan a los de la otra.
Dada dos variables estadísticas X e Y, la condición necesaria y suficiente para que sean independientes es:

$$\frac{n_{ij}}{N} = \frac{n_{i.}}{N}\frac{n_{.j}}{N} \forall i, j$$

Propiedades:

1ª) Si X es independiente de Y, las distribuciones condicionadas de X/Y_j son idénticas a la distribución marginal de X.
2ª) Si X es independiente de Y, Y es independiente de X.
3ª) Si X e Y son 2 variables estadísticamente independientes, su covarianza es cero. La recíproca de esta propiedad no es cierta, es decir, la covarianza de 2 variables puede tomar valor cero, y no ser independientes.

3.3.2.Dependencia funcional (existe una relación matemática exacta entre ambas variables)

El carácter X depende del carácter Y, si a cada modalidad y_j de Y corresponde una única modalidad posible de X. Por lo tanto cualquiera que sea j, la frecuencia absoluta n_{ij} vale cero salvo para un valor de i correspondiente a una columna j tal que $n_{ij} = n.j$

Cada columna de la tabla de frecuencias tendrá, por consiguiente, un único término distinto de cero. Si a cada modalidad x_i de X corresponde una única modalidad posible de Y, será Y dependiente de X. La dependencia de X respecto de Y no implica que Y dependa de X.

Para que la dependencia sea recíproca, los caracteres X e Y deben presentar el mismo número de modalidades (debe ser n=m) y en cada fila como en cada columna de la tabla debe haber uno y solo un término diferente de cero.

Sea X el salario de un empleado e Y la antigüedad del mismo en la empresa

X \ Y	**1**	**3**	**5**	**7**	**9**
100	15	0	0	0	0
120	0	20	0	0	0
140	0	0	30	0	0
160	0	0	0	25	0
180	0	0	0	0	10

Dependencia funcional recíproca: X depende de Y e Y depende de X

X \ Y	**1**	**3**	**5**	**7**	**9**	**10**
100	15	0	0	0	0	0
120	0	20	0	0	0	0
140	0	0	30	0	12	0
160	0	0	0	25	0	0
180	0	0	0	0	0	9

Y depende de X pero X no depende de Y

3.3.3.Dependencia estadística (existe una relación aproximada)

Existen caracteres que ni son independientes, ni se da entre ellos una relación de dependencia funcional, pero si se percibe una cierta relación de dependencia entre ambos; se trata de una dependencia estadística.
Cuando los caracteres son de tipo cuantitativo, el estudio de la dependencia estadística se conoce como el problema de " regresión ", y el análisis del grado de dependencia que existe entre las variables se conoce como el problema de correlación.

3.4.Regresión y correlación lineal simple

3.4.1.Introducción a la regresión lineal simple

Cuando se estudian dos características simultáneamente sobre una muestra, se puede considerar que una de ellas influye sobre la otra de alguna manera. El objetivo principal de la regresión es descubrir el modo en que se relacionan.

Por ejemplo, en una tabla de pesos y alturas de 10 personas

Altura	175	180	162	157	180	173	171	168	165	165
Peso	80	82	57	63	78	65	66	67	62	58

se puede suponer que la variable "Altura" influye sobre la variable "Peso" en el sentido de que pesos grandes vienen explicados por valores grandes de altura (en general).

De las dos variables a estudiar, que vamos a denotar con X e Y, vamos a llamar a la X VARIABLE INDEPENDIENTE o EXPLICATIVA, y a la otra, Y, le llamaremos VARIABLE DEPENDIENTE o EXPLICADA.

En la mayoría de los casos la relación entre las variables es mutua, y es difícil saber qué variable influye sobre la otra. En el ejemplo anterior, a una persona que mide menos le supondremos menor altura y a una persona de poca altura le supondremos un peso más bajo. Es decir, se puede admitir que cada variable influye sobre la otra de forma natural y por igual. Un ejemplo más claro donde distinguir entre variable explicativa y explicada es aquel donde se anota, de cada alumno de una clase, su tiempo de estudio (en horas) y su nota de examen. En este caso un pequeño tiempo de estudio tenderá a obtener una nota más baja, y una nota buena nos indicará que tal vez el alumno ha estudiado mucho. Sin embargo, a la hora de determinar qué variable explica a la otra, está claro que el "tiempo de estudio" explica la "nota de examen" y no al contrario, pues el alumno primero estudia un tiempo que puede decidir libremente, y luego obtiene una nota que ya no decide arbitrariamente. Por tanto,

X = Tiempo de estudio (variable explicativa o independiente)
Y = Nota de examen (variable explicada o dependiente)

El problema de encontrar una relación funcional entre dos variables es muy complejo, ya que existen infinidad de funciones de formas distintas. El caso más sencillo de relación entre dos variables es la relación LINEAL, es decir que

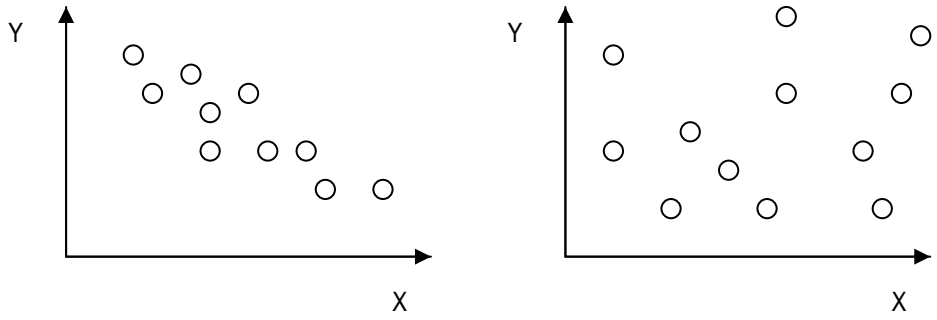
$$\mathbf{Y = a + b\,X}$$

(es la ecuación de una recta) donde a y b son números, que es el caso al que nos vamos a limitar.

Cualquier ejemplo de distribución bidimensional nos muestra que la relación entre variables NO es EXACTA (basta con que un dato de las X tenga dos datos distintos de Y asociados, como en el ejemplo de las Alturas y Pesos, que a 180 cm. de altura le correspondía un individuo de 82 kg. y otro de 78 kg.).

- Diagrama de dispersión o nube de puntos

En un problema de este tipo, se observan los valores (x_i,y_j) y se representan en un sistema de ejes coordenados, obteniendo un conjunto de puntos sobre el plano, llamado " diagrama de dispersión o nube de puntos ".

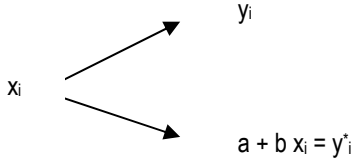


En los diagramas de arriba se puede observar cómo en el de la izquierda, una línea recta inclinada puede aproximarse a casi todos los puntos, mientras que en el otro, cualquier recta deja a muchos puntos alejados de ella. Así pues, el hacer un análisis de regresión lineal sólo estaría justificado en el ejemplo de la izquierda.

Como se puede ver en ambos diagramas, ninguna recta es capaz de pasar por todos los puntos, y seguir siendo recta. De todas las rectas posibles, la RECTA DE REGRESIÓN DE Y SOBRE X es aquella que minimiza un cierto error, considerando a X como variable explicativa o independiente y a Y como la explicada o dependiente.

- Recta de mínimos cuadrados o recta de regresión de Y sobre X ($y^* = a + b\,x$)

Sea $y = a + b\,x$ una recta arbitraria. Para cada dato de X, es decir, para cada x_i de la tabla tenemos emparejado un dato de Y llamada y_i, pero también tenemos el valor de sustituir la x_i en la ecuación de la recta, al que llamaremos y^*_i.



Cuando se toma el dato x_i, el error que vamos a considerar es el que se comete al elegir y^*_i en lugar del verdadero y_i .Se denota con e_i y vale

$e_i = y_i - y^*_i$

Esos errores pueden ser positivos o negativos, y lo que se hace es escoger la recta que minimice la suma de los cuadrados de todos esos errores, que es la misma que la que minimiza la varianza de los errores.

Usando técnicas de derivación se llega a que, de todas las rectas y = a + b x, con a y b números arbitrarios, aquella que minimiza el error elegido es aquella que cumple

$$a = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \qquad \text{y} \qquad b = \frac{s_{xy}}{s_x^2} \qquad \text{por lo tanto} \quad a = \bar{y} - b\bar{x}$$

Así pues, sustituyendo en y = a + b x, la ecuación de la recta de regresión de Y sobre X es

$$y* = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \right) + \left(\frac{s_{xy}}{s_x^2} \right) \cdot x \qquad \text{es decir} \qquad y^* = a + b\bar{x}$$

y recolocando los términos se puede escribir de la forma

$$\boxed{y - \bar{y} = \frac{s_{xy}}{s_x^2} \cdot (x - \bar{x})}$$

- Recta de regresión de X sobre Y

Si se hubiese tomado Y como variable independiente o explicativa, y X como dependiente o explicada, la recta de regresión que se necesita es la que minimiza errores de la X. Se llama RECTA DE REGRESIÓN DE X SOBRE Y y se calcula fácilmente permutando los puestos de x e y, obteniéndose

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} \cdot (y - \bar{y}) \quad \textbf{es decir} \qquad x^* = a' + b' y$$

Sabiendo que : $b' = \frac{S_{xy}}{S_y^2}$ y que $a' = \bar{x} - b'\bar{y}$

PROPIEDADES:

- Ambas rectas de regresión pasan por el punto ($\bar{x}, \bar{y}$)
- La pendiente de la recta de regresión de Y sobre X es " b " y la de X sobre Y es " b ´ ". Dado que las varianzas son positivas por definición, el signo de las pendientes será el mismo que el de la covarianza, y así, las rectas serán ambas crecientes o decrecientes, dependiendo de si la covarianza es positiva o negativa, respectivamente, es decir b y b´ tendrán el mismo signo.

- Los términos de las rectas a y a´ constituyen los orígenes de las rectas, es decir, son los valores que adoptan respectivamente y* ó x* cuando x o y toman el valor cero en sus correspondientes rectas de regresión.
- Las rectas de regresión las emplearemos para realizar predicciones acerca de los valores que adoptaran las variables.
- Puede darse el caso, de no existencia de correlación lineal entre las variables, lo cual no implica que no existan otro tipo de relaciones entre las variables estudiadas: relación exponecial, relación parabólica, etc.

3.4.2.Correlación lineal simple (r ó R)

Para ver si existe relación lineal entre dos variables X e Y, emplearemos un parámetro que nos mida la fuerza de asociación lineal entre ambas variables. La medida de asociación lineal mas frecuentemente utilizada entre dos variables es " r " o coeficiente de correlación lineal de Pearson; este parámetro se mide en términos de covarianza de X e Y.

$$\boxed{R = \frac{S_{xy}}{S_x S_y}} \qquad -1 \le R \le 1$$

- Si R = 1, existe una correlación positiva perfecta entre X e Y
- Si R = -1, existe una correlación negativa perfecta entre X e Y
- Si R = 0, no existe correlación lineal, pudiendo existir otro tipo de relación
- Si $-1 \prec R \prec 0$, existe correlación negativa y dependencia inversa, mayor cuanto más se aproxime a - 1.
- Si $0 \prec R \prec 1$, existe correlación positiva, y dependencia directa, mayor cuanto más se aproxime a 1.

- **<u>Varianza residual y varianza explicada por la regresión. Coeficiente de determinación lineal (R^2)</u>**

Si tenemos dos variables X e Y relacionadas linealmente, parte de la variabilidad de la variable Y, vendrá explicada por variaciones de X (variabilidad explicada por el modelo) , mientras que el resto responderá a variaciones de fenómenos relacionados con la variable Y o con el azar (variabilidad no explicada por el modelo) .

Por tanto nos conviene disponer de una medida que indique el porcentaje de la variabilidad de la variable explicada que se debe a la variabilidad de la variable explicativa. Esta medida es el coeficiente de determinación lineal (R^2) , y si su valor es alto nos indicará que el ajuste lineal efectuado es bueno.

En la regresión lineal de Y sobre X, la varianza de la variable Y, puede descomponerse en la suma de 2 varianzas:

$$S_y^2 = S_r^2 + S_e^2$$

donde:

S_y^2 es la varianza total de la variable Y

S_r^2 es la varianza explicada o variabilidad de Y explicada por la **r**egresión. $\boxed{S_r^2 = bS_{xy}}$

S_e^2 es la varianza residual (**e**) o variabilidad de Y no explicada por la **r**egresión.

$$\boxed{S_e^2 = S_y^2 - bS_{xy}}$$

$$\boxed{R^2 = \frac{S_r^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2}} \qquad 0 \le R^2 \le 1$$

$$\boxed{R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}}$$ también podemos afirmar que $$\boxed{R^2 = b \cdot b'}$$

Es una medida de la bondad del ajuste lineal efectuado. Si lo expresamos en porcentaje, dicho coeficiente nos indica el % de la varianza de la variable explicada (Y) que se ha conseguido explicar mediante la regresión lineal.

Si $R^2 = 1$, existe dependencia funcional; la totalidad de la variabilidad de Y es explicada por la regresión.
Si $R^2 = 0$, dependencia nula; la variable explicativa no aporta información válida para la estimación de la variable explicada.
Si $R^2 \ge 0.75$, se acepta el modelo ajustado

- Relación existente entre los coeficientes de determinación y correlación lineal:

$$\boxed{R = \pm\sqrt{R^2}}$$

El signo del coeficiente de correlación lineal será el mismo que el de la covarianza.

3.5. **Estudio de la asociación entre variables cualitativas**.

En el estudio visto de regresión y correlación se ha tratado solo el caso de variables cuantitativas (ingresos, salarios, precios, etc) Con variables de tipo cualitativo se puede construir tablas de contingencia, a través de las cuales se puede estudiar la independencia estadística entre los distintos atributos.

Si dos atributos son dependientes, se pueden construir una serie de coeficientes que nos midan el grado asociación o dependencia entre los mismos.

Partimos de la tabla de contingencia en la que existen r modalidades del atributo A y s del atributo B. El total de observaciones será:

$$N = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

La independencia estadística se dará entre los atributos si : $\frac{n_{ij}}{N} = \frac{n_{i.}}{N} \cdot \frac{n_{.j}}{N}$ para todo i , j ; si esta expresión no se cumple, se dirá que existe un grado de asociación o dependencia entre los atributos.

$$\frac{n_{ij}}{N} \neq \frac{n_{i.}}{N} \cdot \frac{n_{.j}}{N} \quad \text{----} \rightarrow n_{ij}' = \frac{n_{i.} \cdot n_{.j}}{N}$$

El valor n'_{ij} es la frecuencia absoluta conjunta teórica que existiría si los 2 atributos fuesen independientes.
El valor n_{ij} es la frecuencia absoluta conjunta observada.
El coeficiente de asociación o contingencia es el llamado Cuadrado de Contingencia, que es un indicador del grado de asociación:

$$\chi^2 = \sum_i \sum_j \frac{(n'_{ij} - n_{ij})^2}{n'_{ij}} \text{ siendo } n'_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N}$$

El campo de variación va desde cero (cuando existe independencia y $n'_{ij} = n_{ij}$), hasta determinados valores positivos, que dependerá de las magnitudes de las frecuencias absolutas que lo componen.

Este inconveniente de los límites variables se eliminará con el empleo del Coeficiente de contingencia de Pearson:

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

Varía entre cero y uno. El valor cero se dará en el caso de independencia ($n'_{ij} = n_{ij}$). Cuanto más se aproxime a 1 más fuerte será el grado de asociación entre los dos atributos.

- Estudio de la asociación entre dos atributos

- Para tablas de contingencia 2 x 2

Sean A y B dos variables cualitativas o atributos tales que presentan 2 modalidades cada una. La tabla de contingencia correspondiente es la siguiente:

A \ B	B_1	B_2	
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
	$n_{,1}$	$n_{,2}$	N

A y B son independientes si: $n_{11} = \frac{n_{1.} n_{.1}}{N}$

A las expresiones $\frac{n_{i.} n_{.j}}{N}$ se les denomina frecuencias esperadas y se denotan por n'_{ij} o por E_{ij}

Si finalmente podemos concluir que los dos atributos están asociados, se pueden plantear dos preguntas:
1ª) ¿ Cual es la intensidad de la asociación entre los dos atributos ?
2ª) ¿ Cual es la dirección de la asociación detectada ?

- Asociación perfecta entre dos atributos

Ocurre cuando, al menos, una de las modalidades de uno de los atributos queda determinada por una de las modalidades del otro atributo. Esto ocurre cuando existe algún cero en la tabla 2 x 2. La asociación perfecta puede ser:
a) Asociación perfecta y estricta.
Ocurre cuando dada modalidad de uno de los atributos queda inmediatamente determinada la modalidad del otro. Es decir, cuando $n_{11} = n_{22} = 0$ ó $n_{12} = n_{21} = 0$

Ejemplo:

A= tipo de trabajo (temporal ó indefinido)
B= Sexo (hombre ó mujer)

Sexo \ Tipo trabajo	Temporal	Indefinido
Hombre	20	0
Mujer	0	80

Con estos datos sabemos que si un individuo es hombre el tipo de trabajo sera temporal y si es mujer su contrato será indefinido.

- Asociación perfecta e implicita de tipo 2

Ocurre cuando:
1º) Si se toma la modalidad de un atributo queda determinada la modalidad del otro atributo al que pertenece la observación.

2º) Si se toma la otra modalidad, no queda determinada la modalidad del otro atributo al que pertenece la observación.

Es decir, esta asociación se produce cuando alguna de las frecuencias observada es cero.

Ejemplo:

A= tipo de trabajo (temporal ó indefinido)
B= Sexo (hombre ó mujer)

Sexo \ Tipo trabajo	Temporal	Indefinido
Hombre	5	15
Mujer	0	80

- Si la persona observada es mujer sabremos que su contrato es indefinido; si es varón puede ser indefinido o temporal.
- Si el contrato analizado es temporal pertenecerá a un hombre; si es un contrato indefinido, podrá ser de un hombre o una mujer.

- También podemos delimitar si la asociación es positiva o negativa:

- Asociación positiva

Cuando se verifica que :

a) La modalidad 1 del atributo A está asociada a la modalidad 1 del atributo B
b) La modalidad 2 del atributo A está asociada a la modalidad 2 del atributo B.

- Asociación negativa:

Cuando se verifica que:

a) La modalidad 1 del atributo A está asociada a la modalidad 2 del atributo B
b) La modalidad 2 del atributo A esta asociada a la modalidad 1 del atributo A.

Para medir el sentido de la asociación entre dos atributos emplearemos el indicador Q de Yule:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

$$\boxed{-1 \le Q \le 1}$$

Si Q = 0, entonces existe independencia
Si Q > 0, entonces existe asociación positiva
Si Q < 0, entonces existe asociación negativa

- Tablas de contingencia R x S

Para determinar la intensidad de dicha asociación, calculamos la V de Cramer, que se define como:

$$V = \sqrt{\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}}{Nm}}$$

m = min (r-1 , s-1)

$V \in (0,1)$

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

Existirá una mayor intensidad en la asociación entre 2 variables a medida que el indicador adopte valores próximos a 1.

Capítulo IV	NÚMEROS ÍNDICES

Aplicación

Existe un gran número de fenómenos económicos cuyo significado y estudio alcanza distintos niveles de complejidad (son los que se conocen como coyuntura económica, nivel de inflacción, nivel de desarrollo, etc.).
Los números índice constituyen el instrumental más adecuado para estudiar la evolución de una serie de magnitudes económicas que nos den respuesta a cuestiones tales como: ¿Es la coyuntura económica positiva o negativa? ¿Es el nivel de inflacción el adecuado o no? etc.

Definición

Un número índice puede definirse como una medida estadística que nos proporciona la variación relativa de una magnitud (simple o cmpleja) a lo largo del tiempo o el espacio.

Sea ***X*** una variable estadística cuya evolución se pretende estudiar.

Llamaremos:

- **Periodo inicial o base**, es aquel momento del tiempo sobre el que se va comparando la evolución de la magnitud o variable estadística X_0.
- **Periodo de comparación,** es aquel momento del tiempo en el que el valor de la magnitud X_t se compara con el del periodo base.

El indice de evolución de ***0*** a ***t*** expresado en %: $$I_0^t = \frac{X_t}{X_0}\ 100$$

- I_0^t toma el valor 100 en el periodo base
- I_0^t < 100 implica que $X_t < X_0$ (indicando una evolución negativa del periodo ***0*** al ***t***)
- I_0^t > 100 implica que $X_t > X_0$ (indicando una evolución negativa del periodo ***0*** al ***t***)

- La elaboración de números índices tiene sentido en variables de naturaleza cuantitativa
- El índice, por estar definido por un cociente, es independiente de las unidades de medida en las que venga expresada la variable, con lo que se puedan efectuar agregaciones de distintos índices, construyéndose indicadores de evolución general de fenómenos económicos.

Los números índices se clasifican atendiendo a:

• La naturaleza de las magnitudes que miden	**Simples** **Complejas**
• En el segundo caso, a la importancia relativa de cada componente	**Sin ponderar** **Ponderados**

2. Tipos de Indices

Números Indices Simples

Estudian la evolución en el tiempo de una magnitud que sólo tiene un componente (sin desagregación). Se emplean con gran difusión en el mundo de la empresa a la hora de estudiar las producciones y ventas de los distintos artículos que fabrican y lanzan al mercado.

I_0^t Número índice en el periodo ***t*** de la magnitud ***X***
X_t Valor de la magnitud ***X*** en el periodo ***t***
X_0 Valor de la magnitud ***X*** en el periodo base ***0***

$$I_0^t = \frac{X_t}{X_0}\ 100$$

Números Indices Complejos Sin Ponderar

Estudian la evolución en el tiempo de una magnitud que tiene varios componentes y a los cuales se asigna la misma importancia o peso relativo (siendo esta última hipótesis nada realista).
Por su naturaleza son de poco uso en el mundo de la economía.

- Sea una magnitud compleja agregada de ***N*** componentes (1,2,..., ***N***)
- Sean I_{i0}^t los números índices simples de cada componente ***i*** en el periodo ***t***

I_0^t Número índice total en el periodo ***t*** de la magnitud agregada
I_{i0}^t Número índice simple del componente ***i*** en el periodo ***t***
X_{it} Valor del componente ***i*** en el periodo ***t***
X_{i0} Valor del componente ***i*** en el periodo base ***0***

$$I_0^t = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} I_{i0}^t = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \frac{X_{it}}{X_{i0}}\ 100$$

Números Indices Complejos Ponderados

Estudian la evolución en el tiempo de una magnitud que tiene varios componentes y a los cuales se asigna un determinado coeficiente de ponderación w_i. Son los que realmente se emplean en el análisis de la evolución de fenómenos complejos de naturaleza económica (IPC, IPI, etc.)

- Sea una magnitud compleja agregada de ***N*** componentes (1,2,..., ***N***)
- Sean I_{i0}^t los números índices simples de cada componente ***i*** en el periodo ***t***
- Sean (w_1, w_2, ...,w_N) los coeficientes de ponderación de los componentes

I_0^t Número índice total en el periodo ***t*** de la magnitud agregada
I_{i0}^t Número índice simple del componente ***i*** en el periodo ***t***
X_{it} Valor del componente ***i*** en el periodo ***t***
X_{i0} Valor del componente ***i*** en el periodo base ***0***
w_i Coeficiente de ponderación del componente ***i***

$$I_0^t = \frac{\sum_{i=1}^{\mathrm{N}} I_{i0}^t\ w_i}{\sum_{i=1}^{\mathrm{N}} w_i} = \frac{\sum_{i=1}^{\mathrm{N}} \frac{X_{it}}{X_{i0}}\ 100\ w_i}{\sum_{i=1}^{\mathrm{N}} w_i}$$

Propiedades que cumplen en general los índices simples (pero no todos los complejos)	
Existencia: el índice debe concretarse en un valor real y finito y:	$I_0^t \neq 0$
Identidad: si coinciden el periodo base y el de comparación, entonces:	$I_0^t = 100$
Inversión: el producto de dos índices invertidos de dos periodos es:	$I_0^t \, I_t^0 = 1$
Circular: la generalización de la inversión para varios periodos	$I_0^{t'} \, I_{t'}^t \, I_t^0 = 1$
Proporcionalidad: Si la magnitud varía en proporcion ***1+K***, el índice:	$X'^t_0 = X_0^t + \mathrm{K}\, X_0^t$ $I'^t_o = I_o^t + \mathrm{K}\, I_o^t$

3. Indices de Precios

Estos índices miden la evolución de los precios a lo largo del tiempo.
Los índices de precios se clasifican en:

	Simples		Sauerbeck
Números índice de precios		Sin ponderar	Bradstreet-Dutot
	Complejos		
		Ponderados	Laspeyres
			Paasche
			Edgeworth
			Fisher

3.1. Indices Simples de Precios

P_t Indice simple de precios en el periodo ***t***
p_0 Precio en el periodo base ***0***
p_t Precio en el periodo ***t***

$$P_0^t = \frac{p_t}{p_0} \, 100$$

3.2. Indices Complejos de Precios Sin Ponderar

a) Indice media aritmética de índices simples o de Sauerbeck
Es la media aritmética de los índices de precios simples para cada componente.

P_S Indice de Sauerbeck

$$P_S = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} P_{i0}^t = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \frac{p_{it}}{p_{i0}} \, 100$$

b) Indice media agregativa simple o de Bradstreet-Dutot
Es el cociente entre la media aritmética simple de los N precios en el periodo ***t*** y en el ***0***.

P_{BD} Indice de Bradstreet-Dutot $$P_{BD} = \frac{\frac{1}{N}\sum_{i=1}^{N} p_{it}}{\frac{1}{N}\sum_{i=1}^{N} p_{i0}} 100 = \frac{\sum_{i=1}^{N} p_{it}}{\sum_{i=1}^{N} p_{i0}} 100$$

3.3. Indices Complejos de Precios Ponderados

Cada uno de los métodos apunta una forma distinta para establecer los coeficientes de ponderación w_i $$P_X = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} w_i}{\sum_{i=1}^{N} w_i} 100$$

a) Indice de precios de Laspeyres P_L
Utiliza como coeficientes de ponderación el valor de las transacciones en el periodo base.
Tiene la ventaja de que las ponderaciones del periodo se mantienen fijas para todos los periodos pero por contra el inconveniente de que su representatividad disminuye según nos alejamos.
$w_i = p_{i0}\ q_{i0}$

b) Indice de precios de Paasche P_P
Utiliza como coeficientes de ponderación el valor de las transacciones, con las cantidades del periodo de comparación y los precios del periodo base. Las ponderaciones son por ello variables.
Tiene la ventaja de que los pesos relativos de los distintos componentes se actualizan cada periodo con el agravante de complejidad y costes derivados de este cálculo.
$w_i = p_{i0}\ q_{it}$

c) Indice de precios de Edgeworth P_E
Utiliza como coeficientes de ponderación la suma de los dos anteriores.
$w_i = p_{i0}\ q_{i0} + p_{i0}\ q_{it}$

d) Indice de precios de Fisher P_F
Se define como la media geométrica de los indices de Laspeyres y Paasche
$P_F = \sqrt{P_L\ P_P}$

4. Indices de Cantidades o Cuánticos

Los índices cuánticos miden la evolución de cantidades a lo largo del tiempo.
Para cualquier magnitud y por supuesto las cantidades, siempre se pueden elaborar números índices simples, complejos sin ponderar y complejos ponderados empleando las mismas consideraciones y la misma formulación que la vista para índices de precios.

Cada uno de los métodos apunta una forma distinta para establecer los coeficientes de ponderación w_i

$$Q_X = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}} w_i}{\sum_{i=1}^{N} w_i} 100$$

a) Indice cuántico de Laspeyres Q_L $w_i = q_{i0}\ p_{i0}$

b) Indice cuántico de Paasche Q_P $w_i = q_{i0}\ p_{it}$

c) Indice cuántico de Edgeworth Q_E $w_i = q_{i0}\ p_{i0} + q_{i0}\ p_{it}$

d) Indice cuántico de Fisher Q_F $Q_F = \sqrt{Q_L\ Q_P}$

5. Propiedades de los Indices Complejos

Los números índices deben cumplir una serie de propiedades ideales: existencia, identidad, inversión, circular y de proporcionalidad. Así como los índices simples las cumplen en su mayoría, los complejos y ponderados no cumplen algunas de ellas.

Sobre los dos índices más importantes, los de Laspeyres y Paasche podemos decir:
- **Cumplen**: la existencia, identidad y proporcionalidad
- **No cumplen**: la inversión y por tanto tampoco la circular

El índice de Laspeyres (tanto de precios como cuántico) es el más utilizado en los indicadores generales de precios y producción. Su diseño y posterior cálculo requiere una rigurosa selección de sus componentes y ponderaciones.
Ahora bien, a medida que nos alejamos del periodo base, la estructura de coeficientes de ponderación de este índice (y de los demás) es cada vez menos representativa con lo que es necesario fijar un nuevo periodo base y establecer una nueva estructura de ponderaciones.

5.1. Cambio de Base en una Serie de Números Indices

En los enlaces de series de números índices que tienen distinta base, nos apoyamos en la propiedad de inversión que, como se ha indicado, no la cumple el índice de Laspeyres, pero que se actua en la práctica como si se cumpliera, ante la necesidad de efectuar dichos enlaces.

Sea una serie de números índices cuyo periodo base es ***o***: $I_0^0, I_0^1, I_0^2, \ldots, I_0^n$
Puede interesar cambiar la base ***o*** si está muy alejada en el periodo ***t*** de comparación.

Para ello no es necesario efectuar un profundo estudio para determinar nuevos coeficientes de ponderación (en el caso de índices complejos) sino únicamente apoyarnos en la propiedades de inversión y circular que nos permiten obtener el coeficiente técnico que transforma la serie dada en una nueva con un periodo base distinto ***t'***.

a) Valor de cada índice en la nueva base *t'*

Para el periodo ***t'*** existirá un índice $I_0^{t'}$ que sirve de enlace técnico para transformar la serie.
Se cumple que para expresar cada índice antiguo en la nueva base:

$$I_0^{t'}\, I_{t'}^{t}\, I_t^0 = 1 \quad \Rightarrow \quad I_0^{t'}\, I_{t'}^{t} = \frac{1}{I_t^0} = I_0^t \quad \Rightarrow \qquad\qquad I_{t'}^{t} = \frac{I_0^t}{I_0^{t'}}\ 100$$

b) Cálculo del coeficiente de transformación

Este coeficiente, multiplicado por cada índice antiguo permite obtener el índice en la nueva base

$$I_{t'}^0 = \frac{I_0^0}{I_0^{t'}} \ \Rightarrow$$

Coeficiente de Transformación: $I_{t'}^0 = \dfrac{100}{I_0^{t'}}$

5.2. Renovación o Enlace de Series de Números Indices con Distintas Bases

La necesidad de la renovación periódica de una serie de números índices nos puede llevar a contar con dos series que tienen períodos base distintos y hay que enlazarlos o empalmarlos para poder estudiar el fenómeno, comparando su evolución con una única base.

El periodo base que se mantiene es el de la serie que lo tiene más cercano al momento actual, aplicando el **coeficiente de enlace o empalme** a la serie más antigua.
El concepto o definición de este coeficiente de enlace es el mismo empleado en los cambios de base dentro de la misma serie, aplicándose ahora sólo a los elementos de la serie que tengan la base más antigua.

Sea la serie de números índices más antigua en base ***o***: $I_0^0, I_0^1, I_0^2, \ldots, I_0^t, \ldots, I_0^n$
Sea ***t'*** el periodo de la serie antigua al que se quieren actualizar los índices de esta serie.
El coeficiente de enlace, multiplicado por cada índice de la serie antigua permite obtener el índice en la nueva base.

$$I_{t'}^0 = \frac{I_0^0}{I_0^{t'}} \ \Rightarrow$$

Coeficiente de Enlace : $I_{t'}^0 = \dfrac{100}{I_0^{t'}}$

6. Indices de Valor y Deflactación de Series Económicas

En economía, los bienes y servicios producidos son adquiridos por las familias, empresas, etc. Estos bienes presentan gran heterogeneidad y para agregarlos hay que someterlos a un proceso de homogeneización a través de la obtención de su **valor**, aplicando un sistema de precios.

El proceso de multiplicar (cantidades ***x*** precios respectivos) de los distintos componentes transforma cantidades físicas heterogéneas (leche, pescado, etc.) en **valores económicos** homogéneos (pesetas, dolares, etc.)

Los índices de valor nos permiten estudiar la evolución a lo largo del tiempo de la cuantificación monetaria de un conjunto de bienes. Este valor se llama nominal o **en pesetas corrientes** o de cada año cuando los precios son los del periodo de comparación.

Valor en el periodo base ***0*** : $V_0 = \sum_{i=1}^{N} p_{i0}\, q_{i0}$

Valor en el periodo ***t*** : $V_t = \sum_{i=1}^{N} p_{it}\, q_{it}$

El índice complejo de valor, para ***N*** componentes: $I_{V_t} = \dfrac{\sum_{i=1}^{N} p_{it}\ q_{it}}{\sum_{i=1}^{N} p_{i0}\ q_{i0}}$

La evolución de los índices a lo largo del tiempo está motivado, según la expresión anterior, por variaciones conjuntas de precios y cantidades, no pudiendo aislarse la influencia de cada una.
En economía interesa analizar la evolución del conjunto de ***N*** mercancías bajo lo que se denomina **a precios constantes**, es decir, sin que se produzcan variaciones en los precios de los distintos componentes. Para hacerlo se realiza la operación denominada **deflactación** de series de valores expresadas en precios o pesetas corrientes de cada año.

Para comparar el valor de un conjunto de bienes en dos periodos distintos interesa aislarlo de la subida, **inflación**, o de la bajada, **deflación**, de sus respectivos precios. De esta manera, se consigue aislar el cálculo de la distorsión que las subidas de precios, que no sean debidas a una mejora en la calidad de los bienes y los servicios.

Para poder efectuar un análisis comparativo de una serie de valor entre distintos periodos, hay que pasarla de pesetas corrientes o de cada año a pesetas constantes o del periodo que se considere como base. Esto es lo que se denomina **deflactar** la serie, dividiéndola por el índice de precios que se considere más adecuado llamado **deflactor** de la serie.

Los índices de Laspeyres y Paasche son los que más se utilizan como deflactores de series.

a) Deflacción por un Indice de Laspeyres

Si el valor a precios corrientes se divide por un índice de precios de Laspeyres tendremos:

$$\frac{V_t}{P_L} = \frac{\sum_{i=1}^{N} p_{it}\ q_{it}}{\frac{\sum_{i=1}^{N} p_{it}\ q_{i0}}{\sum_{i=1}^{N} p_{i0}\ q_{i0}}} = \sum_{i=1}^{N} p_{it}\ q_{i0} \frac{\sum_{i=1}^{N} p_{it}\ q_{it}}{\sum_{i=1}^{N} p_{it}\ q_{i0}} = V_0\ Q_P$$

Al deflactar una serie de valor a precios corrientes por un índice de precios de Laspeyres no se obtiene una serie de valor a precios constantes sino $V_0\ Q_p$ que es el producto del valor en el año base por un índice cuántico de Paasche. El P_L no es por tanto un verdadero deflactor, aunque en la práctica se considera como tal por ser el índice que se suele elaborar.

b) Deflacción por un Indice de Paasche

Este si es un verdadero deflactor ya que la expresión nos da el valor actual de un conjunto de N mercancías a precios constantes del año p_{i0} base:

$$\frac{V_t}{P_P} = \frac{\sum_{i=1}^{N} p_{it}\ q_{it}}{\frac{\sum_{i=1}^{N} p_{it}\ q_{it}}{\sum_{i=1}^{N} p_{i0}\ q_{it}}} = \sum_{i=1}^{N} p_{i0}\ q_{it}$$

Según sea la serie económica que se desea deflactar, así habrá que elegir el índice de precios más adecuado:

- Para deflactar la renta disponible de una familia en pesetas constantes de un determinado año, el deflactor adecuado será el IPC (Indice de Precios de Consumo)
- Para deflactar una serie del valor de un conjunto de productos industriales, su deflactor adecuado será el IPI (Indice de Precios Industriales)

7. Indice de Precios de Consumo (IPC)

- FICHA TÉCNICA

- Tipo de encuesta: continua de periodicidad mensual
- Período base: 2001
- Periodo de referencia de las ponderaciones: desde el 2º trimestre de 1999 hasta el 1º de 2001
- Muestra de municipios: 141 para alimentación y 97 para el resto

- Número de artículos: 484
- Número de observaciones: aproximadamente 200.000 precios mensuales
- Clasificación funcional: 12 grupos, 37 subgrupos, 80 clases y 117 subclases; 57 rúbricas y 37 grupos especiales
- Método general de cálculo: Laspeyres encadenado
- Método de recogida: agentes entrevistadores en establecimientos y recogida centralizada para artículos especiales

Características principales

INDICE DE PRECIO DE CONSUMO (I.P.C.)

El **Indice de Precios de Consumo** es una medida estadística de la evolución del conjunto de precios de los bienes y servicios que consume la población residente en viviendas familiares en España.

El **Indice de Precios de Consumo Armonizado** es un indicador estadístico cuyo objetivo es proporcionar una medida común de la inflación que permita realizar comparaciones entre los países de la Unión Europea. Este indicador de cada país cubre las parcelas que superan el uno por mil del total de gasto de la cesta de la compra nacional.

Hacia un indicador más moderno y dinámico

Desde enero de 1993 ha estado vigente en España el Índice de Precios de Consumo **(IPC) Base 92**. La cesta de la compra, a partir de la cual se calcula el IPC, se obtiene básicamente del consumo de las familias en un momento determinado y debe actualizarse cada cierto tiempo. En este caso no sólo se renueva la cesta sino que se introducen novedades en la forma de calcular el IPC por lo que no estamos ante un «cambio de base» sino ante un «cambio de sistema», **Sistema de Indices de Precios Base 2001**, que entra en vigor con la publicación del IPC de enero de 2002.
Se ha conseguido un **indicador más dinámico**, ya que se podrán **actualizar las ponderaciones más frecuentemente** y se **adaptará mejor a la evolución del mercado**. Además, se podrán **incluir nuevos productos** en la cesta de la compra en el momento en que su consumo comience a ser significativo.
El nuevo Sistema también será técnicamente más moderno, ya que permitirá la inclusión inmediata de mejoras en la metodología que ofrezcan los distintos foros académicos y de organismos nacionales e internacionales, especialmente las decisiones que favorezcan la armonización de los IPC de los países de la Unión Europea.

A partir del **22 de febrero de 2002**, el INE ha llevado a cabo la implantación definitiva del nuevo sistema de Índices de Precios de Consumo Base 2001.
El nuevo Sistema se ha llevado a efecto en **dos fases**, a lo largo de dos años. En la primera (hasta 01-2001) se introdujeron algunas mejoras, que se han completado en 01-2002. Se ha conseguido un indicador más dinámico, ya que se podrán actualizar las ponderaciones más frecuentemente y se adaptará mejor a la evolución del mercado. Además, se podrán incluir nuevos productos en la cesta de la compra en el momento en que su consumo comience a ser significativo.

Introducción

El Índice de Precios de Consumo (IPC) requiere para su elaboración la selección de una muestra de bienes y servicios representativa de los distintos comportamientos de consumo de la población, así como la estructura de ponderaciones que defina la importancia de cada uno de estos productos. Como en la mayoría de los países, el IPC español obtiene esta información de la Encuesta de Presupuestos Familiares (EPF), que fue realizada por última vez en el periodo comprendido entre abril de 1990 y marzo de 1991; esta encuesta es la que se utilizó para llevar a cabo el anterior cambio de base del IPC.

Desde entonces, el comportamiento de los consumidores ha cambiado considerablemente, ya sea porque variaron los gustos o las modas, su capacidad de compra, o porque han aparecido nuevos productos en el mercado hacia los que se desvía el gasto. Todos estos cambios deben reflejarse en la composición del IPC y en su estructura de ponderaciones; es por ello por lo que se hace preciso realizar un cambio de base que permita una mejor adaptación de este indicador a la realidad económica actual.

A partir del 2º trimestre de 1997 se implantó la nueva Encuesta Continua de Presupuestos Familiares (ECPF), con el fin de sustituir a la que se venía realizando de forma trimestral y a la Encuesta Básica que se hacía en periodos de entre ocho y nueve años, que era la utilizada para los distintos cambios de base del IPC.

Esta nueva encuesta permite disponer de información sobre el gasto de las familias de forma más detallada que su predecesora y con una periodicidad mayor que la Encuesta Básica. Esto hace que el nuevo Sistema del IPC, cuyas líneas generales se presentan en este documento, parta de un planteamiento conceptual diferente a todos los Sistemas anteriores.

Por un lado, destaca su **dinamismo**, ya que se podrán actualizar las ponderaciones en periodos cortos de tiempo, lo que sin duda redundará en una mejor y más rápida adaptación a la evolución del mercado. Además, esta adaptación a la evolución del mercado y al comportamiento de los consumidores se conseguirá también con la posibilidad de **incluir nuevos productos** en el momento en que su consumo comience a ser significativo.

Por otro lado, el nuevo Sistema será técnicamente más **moderno**, ya que permitirá la inclusión inmediata de mejoras en la metodología que ofrezcan los distintos foros académicos y de organismos nacionales e internacionales. En este sentido, se valorarán especialmente las decisiones provenientes del Grupo de Trabajo para la armonización de los IPC de la Unión Europea (UE).

Con este propósito, se creará un **proceso de actualización continua de la estructura de consumo**, basado en un flujo continuo de información entre el IPC y la ECPF, como fuente fundamental de información.

Características más importantes

Período base

El período base es aquél para el que la media aritmética de los índices mensuales se hace igual a 100. El año 2001 es el periodo base del nuevo Sistema, esto quiere decir que todos los índices que se calculen estarán referidos a este año.

Período de referencia de las ponderaciones

Es el período al que están referidas las ponderaciones que sirven de estructura del Sistema; dado que éstas se obtienen de la ECPF, el período de referencia del IPC es el período durante el cual se desarrolla esta encuesta.

El actual cambio de Sistema se ha realizado con la información proveniente de la ECPF, que proporciona la información básica sobre gastos de las familias en bienes y servicios de consumo. Así, el periodo de referencia del nuevo Sistema es el comprendido entre el 2º trimestre de 1999 y el 1º trimestre de 2001.

Para el cálculo de las ponderaciones se ha dado más importancia a la información correspondiente a los trimestres más cercanos al momento de la actualización.

Muestra

Para obtener indicadores significativos para todos los niveles de desagregación funcional y geográfica para los que se publica el IPC, se ha estructurado el proceso de selección de la muestra en tres grandes apartados, cada uno de los cuales tiene como objetivo la selección de los diferentes componentes de la misma. Estos son los siguientes :

- Selección de municipios.
- Selección de zonas comerciales y establecimientos.
- Determinación del número de observaciones.

Para la selección de municipios, como en bases anteriores, se han utilizado criterios poblacionales, y se ha tenido en cuenta la situación de las principales zonas comerciales en cada una de las provincias. La muestra de municipios ha aumentado respecto a la base 92, pasando de 130 a 141 (para alimentación) y de 70 a 97 (para resto).

Por otro lado, se ha prestado especial atención a los distintos tipos de establecimiento existente, así como a la recogida de precios de los artículos perecederos en municipios no capitales. Con todo ello, se ha aumentado el número de precios procesados respecto a la base anterior, pasando ahora a ser aproximadamente 180.000 precios mensuales.

Campo de consumo

Es el conjunto de los bienes y servicios que los hogares destinan al consumo; no se consideran, pues, los gastos en bienes de inversión ni los autoconsumos, autosuministros ni alquileres imputados. En la ECPF los bienes y servicios han sido clasificados según la clasificación internacional de consumo COICOP (en inglés, Classification Of Individual Consumption by Purpose). Así, cada parcela de consumo de la ECPF debe estar representada por uno o más artículos en el IPC, de forma que la evolución de sus precios represente la de todos los elementos que integran dicha parcela.

Cesta de la compra

Es el conjunto de bienes y servicios seleccionados en el IPC cuya evolución de precios representan la de todos aquellos que componen la parcela COICOP a la que pertenece.

El proceso para determinar la composición de la cesta de la compra y su estructura de ponderaciones utiliza como fuente fundamental de información la ECPF; así, en función de la importancia de cada parcela se han seleccionado uno o más artículos para el IPC. El número total de artículos que componen la nueva cesta de la compra es 484.

Para cada uno de los artículos se elabora su descripción o especificación con el fin de facilitar su identificación por parte del encuestador y permitir la correcta recogida de los precios. Estas especificaciones tienen en cuenta las particularidades propias de cada región.

Clasificación funcional

El IPC base 2001 se adapta completamente a la clasificación internacional de consumo COICOP.

Así, la estructura funcional del IPC constará de 12 grupos, 37 subgrupos, 80 clases y 117 subclases. Además, se mantienen las 57 rúbricas existentes y se amplía el número de grupos especiales.

Método general de cálculo

Hasta ahora, todos los sistemas españoles anteriores utilizaron lo que se denomina un índice tipo Laspeyres con base fija, al igual que otros muchos países de la Unión Europea. La ventaja fundamental de un índice de este tipo es que permite la comparabilidad de una misma estructura de artículos y ponderaciones a lo largo del tiempo que esté en vigor el Sistema; sin embargo, tiene un inconveniente y es que la estructura de ponderaciones pierde vigencia a medida que pasa el tiempo y evolucionan las pautas de consumo de los consumidores.

El nuevo Sistema utilizará la fórmula de "Laspeyres encadenado", que consiste en referir los precios del periodo corriente a los precios del año inmediatamente anterior; además, con una periodicidad que no superará los dos años se actualizarán las ponderaciones de las parcelas con información proveniente de la ECPF.

La utilización de las ponderaciones provenientes de la ECPF para calcular los índices encadenados evita la auto-ponderación de las parcelas del IPC por medio del nivel de los índices, es decir, las parcelas no irán ganando peso en la cesta de la compra a medida que vaya alcanzando mayor magnitud su índice.

Por otro lado, la actualización anual de ponderaciones tiene las siguientes ventajas:

- El IPC se adapta a los cambios del mercado y de los hábitos de consumo en un plazo muy breve de tiempo;
- Se puede detectar la aparición de nuevos bienes o servicios en el mercado para su inclusión en el IPC, así como la desaparición de los que se consideren poco significativos.

Básicamente, el proceso de cálculo es el mismo que el de un Laspeyres: se calculan medias ponderadas de los índices de los artículos que componen cada una de las agregaciones funcionales para las cuales se obtienen índices, y se compraran con los calculados el mes anterior. En este caso las ponderaciones utilizadas no permanecen fijas durante todo el período de vigencia del sistema.

Otra novedad importante en el nuevo Sistema es la utilización de la **media geométrica** para el cálculo de los precios medios provinciales de todos los artículos de la cesta de la compra, que intervienen en la elaboración del índice mensual.

Cambios de calidad

El tratamiento de los cambios de calidad es uno de los temas que más afectan a cualquier índice de precios.

Un cambio de calidad ocurre cuando cambia alguna de las características de la variedad para la que se recoge el precio y se considera que este cambio implica un cambio en la utilidad que le reporta al consumidor.

Para la correcta medición de la evolución de los precios es preciso estimar en qué medida la variación observada del precio es debida al cambio en la calidad del producto y qué parte de esta variación es achacable al precio, independientemente de su calidad.

Los métodos más utilizados en el IPC son la **consulta a expertos**, que consiste en solicitar a los propios fabricantes o vendedores la información para poder estimar el cambio de calidad; **los precios de las opciones**, que analiza los elementos componentes del antiguo producto y del nuevo para establecer el coste de las diferencias entre ambos; y el **precio de solapamiento**, basado en suponer que el valor de la diferencia de calidad entre el producto que desaparece y el nuevo es la diferencia de precio entre ellos en el periodo de solapamiento, es decir, en el periodo que estén en vigencia los precios de ambos.

El nuevo Sistema introduce una novedad en los métodos de ajustes de calidad que se vienen utilizando hasta ahora en el IPC, y es la utilización de la **regresión hedónica** para realizar ajustes de calidad en determinados grupos de productos, como los electrodomésticos. Dicho método se utiliza como complemento a los citados anteriormente.

Durante el año 2002 se continuarán los estudios que permitirán determinar la viabilidad de aplicar este nuevo método de ajuste de calidad a otros artículos de la cesta de la compra, en función de la información disponible.

Inclusión de las ofertas y rebajas

Uno de los cambios más importantes que se producirán en el IPC con la entrada en vigor del nuevo Sistema, base 2001, es la inclusión de los precios rebajados.

El IPC, base 1992, no contemplaba la recogida de estos precios por lo que su inclusión en el nuevo Sistema producirá una ruptura en la serie de este indicador, que no es posible solucionar totalmente con el método de los enlaces legales, utilizado cada vez que se lleva a cabo un cambio de base.

No obstante el INE facilitará los datos correspondientes a las tasas de variación a largo plazo, para evitar la falta comparación por la ruptura de la serie.

Enlace de series

Como en los anteriores cambios de base el INE publicará las series enlazadas, así como los coeficientes de enlace que permiten su cálculo, en aquellos casos en los que sea posible dar continuidad a la serie. En los casos en los que no exista una equivalencia exacta se indicará oportunamente.

El coeficiente de enlace legal se calcula como el cociente entre el índice del mes de diciembre de 2001 en base 2001 y en base 92. Es por este coeficiente por el que se multiplican los índices en base 92 y se obtienen los índices en base 2001.

Periodicidad del cambio de Sistema

Debido a la disponibilidad de datos anuales sobre ponderaciones provenientes de la Encuesta Continua de Presupuestos Familiares, una de las modificaciones más importantes en este nuevo proceso de cambio de sistema es la actualización continua de ponderaciones.

Una vez establecido el nuevo Sistema de IPC, el proceso constará de dos partes:

A) Adaptación continua del IPC.

Consistirá en la **revisión anual** de las ponderaciones para determinados niveles de desagregación geográfica y funcional; en ella se estudiarán cada año la conveniencia o no de ampliar la composición de la cobertura de productos así como la posibilidad de modificar alguno de los tratamientos empleados en el cálculo del índice.

B) Revisión estructural del IPC.

Cada cinco años se realizará un completo cambio de base; por tanto, las operaciones a realizar consistirán en determinar la composición de la cesta de la compra, las ponderaciones para los niveles más desagregados y la selección de la muestra. También vendrá acompañada de una revisión mucho más profunda de todos los aspectos metodológicos que definen el IPC.

De esta forma, se conseguirá un indicador más dinámico y que se adapte de forma más rápida a los movimientos del mercado y a la aparición de innovaciones metodológicas. Además, se cumplirá con las exigencias de la UE a través de Eurostat.

Se establece a partir de ahora un marco de actuación totalmente distinto al existente hasta el momento, al no tratarse de un mero cambio de base y sí de un proceso mucho más amplio.

Ponderaciones

El IPC es un índice de precios de Laspeyres (luego complejo y ponderado). Según la metodología del INE, la ponderación del artículo, w_i, se obtiene como cociente del gasto realizado en las parcelas representadas por dicho artículo (durante el periodo de referencias de la EPF 1990/1991) y el gasto total realizado en ese mismo periodo.

Las ponderaciones permanecen fijas a lo largo del periodo de vigencia del sistema de indices de precios al consumo. Un mismo artículo puede tener ponderaciones diferentes en las distintas agrupaciones geográficas.

Enlaces de series

Hasta 1976, el IPC elaborado por el INE se denominaba **Indice de Coste de la Vida**. Con este nombre se creó y se mantuvo en sus diferentes revisiones:

- Desde 1939 a 1960 (base 1936)
- Desde 1961 a 1968 (base 1958)
- Desde 1969 a 1976 (base 1968).

Con el nombre de **Indice de Precios al Consumo (IPC)**:

- Desde 1977 a 1985 (base 1976)
- Desde 1985 a 1992 (base 1983)

- Desde 1993 hasta nuestros días (base 1992)

La citada Metodología del INE (1992) propone los siguientes coeficientes de enlace de series:

$K_{92/58}$ = 0.044699 $K_{92/68}$ = 0.082113 $K_{92/76}$ = 0.184347 $K_{92/83}$ = 0.545261

8. Otros Indicadores de Coyuntura en España

Indices de Producción Industrial (IPI)

Es un índice de naturaleza cuántica que estudia la evolución de los volúmenes de producción física de los distintos sectores industriales. Se elabora con periodicidad mensual a través de 600 productos industriales representativos del conjunto. El organismo responsable es el INE.

Indice de Precios Industriales

Completa con el anterior la panorámica coyuntural de la industria en nuestro pais. El precio que se mide es el de salida de fabrica y cubre las mismas ramas que el IPI. Es el deflactor que debe utilizarse para obtener la evolución del valor real de los bienes de producción intermedia.

Indice de Ventas al Por Menor

El INE está construyendo un nuevo indicador de las ventas al por menor tomando datos en las grandes superficies.

Encuestas de coyuntura

- Encuesta de coyuntura Industrial del Ministerio de Industria y Energía
- Encuesta de coyuntura de la construcción elaborada por el MOPTMA
- Encuesta de coyuntura laboral del Ministerio de Trabajo

Indices de Cotización Bursatil

Mide las fluctuaciones de la cotizaciones de las acciones en los distintos mercados bursátiles de forma diaria. A partir de las cotizaciones de cada valor se elaboran índices de grupo (banca, comunicaciones, etc.) que convenientemente ponderados dan lugar al índice general de cada mercado.

EL EFECTO DE LA INCLUSIÓN DE LAS REBAJAS EN EL IPC2001 SOBRE EL CALCULO DE LAS TASAS

Cálculo de tasas de variación entre periodos de distintas bases

Cuando el periodo inicial sea anterior al año 2001, los únicos índices disponibles no incluyen precios rebajados, por ello, es necesario utilizar un método de cálculo de

variaciones que solucione el problema citado anteriormente: las tasas de variación de precios se calcularán por partes. Es decir, si se quiere saber cuál ha sido la variación del IPC desde el mes m del año t (t < 2001) hasta el mes m' del año t' (t' >= 2002) se considerarán los siguientes tramos temporales:

$\dot{y}^{m'\,2001}_{m\,t}$ $\Delta\ m\ t\ /\ m'\ 2001$	Compara meses diferentes, dado que se trata de periodos anteriores a enero de 2002 y ninguno viene afectado por el efecto rebajas. Calculada con datos públicos
$\dot{y}^{m'\,2002}_{m'\,2001}$ $\Delta\ m'\ 2001\ t\ /\ m'\ 2002$	Compara el mismo mes de 2001 con el de 2002. Las variaciones reales para estos periodos serán publicadas por el INE pero no se podrán calcular a partir de los índices publicados.
$\dot{y}^{m'\,t'}_{m'\,2002}$ $\Delta\ m'\ 2002\ t\ /\ m'\ t'$	Compara el mismo mes de distintos años, ambos posteriores a 2001 (incluyen precios rebajados), pero al tratarse del mismo mes la tasa no se ve afectada. Calculada con datos públicos

Ejemplo del efecto de la inclusión de precios rebajados en el IPC

Se ha considerado una serie de IPC - base 1992 desde enero de 2000 hasta diciembre de 2001 (en la tabla, se ha denominado Serie Publicada – base 1992). Durante 2001 han sido recogidos los precios y calculados los índices medidos según el nuevo Sistema aunque no se publican (en la tabla, a la serie de estos índices se le ha denominado Serie no Publicada – base 2001, y está sombreada, incluye los precios rebajados).
Debido a que entre estas dos series se ha producido un cambio de base, es necesario enlazarlas para obtener una única serie continua medida en la base 2001. Así, la Serie Publicada – base 92 se transformará en una serie en base 2001 a través de un enlace legal.

	Serie Publicada		Serie NO Publicada	Serie Publicada			
	Base 1992=100		Base 2001=100 (incluye rebajas)	Base 2001=100 (enlazada)			
Año	2000	2001	2001	2000	2001	2002	2003
Enero	121,26	124,02	95,21	98,81	101,05	99,22	103,29
Febrero	121,3	124,11	95,02	98,84	101,13	99,32	103,29
Marzo	121,5	124,47	99,67	99	101,42	104,38	
Abril	122,03	125,18	99,97	99,43	101,99	104,59	
Mayo	122,23	125,47	100,17	99,59	102,24	104,59	
Junio	122,3	125,56	100,47	99,65	102,31	105,01	
Julio	122,34	125,64	96,76	99,68	102,37	100,91	
Agosto	122,36	125,67	96,56	99,7	102,39	100,91	
Septiembre	122,65	126,02	103,81	99,93	102,68	108,18	
Octubre	123,23	127,09	103,91	100,4	103,55	108,4	
Noviembre	123,84	127,77	104,22	100,91	104,11	108,5	

Diciembre	123,94	127,91	104,22	100,99	104,22	108,72	

El enlace legal consiste en calcular un coeficiente que relacione los índices en ambas bases en el mes de diciembre de 2001. Su cálculo se realiza como el cociente del índice de diciembre de 2001, en base 2001 (Serie No Publicada – base 01 (incluye rebajas) y el índice de diciembre de 2001 en base 1992 (Serie Publicada – base 1992). En el ejemplo, este coeficiente sería:

$$1104{,}22 \Big/ 127{,}91 = 0{,}814797$$

Los índices enlazados se obtienen como producto de los índices en base 92 por el coeficiente de enlace legal. Así, por ejemplo:

- Índice enlazado enero 2000 = 121,26 x 0,81479 = 98,81
- Índice enlazado febrero 2000 = 121,30 x 0,81479 = 98,84
- …

La serie que aparece en la Tabla con el nombre Serie Publicada – Base 2001 (serie enlazada) contiene los índices base 1992 transformados con el coeficiente de enlace legal, hasta diciembre de 2001 y los índices calculados en la nueva base a partir de enero de 2002.

Cálculo de la tasa de variación entre marzo de 2000 y febrero de 2003.

Se consideran tres tramos temporales para los que se obtendrán las tasas de variación del IPC:

Tramo 1: marzo 2000 – febrero 2001 (a partir de los índices publicados)

Tramo 2: febrero 2001 – febrero 2002 (variación publicada por el INE)

Tramo 3: febrero 2002 – febrero 2003 (a partir de los índices publicados)

La variación de precios para el primer y tercer tramo se calculará a partir de los índices publicados, mientras que para el segundo tramo se tomará la (es una tasa anual).

Tramo 1:

La variación para el primer tramo se puede obtener a partir de los índices publicados en base 1992 o a partir de los índices enlazados en base 2001 al coincidir las tasas de variación en ambos casos.

(Marzo 2000 - Febrero 2001)

Así:

$I_{2001}^{mar-00} = 99{,}00 \qquad I_{2001}^{feb-01} = 101{,}13 \qquad I_{1992}^{mar-00} = 121{,}50 \qquad I_{1992}^{feb-01} = 124{,}11$.

$$\dot{y}_{mar-00}^{feb-01} = \frac{101{,}13 - 99{,}00}{99{,}00} \times 100 = 2{,}1\% \qquad \text{o lo que es lo mismo} \qquad \dot{y}_{mar-00}^{feb-01} = \frac{124{,}11 - 121{,}50}{121{,}50} \times 100 = 2{,}1\%$$

Tramo 2:

El cálculo de la tasa de este tramo se realiza con los índices no publicados para el período inicial y con los índices de la serie publicada con base 2001 para el periodo final. Esta tasa será publicada por el INE. La forma de calcularla es la siguiente:

(Febrero 2001 - febrero 2002)

$$I*_{2001}^{feb-01} = 95{,}02 \qquad I_{2001}^{feb-02} = 99{,}32 \qquad \dot{y}_{feb-01}^{feb-02} = \frac{99{,}32 - 95{,}02}{95{,}02} \times 100 = 4{,}3\%$$

Tramo 3:
La variación para el tercer tramo se obtiene a partir de los índices publicados en base 2001.
(Febrero 2002 - febrero 2003)

$$I_{2001}^{feb-02} = 99{,}32 \qquad I_{2001}^{feb-03} = 103{,}29 \qquad \dot{y}_{feb-02}^{feb-03} = \frac{103{,}29 - 99{,}32}{99{,}32} \times 100 = 4{,}0\%$$

Con las tasas de variación calculadas para los tres tramos, se obtiene la tasa de variación para el periodo completo:

$$\dot{y}_{mar-00}^{feb-03} = \left[\left(1 + \frac{\dot{y}_{mar-00}^{feb-01}}{100} \right) \times \left(1 + \frac{\dot{y}_{feb-01}^{feb-02}}{100} \right) \times \left(1 + \frac{\dot{y}_{feb-02}^{feb-03}}{100} \right) - 1 \right] \times 100$$

$$= \left[\left(1 + \frac{2{,}1}{100} \right) \times \left(1 + \frac{4{,}3}{100} \right) \times \left(1 + \frac{4}{100} \right) - 1 \right] \times 100 = 10{,}7\%$$

Si en lugar de realizar el cálculo por tramos se hubiera hecho como es habitual, es decir, con el índice enlazado de marzo de 2000 en base 2001 y el índice de febrero de 2003 en base 2001, la tasa hubiera sido:

$$I_{2001}^{mar-00} = 99{,}00 \qquad I_{2001}^{feb-03} = 103{,}29 \qquad \dot{y}_{feb-02}^{feb-03} = \frac{103{,}29 - 99{,}00}{99{,}00} \times 100 = 4{,}3\%$$

Como se observa el valor obtenido por el método habitual (4,3%) es muy diferente de la real (10,7%).

Capítulo V — SERIES TEMPORALES

5.1. Introducción

Toda institución, ya sea la familia, la empresa o el gobierno, necesita realizar planes para el futuro si desea sobrevivir o progresar.
La planificación racional exige prever los sucesos del futuro que probablemente vayan a ocurrir.
La previsión se suele basar en lo ocurrido en el pasado.
La técnica estadística utilizada para hacer inferencias sobre el futuro teniendo en cuenta lo ocurrido en el pasado es el ANÁLISIS DE SERIES TEMPORALES.

NUMEROS ÍNDICES →Tratamos de estudiar la evolución de una determinada magnitud a lo largo del tiempo.
SERIES TEMPORALES →Tratamos de hacer predicciones sobre esa magnitud, teniendo en cuenta sus características históricas o del pasado.

5.2. CONCEPTO DE SERIE TEMPORAL Y DEFINICIÓN DE SUS COMPONENTES.-

Se define una serie temporal (también denominada histórica, cronológica o de tiempo) como un conjunto de datos, correspondientes a un fenómeno económico, ordenados en el tiempo.

Ejemplos

- Nº de accidentes laborales graves en las empresas de más de 500 empleados de Sevilla, durante los últimos 5 años.
- Ventas de nuestra empresa en los últimos 10 años.
- Cantidad de lluvia caída al día durante el último trimestre.

Los datos son de la forma (y_t, t) donde:

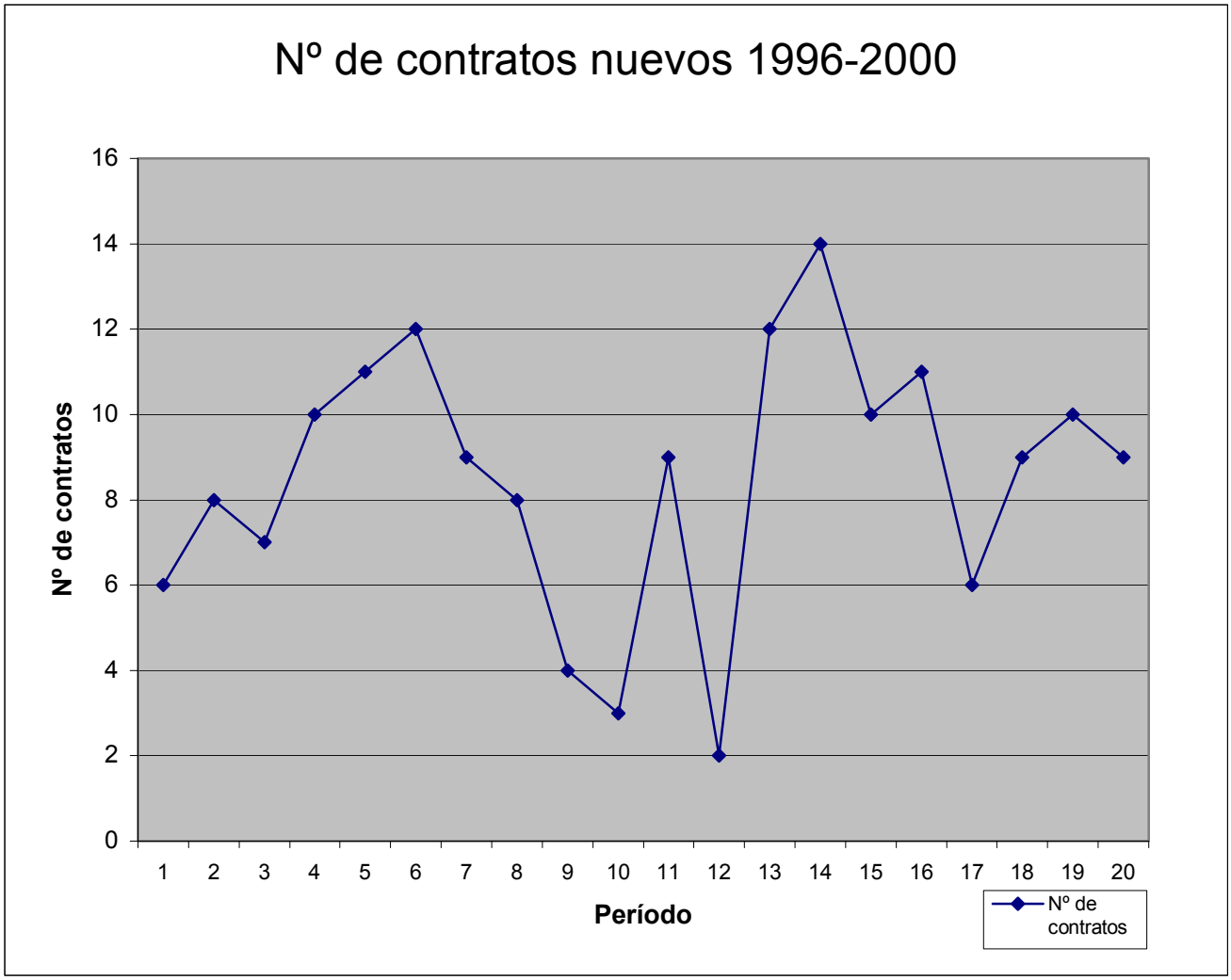
$y_t \longrightarrow$ Variable endógena o dependiente
$t \longrightarrow$ Variable exógena o independiente

Nota: realmente sólo hay una variable a estudiar que es y_t. En el análisis de regresión teníamos dos variables (explicábamos una variable a partir de la otra). Aquí sólo hay una variable (explicamos una variable a partir de su pasado histórico).

Ejemplo

Los datos siguientes corresponden al número de contratos nuevos realizados por las empresas de menos de 10 empleados, en Sevilla, durante el período 1996-2000.

	1996	1997	1998	1999	2000
1º Trimestre	6	11	4	12	6
2º Trimestre	8	12	3	14	9
3º Trimestre	7	9	9	10	10
4º Trimestre	10	8	2	11	9



Componentes de una serie temporal:

- La tendencia.
- Las variaciones cíclicas.
- Las variaciones estacionales.
- Las variaciones accidentales.

LA TENDENCIA (T)

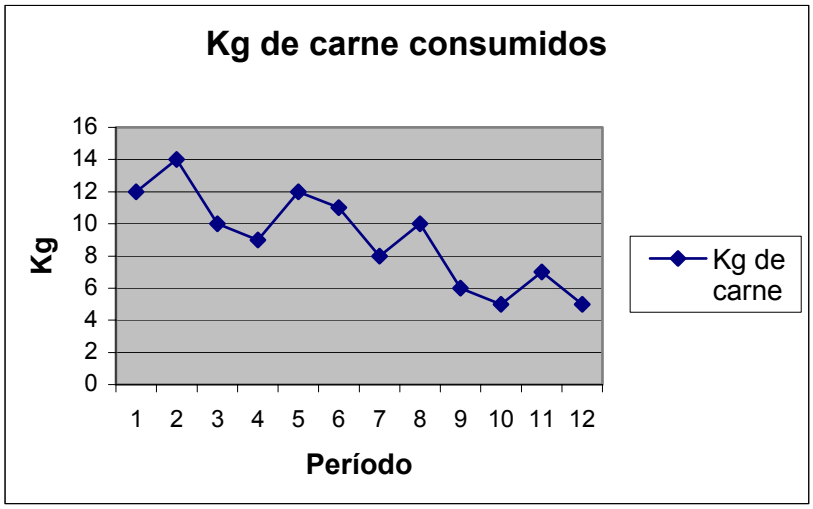
Es una componente de la serie temporal que refleja su evolución a largo plazo.

Puede ser de naturaleza estacionaria o constante (se representa con una recta paralela al eje de abcisas), de naturaleza lineal, de naturaleza parabólica, de naturaleza exponencial, etc.

Ejemplo para la tendencia

Supongamos que tenemos el número de kg de carne de vaca consumidos por trimestre durante los últimos años en unos grandes almacenes.

	1998	1999	2000
1º Trimestre	12	12	6
2º Trimestre	14	11	5
3º Trimestre	10	8	7
4º Trimestre	9	10	5

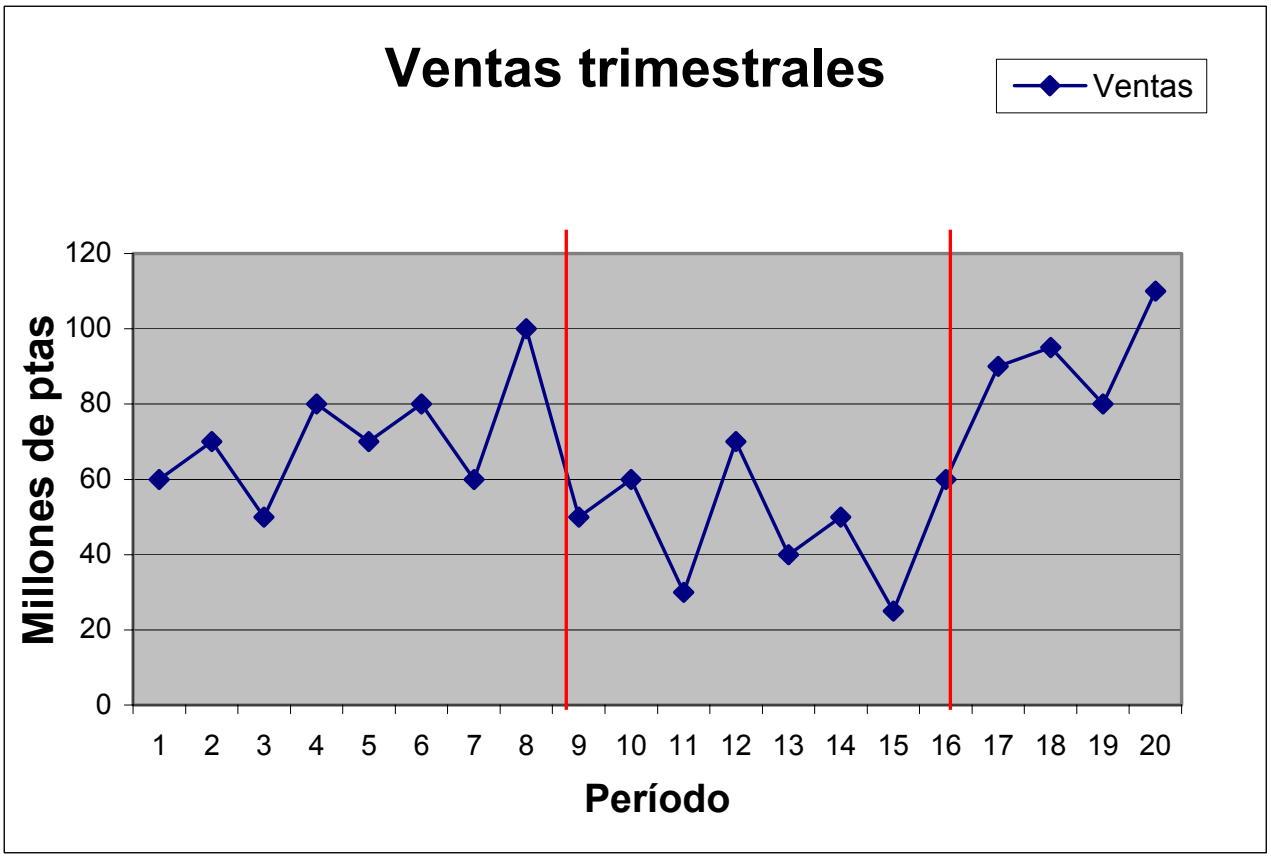


LAS VARIACIONES CÍCLICAS (C)

Es una componente de la serie que recoge oscilaciones periódicas de amplitud superior a un año. Estas oscilaciones periódicas no son regulares y se presentan en los fenómenos económicos cuando se dan de forma alternativa etapas de prosperidad o de depresión.

Ejemplo para las variaciones cíclicas
Supongamos que tenemos las ventas trimestrales de un supermercado en el período 1990-1994, expresadas en millones de pesetas constantes del año 1990.

	1990	1991	1992	1993	1994
1º Trimestre	60	70	50	40	90
2º Trimestre	70	80	60	50	95
3º Trimestre	50	60	30	25	80
4º Trimestre	80	100	70	60	110



LAS VARIACIONES ESTACIONALES (E)

Es una componente de la serie que recoge oscilaciones que se producen alrededor de la tendencia, de forma repetitiva y en períodos iguales o inferiores a un año.

Su nombre proviene de las estaciones climatológicas: primavera, verano, otoño e invierno.

Ejemplos de variaciones estacionales

- En Navidad las ventas de establecimientos se suelen incrementar.
- El consumo de gasolina aumenta la primera decena del mes y disminuye en la última.
- El clima afecta a la venta de determinados productos: los helados se venden fundamentalmente en verano y la ropa de abrigo en invierno.

LAS VARIACIONES ACCIDENTALES (A)

Es una componente de la serie que recoge movimientos provocados por factores imprevisibles (un pedido inesperado a nuestra empresa, una huelga, una ola de calor, etc). También reciben el nombre de variaciones irregulares, residuales o erráticas.

¿Cómo actúan estas 4 componentes?

- Modelo Aditivo : $y_t=T+C+E+A$
- Modelo Multiplicativo: $y_t=T\cdot C\cdot E\cdot A$
- Modelo Mixto : $y_t=T\cdot C\cdot E+A$

¿ Cómo detectamos el modo en que interactúan las componentes de una serie temporal?
¿ Esquema aditivo o multiplicativo?

1º) Calculamos 2 tipos de indicadores:

Ci= Y(i,t+1) / Y(i,t)

di=Y(i,t+1) / Y (i,t)

2º) Calculamos los coeficientes de variación para las series formadas por los dos indicadores, y si:

CV Ci < CV di → Esquema multiplicativo
CV di < CV Ci → Esquema aditivo

EJEMPLO:

Según la ECL, las horas no trabajadas por trimestre y trabajador entre 1992 y 1997 son:

	92	93	94	95	96	97
TRIM. I	46	45,7	44,6	45,4	35,6	53,8
TRIM. II	51,7	49,8	38,8	50,8	49,3	47
TRIM. III	106,4	100,7	112,4	96,7	100,2	84,8
TRIM IV	68,7	72,1	67,6	62,6	69,5	52

¿ Qué esquema de agregación es el más apropiado?

1º) Calculamos los 2 tipos de indicadores:

di	92-93	93-94	94-95	95-96	96-97
TRIM. I	-0,3	-1,1	0,8	-9,8	18,2
TRIM. II	-1,9	-11	12	-1,5	-2,3

TRIM. III	-5,7	11,7	-15,7	3,5	-15,4
TRIM IV	3,4	-4,5	-5	6,9	-17,5

ci	92-93	93-94	94-95	95-96	96-97
TRIM. I	0,993	0,976	1,018	0,784	1,511
TRIM. II	0,963	0,779	1,309	0,970	0,953
TRIM. III	0,946	1,116	0,860	1,036	0,846
TRIM IV	1,049	0,938	0,926	1,110	0,748

2º) Calculamos los Coeficientes de variación de ambas distribuciones:

Sd = 9.271;
CV d = 5.267

Sc = 0.174;
CV c = 0.174

METODOS PARA DETERMINAR LA TENDENCIA

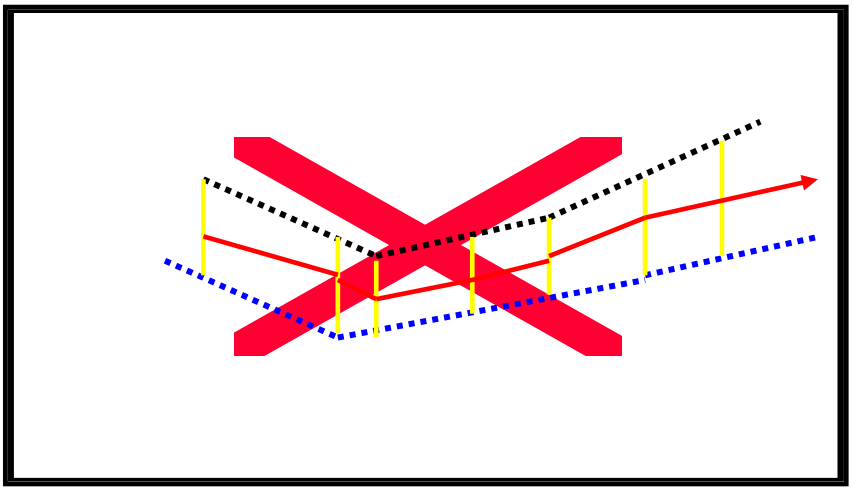
1º) METODO GRAFICO

a) Se efectúa la representación gráfica de la serie ordenada Yt.
b) Se unen mediante segmentos rectilíneos todos los puntos altos de la serie, obteniéndose una poligonal de cimas.
c) Se realiza lo mismo con los puntos bajos, obteniéndose la línea poligonal de fondos.
d) Se trazan perpendiculares al eje de abscisas por los puntos cimas y fondos.
e) La tendencia viene dada por la línea amortiguada que une los puntos medios de los segmentos.

EJEMPLO:

Dada la siguiente serie trimestral de ventas de un supermercado, representa gráficamente la tendencia:

Trimestres / Años	1998	1999	2000	2001
TRIM. I	50	20	50	70
TRIM. II	80	50	70	100
TRIM. III	70	40	50	90
TRIM IV	60	30	40	60



2º) METODO DE LAS MEDIAS MOVILES

*** *Empleando 3 observaciones*

a) Partimos de la serie temporal observada Yt.

Trimestres / Años	1999	2000	2001
TRIM. I	150	155	160
TRIM. II	165	170	180
TRIM. III	125	135	140
TRIM IV	170	165	180

b) Se obtienen sucesivas medias aritméticas para cada Yt, con un número de observaciones anteriores y posteriores fijado de antemano.
   - *Si el número de observaciones es impar la media Yt, está centrada y coincide con el período t.*

$\square_2$ = (y1+y2+y3) / 3= (150+155+125)/3= 146.6

$\square_3$ = (y2+y3+y4) / 3= (165+125+170)/3= 153.3

$\square_4$ = (y3+y4+y5) / 3= (125+170+155)/3= 150

$\square_5$ = (y4+y5+y6) / 3= (170+155+170)/3=165

$\square_6$ = (y5+y6+y7) / 3= (155+170+135)/3=153.3

$\square_7$ = (y6+y7+y8) / 3= (170+135+165)/3=156.6

$\square_8$ = (y7+y8+y9) / 3= (135+165+160)/3=153.3

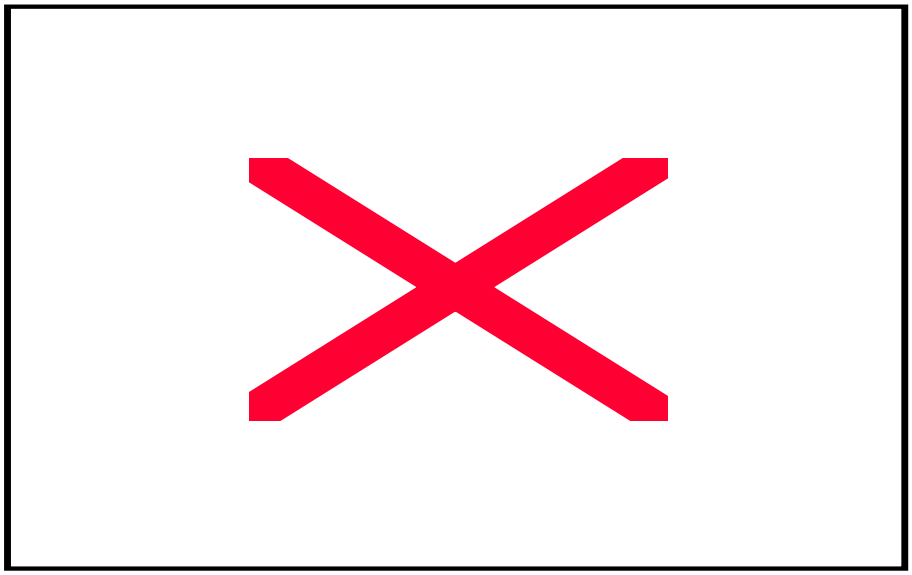
$\square_9$ = (y8+y9+y10)/ 3= (165+160+180)/3=168.3

$\square_{10}$ = (y9+y10+y11)/3= (160+180+140)/3=160

$\square_{11}$ = (y10+y11+y12)/3=(180+140+180)/3=166.6

X

c) La serie formada por $\square_t$, nos indica la línea amortiguada de la tendencia



*** *Empleando 4 observaciones*

a) Partimos de la serie temporal observada Yt.

Trimestres / Años	1999	2000	2001
TRIM. I	150	155	160
TRIM. II	165	170	180
TRIM. III	125	135	140
TRIM IV	170	165	180

b) Se obtienen las sucesivas medias aritméticas
Si el número de observaciones empleados para obtener la media es par, yt no está centrada y no coincide con el período t, y habrá que volver a calcular una nueva media aritmética yt, utilizando los yt, obteniendo de esta manera una nueva serie de medias móviles centradas.

$\square_{2.5}$ = (y1+y2+y3+y4) / 4 = (150+165+125+170)/4= 152.5

$\square_{3.5}$ = (y2+y3+y4+y5) / 4 = (165+125+170+155)/4= 153.75

$\square_{4.5}$ = (y3+y4+y5+y6) / 4 = (125+170+155+170)/4= 155

$\square_{5.5}$ = (y4+y5+y6+y7) / 4 = (170+155+170+135)/4= 157.5

$\square_{6.5}$ = (y5+y6+y7+y8) /4 = (155+170+135+165)/4= 156.25

$\square_{7.5}$ = (y6+y7+y8+y9) /4 = (170+135+160+180)/4= 157.5

$\square_{8.5}$ = (y7+y8+y9+y10)/4 = (135+165+160+180)/4= 160

$\square_{9.5}$ = (y8+y9+y10+y11)/4 = (165+160+180+140)/4= 161.25

$\square_{10.5}$ =(y9+y10+y11+y12)/4= (160+180+140+180)/4= 165

Como se puede observar la serie de las medias obtenidas no está centrada, y debemos obtener una nueva serie de medias centradas, a partir de la serie " descentrada "

$\square_3$ = (y2.5+y3.5) / 2 = (152.5 + 153.75)/2= 153.125

$\square_4$ = (y3.5+y4.5) / 2 = (153.75 +155) /2 = 154.375

$\square_5$ = (y4.5+y5.5) / 2 = (155 +157.5) / 2 = 156.25

$\square_6$ = (y5.5+y6.5) / 2 = (157.5+156.25)/2 = 156.875

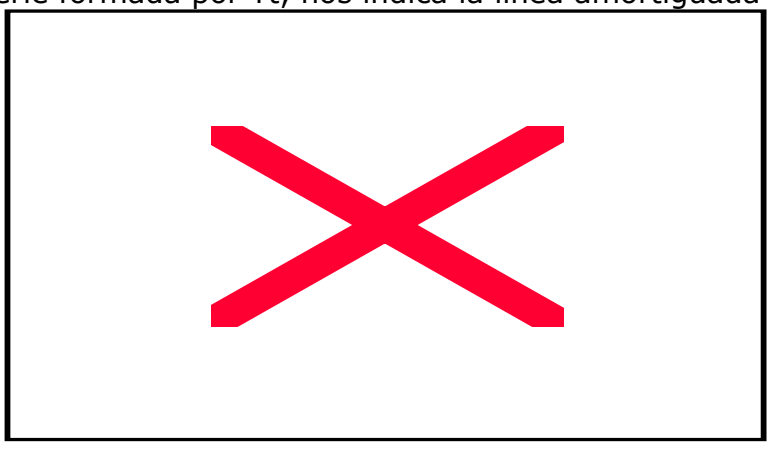
$\square_7$ = (y6.5+y7.5) / 2 = (156.25+157.5) /2 = 156.875

$\square_8$ = (y7.5+y8.5) / 2 = (157.5 +160)/2 = 158.75

$\square_9$ = (y8.5+y9.5) / 2 = (160 + 161.25) /2=160.625

$\square_{10}$ =(y9.5+y10.5) / 2 = (161.25+165) / 2=163.125

c) La serie formada por Yt, nos indica la línea amortiguada de la tendencia



3º) MÉTODO ANALÍTICO DE LOS MÍNIMOS CUADRADOS

a) Obtendremos la tendencia a partir de la recta Yt= a+ bt, siendo Yt, la media anual de las observaciones trimestrales de los casos anteriores.

y1=(y1+y2+y3+y4) / 4 = 152.5
y2=(y5+y6+y7+y8) / 4 = 156.25
y3=(y9+y10+y11+y12) / 4 = 165.00

n = numero de observaciones y coincide con el nº de períodos (en nuestro caso 3)

t´= t – Ot →Si el numero de observaciones es impar Ot es el valor que ocupa el lugar central(2000)

t´= 2 (t – O´t) → Si el número de observaciones es par O´t es la media de los 2 lugares centrales de la serie de periodos t.

b) Calculamos los coeficientes "a" y "b" de la recta de regresión.
a= Σ yt / n

b= Σyt * t´ / Σ t´²

T	Yt	t´	Yt* t´	t´²	Yt²
1999	152,5	-1	-152,5	1	23256,25
2000	156,25	0	0	0	24414,06
2001	165	1	165	1	27225
TOTAL	473,75	0	12,5	2	74895,31

a= 473.75 /3 = 157.92

b= 12.5 / 2 =6.25

Deshaciendo el cambio de variable, tendremos la siguiente predicción de la tendencia:

- Predicción de las ventas para el año 2002 → y2002= -12342.08 + 6.25 * 2002 = 170.42
- Hemos calculado la media trimestral → media anual = 170.42 * 4 = 681.68 millones u.m.

¿ Es fiable la predicción realizada ?

$R^2 = (S_{t' yt})^2 / S_{t'}^2 * S_{yt}^2$

$R^2 = (4.16)^2 / 0.66 * 32.69 = 0.7939$

5.4 DETERMINACION DE LAS VARIACIONES ESTACIONALES

- Método de la razón a la media móvil para determinar la componente estacional en una serie temporal.

1º) Se determina la tendencia por el método de las medias centradas en los períodos (Yt)
(estamos aplicando cuatro observaciones para el cálculo de la media aritmética)

Trimestres / Años	1999	2000	2001
TRIM. I		156,25	160,625
TRIM. II		156,875	163,125
TRIM. III	153,125	156,875	
TRIM IV	154,375	158,75	

2º) Cómo este método se basa en la hipótesis multiplicativa, si dividimos la serie observada Yt, por su correspondiente media móvil centrada, eliminamos de forma conjunta las componentes del largo plazo (tendencia y ciclo), pero la serie seguirá manteniendo el efecto de la componente estacional.

Trimestres / Años	1999	2000	2001
TRIM. I		155/156,25	160/160,625
TRIM. II		170/156,875	180/163,125
TRIM. III	125/153,125	135/156,875	
TRIM IV	170/154,375	165/158,75	

Trimestres / Años	1999	2000	2001
TRIM. I		0,9921	0,997
TRIM. II		1	1,1035
TRIM. III	0,8163	0,8606	
TRIM IV	1,1012	1,04	

3º) Para eliminar el efecto de la componente estacional, calcularemos las medias aritméticas a nivel de cada estación (cuatrimestre). Estas medias representan de forma aislada la importancia de la componente estacional.

M1= (0.9921 + 0.9970) /2 =0.9945
M2= (1.0837 + 1.1035) / 2=1.0936
M3= (0.8136 + 0.8606) / 2 = 0.8385
M4 =(1.1012 + 1.0400) / 2 = 1.0706

4º) Calcularemos los índices de variación estacional, para lo que previamente calcularemos la media aritmética anual de las medias estacionales (M1, M2, M3, M4) , que será la base de los índices de variación estacional.

MA = (M1+M2+M3+M4) / 4 = (0.9946+1.0936+0.8385+1.0706) / 4 = 0.9993

I1= M1 / MA = 0.9945/0.9993 =0.9953
I2= M2 / MA = 1.0936/0.9993 =1.0944

I3= M3 / MA = 0.8385/0.9993 =0.8391
I4 =M4 / MA= 1.0706/0.9993 =1.0714

Existirán tantos índices como estaciones o medias estacionales tengan las observaciones.

5º) Una vez obtenidos los índices de variación estacional puede desestacionalizarse la serie observada, dividiendo cada valor de la correspondiente estación por su correspondiente índice.

Trimestres / Años	1999	2000	2001
TRIM. I	150/0,9953	155/09953	160/0,9953
TRIM. II	165/1,0944	170/1,0944	180/1,0944
TRIM. III	125/0,8391	135/0,8391	140/0,8391
TRIM IV	170/1,0714	165/1,0714	180/1,0714

Trimestres / Años	1999	2000	2001
TRIM. I	150,71	155,73	160,8
TRIM. II	150,77	155,34	164,5
TRIM. III	148,97	160,89	166,9
TRIM IV	158,67	154	168

- **Método de la Tendencia por Ajuste Mínimo-Cuadrático**

El objetivo sigue siendo aislar la componente estacional de la serie por eliminación sucesiva de todos los demás. La diferencia con el método anterior es que, en este caso, las componentes a l/p (tendencia-ciclo) las obtenemos mediante un ajuste mínimo-cuadrático de las medias aritméticas anuales $\bar{y}_t$ calculándose bajo la hipótesis aditiva. Sigue los siguientes pasos:

- Se calculan las medias anuales de los datos observados $y_t : \bar{y}_1, \bar{y}_2, \bar{y}_3, \ldots$ Si las observaciones son trimestrales estas medias se obtienen con 4 datos, si son mensuales con 12 datos, etc. para el caso de que el periodo de repetición sea el año
- Se ajusta una recta por mínimos cuadrados $\bar{y}_t = a + b\,t$ que nos representa, como sabemos, la tendencia, siendo el coeficiente angular de la recta el incremento medio anual de la tendencia, que influirá de forma distinta al pasar de una estación a otra
- Se calculan, con los datos observados, las medias estacionales (M_1, M_2, M_3, ...) con objeto de eliminar la componente accidental. Estas medias son brutas pues siguen incluyendo los componentes a l/p (tendencia-ciclo) que deben someterse a una corrección
- Empleando el incremento medio anual dado por el coeficiente, se obtienen las medias estacionales corregidas de las componentes a largo plazo (M'_1, M'_2, M'_3, ...) bajo el esquema aditivo:

Para la r-ésima estación,
la media estacional corregida
de la tendencia interestacional será

$$M'_r = M_r - \frac{(r-1)b}{n^o\ estaciones}$$

- Los índices de variación estacional se obtienen con la misma sistemática del método anterior: con las medias estacionales corregidas se obtiene la media aritmética anual $M'A$ que sirve de base para calcular los índices:

$$I_1 = \frac{M'_1}{M'A}100 \quad I_2 = \frac{M'_2}{M'A}100 \ldots \text{(expresados en \%)}$$

- Obtenidos estos índices, podemos desestacionalizar la serie como en el método anterior.

5.4. DETERMINACIÓN DE LAS VARIACIONES CÍCLICAS

Cuando hemos definido esta componente se ha dicho que recoge las oscilaciones periódicas de larga duración. El problema es que estos movimientos no suelen ser regulares como los estacionales y su determinación encierra dificultades de forma que como se ha apuntado en los casos prácticos se suelen tratar conjuntamente con la tendencia llamando componente extraestacional al efecto (TxC) si estamos en el marco multiplicativo o (T+C) si es el aditivo.
A pesar de estas dificultades se puede tratar de aislar el ciclo bajo la hipótesis multiplicativa dejándola como residuo con la eliminación de la tendencia y la variación estacional.

Pasos a seguir:

- Estimar la tendencia.
- Calcular los índices de variación estacional.
- Se desestacionaliza la serie observada.
- Se elimina la tendencia dividiendo cada valor desestacionalizado por la serie de tendencia.

  Expresando el proceso en forma de cociente sería:

$$\frac{y_t}{TxE} = \frac{TxExCxA}{TxE} = CxA$$

El proceso finalizaría intentando eliminar la componente accidental A y determinando el periodo de los ciclos que nos llevaría a un tratamiento de análisis armónico que superaría el nivel descriptivo que estamos dado al tratamiento clásico de las Series temporales.

VI DISTRIBUCIONES ALEATORIAS

1. Introducción

Dentro de la estadística se pueden considerar dos ramas perfectamente diferenciadas por sus objetivos y por los métodos que utilizan:

Estadística Descriptiva o Deductiva
Inferencia Estadística o Estadística Inductiva

- Se utiliza cuando la observación de la población no es exhaustiva sino sólo de un subconjunto de la misma de forma que, los resultados o conclusiones obtenidos de la muestra, los generalizamos a la población
- La muestra se toma para obtener un conocimiento de la población pero nunca nos proporciona información exacta, sino que incluye un cierto nivel de incertidumbre
- Sin embargo sí será posible, a partir de la muestra, hacer afirmaciones sobre la naturaleza de esa incertidumbre que vendrá expresada en el lenguaje de la probabilidad, siendo por ello un concepto muy necesario y muy importante en la inferencia estadística
- Según V. Barnett (1982): -La estadística es la ciencia que estudia como debe emplearse la información y como dar una guía de acción en situaciones prácticas que envuelven incertidumbre-. Las "situaciones prácticas que envuelven incertidumbre" son lo que nosostros llamaremos experimentos aleatorios.

2. Fenómenos Aleatorios

Un **experimento** es cualquier situación u operación en la cual se pueden presentar uno o varios resultados de un conjunto bien definido de posibles resultados.

Los experimentos pueden ser de dos tipos según si, al repetirlo bajo idénticas condiciones:

Determinístico	**Aleatorio**
Se obtienen siempre los mismo resultados. Ej: medir con la misma regla e identicas condiciones la longitud de una barra	No se obtienen siempre los mismo resultados. Ej: el lanzamiento de una moneda observando la sucesión de caras y cruces que se presentan

Las siguientes son **características de un experimento aleatorio**:

- El experimento se puede repetir indefinidamente bajo idénticas condiciones
- Cualquier modificación a las condiciones iniciales de la repetición puede modificar el resultado
- Se puede determinar el conjunto de posibles resultados pero no predecir un resultado particular
- Si el experimento se repite gran número de veces entonces aparece algún modelo de regularidad estadística en los resultados obtenidos

3. Espacio Muestral

Se denomina **resultado básico o elemental**, **comportamiento individual** o **punto muestral** a cada uno de los posibles resultados de un experimento aleatorio. Los resultados básicos elementales serán definidos de forma que no puedan ocurrir dos simultáneamente pero si uno necesariamente.

Se denomina **conjunto universal**, **espacio muestral** o **espacio de comportamiento** ***E*** al conjunto de todos los resultados elementales del experimento aleatorio. Pueden ser de varios tipos:

Espacio Muestral Discreto

Espacio muestral finito

Tiene un número finito de elementos.

Ejemplo:
Experimento aleatorio consistente en lanzar un dado. El espacio muestral es ***E***={1,2,3,4,5,6}

Espacio muestral infinito numerable

Tiene un número infinito numerable de elementos es decir, se puede establecer una aplicación biyectiva entre ***E*** y ***N***.

Ejemplo:
Experimento aleatorio consistente en lanzar un dado hasta que sea obtenido el número 1
E={{1},{2,1},{3,1} ... {2,2,1},{2,3,1},...}

Espacio Muestral Continuo

Si el espacio muestral contiene un número infinito de elementos, es decir, no se puede establecer una correspondencia biunívoca entre ***E*** y ***N***.

Ejemplo:
Experimento aleatorio consistente en tirar una bola perfecta sobre un suelo perfecto y observar la posición que ocupará esa bola sobre la superficie. ***E***={Toda la superficie del suelo}

4. Sucesos

Un **suceso** ***S*** es un subconjunto del espacio muestral, es decir, un subconjunto de resultados elementales del experimento aleatorio.
Diremos que **ocurre** o **se presenta el suceso** cuando al realizarse el experimento aleatorio, da lugar a uno de los resultados elementales pertenecientes al subconjunto ***S*** que define el suceso

Se pueden considerar cuatro tipos de sucesos según el nº de elementos que entren a formar parte:

- **Suceso elemental**, **suceso simple** o **punto muestral** es cada uno de los resultados posibles del experimento aleatorio luego los sucesos elementales son subconjuntos de ***E*** con sólo un elemento
- **Suceso compuesto** es aquel que consta de dos o más sucesos elementales
- **Suceso seguro**, **cierto** o **universal** es aquel que consta de todos los sucesos elementales del espacio muestral ***E***, es decir, coincide con ***E***. Se le denomina seguro o cierto porque ocurre siempre.
- **Suceso imposible** es aquel que no tiene ningún elemento del espacio muestral ***E*** y por tanto no ocurrirá nunca. Se denota por ∅.

5. Operaciones con sucesos

Con los sucesos se opera de manera similar a como se hace en los conjuntos y sus operaciones se definen de manera análoga. Los sucesos a considerar serán los correspondientes a un experimento aleatorio y por tanto serán subconjuntos del espacio muestral ***E***.

Suceso Contenido en Otro

Dados dos sucesos ***A*** y ***B*** de un experimento aleatorio:

Diremos que ***A*** está incluido en ***B*** si:
Cada suceso elemental de ***A*** pertenece también a ***B***, es decir, siempre que ocurre el suceso ***A***, también ocurre el suceso ***B***.
Diremos también que ***A*** implica ***B***.

$A \subset B$ ó $A \Rightarrow B$

Igualdad de Sucesos

Dados dos sucesos ***A*** y ***B*** de un experimento aleatorio:

Diremos que ***A*** y ***B*** son iguales si:
Siempre que ocurre el suceso ***A*** también ocurre ***B*** y al revés.

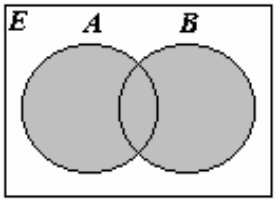
$$A = B \Leftrightarrow \left\{ \begin{matrix} A \subset B \\ B \subset A \end{matrix} \right\}$$

Unión de Sucesos

Dados dos sucesos ***A*** y ***B*** de un experimento aleatorio:

La union de ambos sucesos ***A*** y ***B*** es:
Otro suceso compuesto por los resultados o sucesos elementales pertenecientes a ***A***, a ***B***, o a los dos a la vez (intersección).

$A \cup B$



En general, dados ***n*** sucesos A_1, A_2, A_3,..., A_n, su unión es

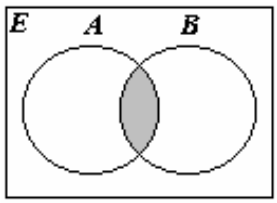
otro suceso formado por los resultados o sucesos elementales que pertenecen al menos a uno de los sucesos A_i.

$$\bigcup_{i=1}^{n} A_i$$

Intersección de Sucesos

Dados dos sucesos ***A*** y ***B*** de un experimento aleatorio:

$A \cap B$



La intersección de ambos sucesos ***A*** y ***B*** es:
Otro suceso compuesto por los resultado o sucesos elementales que pertenecen a ***A*** y a ***B***, simultáneamente.

En general, dados ***n*** sucesos A_1, A_2, A_3,..., A_n, su intersección es otro suceso formado por los resultados o sucesos elementales que pertenecen a todos los sucesos A_i.

$$\bigcap_{i=1}^{n} A_i$$

Sucesos Disjuntos, Incompatibles o Excluyentes

Dados dos sucesos ***A*** y ***B*** de un experimento aleatorio:

$$A \cap B = \varnothing$$

Diremos que estos sucesos ***A*** y ***B*** son disjuntos, incompatibles o mutuamente excluyentes cuando:
No tienen ningún suceso elemental en común o dicho de otra forma, si al verificarse ***A*** no se verifica ***B***, ni al revés.

$$\bigcap_{i=1}^{n} A_i = \varnothing$$

Sistema Exhaustivo de Sucesos

Dados ***n*** sucesos A_1, A_2, A_3,..., A_n de un experimento aleatorio:

Diremos que estos forman una colección o sistema exhaustivo de sucesos si la unión de todos ellos es igual al espacio muestral ***E***.

$$A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A_i = E$$

Diremos que estos forman un sistema completo de sucesos o una partición de ***E*** si, además de la anterior, condición, se cumple que son disjuntos dos a dos, es decir, son mutuamente excluyentes, disjuntos o incompatibles.

$$A_i \cap A_j = \varnothing \quad \forall i \neq j$$

El conjunto de todos los sucesos elementales que constituyen un espacio muestral forman una colección de sucesos mutuamente excluyente y exhaustivo ya que, de todos ellos, sólo uno debe ocurrir y no pueden ocurrir dos simultáneamente.

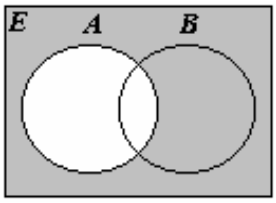
Suceso Complementario o Contrario

Dado un suceso ***A*** de un experimento aleatorio:

$$\overline{A}$$

Se define como suceso complementario o contrario de ***A*** a:
Otro suceso que ocurre cuando no ocurre el suceso ***A***, o bien, es el suceso constituido por todos los sucesos elementales del espacio muestral ***E*** que no pertenecen a

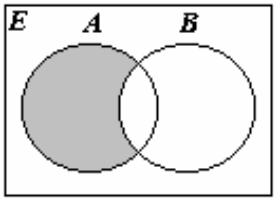
A.



Diferencia de Sucesos

Dados dos sucesos ***A*** y ***B*** de un experimento aleatorio:

$$A - B = A \cap \overline{B}$$



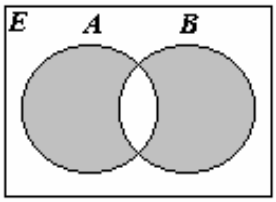
Se define como la <u>diferencia de ambos sucesos ***A*** y ***B***</u> a:
Otro suceso constituido por los sucesos elementales que pertenecen a ***A***, pero no a ***B***.

Diferencia Simétrica de Sucesos

Dados dos sucesos ***A*** y ***B*** de un experimento aleatorio:

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

$$A \Delta B = (A \cap \overline{B}) \cup (B \cap \overline{A})$$



Se define como <u>diferencia simétrica de ambos sucesos ***A*** y ***B***</u> a:
Otro suceso constituido por los sucesos elementales que pertenecen a ***A***, o a ***B***, pero que no simultáneamente a ambos.

6. Propiedades de las Operaciones con Sucesos

Los sucesos asociados a un experimento aleatorio verifican las siguientes propiedades:

$\overline{E} = \varnothing \quad \overline{\varnothing} = E \quad \overline{\overline{A}} = A$

$E \cup A = E \quad \varnothing \cup A = A \quad A \cup \overline{A} = E \quad E \cap A = A \quad \varnothing \cap A = \varnothing \quad A \cap \overline{A} = \varnothing$

Propiedad idempotente: $A \cup A = A \quad A \cap A = A$

Propiedad conmutativa: $A \cup B = B \cup A \quad A \cap B = B \cap A$

Propiedad asociativa: $A_1 \cup (A_2 \cup A_3) = (A_1 \cup A_2) \cup A_3 \quad A_1 \cap (A_2 \cap A_3) = (A_1 \cap A_2) \cap A_3$

Propiedad distributiva:
$$A_1 \cup (A_2 \cap A_3) = (A_1 \cup A_2) \cap (A_1 \cup A_3)$$
$$A_1 \cap (A_2 \cup A_3) = (A_1 \cap A_2) \cup (A_1 \cap A_3)$$

Propiedad simplificativa: $A \cup (A \cap B) = A \qquad A \cap (A \cup B) = A$

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

7. Sucesión de Sucesos

Llamaremos **sucesión de sucesos** a una familia de sucesos $A_1, A_2, A_3, \ldots, A_n$ en la que éstos aparecen ordenados por el subíndice n. La representaremos por $\{ A_n \}$ $n=1,2,3,\ldots$

Sucesión Creciente

Una sucesión de sucesos $\{ A_n \}$ $A_1 \subset A_2 \subset A_3 \subset \ldots$ diremos que es creciente si se verifica: la representamos por $\{ A_n \uparrow \}$

Sucesión Decreciente

Una sucesión de sucesos $\{ A_n \}$ $A_1 \supset A_2 \supset A_3 \supset \ldots$ diremos que es decreciente si se verifica: la representamos por $\{ A_n \downarrow \}$

Límite de una Sucesión

El límite de una sucesión creciente / decreciente de sucesos $\{ A_n \uparrow \}$ / $\{ A_n \downarrow \}$: $\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$ / $\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n$

Límites Inferior y Superior de una Sucesión de Sucesos

$$A_0 = \lim_{n\to\infty} \inf A_n = \bigcup_{n=1}^{\infty} \left(\bigcap_{k=n}^{\infty} A_k \right) \qquad A^0 = \lim_{n\to\infty} \sup A_n = \bigcap_{n=1}^{\infty} \left(\bigcup_{k=n}^{\infty} A_k \right)$$

El límite inferior de la sucesión es un suceso formado por los resultados o sucesos elementales que pertenecen a todos los sucesos de la sucesión excepto quizá a un número finito de sucesos

El límite superior de la sucesión es un suceso formado por todos los resultados o sucesos elementales que pertenecen a una infinidad de sucesos de la sucesión

En el supuesto que se verifique diremos que la sucesión es convergente y se expresa como:

$$A_0 = \lim_{n\to\infty} \inf A_n = \lim_{n\to\infty} \sup A_n = A^0 = A \qquad A_n \to A \quad \lim_{n\to\infty} A_n = A$$

8. Algebra de Sucesos

Como se ha venido observando, los sucesos los consideramos como conjuntos, siendo válido para éstos todo lo estudiado en la teoría de conjuntos. Para llegar a la construcción axiomática del Cálculo de Probabilidades, necesitamos dar unas estructuras algebraicas básicas construidas sobre los sucesos, de la misma manera que se construyen sobre los conjuntos.

Colección de Conjuntos o Sucesos

Es otro conjunto cuyos elementos son conjuntos y lo llamaremos **conjunto de las partes de *E***, es decir, ***A=P(E)*** es el conjunto formado por todos los subconjuntos de ***E*** o por todos los sucesos contenidos en el espacio muestral ***E***.

Algebra de Sucesos o Algebra de Boole

Sea ***A=P(E)*** una colección de sucesos donde se han definido las operaciones:
- Unión de sucesos - Intersección de sucesos - Complementario de un suceso
y que además verifican las propiedades definidas al exponer las operaciones con sucesos.
Diremos que la colección de sucesos no vacia ***A*** **tiene estructura de Algebra de Boole** si ***A*** es una clase cerrada frente a las operaciones de complementario, unión e intersección de sucesos en número finito, es decir si se verifican las condiciones siguientes:

- $\forall A \in A = P(E)$ se verifica que su complementario $\overline{A} \in A = P(E)$
- $\forall A_1, A_2 \in A = P(E)$ se verifica que $A_1 \cup A_2 \in A = P(E)$
- Lo relativo a que la intersección sea cerrada y que el número de sucesos sea finito se obtiene como consecuencia de las condiciones anteriores, como ahora se indicará

De las dos condiciones iniciales, se deducen las siguientes consecuencias:

- El espacio muestral $E \in A = P(E)$
- Si los sucesos $A, B \in A = P(E)$ se verifica que $A \cap B \in A = P(E)$
- El suceso imposible $\varnothing \in A = P(E)$
- Si $A_1, A_2, A_3, \ldots, A_n \in A = P(E)$ se verifica que $\bigcup_{i=1}^{n} A_i \in A = P(E)$ y $\bigcap_{i=1}^{n} A_i \in A = P(E)$

Si hacemos la extensión al caso de un número infinito numerable de sucesos, entonces aparece una nueva estructura algebraica que recibe el nombre de **α-Algebra** o **Campo de Borel**, que es una generalización de la anterior.

Cuando el espacio muestral ***E*** es finito, todos los subconjuntos de ***E*** se pueden considerar como sucesos. Esto no ocurre cuando el espacio muestral es infinito (no numerable), pues es difícil considerar el conjunto formado por todos los subconjuntos posibles, existiendo subconjuntos que no pueden considerarse como sucesos.

En resumen, podemos decir que a partir del espacio muestral ***E***, hemos llegado a definir la colección de sucesos ***A=P(E)*** que tiene estructura de **Algebra de Sucesos o Algebra de Boole** si el espacio muestral es finito o bien tiene la estructura de **α-Algebra** si el espacio muestral es infinito.

Al par (***E, A***) en donde ***E*** es el espacio muestral y ***A*** una **α-Algebra** sobre ***E***, le llamaremos espacio o conjunto medible, en el cual será posible establecer una medida o probabilidad, como se verá después.

9. Métodos de Enumeración o Conteo

Las siguientes son algunas técnicas útiles para contar el número de resultados o sucesos de un experimento aleatorio.

Tablas de Doble Entrada

Es útil para relacionar dos pruebas, indicándonos los resultados que integran el espacio muestral, pudiendo indicar sobre la tabla determinados sucesos en los que estemos interesados. En general con m elementos $a_1, a_2, a_3,..., a_m$ y n elementos $b_1, b_2, b_3,..., b_n$ es posible formar $m \times n$ pares (a_r , b_s) tales que cada par tiene al menos algún elemento diferente de cada grupo.

Principio de Multiplicación

Sean los conjuntos $C_1, C_2, C_3,..., C_k$ que tienen respectivamente $n_1, n_2, n_3,..., n_k$ k-uplas donde, en cada k-upla, el primer elemento pertenece a C_1, el segundo a C_2, etc.

- En el caso particular de que $n_1 = n_2 = n_3 = ... = n_k$ el número posible de k-uplas sera n^k .
- En el caso general, el número de posibles resultados será $n_1 \times n_2 \times n_3 \times ... \times n_k$. Este principio es de utilidad en el caso de un experimento aleatorio compuesto por otros k experimentos.

Diagramas de árbol

Este diagrama nos permite indicar de manera sencilla el conjunto de posibles resultados en un experimento aleatorio siempre y cuando los resultados del experimento puedan obtenerse en diferentes fases sucesivas.
Ej: Experimento aleatorio consistente en lanzar al aire un dado y después 3 veces consecutivas una moneda.

Combinaciones, Variaciones y Permutaciones

Combinaciones Llamaremos <u>combinaciones de m elementos</u> <u>tomados de n en n</u> al número de subconjuntos diferentes de n elementos que se pueden formar con los m elementos del conjunto inicial	$C_{m,n} = \binom{m}{n} = \frac{m!}{n!(m-n)!}$
Combinaciones con repetición Si en los subconjuntos anteriores se pueden repetir los elementos	$CR_{m,n} = \binom{m+n-1}{n} = \frac{(m+n-1)!}{n!(m-1)!}$
Variaciones Llamaremos <u>variaciones de m elementos</u> <u>tomados de n en n</u> a los distintos subconjuntos diferentes de n elementos que se pueden formar con los m elementos, influyendo el orden en el que se toman	$V_{m,n} = m(m-1)(m-2)...(m-m+1) = \frac{m!}{(m-n)!}$
Variaciones con repetición Si en los subconjuntos anteriores se pueden repetir los elementos	$VR_{m,n} = m^n$

Permutaciones Llamaremos permutaciones de n elementos a las variaciones de n elementos tomados de n en n	$P_n = n!$
Permutaciones con repetición Llamaremos permutaciones con repeticion de n elementos k-distintos que se repiten uno x_1 veces, otro x_2 veces, ... y el último x_k veces	$P_n^{x_1,x_2,...,x_k} = \frac{n!}{x_1!\, x_2! \ldots x_k!}$ $x_1 + x_2 + \ldots + x_k = n$

Capítulo VII	PROBABILIDAD

1. Introducción

- Se indicaba en el capítulo anterior que cuando un experimento aleatorio se repite un gran número de veces, los posibles resultados tienden a presentarse un número muy parecido de veces, lo cual indica que la frecuencia de aparición de cada resultado tiende a estabilizarse.
- El concepto o idea que generalmente se tiene del término probabilidad es adquirido de forma intuitiva, siendo suficiente para manejarlo en la vida corriente.

Nos interesa ahora la medida numérica de la posibilidad de que ocurra un suceso ***A*** cuando se realiza el experimento aleatorio. A esta medida la llamaremos **probabilidad del suceso *A*** y la representaremos por ***p(A)***.

La probabilidad es una medida sobre la **escala 0 a 1** de tal forma que:
- Al suceso imposible le corresponde el valor 0
- Al suceso seguro le corresponde el valor 1
- El resto de sucesos tendrán una probabilidad comprendida entre 0 y 1

El concepto de probabilidad no es único, pues se puede considerar desde distintos puntos de vista:
- El punto de vista objetivo
  - Definición clásica o a priori
  - Definición frecuentista o a posteriori
- El punto de vista subjetivo

2. Definición Clásica de la Probabilidad

Sea un experimento aleatorio cuyo correspondiente espacio muestral ***E*** está formado por un número ***n*** finito de posibles resultados distintos y con la misma probabilidad de ocurrir $\{e_1, e_2, \ldots, e_n\}$.
Si n_1 resultados constituyen el subconjunto o suceso A_1, n_2 resultados constituyen el subconjunto o suceso A_2 y, en general, n_k resultados constituyen el subconjunto o suceso A_k de tal forma que:

$$n_1 + n_2 + \ldots + n_k = n$$

Las probabilidades de los sucesos A_1, A_1,..., A_n son:

$$p(A_1) = \frac{n_1}{n} \quad p(A_2) = \frac{n_2}{n} \quad \ldots \quad p(A_k) = \frac{n_k}{n}$$

es decir, que la probabilidad de cualquier suceso ***A*** es igual al cociente entre el número de casos favorables que integran el suceso ***A***

Regla de Laplace para ***E*** finitos

y el número
de casos posibles del espacio muestral ***E***. $$p(A) = \frac{N^o \text{ de casos favorables de } A}{N^o \text{ de casos posibles de } E}$$

- Para que se pueda aplicar la regla de Laplace es necesario que todos los sucesos elementales sean equiprobables, es decir: $p(e_1) = p(e_2) = \ldots = p(e_n)$ y por tanto $p(e_i)=1/n \quad \forall i=1,2,\ldots,n$

- Siendo $A=\{e_1, e_2, \ldots, e_k\}$ el suceso formado por k sucesos elementales siendo $k \leq n$ tendremos: $$p(A) = \sum_{j=1}^{n} p(e_j) = \frac{k}{n} = \frac{N^o \text{ casos favorables}}{N^o \text{ casos posibles}}$$

La probabilidad verifica las siguientes condiciones:

- La probabilidad de cualquier suceso es siempre un número no negativo entre 0 y 1 $$p(A) = \frac{n_i}{n} \qquad n_i \leq n \qquad n_i, n \geq 0$$

- La probabilidad del suceso seguro ***E*** vale 1
- La probabilidad del suceso inmposible es 0 $$p(E) = \frac{n}{n} = 1 \qquad p(\varnothing) = \frac{0}{n} = 0$$

- La probabilidad de la unión de varios sucesos incompatibles o excluyentes $A_1, A_1, \ldots, A_r$ es igual a la suma de probabilidades de cada uno de ellos $$p(A_1 \cup \ldots \cup A_r) = p(A_1) + p(A_2) + \ldots + p(A_r)$$

Esta definición clásica de probabilidad fue una de las primeras que se dieron (1900) y se atribuye a Laplace; también se conoce con el nombre de **probabilidad a priori** pues, para calcularla, es necesario conocer, antes de realizar el experimento aleatorio, el espacio muestral y el número de resultados o sucesos elementales que entran a formar parte del suceso.

La aplicación de la definicion clásica de probabilidad puede presentar dificultades de aplicación cuando el espacio muestral es infinito o cuando los posibles resultados de un experimento no son equiprobables. Ej: En un proceso de fabricación de piezas puede haber algunas defectuosas y si queremos determinar la probabilidad de que una pieza sea defectuosa no podemos utilizar la definición clásica pues necesitaríamos conocer previamente el resultado del proceso de fabricación.
Para resolver estos casos, se hace una extensión de la definición de probabilidad, de manera que se pueda aplicar con menos restricciones, llegando así a la definición frecuentista de probabilidad.

3. Definición Frecuentista de la Probabilidad

La definición frecuentista consiste en definir la probabilidad como el límite cuando n tiende a infinito de la proporción o frecuencia relativa del suceso.

Sea un experimento aleatorio cuyo espacio muestral es ***E***
Sea ***A*** cualquier suceso perteneciente a ***E***
Si repetimos ***n*** veces el experimiento en las mismas condiciones, la frecuencia relativa del suceso ***A*** sera:

$$\frac{n(A)}{n}$$

Cuando el número ***n*** de repeticiones se hace muy grande
la frecuencia relativa converge hacia un valor que llamaremos **probabilidad del suceso *A***.

$$p(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

Es imposible llegar a este límite, ya que no podemos repetir el experimiento un número infinito de veces, pero si podemos repetirlo muchas veces y observar como las frecuencias relativas tienden a estabilizarse.

Esta definición frecuentista de la probabilidad se llama también **probabilidad a posteriori** ya que sólo podemos dar la probabilidad de un suceso después de repetir y observar un gran número de veces el experimento aleatorio correspondiente. Algunos autores las llaman **probabilidades teóricas**.

4. Definición Subjetiva de la Probabilidad

Tanto la definición clásica como la frecuentista se basan en las repeticiones del experimento aleatorio; pero existen muchos experimentos que no se pueden repetir bajo las mismas condiciones y por tanto no puede aplicarse la interpretación objetiva de la probabilidad.
En esos casos es necesario acudir a un punto de vista alternativo, que no dependa de las repeticiones, sino que considere la probabilidad como un concepto **subjetivo** que exprese el grado de creencia o confianza individual sobre la posibilidad de que el suceso ocurra.
Se trata por tanto de un juicio personal o individual y es posible por tanto que, diferentes observadores tengan distintos grados de creencia sobre los posibles resultados, igualmente válidos.

5. Definición Axiomática de la Probabilidad

La definición axiomática de la probabilidad es quizás la más simple de todas las definiciones y la menos controvertida ya que está basada en un conjunto de axiomas que establecen los requisitos mínimos para dar una definición de probabilidad.

La ventaja de esta definición es que permite un desarrollo riguroso y matemático de la probabilidad. Fue introducida por A. N. Kolmogorov y aceptada por estadísticos y matemáticos en general.

Definición

Dado el espacio muestral ***E*** y la α**-Algebra *A=P(E)*** diremos que una función ***p: A*** $\rightarrow$ ***[0,1]*** es una probabilidad si satisface los siguientes **axiomas de Kolmogorov**:

- $p(A) \geq 0$ para cualquier suceso $A \in A=P(A)$
- $p(E) = 1$
- Dada una sucesión numerable de sucesos incompatibles $A_1, A_2, \ldots \in A$, se verifica que

$$p(A_1 \cup A_2 \cup \ldots) = p\left(\bigcup_{i=1}^{\infty} A_i\right) = p(A_1) + p(A_2) + \ldots$$

A la función $p: A \rightarrow [0,1]$
$A \rightarrow p = p(A)$ se denomina **probabilidad del suceso *A***.

La terna (***E***, ***A***, ***p***) formada por el espacio muestral ***E***, la **α-Algebra *A=P(E)*** y la probabilidad ***p*** se denomina **espacio probabilístico**.

6. Teoremas Elementales o Consecuencias de los Axiomas

Los siguientes resultados se deducen directamente de los axiomas de probabilidad.

Teorema I

La probabilidad del suceso imposible es nula $p(\varnothing) = 0$

- Si para cualquier suceso ***A*** resulta que ***p(A)=0*** diremos que ***A*** es el suceso **nulo**, pero esto no implica que $A = \varnothing$
- Si para cualquier suceso ***A*** resulta que ***p(A)=1*** diremos que ***A*** es el suceso **casi seguro**, pero esto no implica que $A = E$

Teorema II

Para cualquier suceso $A \in A=P(A)$ se verifica que:
La probabilidad de su suceso complementario es $p(\overline{A}) = 1 - p(A)$

Teorema III

La probabilidad ***P*** es monótona no decreciente, es decir:
$\forall A,B \in A=P(A)$ con $A \subset B \Rightarrow p(A) \leq p(B)$ y además $p(B - A) = p(B) - p(A)$

Teorema IV

Para cualquier suceso $A \in A=P(A)$ se verifica que: $p(A) \leq 1$

Teorema V

Para dos sucesos cualesquiera $A,B \in A=P(A)$ se verifica que: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Esta propiedad es generalizable a n sucesos:

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{n} p(A_i) - \sum_{i<j}^{n} p(A_i \cap A_j) + \sum_{i<j<k}^{n} p(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} p\left(\bigcap_{i=1}^{n} A_i\right)$$

Teorema VI

Para dos sucesos cualesquiera $A,B \in A=P(A)$ se verifica que: $p(A \cup B) \leq p(A) + p(B)$

Esta propiedad es generalizable a n sucesos:

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{n} p(A_i)$$

Teorema VII

Dada una sucesión creciente de sucesos $A_1, A_2, \ldots, A_n$ (abreviadamente representado por $\{A_n \uparrow\}$)

se verifica que:

$$\lim_{n\to\infty} p(A_n) = p\left(\lim_{n\to\infty} A_n\right) = p\left(\bigcup_{n=1}^{\infty} A_n\right)$$

Teorema VIII

Dada una sucesión decreciente de sucesos $A_1, A_2, \ldots, A_n$ (abreviadamente representado por $\{A_n \downarrow\}$)

se verifica que:

$$\lim_{n\to\infty} p(A_n) = p\left(\lim_{n\to\infty} A_n\right) = p\left(\bigcap_{n=1}^{\infty} A_n\right)$$

7. Probabilidad Condicionada

Hasta ahora hemos introducido el concepto de probabilidad considerando que la única información sobre el experimiento era el espacio muestral. Sin embargo hay situaciones en las que se incorpora información suplementaria respecto de un suceso relacionado con el experimento aleatorio, cambiando su probabilidad de ocurrencia.

El hecho de introducir más información, como puede ser la ocurrencia de otro suceso, conduce a que determinados sucesos no pueden haber ocurrido, variando el espacio de resultados y cambiando sus probabilidades.

Definición

Dado un espacio probabilístico (E, A, p) asociado a un experimento aleatorio.
Sea A un suceso tal que $A \in A=p(A)$ y $p(A) \geq 0$
Sea B un suceso tal que $B \in A=p(A)$

Se define la **probabilidad condicionada de B dado A** o
probabilidad de B condicionada a A como:

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

La probabilidad condicionada cumple los **tres axiomas de Kolmogorov**:

- $\forall B \in A = P(E) \quad p(B/A) = \frac{p(A \cap B)}{p(A)} \geq 0$
- $p(E/A) = \frac{p(A \cap E)}{p(A)} = 1$
- Sea { A_i } una sucesión de sucesos disjuntos dos a dos, entonces: $p\left(\bigcup_{i=1}^{\infty} A_i / A\right) = \sum_{i=1}^{\infty} p(A_i / A)$

Regla de Multiplicación de Probabilidades o Probabilidad Compuesta

Partiendo de la definición de la probabilidad condicionada ***p(B/A)*** podemos escribir: $p(A \cap B) = p(A)\ p(B/A)$

8. Teorema de la Probabilidad Compuesta o Producto

Sean ***n*** sucesos $A_1, A_2, \ldots, A_n \in A = P(A)$ y tales que $p\left(\bigcap_{i=1}^{n-1} A_i\right) > 0$. Se verifica que:

$$p(A_1 \cap A_2 \cap \ldots \cap A_n) = p(A_1)\, p(A_2/A_1)\ p(A_3/A_1 \cap A_2) \ldots p(A_n/A_1 \cap \ldots \cap A_{n-1})$$

9. Teorema de la Probabilidad Total

Sean ***n*** sucesos disjuntos $A_1, A_2, \ldots, A_n \in A = P(A)$ tales que $p(A_i) > 0 \quad i=1,2,\ldots,n$ y tales que forman un sistema completo de sucesos. Para cualquier suceso $B \in A = P(A)$ cuyas probabilidades condicionadas son conocidas $p(B/A_i)$, se verifica que:

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

10. Teorema de Bayes

Sean ***n*** sucesos disjuntos $A_1, A_2, \ldots, A_n \in A = P(A)$ tales que $p(A_i) > 0 \quad i=1,2,\ldots,n$ y tales que forman un sistema completo de sucesos. Para cualquier suceso $B \in A = P(A)$ se verifica que:

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

y aplicando el teorema de la probabilidad total:

$$p(A_i/B) = \frac{p(A_i)\ p(B/A_i)}{p(B)}$$

Sistema completo de sucesos $A_1, A_2, \ldots, A_n$ Se denominan **hipótesis**

Las probabilidades $p(A_i) > 0$ Se denominan **probabilidades a priori** ya

i=1,2,...,n	que son las que se asignan inicialmente al los sucesos A_i
Las probabilidades $p(B/A_i)>0$ ***i=1,2,...,n***	Se denominan **verosimilitudes** del suceso ***B*** admitiendo la hipótesis A_i

Las verosimilitudes $p(B/A_i)$ nos permiten modificar nuestro grado de creencia original $p(A_i)$ obteniendo la probabilidad a posteriori $p(A_i/B)$.

El teorema de Bayes, además de ser una aplicación de las probabilidades condicionadas, es fundamental para el desarrollo de la estadística bayesiana, la cual utiliza la interpretación subjetiva de la probabilidad.

11. Independencia de Sucesos

Teniendo en cuenta la definición de la probabilidad del suceso ***B*** condicionada a ***A*** se puede decir:

- Cuando $p(B/A) > P(B)$ entonces el suceso ***A* favorece al *B***
- Cuando $p(B/A) < P(B)$ entonces el suceso ***A* desfavorece al *B***
- Cuando $p(B/A) = P(B)$ entonces la ocurrencia de ***A*** no tiene ningún efecto sobre la de ***B***

Diremos que dos sucesos ***A*** y ***B* son independientes** si se verifica una cualquiera de las siguientes condiciones equivalentes:

- $p(B/A) = p(B)$ si $P(A)>0$
- $p(A/B) = p(B)$ si $P(B)>0$
- $p(A \cap B) = p(A)\ p(B)$

Podemos decir por tanto que si el suceso ***B*** es independiente del suceso ***A***, entonces el suceso ***A*** también es independiente del suceso ***B***, lo que equivale a decir que ambos sucesos son **mutuamente independientes**.

La independencia de sucesos puede extenderse a más de dos sucesos: $p(A \cap B \cap C) = p(A)\ p(B)\ p(C)$
Además, se cumple el siguiente teorema: Si ***A*** y ***B*** son dos sucesos independientes, entonces también lo son los sucesos $\overline{A}$ y B A y $\overline{B}$ $\overline{A}$ y $\overline{B}$

Capítulo VIII VARIABLES ALEATORIAS Y SUS DISTRIBUCIONES

1. Variable Aleatoria Unidimensional

La relación entre los sucesos del espacio muestral y el valor numérico que se les asigna se establece a través de variable aleatoria.

Definición
Una variable aleatoria es una función que asigna un valor numérico a cada suceso elemental del espacio muestral.
Es decir, una variable aleatoria es una variable cuyo valor numérico está determinado por el resultado del experimento aleatorio. La variable aleatoria la notaremos con letras en mayúscula X, Y, ... y con las letras en minúscula x, y, ... sus valores.

La v.a. puede tomar un número numerable o no numerable de valores, dando lugar a dos tipos de v.a.: discretas y continuas.

Definición
Se dice que una variable aleatoria X es discreta si puede tomar un número finito o infinito, pero numerable, de posibles valores.

Definición
Se dice que una variable aleatoria X es continua si puede tomar un número infinito (no numerable) de valores, o bien, si puede tomar un número infinito de valores correspondientes a los puntos de uno o más intervalos de la recta real.

1.1. DISTRIBUCIÓN DE PROBABILIDAD DE VARIABLES ALEATORIAS DISCRETAS

Sea X una v.a. discreta que toma un número finito de valores, r en total, x_1, x_2, ..., x_i, ..., x_r.

La probabilidad de que la v.a. X tome un valor particular x_i se representará por:

$$p_i = P(x_i) = P(X = x_i).$$

Definición
La distribución de probabilidad o función de probabilidad de una variable aleatoria X, P(x), es una función que asigna las probabilidades con que la v.a. toma los posibles valores, de forma que las probabilidades verifiquen:

1) $0 \le P(x_i) \le 1, \qquad i = 1,2,...,r$

2) $\sum_{i=1}^{r} P(x_i) = 1$

Si X es una v.a. discreta, para determinar la $P[a \le X \le b]$, siendo $a \le b$, tan sólo hay que sumar las probabilidades correspondientes a valores de X comprendidos entre a y b, es decir: $P[a \le X \le b] = \sum_{x=a}^{x=b} P(X = x) = \sum_{x=a}^{x=b} P(x)$

Definición

Sea una v.a. X de tipo discreto que toma un nº finito o infinito numerable de valores x_1, x_2, ..., x_r y cuya distribución de probabilidad es P(x). Se define la función de distribución acumulativa de la v.a. X, F(x), como la probabilidad de que la v.a. X tome valores menores o iguales que x, es decir:

$$F(x) = P(X \le x) = \sum_{x_i \le x} P(x_i) = \sum_{x_i = x_1}^{x} P(X = x_i)$$

Representando la suma de las probabilidades puntuales hasta el valor x inclusive de la v.a. X.

La función de distribución de una v.a. discreta es una función que verifica:

- $0 \le F(x) \le 1, \quad \forall x$
- Es no decreciente, es decir, si $x_i \le x_j$, entonces $F(x_i) \le F(x_j)$.

Se puede decir por tanto que una v.a. X discreta, está caracterizada por su función de probabilidad o distribución de probabilidad P(x) y también por su función de distribución F(x).

1.2. DISTRIBUCIÓN DE PROBABILIDAD DE VARIABLES ALEATORIAS CONTINUAS

La v.a. de tipo continuo se tratará de forma diferente a como se ha visto en el caso de v.a. discreta, ya que en el caso continuo no es posible asignar una probabilidad a cada uno de los infinitos posibles valores de la v.a. y que estas probabilidades sumen uno (como en el caso discreto), teniendo por tanto que utilizar una aproximación diferente para llegar a obtener la distribución de probabilidad de una v.a. continua.

Definición

Sea X una v.a. continua. Si existe una función f(x) tal que verifica:

I. $f(x) \ge 0, \quad -\infty < x < +\infty$

II. $\int_{-\infty}^{+\infty} f(x)dx = 1$

Diremos que f(x) es la función de densidad de la v.a. continua X.

En el caso continuo la suma de las densidad o el área bajo la curva f(x) debe ser igual a la unidad, y como consecuencia la probabilidad de que la v.a. continua X tome valores en el intervalo [a,b] será igual al área bajo la curva f(x) acotada entre a y b, es decir:

$$P(a \le X \le b) = \int_a^b f(x)dx$$

Si la v.a. X es discreta, existen las probabilidades puntuales $P(X=x_i)$. En el caso continuo, si x_i es un punto interior al intervalo de definición de la v.a. continua X, entonces $P(X=x_i) = 0$: $P(X = x_i) = P(x_i \le X \le x_i) = 0$
es decir, el área entre x_i y x_i es nula.

En el caso continuo $P(X=x_i) = 0$, con lo que se verifica que:
$P(a \le X \le b) = P(a < x < b)$ ya que $P(X = a) = 0$ y $P(X = b) = 0$ lo cual no sucede en el caso discreto.

Definición
Sea X una v.a. de tipo continuo que toma un número infinito de valores sobre la recta real y con función de densidad f(x). Se define la función de distribución acumulativa de la v.a. X, F(x), como la probabilidad de que la v.a. continua X tome valores menores o iguales a x, es decir:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx$$

y representa el área limitada por la curva f(x), a la izquierda de la recta X=x.

Consideramos una masa unidad distribuida sobre el intervalo $(-\infty,+\infty)$, y la función de distribución F(x) para cada valor x de la v.a. nos da la cantidad de masa que hay en el intervalo $(-\infty, x]$, es decir, la masa que hay en el punto x y a la izquierda de x, aunque nos dará igual considerar el intervalo $(-\infty, x)$, de forma que:
$F(x) = P(X \le x) = P(X < x)$ ya que $P(X = x) = 0$
y también $1 - F(x) = P(X \ge x) = P(X > x)$

La función de distribución de una v.a. continua es una función que verifica:

1. $F(-\infty) = 0$
2. $F(+\infty) = 1$
3. Es no decreciente, es decir, si $x_i < x_j$ entonces $F(x_i) \le F(x_j)$
4. $P(a \le X \le b) = \int_a^b f(x)dx = F(b) - F(a)$
5. La derivada de la función de distribución es la función de densidad

$$\frac{dF(x)}{dx} = f(x)$$

Como la $\int_a^b f(x)dx$ representa gráficamente el área encerrada bajo la curva f(x) y los valores a y b del eje de abscisas, entonces a la probabilidad

$$P(a \le X \le b) = F(b) - F(a)$$

se le puede dar la misma interpretación.

2. VARIABLE ALEATORIA BIDIMENSIONAL

En determinadas ocasiones hay que trabajar en espacios de más de una dimensión, estableciendo aplicaciones que transforman los sucesos elementales del experimento aleatorio en puntos del espacio n-dimensional (R^n), estas aplicaciones se hacen utilizando v.a. bidimensionales o n-dimensionales.

En muchas ocasiones nos puede interesar estudiar conjuntamente dos características del fenómeno aleatorio, es decir, estudiar el comportamiento conjunto de dos v.a. para intentar explicar la posible relación entre ellas. Para poder estudiar conjuntamente las dos v.a., es necesario conocer la distribución de probabilidad conjunta.

2.1. DISTRIBUCIÓN DE PROBABILIDAD BIDIMENSIONAL

Sea la X una v.a. discreta que toma un número finito de valores $x_1, x_2, ..., x_r$ y sea Y una v.a. de tipo discreto, que toma valores $y_1, y_2, ..., y_s$.
La probabilidad de que la v.a. X tome el valor x_i, y la v.a. Y toma el valor y_j, la designaremos por:

$$p_{ij} = P(x_i, y_j) = P(X = x_i, Y = y_j)$$

Definición
La distribución de probabilidad bidimensional o distribución de probabilidad conjunta de una v.a. discreta bidimensional es una función $P(x_i,y_j)$ que asigna las probabilidades a los diferentes valores conjuntos de la v.a. bidimensional (X,Y), de tal manera que se verifiquen las dos condiciones siguientes:

I. $0 \le P(x_i, y_j) \le 1, \quad i = 1, 2, ..., r; \quad j = 1, 2, ..., s$

II. $\sum_{i=1}^{r} \sum_{j=1}^{s} P(x_i, y_j) = 1$

Definición
Sea una v.a. bidimensional (X,Y) de tipo discreto cuya distribución de probabilidad es $p_{ij} = P(x_i,y_j)$, i=1, 2, ..., r y j = 1, 2, ..., s. Se define la función de distribución conjunta, F(x,y) como:

$$F(x,y) = P(X \le x, Y \le y) = \sum_{x_i \le x} \sum_{y_j \le y} P(x_i, y_j) = \sum_{x_i = x_1}^{x} \sum_{y_j = y_1}^{y} P(X = x_i, Y = y_j)$$

y representa la suma de las probabilidades puntuales $P(x_i,y_j)$ hasta el valor (x , y) inclusive de la v.a. bidimensional (X,Y).

Consideremos ahora una v.a. bidimensional (X,Y) de tipo continuo:

Definición

Sea (X,Y) una v.a. bidimensional de tipo continuo, si existe una función f(x,y) tal que verifica:

I. $f(x,y) \geq 0, \quad -\infty < x < +\infty, \quad -\infty < y < +\infty$

II. $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)\,dx\,dy = 1$

diremos que f(x,y) es la función de densidad de la v.a. bidimensional continua (X,Y).

Esta función de densidad conjunta, se puede interpretar como un histograma de frecuencias relativas conjuntas para X e Y, pues la función de densidad f(x,y) representa una superficie de densidad de probabilidad en el espacio tridimensional , y el volumen por debajo de esta superficie y por encima del rectángulo $x_1 \leq X \leq x_2$ e $y_1 \leq Y \leq y_2$, es igual a la probabilidad de que las v.a. tienen valores dentro del rectángulo indicado, es decir:

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} f(x,y)dxdy$$

Definición

Sea una v.a. bidimensional (X,Y) de tipo continuo que toma valores sobre el espacio bidimensional R^2 y cuya función de densidad es f(x,y). Se define la función de distribución de la v.a. bidimensional, F(x,y) como:

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy$$

La función de distribución bidimensional satisface una serie de propiedades:

1. $F(-\infty,-\infty) = F(-\infty,y) = F(x,-\infty) = 0$
2. $F(+\infty,+\infty) = 1$
3. Es monótona no decreciente respecto a las dos variables, es decir :

Si $x_1 < x_2 \Rightarrow F(x_1,y) \leq F(x_2,y)$

Si $y_1 < y_2 \Rightarrow F(x,y_1) \leq F(x,y_2)$

4. Si $a_1 < a_2$ y $b_1 < b_2$, siendo $a_1, a_2, b_1, b_2 \in R$ entonces :

$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) \geq 0$

5. La derivada parcial segunda respecto de x e y de la función de distribución F(x,y) es la función de densidad

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$$

3. DISTRIBUCIÓNES MARGINALES

Ahora nos va a interesar conocer la distribución de alguna o de ambas v.a. por separado, a partir de la información que nos proporcionaba la distribución conjunta de (X,Y), lo cual nos lleva al concepto de distribución marginal.

En el caso de la v.a. bidimensional (X,Y), teniendo en cuenta que todos los sucesos bivariantes, correspondientes a los diferentes valores que puede tomar la v.a. bidimensional, es decir: $(X=x_i, Y=y_j)$ i, j = 1, 2, 3, ...

eran sucesos mutuamente excluyentes, entonces podremos decir que el suceso univariante $X=x_i$ es la unión de todos los sucesos bivariantes $(X=x_i, Y=y_j)$ para todos los valores posibles de y_j.

Definición
Sea una v.a. bidimensional (X,Y) de tipo discreto cuya distribución de probabilidad es $p_{ij}=P(x_i, y_j)$. Entonces las distribuciones de probabilidad marginales, de X e Y, respectivamente vienen dadas por

$$P_{i.} = \sum_j p_{ij} = P(x_i) = \sum_{y_j} P(x_i, y_j) = \sum_{y_j} P(X = x_i, Y = y_j)$$

$$P_{.j} = \sum_i p_{ij} = P(y_j) = \sum_{x_i} P(x_i, y_j) = \sum_{x_i} P(X = x_i, Y = y_j)$$

La distribución de probabilidad marginal de X es la probabilidad de que $X=x_i$ con independencia del valor que tome Y, con lo que se puede escribir:

$$P_{i.} = P(X = x_i) = \sum_j p_{ij}$$

De forma análoga la distribución de probabilidad marginal de Y será:

$$P_{.j} = P(Y = y_j) = \sum_i p_{ij}$$

Definición
Sea una v.a. bidimensional (X,Y) de tipo discreto cuya distribución de probabilidad es $p_{ij} = P(x_i,y_j)$. Entonces las funciones de distribución marginales de X e Y, se definen:

$$F_1(x) = F(x,+\infty) = P(X \le x, Y < +\infty) = \sum_{x_i \le x} \sum_{y_j} P(x_i, y_j) = \sum_{x_i \le x} P_{i.}$$

$$F_2(y) = F(+\infty,y) = P(X < +\infty, Y \le y) = \sum_{y_j \le y} \sum_{x_i} P(x_i, y_j) = \sum_{y_j \le y} P_{.j}$$

Si consideramos una v.a. bidimensional de tipo continuo (X,Y) se obtiene:

Definición. Sea una v.a. bidimensional (X,Y) de tipo continuo, cuya función de densidad conjunta es f(x,y). Entonces las funciones de densidad marginales de X e Y, se definen:

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

Definición. Sea una v.a. bidimensional (X,Y) de tipo continuo, cuya función de densidad conjunta es f(x,y). Entonces se definen las funciones de distribución marginales de X e Y como:

$$F_1(x) = F(x,+\infty) = P(X \le x, Y < +\infty) = \int_{-\infty}^{x} \left(\int_{-\infty}^{+\infty} f(x,y)dy \right) dx = \int_{-\infty}^{x} f_1(x)dx$$

$$F_2(y) = F(+\infty,y) = P(X < +\infty, Y \le y) = \int_{-\infty}^{y} \left(\int_{-\infty}^{+\infty} f(x,y)dx \right) dy = \int_{-\infty}^{y} f_2(y)dy$$

4. DISTRIBUCIÓNES CONDICIONADAS

Queremos conocer cómo se distribuye una de las variables cuando se imponen condiciones a la otra variable. En el tema anterior definiamos

$$P(B/A)=\frac{P(A\cap B)}{P(A)},\qquad P(A)>0\,.$$

Ahora consideramos la v.a. bidimensional (X,Y), con su correspondiente distribución de probabilidad.

Se define la probabilidad condicionada de la v.a. discreta X dado $Y=y_j$

$$P(x_i/y_j)=P(X=x_i/Y=y_j)=\frac{P(X=x_i,Y=y_j)}{P(Y=y_j)}=\frac{p_{ij}}{p_{.j}},$$

siempre que $P(Y=y_j)>0$

En este caso y_j es fijo y x_i varía sobre todos los posibles valores de la v.a. X, obteniendo así la distribución de probabilidad de la v.a. discreta X bajo la condición $Y=y_j$.

Análogamente la distribución de probabilidad condicionada de la v.a. discreta Y dado $X=x_i$ sería:

$$P(y_j/x_i)=P(Y=y_j/X=x_i)=\frac{P(X=x_i,Y=y_j)}{P(X=x_i)}=\frac{p_{ij}}{p_{i.}},$$

siempre que $P(X=x_i)>0$

En este caso x_i es fijo e y_j varía sobre todos los posibles valores de la variable aleatoria Y.

Lógicamente :

$P(x_i/y_j)\geq 0$
$P(y_j/x_i)\geq 0$ y además $\sum_i P(X=x_i/Y=y_j)=\frac{\sum_i p_{ij}}{p_{.j}}=\frac{p_{.j}}{p_{.j}}=1$

$$\sum_j P(Y=y_j/X=x_i)=\frac{\sum_j p_{ij}}{p_{i.}}=\frac{p_{i.}}{p_{i.}}=1$$

Función de distribución condicionada de la v.a. discreta X dado $Y=y_j$:

$$F(x/y_i)=P(X\leq x/Y=y_j)=\frac{\sum_{x_i\leq x}P(X=x_i,Y=y_j)}{P(Y=y_j)}=\frac{\sum_{x_i\leq x}p_{ij}}{p_{.j}},\quad \text{siempre que}$$

$P(Y=y_j)=p_{.j}>0$

Análogamente la función de distribución condicionada de la v.a. discreta Y dado $X=x_i$.

$$F(y/x_i)=P(Y\leq y/X=x_i)=\frac{\sum_{y_j\leq y}P(X=x_i,Y=y_j)}{P(X=x_i)}=\frac{\sum_{y_j\leq y}p_{ij}}{p_{i.}},\quad \text{siempre que}$$

$P(X=x_i)=p_{i.}>0$

Función de densidad condicionada de la v.a. continua *X* dado *Y=y*:

$$f(x/y)=\begin{cases}\dfrac{f(x,y)}{f_2(y)}, & f_2(y)>0\\ 0, & \text{en el resto}\end{cases}$$

Función de densidad condicionada de la v.a. continua *Y* dado *X=x*:

$$f(y/x)=\begin{cases}\dfrac{f(x,y)}{f_1(x)}, & f_1(x)>0\\ 0, & \text{en el resto}\end{cases}$$

Función de distribución condicionada de la v.a. continua X dado Y=y:

$$F(x/y)=\int_{-\infty}^{x} f(x/y)dx=\frac{\int_{-\infty}^{x} f(x,y)dx}{f_2(y)}$$

Función de distribución condicionada de la v.a. continua Y dado X=x:

$$F(y/x)=\int_{-\infty}^{y} f(y/x)dy=\frac{\int_{-\infty}^{y} f(x,y)dy}{f_1(x)}$$

4. INDEPENDENCIA DE VARIABLES ALEATORIAS

Independencia de sucesos: $P(A\cap B)=P(A)\cdot P(B)$

Definición
Consideremos una v.a. bidimensional (X,Y) cuya función de distribución conjunta es F(x,y) y sus funciones de distribución marginales $F_1(x)$ y $F_2(y)$ respectivamente.
Se dirá que las v.a. X e Y son independientes si y sólo si se verifica que la función de distribución conjunta F(x,y) es igual al producto de sus distribuciones marginales, es decir:

X e Y son independientes $\Leftrightarrow F(x,y)=F_1(x)\cdot F_2(y), \quad \forall(x,y)$

O bien de manera equivalente en el caso discreto :

$P(x_i/Y=y_j)=p_{i.}, \quad \forall j$

$P(y_j/X=x_i)=p_{.j}, \quad \forall i$

y en el caso continuo si $f(x/y)=f_1(x), \quad \forall y$

$f(y/x)=f_2(y), \quad \forall x$

Si X e Y son v.a. discretas, con distribución de probabilidad conjunta $p_{ij}=P(x_i,y_j)$, y con distribuciones de probabilidad marginales $P_{i.}=P(x_i)$ y $P_{.j}=P(y_j)$ respectivamente, entonces diremos que las v.a. son independientes si y sólo si se verifica:

$P(x_i,y_j)=P(x_i).P(y_j) \quad \forall(x,y)$, o bien $p_{ij}=p_{i.}p_{.j.}$

Si X e Y son v.a. continuas, con función de densidad conjunta *f(x,y)* y funciones de densidad marginales $f_1(x)$ y $f_2(y)$ respectivamente, se dirá que las v.a. son independientes si y sólo si se verifica:

$$f(x,y)=f_1(x).f_2(y) \quad \forall(x,y)$$

Capítulo IX CARACTERÍSTICAS DE LAS VARIABLES

En este tema estudiaremos algunas características, tanto para v.a. discretas como para v.a. continuas, como pueden ser la media, varianza, etc, así como sus relaciones.

1. VALOR ESPERADO DE UNA VARIABLE ALEATORIA UNIDIMENSIONAL

Definición

Sea X una v.a. discreta que toma los valores x_1, x_2,...,x_r y cuya función de probabilidad es $P(x_i)$, i = 1,2,3,...

Se define el valor esperado de X, E(X) como:

$$E(X) = \sum_i x_i P(x_i)$$

Si la v.a. X es de tipo continuo, con función de densidad f(x), definimos el valor esperado E(X), como:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

La expresión de E(X) en el caso que X sea una v.a. discreta, este valor es la media ponderada de los posibles valores que puede tomar la variable X, en donde los pesos o ponderaciones son las probabilidades, $P(x_i) = P(X = x_i)$, de ocurrencia de los posibles valores de X. Luego el valore esperado de X se interpreta como una media ponderada de los posibles valores de X, y no como el valor que se espera que tome X, pues puede suceder que E(X) no sea uno de los posibles valores de X. En el caso de v.a. continua, E(X) nos indica el centro de la función de densidad, es decir, nos indica el centro de gravedad de la distribución.

Sea X una v.a. de tipo discreto, con función de probabilidad $P(x_i)$, y sea g(X) una función de la v.a. X, entonces definimos E[g(X)] como:

$$E[g(X)] = \sum_i g(x_i)P(x_i)$$

Si la v.a. X es de tipo continuo, con función de densidad f(x), entonces definimos el valor esperado de g(X), E[g(X)] como:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

PROPIEDADES:

1. La esperanza de una constante es la propia constante. Es decir si *K* es una constante entonces:

$E[K] = k$

2. Si una v.a. X esta acotada, es decir existen dos valores *a* y *b* tales que $a \leq X \leq b$, entonces se verifica que:

$$a \leq E(X) \leq b$$

3. Sea X una v.a. y sean g(X) y h(X) dos funciones de X que, a su vez son v.a., cuyos valores esperados existen y sean *a* y *b* dos constantes cualesquiera, entonces:

$E[a \cdot g(X) + b \cdot h(X)] = a \cdot E[g(X)] + b \cdot E[h(X)]$

4. Sea X una v.a y sean g(X) y h(X) dos funciones de X que a su vez son v.a cuyos valores esperados existen; si $g(X) \leq h(X)$, entonces:

$E[g(X)] \leq E[h(X)]$

5. Sea X una v.a. y sea g(X) una función de X que, a su vez es una v.a., cuyo valor existe, entonces:

$$|E[g(X)]| \leq E[|g(X)|]$$

6. Si X es una v.a. con distribución simétrica respecto del valor c, entonces, si existe la E[X], su valor será igual a c, es decir, $E[X] = c$.

2. MOMENTOS DE UNA VARIABLE ALEATORIA UNIDIMENSIONAL

Los momentos nos proporcionan información sobre las propiedades de la distribución de la variable aleatoria. Nos permitirán hacer comparaciones de dos distribuciones y determinar cuál es más dispersa respecto a la media, o cuál es más apuntada, etc.

Proporcionan valores numéricos sobre características de la distribución de una v.a.

Definición
Sea X una v.a. y *r* un número entero positivo. Se define el momento de orden *r* de X, y se notará por α_r como:

$$\alpha_r = E[X^r], \text{ para } r = 1, 2, ...$$

Definición
Sea X una v.a. y *r* un número entero y positivo. Se define el momento central de orden *r* de X, y se notará por μ_r como:

$$\mu_r = E[(X - E[X])^r] \text{ , para } r = 1, 2, ...$$

En ambos casos los momentos se definen como valores esperados, luego para que existan los momento tendrán que existir los correspondientes valores esperados .

Los momentos, si existen, son números calculados como sumas o integrales, dependiendo de si se trata de v.a. discretas o continuas, de forma que se puede escribir el momento de orden r, como:

$$\alpha_r = E[X^r] = \sum_i x_i^r P(x_i), \qquad r = 1, 2, \ldots$$
$$\alpha_r = E[X^r] = \int_{-\infty}^{+\infty} x^r f(x)dx, \qquad r = 1, 2, \ldots$$

Y el momento central de orden r será:

$$\mu_r = E[(X - E[X])^r] = \sum_i (x_i - E[X])^r P(x_i), \qquad r = 2, \ldots$$
$$\mu_r = E[(X - E[X])^r] = \int_{-\infty}^{+\infty} (x - E[X])^r f(x)dx, \qquad r = 2, \ldots$$

Caso especial:
$\alpha_1 = E[X] = \mu$
Con lo que el momento central de orden r, se puede expresar:

$$\mu_r = E[(X - E[X])^r] = E[(X - \mu)^r]$$

Nota: En la definición del momento central de orden r, indicamos desde r = 2,3,..., ya que el momento central de orden uno es igual a cero, $\mu_1 = 0$.

3. VARIANZA

Se define como el momento central de orden dos, y se expresa como

$$Var(X) = \sigma_X^2 = \mu_2 = E[(X - E[X])^2]$$

y se le llama varianza de la distribución.
Es una medida de dispersión de los valores de la variable respecto de su media, y nos permite conocer el grado de dispersión de los valores de la distribución, pudiendo realizar comparaciones con otras distribuciones. También se puede utilizar como una cierta medida de cómo representa la media a la distribución.

La varianza se expresa en las mismas unidades que la variable X, pero al cuadrado. La desviación estándar o desviación típica (raíz cuadrada positiva de la varianza), se expresa en las mismas unidades de medida que la variable X. Se notará por σ_X.

Propiedades de la Varianza:

1. La varianza se puede expresar como:
$$Var(X) = E[X^2] - (E[X])^2 = \alpha_2 - \alpha_1^2 = \alpha_2 - \mu^2$$

2. La varianza de una constante *c* es cero.

3. Sea X una v.a. cuya varianza existe, y *a* una constante cualquiera. Entonces:
$$Var(a.X) = a^2 Var(X)$$

4. Sea X una v.a. cuya varianza existe y *a, b* dos constantes cualesquiera. Entonces:

$$Var(a.X + b) = a^2 Var(X)$$

5. Sean X e Y dos v.a. independientes cuyas varianzas existen, entonces se verifica que la varianza de la suma o de la diferencia de ambas v.a. independientes es igual a la suma de las varianzas. Es decir:

$$Var(X \pm Y) = Var(X) + Var(Y)$$

Si las v.a. no son independientes entonces:

$$E\left[(X - E(X))(Y - E(Y))\right] = Cov(X,Y)$$

y se verificará que:

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$$

6. Si X_1, X_2, ..., X_n, son v.a. independientes, cuyas varianzas existen, entonces para a_1, a_2, ..., a_n, constantes cualesquiera se verifica que:

$$Var(\sum_{i=1}^{n} a_i.X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

7. La varianza de una v.a. X es cero si y sólo si X es una v.a. degenerada o constante.

4. COEFICIENTE DE VARIACIÓN

La varianza es una medida de dispersión que está influenciada por el tamaño de los valores de dicha variable y por la media. Para evitar esta influencia damos una medida relativa de dispersión que exprese la dispersión de la variable respecto del tamaño de la variable, midiendo el tamaño de la variable por su valor esperado. Esta medida de dispersión es el coeficiente de variación:

$$CV_X = \frac{\sigma_X}{E[X]} = \frac{\sigma_X}{\mu_X}$$

Este coeficiente de variación expresa la dispersión de una v.a. respecto a su media y es muy útil para comparar dos distribuciones de probabilidad.

El coeficiente de variación no tendrá sentido cuando la variable X tome valores positivos y negativos (pues puede ocurrir que la media quede compensada por los valores positivos y negativos y no refleje el tamaño de X). Con lo que el coeficiente de variación sólo tendrá sentido cuando X sea una variable que tome sólo valores positivos.

5. CAMBIO DE ORIGEN Y DE ESCALA

Dada una v.a. X, se dirá que se ha realizado un cambio de origen 0' cuando se realice la siguiente transformación:

$$T = X - 0'$$

A través de esta transformación todos los valores de la v.a. X se desplazan 0' unidades del eje de ordenadas manteniendo la misma posición relativa y la misma distancia entre ellos.
Siendo el valor esperado de la nueva variable T:
$E[T] = E[X] - 0'$, o lo que es lo mismo $\mu_T = \mu_X - 0'$

Si además realizamos un cambio de escala *e* ($e > 0$) cuando se realiza la siguiente transformación:

$$Y = \frac{X - 0'}{e}$$

Si $e < 1$ entonces los nuevos valores transformados se alejan unos de otros y aumenta la dispersión respecto de X ó T; si
$e > 1$ entonces los nuevos valores transformados se acercan unos a otros, es decir, se concentran y disminuye la dispersión respecto de X ó T.

$$E[Y] = E\left[\frac{X - 0'}{e}\right] = \frac{E[X] - 0'}{e} \text{ , es decir : } \quad \mu_Y = \frac{\mu_X - 0'}{e}$$

Lo cual nos indica que se produce también el mismo cambio de origen y de escala en el valor esperado o media de la variable.

El cambio de origen no le afecta a la varianza de la variable transformada, ya que lo único que hace el cambio de origen es desplazar los valores de la variable inicial X manteniendo la misma posición relativa y en consecuencia no modifica la dispersión, es decir:

$$\sigma_Y^2 = Var(Y) = Var\left(\frac{X - 0'}{e}\right) = \frac{Var(X)}{e^2} = \frac{\sigma_X^2}{e^2},$$

$$\text{siendo la desviación típica } \sigma_Y = \frac{\sigma_X}{e}$$

Es decir, la varianza es invariante frente a cambios de origen pero no lo es frente a cambios de escala, ya que al dividir todos los valores de la variable X por el cambio de escala, éstos se dispersan si $e < 1$, y se concentrn si $e > 1$.

Veamos como afecta el cambio de origen y de escala al coeficiente de variación:

$$CV_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\frac{\sigma_X}{e}}{\frac{\mu_X - 0'}{e}} = \frac{\sigma_X}{\mu_X - 0'}, \text{ siendo } CV_X = \frac{\sigma_X}{\mu_X}$$

Luego $CV_Y \neq CV_X$, salvo cuadno $0' = 0$, es decir, cuando no haya cambio de origen.
Pudiendo decir que el coeficiente de variación es invariante frente a cambios de escala, pero no frente a cambios de origen, ya que cuando se produce un cambio de origen la media de la variable cambia, y por tanto el coeficiente de variación.

6. TIPIFICACIÓN DE UNA VARIABLE

Puede ocurrir que la forma de distintas distribuciones sean análogas, diferenciándose tan sólo por el hecho de estar representadas o medidas en sistemas con distinto origen o distintas escalas. Entonces decimos que todas estas distribuciones pertenecen a la misma familia de distribuciones.

Es frecuente el estudiar las propiedades de las distribuciones no en particular para cada una de las posibilidades existentes sino para alguna distribución que podría considerarse estándar o más representativa de toda la familia, y luego mediante la aplicación del correspondiente cambio de origen y de escala, intentar deducir las propiedades para la distribución particular considerada.

Cuando se pretende comparar magnitudes expresadas en distintas unidades o en distintas situaciones, para poder compararlas, tendrán que ser homogéneas. Al proceso de homogeneización de las diferentes magnitudes se le llama tipificación o normalización.
Se dirá que una variable Z, está tipificada cuando su media es cero y su desviación típica sea uno. Es decir, dada una variable X, de media μ_X y varianza σ^2_X, la variable Z tipificada será:

$$Z = \frac{X - \mu_X}{\sigma_X}, \text{ siendo } \sigma_X > 0$$

Es decir, para tipificar una variable, sólo hay que transformarla mediante un cambio de origen $0' = \mu_X$, y un cambio de escala $e = \sigma_X$.
La nueva variable Z es adimensional (no tiene asociada unidad de medida), y por tanto se podrá comparar directamente con otras variables tipificadas.

Teniendo en cuenta el efecto que produce el cambio de escala y de origen, se verifica:

$$\mu_Z = E[Z] = E\left[\frac{X - \mu_X}{\sigma_X}\right] = \frac{\mu_X - \mu_X}{\sigma_X} = 0$$

$$\sigma_Z^2 = Var(Z) = Var\left(\frac{X - \mu_X}{\sigma_X}\right) = \frac{\sigma_x^2}{\sigma_x^2} = 1$$

7. OTRAS MEDIDAS DE POSICIÓN Y DISPERSIÓN

Las principales medidas de posición y de dispersión son la media y la varianza, pero hay otras que también se utilizan como son: los cuantiles (mediana, cuartiles, deciles, percentiles), la moda, desviación absoluta media respecto de la mediana, recorrido intercuartílico, etc. Veamos a continuación alguna de ellas.

Definición
Dada una v.a. X, se define el cuantil de orden q, para $0 \le q \le 1$, como aquel valor x_q tal que:

$$P(X \le x_q) \ge q, \quad \text{y} \quad P(X \ge x_q) \ge 1 - q \quad \text{si X es discreta}$$

$$P(X \le x_q) = q, \quad \text{o} \quad F(x_q) = q \quad \text{si X es continua}$$

Los cuantiles x_q dividen la distribución en dos partes, así pues a la izquierda quedan todos los valores de la v.a. que son menores o iguales que x_q y a la derecha quedan los valores que son mayores o iguales que x_q.

Los cuantiles en el caso de una v.a. continua se calcularán simplemente resolviendo la ecuación: $F(x_q) = q$, cuya solución no será siempre única, y la solución será el cuantil que estamos buscando.
En el caso discreto el cuantil tampoco tiene por qué ser único, obteniéndose generalmente por interpolación, ya que no siempre se podrá obtener una solución exacta única.
Dentro de los cuantiles nos interesan: la Mediana (M_e), los Cuartiles (Q_1, Q_2, Q_3), Deciles (D_1, D_2, ..., D_9)y Percentiles (P_1, P_2, ..., P_{99}), los cuales dividen los valores de la variable en mitades, cuartas, décimas y centésimas partes, respectivamente.
Otra medida de posición central es la Moda.

Definición
Dada una v.a. X, con su correspondiente función de probabilidad o con su función de densidad, según que estemos en el caso discreto o continuo, se define la Moda, como el valor más probable de la variable, que será aquel valor de la variable para el cual la función de probabilidad o la función de densidad se hace máxima.
En el caso continuo la Moda será aquel valor M_o tal que si $f(x)$ es la función de densidad de la v.a. X, entonces:

$$f'(M_o) = 0 \;\; ; f''(M_o) < 0$$

Definición
La desviación absoluta media respecto a la mediana se define como un momento absoluto de primer orden.
En el caso continuo será:

$$E\left[\left|X - M_e\right|\right] = \int_{-\infty}^{+\infty} \left|x - M_e\right| f(x)dx$$

En el caso discreto será:

$$E\left[\left|X - M_e\right|\right] = \sum_i \left|x_i - M_e\right| P(X = x_i)$$

Definición
Una medida de dispersión asociada a los cuartiles es el Recorrido intercuartílico:

$$R_Q = Q_3 - Q_1$$

Dentro de este intervalo intercuartílico se encuentran el 50% de los valores centrales de la variable (prescindiendo del 25% de los valores más pequeños y del 25% de los valores más grandes).

8. MEDIDAS DE FORMA

Necesitamos definir una serie de medidas que nos permitan cuantificar, en la medida de lo posible, la forma de la distribución, es decir, tendremos que definir unas medidas que nos den información sobre la función de probabilidad o sobre la función de densidad, de forma que podamos cuantificar el grado de simetría y el de apuntamiento o de aplastamiento, bien de la función de probabilidad, o bien de la función de densidad.
Las medidas de forma serán adimensionales e invariantes ante cambios de origen y de escala.

- Coeficiente de Asimetría

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

de manera que:

$\gamma_1 < 0$, distribución asimétrica a la izquierda

$\gamma_1 = 0$, distribución simétrica o casi simétrica respecto a la Mediana (punto de simetría)

$\gamma_1 > 0$, distribución asimétrica a la derecha

Existen otros coeficientes de asimetría, de forma que cuando la distribución es unimodal, se define el índice de simetría de Pearson:

$$\rho_1 = \frac{\mu - M_o}{\sigma}$$

- Coeficiente de curtosis o apuntamiento

Este coeficiente trata de medir el grado de agrupamiento de los valores centrales en torno de la media aritmética. Es decir, mide el grado de aplastamiento de la gráfica correspondiente a la función de densidad.

Fisher define este coeficiente de curtosis como:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

tomando como referencia la función de densidad de una distribución normal . En este caso tan sólo hacemos referencia al caso continuo, ya que comparamos con la función de densidad de la distribución normal.

$\gamma_2 = 0$, distribución Mesocúrtica (mismo grado de concentración que la distribución normal).

$\gamma_2 < 0$, distribución Platicúrtica (la distribución tiene los datos centrales menos concentrados que la distribución normal, es decir, menos apuntada que la distribución normal).

$\gamma_2 > 0$, distribución Leptocúrtica (la distribución tiene los datos centrales más concentrados que la distribución normal, es decir, será más apuntada que la distribución normal).

9. TEOREMA DE MARKOV Y DESIGUALDAD DE CHEBYCHEV

El problema al que nos enfrentamos es que ahora no conocemos la distribución y tan sólo conocemos la media y la varianza de una distribución desconocida y queremos calcular cotas superiores de ciertas probabilidades e incluso la probabilidad para algún intervalo relativo a la media. Estos resultados vendrán dados por medio del teorema de Markov y por la desigualdad de Chebychev.

Teorema de Markov

Sea X una v.a., no negativa (es decir, $P(X \geq 0) = 1$), y cuya media, $E[X]$, existe. Entonces para cualquier $k > 0$, se verifica

$$P(X \geq k) \leq \frac{E[X]}{k}$$

Una consecuencia inmediata del teorema de Markov viene dada por la desigualdad de Chebychev.

Desigualdad de Chebychev

Sea X una v.a. con media μ y varianza σ^2 finita. Entonces para cualquier $k > 0$, se verifica:

$$P\left[\left|X - \mu\right| \geq k\right] \leq \frac{\sigma^2}{k^2}$$

Otra de expresar esta desigualdad será:

$$P\left[\left|X - \mu\right| < k\right] \geq 1 - \frac{\sigma^2}{k^2}$$

Estas expresiones de la desigualdad de Chebychev nos indican que podemos expresar en términos de probabilidad la dispersión de los valores de la variable alrededor de su media, utilizando para ello la varianza como medida de dispersión y no siendo necesario el conocimiento de la distribución de la v.a. X.
Esto nos indica que la varianza tiene un efecto muy significativo sobre las probabilidades asociadas a estos intervalos.

En general y para cualquier $\lambda > 0$, y haciendo $k = \lambda\sigma$ en la desigualdad de Chebychev resulta que:

$$P\left[\left|X - \mu\right| \geq \lambda\sigma\right] \leq \frac{\sigma^2}{\lambda^2\sigma^2} = \frac{1}{\lambda^2}, \text{ o bien}$$

$$P\left[\mu - \lambda\sigma < X < \mu + \lambda\sigma\right] \geq 1 - \frac{1}{\lambda^2}$$

10. FUNCIÓN GENERATRIZ DE MOMENTOS

Un valor esperado útil, cuando existe, es $E[e^{tX}]$, donde *t* es un número real, y se llama función generatriz de momentos de X. Se puede utilizar para calcular los momentos de la distribución de una v.a. y es también utilizada para obtener la distribución de una función de v.a.

Definición
Sea X una v.a. Se define la función generatriz de momentos de X, $g_X(t)$, como:

$$g_X(t) = E[e^{tX}]$$

Caso discreto :

$$g_X(t) = E\left[e^{tX}\right] = \sum_i P(x_i) \cdot e^{tx_i}$$

Caso continuo:

$$g_X(t) = E\left[e^{tX}\right] = \int_{-\infty}^{+\infty} e^{tx} f(x)dx$$

Para que exista la función generatriz de momentos tendrá que existir el correspondiente valor esperado.
Si existe la función generatriz de momentos, se puede obtener cualquier momento respecto al origen α_r.

Teorema

Si existe el momento de orden r, respecto al origen, α_r, entonces para cualquier valor entero y positivo r,

$$\alpha_r = g_x^{(r)}(0) = \left.\frac{d^r g_X(t)}{dt^r}\right]_{t=0}$$

Es decir, el momento respecto al origen de orden *r*, α_r, se puede obtener derivando la función generatriz de momentos *r* veces, respecto a *t*, y particularizando para *t* = 0.

CAMBIOS DE VARIABLE EN LAS FUNCIONES GENERATRICES DE MOMENTOS

Sea X una v.a. cuya función generatriz de momentos es $g_X(t) = E\left[e^{tX}\right]$, y pretendemos calcular la función generatriz de momentos de una v.a. Y = a X + b, donde a y b son constantes reales.
De forma que:

$$g_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(aX+b)}\right] = E\left[e^{taX}e^{tb}\right] =$$
$$= e^{tb}E\left[e^{taX}\right] = e^{tb}g_X(at)$$

FUNCIÓN GENERATRIZ DE MOMENTOS DE UNA COMBINACIÓN LINEAL DE n-VARIABLES ALEATORIAS INDEPENDIENTES

Sean n-variables aleatorias *X1, X2, ..., Xn*, todas ellas independientes, con función generatriz de momentos $g_{X_1}(t), g_{X_2}(t), \ldots, g_{X_n}(t)$, respectivamente, y sea $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n = \sum_{i=1}^{n} a_iX_i$, donde *a1,...,an,* son constantes reales.

La función generatriz de momentos de Y será:

$$g_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(a_1X_1+a_2X_2+\ldots+a_nX_n)}\right] =$$
$$= E\left[e^{ta_1X_1}\right] \cdot E\left[e^{ta_2X_2}\right] \ldots E\left[e^{ta_nX_n}\right] =$$
$$= g_{X_1}(a_1t), g_{X_2}(a_2t), \ldots, g_{X_n}(a_nt) =$$
$$= \prod_{i=1}^{n} g_{X_i}(a_it)$$

Si todas las v.a. *X1, X2, ..., Xn*, son independientes y tienen la misma distribución, y por tanto, la misma función generatriz de momentos g(t), entonces la función generatriz de momentos de la variable Y = *X1 + X2 + ... + Xn* será:

$$g_Y(t) = \left[g(t)\right]^n$$

UNICIDAD DE LA FUNCIÓN GENERATRIZ DE MOMENTOS

Si dos v.a. tienen la misma función generatriz de momentos, entonces también tienen la misma distribución de probabilidad. Es decir, si X e Y son dos v.a. con funciones generatrices $g_X(t)$ y $g_Y(t)$, respectivamente, y si $g_X(t) = g_Y(t)$, entonces las correspondientes distribuciones de probabilidad también son iguales. La inversa también se verifica.

11. VALOR ESPERADO DE UNA VARIABLE ALEATORIA BIDIMENSIONAL

Designemos por *(X,Y)* una v.a. bidimensional y por *g(X,Y)* una función de *(X,Y)* que también será una v.a.

Definición
Llamamos valor esperado o esperanza de la función *g(X,Y)* para el caso en que la v.a. bidimensional *(X,Y)* sea discreta, con distribución de probabilidad $P[X=x_i, Y=y_j]$ al valor:

$$E[g(X,Y)] = \sum_i \sum_j g(x_i, y_j) P[X = x_i, Y = y_j]$$

Si la v.a. bidimensional *(X,Y)* es continua, con función de densidad *f(x,y)* entonces el valor esperado de *g(X,Y)* será:

$$E[g(X,Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f(x,y) dxdy$$

Como sucedía en el caso unidimensional, en ambos casos el resultado es un número y para que exista el valor esperado es necesario que la correspondiente serie o integral de la definición sean absolutamente convergentes, es decir, será necesario que $E[|g(X,Y)|]$ sea finito.

Si la función de la v.a. bidimensional es *g(X,Y) = XY* entonces la esperanza o valor esperado para el caso discreto será:

$$E[X \cdot Y] = \sum_i \sum_j x_i \cdot y_j P[X = x_i, Y = y_j] = \sum_i \sum_j x_i \cdot y_j p_{ij}$$

Para el caso continuo

$$E[X \cdot Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \cdot y f(x,y) dxdy$$

Si la función es *g(X,Y) = X + Y*
Entonces la esperanza o valor esperado para el caso discreto será:

$$E[X + Y] = \sum_i \sum_j (x_i + y_j) P[X = x_i, Y = y_j] = \sum_i \sum_j (x_i + y_j) p_{ij}$$

Caso continuo:

$$E[X + Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y) f(x,y) dxdy$$

PROPIEDADES:

1. Sean X e Y dos v.a. cuyos valores esperados E[X] y E[Y] existen, y sean a y b dos constantes cualesquiera, entonces
   $$E[aX + bY] = aE[X] + bE[Y]$$

2. Sean $X_1, X_2, \ldots, X_n$ v.a. cuyos valores esperados $E[X_i]$, $i = 1,2,...,n$, existen y sean $a_1, a_2, ..., a_n$ constantes cualesquiera, entonces:

   $$E[a_1X_1+a_2X_2+...+a_nX_n] = a_1E[X_1]+ a_2E[X_2]+...+ a_nE[X_n]$$

3. Si X e Y son dos v.a. independientes, cuyos valores esperados E[X] y E[Y] existen, entonces:

$$E[XY] = E[X] \,.\, E[Y]$$

4. Si $X_1, X_2, \ldots, X_n$ son v.a. independientes y sus valores esperados $E[X_i]$ existen, entonces:

$$E[X_1.\, X_2 \ldots X_n] = E[X_1].\, .E[X_2] \ldots E[X_n]$$

5. Si $X_1, X_2, \ldots, X_n$ son v.a. independientes, y existen los valores esperados $E[g_i(X_i)]$, i = 1,2,…, n, entonces:

$$E[g_1(X_1).\, g_2(X_2) \ldots g_n(X_n)] = E[g_1(X_1)].\, E[g_2(X_2)] \ldots E[g_n(X_n)]$$

12. MOMENTOS DE UNA VARIABLE ALEATORIA BIDIMENSIONAL

Definición

Sea (X,Y) una v.a. bidimensional, y *r*, *s* dos números enteros no negativos. Definimos el momento de orden *r* en X y de orden *s* en Y, y lo notaremos por a_{rs}, como el siguiente valor esperado, si existe:

$$a_{rs} = E[X^r.\, Y^s] \text{ , para } r, s = 0, 1, 2, \ldots$$

Si hacemos *s* = 0 ó bien *r* = 0 obtendríamos los correspondientes momentos respecto al origen, de orden r para *X* y de orden s para *Y*. Así pues, los que más se utilizan son:

$$a_{10} = E[X] = \mu_X \qquad a_{20} = E[X^2]$$
$$a_{01} = E[Y] = \mu_Y \qquad a_{02} = E[Y^2]$$
$$a_{11} = E[XY]$$

Definición

Sea (X,Y) una v.a. bidimensional, cuyas esperanzas E[X] y E[Y] existen, y r, s dos enteros no negativos. Definimos el momento central de orden r en X y de orden s en Y, y lo notaremos por μ_{rs} como el siguiente valor esperado, si existe:

$$\mu_{rs} = E[(X - E[X])^r\, (Y - E[Y])^s], \quad \text{para } r, s = 0, 1, 2, \ldots$$

Obsérvese que:

$$\mu_{10} = E[(X - E[X])] = 0 \quad \text{y} \quad \mu_{01} = E[(Y - E[Y])] = 0$$

$$\mu_{20} = E[(X - E[X])^2] = Var(X)$$

$$\mu_{02} = E[(Y - E[Y])^2] = Var(Y)$$

$$\mu_{11} = E[(X - E[X])\,(Y - E[Y])] = \alpha_{11} - \alpha_{10}\ \alpha_{01} = Cov(X,Y)$$

13. COVARIANZA

Uno de los momentos centrales de mayor interés es el momento μ_{11}, que se define:

$$\mu_{11} = E[(X - E[X])\,(Y - E[Y])]$$

A este momento le denominamos covarianza entre X e Y, y lo representaremos por Cov(X,Y), σ_{xy} o μ_{xy}.

La covarianza nos permite dar una medida de la dependencia lineal entre X e Y.
Covarianza positiva: Cuando los valores de X crecen los valores de Y también tiende a crecer.
Covarianza negativa: Cuando los valores de X crecen los valores de Y tiende a decrecer.

Se puede utilizar la covarianza como una medida de la relación entre dos v.a.
La covarianza no es una medida de las relaciones o dependencias entre las dos variables X e Y, sino que únicamente es una medida de la fuerza de la relación lineal entre X e Y.
Un inconveniente para utilizar la covarianza, es que la magnitud o tamaño numérico de la covarianza no es significativa, pues la magnitud de la covarianza depende de las unidades de medida utilizadas en X e Y.
PROPIEDADES:

1. La covarianza se puede expresar en función de los momentos respecto del origen:

$$Cov(X,Y) = E[X.Y] - E[X]\,E[Y] = \alpha_{11} - \alpha_{10}\ \alpha_{01}$$

2. Si X e Y son dos v.a. independientes entonces:

$$Cov(X,Y) = 0$$

3. Sean X e Y v.a. y sean también las v.a. $U = aX$ y $V = bY$, siendo a, b dos números reales cualesquiera. Entonces se verifica:

$$Cov(U,V) = a.b.\ Cov\ (X,Y)$$

4. *La Cov(X,Y) = Cov(Y,X)*
5. La *Cov(X,X) = Var(X)*
6. La *Cov(X,a) = 0* para cualquier número real *a*.
7. Si *X, Y, X* son v.a., entonces:

$$Cov(X + Y,Z) = Cov(Z,\ X + Y\ ) = Cov(X,Z) + Cov(Y,Z)$$

8. Si X e Y son dos v.a., entonces:

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2\ Cov(X,Y)$$

y si X e Y son independientes, entonces:

$$Var(X \pm Y) = Var(X) + Var(Y)$$

9. Si $X_1, X_2, \ldots, X_n$ son v.a. cualesquiera. Entonces:

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} Cov(X_i, X_j)$$

10. Si $X_1, X_2, \ldots, X_n$ son v.a. son v.a. y $a_1, a_2, \ldots, a_n$ números reales cualesquiera. Entonces:

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} a_i a_j \, Cov(X_i, X_j)$$

En estas dos propiedades si las v.a. son independientes lo serán dos a dos y entonces la $Cov(X_i, X_j)$ son nulas.

14. COEFICIENTES DE CORRELACIÓN

Este coeficiente mide la fuerza de la relación lineal entre las v.a.

Definición

Sean X e Y dos v.a. cuyas varianzas existen y son distintas de cero. Entonces definimos el coeficiente de correlación lineal entre X e Y, que notaremos por ρ_{XY} :

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

PROPIEDADES:

1. Si las variables X e Y son independientes el coeficiente de correlación es nulo.
2. Si X e Y son v.a. cuyas varianzas existen y son distintas de cero, entonces:

$$-1 \leq \rho_{XY} \leq +1$$

El coeficiente de correlación nos da una medida de la fuerza de la relación lineal entre las variables, es decir, nos da una medida numérica del grado en que las variables están relacionadas linealmente.

Si ρ_{XY} toma un valor próximo a la unidad entonces existe una fuerte relación lineal entre X e Y.

Si $\rho_{XY} = 1$ entonces existe exactamente una relación lineal entre X e Y.

Si ρ_{XY} toma un valor próximo a cero puede existir una relación no lineal entre X e Y.

Si $\rho_{XY} = 0$ diremos que las variables están incorrelacionadas y si $\rho_{XY} \neq 0$ entonces se dirá que están correlacionadas. La correlación será positiva cuando $\rho_{XY} > 0$ y será negativa cuando $\rho_{XY} < 0$.

15. FUNCIÓN GENERATRIZ DE MOMENTOS DE UNA VARIABLE ALEATORIA BIDIMENSIONAL

Sea (X,Y) una v.a. bidimensional. Definimos la función generatriz de momentos de (X,Y), que notaremos por $g(t_1,t_2)$ como:

$$g(t_1,t_2) = E\left[e^{t_1X+t_2Y}\right]$$

Caso discreto:

$$g(t_1,t_2) = E\left[e^{t_1X+t_2Y}\right] = \sum_{i,j} e^{t_1x_i+t_2y_j}P(x_i,y_j)$$

Caso continuo:

$$g(t_1,t_2) = E\left[e^{t_1X+t_2Y}\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{t_1x+t_2y}f(x,y)dxdy$$

Para que exista la función generatriz de momentos tiene que existir el correspondiente valor esperado, y por tanto la correspondiente serie o integral tendrán que ser convergentes.

Las funciones generatrices marginales de momentos serán:

$$g(t_1,0) = E\left[e^{t_1X}\right]$$
$$g(0,t_2) = E\left[e^{t_2Y}\right]$$

Teorema

Si existe el momento respecto al origen de orden r para X y de orden s para Y, α_{rs}, entonces se verifica que

$$\alpha_{rs} = E\left[X^rY^s\right] = \left.\frac{\partial^{r+s}g(t_1,t_2)}{\partial t_1^r\partial t_2^s}\right]_{t_1=t_2=0}$$

Es decir, los momentos de orden r, s, se pueden obtener derivando la función generatriz de momentos r-veces respecto a t_1 y s-veces respecto a t_2, y particularizando para los valores $t_1 = 0$ y $t_2 = 0$.