

# Apuntes de ESTADÍSTICA

6 de junio de 2011



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento a los profesores del departamento del Matemática Aplicada de la Universidad de Málaga con los que he compartido asignatura en los últimos cursos académicos y, en particular, a los compañeros Carlos Cerezo, Inmaculada Fortes, Carlos Guerrero, José Morones y Agustín Valverde, por sus correcciones y sugerencias en la elaboración de estos apuntes.*



## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

# Índice general

<b>1. Estadística descriptiva</b>	<b>11</b>
1.1. Conceptos elementales . . . . .	11
1.2. Distribuciones de frecuencias de un carácter . . . . .	13
1.2.1. Frecuencias . . . . .	13
1.2.2. Distribuciones discretas . . . . .	15
1.2.3. Distribuciones continuas . . . . .	16
1.3. Representaciones gráficas . . . . .	18
1.3.1. Caracteres cualitativos . . . . .	19
1.3.2. Caracteres cuantitativos . . . . .	20
1.4. Medidas de posición . . . . .	22
1.4.1. Media aritmética . . . . .	22
1.4.2. Moda . . . . .	24
1.4.3. Mediana . . . . .	26
1.4.4. Cuantiles . . . . .	28
1.5. Medidas de dispersión . . . . .	30
1.5.1. Rango . . . . .	30
1.5.2. Desviación media . . . . .	31
1.5.3. Varianzas y desviación típica . . . . .	32
1.5.4. Coeficiente de variación . . . . .	35
1.5.5. Momentos . . . . .	36
1.6. Medidas de forma . . . . .	37
1.6.1. Medidas de asimetría . . . . .	37
1.6.2. Medidas de apuntamiento . . . . .	39
1.7. Relación de problemas . . . . .	41

1.8. Anexo I: Comandos de R . . . . .	49
<b>2. Regresión y correlación</b>	<b>53</b>
2.1. Distribuciones bidimensionales . . . . .	53
2.1.1. Representación tabular . . . . .	53
2.1.2. Representaciones gráficas . . . . .	57
2.1.3. Distribuciones Marginales . . . . .	58
2.1.4. Distribuciones Condicionadas . . . . .	59
2.1.5. Distribuciones conjuntas: Momentos mixtos . . . . .	60
2.2. Regresión y correlación . . . . .	64
2.2.1. Relación entre variables . . . . .	64
2.2.2. Regresión: Método de los mínimos cuadrados . . . . .	67
2.2.3. Correlación . . . . .	71
2.3. El modelo lineal . . . . .	73
2.3.1. Regresión lineal . . . . .	73
2.3.2. Correlación lineal . . . . .	76
2.4. Modelos de regresión no lineal . . . . .	81
2.4.1. Linealización de modelos . . . . .	82
2.4.2. Ajuste parabólico . . . . .	83
2.4.3. Otros ajustes . . . . .	85
2.4.4. Bondad del ajuste . . . . .	87
2.5. Relación de problemas . . . . .	89
2.6. Anexo I: Justificación de algunos resultados . . . . .	97
2.6.1. Descomposición de las varianzas para el modelo lineal de regresión . . . . .	97
2.6.2. El coeficiente de correlación lineal de Pearson ( $r$ ) es un número comprendido entre -1 y 1 . . . . .	97
2.7. Anexo II: Comandos de R . . . . .	99
<b>3. Series estadísticas</b>	<b>103</b>
3.1. Números índice . . . . .	103
3.1.1. Clasificación de números índice . . . . .	104
3.1.2. Propiedades de los números índice . . . . .	104
3.2. Índices simples . . . . .	105

3.2.1. Índices simples elementales (ISE) . . . . .	105
3.2.2. Índices simples en cadena (ISC) . . . . .	107
3.2.3. Relación de precios, cantidades y valores . . . . .	108
3.3. Índices complejos . . . . .	110
3.3.1. Índices complejos sin ponderar . . . . .	111
3.3.2. Índices complejos ponderados . . . . .	112
3.3.3. Índices de precios . . . . .	113
3.4. Series de números índice . . . . .	116
3.4.1. Cambio de periodo base . . . . .	116
3.4.2. Renovación y empalme . . . . .	117
3.4.3. Deflación de series estadísticas . . . . .	118
3.5. Series Temporales o Cronológicas . . . . .	121
3.5.1. Representación gráfica . . . . .	121
3.5.2. Promedios o Medias Móviles . . . . .	121
3.6. Análisis de las series temporales . . . . .	123
3.6.1. Tendencia secular . . . . .	124
3.6.2. Variaciones estacionales o periódicas . . . . .	125
3.6.3. Variaciones cíclicas . . . . .	125
3.6.4. Variaciones aleatorias, irregulares o accidentales . . . . .	125
3.7. Estimación de la tendencia . . . . .	125
3.7.1. Método gráfico . . . . .	125
3.7.2. Método de las medias móviles . . . . .	126
3.7.3. Método de mínimos cuadrados . . . . .	127
3.7.4. Método de semipromedios . . . . .	128
3.8. Estimación de la variación estacional . . . . .	129
3.8.1. Método de la media móvil en porcentajes . . . . .	129
3.8.2. Método del porcentaje medio . . . . .	131
3.8.3. Estimación de la variación estacional para el modelo aditivo . . . . .	133
3.8.4. Desestacionalización de una serie temporal . . . . .	134
3.9. Estimación de las variaciones cíclicas . . . . .	136
3.10. Estimación de las variaciones aleatorias . . . . .	137

3.11. Relación de problemas . . . . .	139
<b>4. Probabilidad</b>	<b>147</b>
4.1. Álgebra de Boole de sucesos . . . . .	148
4.2. Probabilidad . . . . .	149
4.2.1. Definición axiomática de probabilidad . . . . .	149
4.2.2. Relación entre frecuencias y probabilidad . . . . .	151
4.3. Probabilidad condicionada. Sucesos independientes . . . . .	152
4.4. Teorema de la probabilidad total. Teorema de Bayes . . . . .	154
4.4.1. Teorema de la probabilidad total . . . . .	154
4.4.2. Teorema de Bayes . . . . .	155
4.5. ANEXO: Combinatoria . . . . .	156
4.5.1. Identificación del problema . . . . .	157
4.6. Relación de problemas . . . . .	159
<b>5. Variable aleatoria</b>	<b>173</b>
5.1. Variable aleatoria unidimensional . . . . .	174
5.2. Función de distribución . . . . .	174
5.3. Variable aleatoria discreta . . . . .	175
5.3.1. Distribución de probabilidad . . . . .	175
5.3.2. Función de distribución . . . . .	177
5.3.3. Función generatriz de probabilidad . . . . .	178
5.4. Variable aleatoria continua . . . . .	179
5.4.1. Función de densidad . . . . .	179
5.4.2. Función de distribución . . . . .	180
5.5. Esperanza matemática y otras medidas . . . . .	182
5.5.1. Esperanza matemática . . . . .	182
5.5.2. Momentos . . . . .	182
5.5.3. Función generatriz de momentos . . . . .	183
5.5.4. Medidas de posición . . . . .	184
5.5.5. Medidas de dispersión . . . . .	185
5.5.6. Medidas de forma . . . . .	186

5.6. Variable aleatoria bidimensional . . . . .	187
5.6.1. Función de distribución . . . . .	187
5.6.2. Tipos de variables aleatorias bidimensionales . . . . .	188
5.7. Relación de problemas . . . . .	195
<b>6. Distribuciones de probabilidad</b>	<b>207</b>
6.1. Distribuciones uniformes . . . . .	207
6.1.1. Distribución uniforme discreta . . . . .	207
6.1.2. Distribución uniforme continua . . . . .	208
6.1.3. Distribución uniforme bidimensional . . . . .	208
6.2. Distribución Binomial . . . . .	209
6.2.1. Distribución de Bernoulli . . . . .	209
6.2.2. Distribución Binomial . . . . .	210
6.2.3. Distribución Multinomial . . . . .	211
6.2.4. Distribución Hipergeométrica . . . . .	212
6.2.5. Distribución Binomial negativa . . . . .	213
6.3. Distribuciones asociadas a fenómenos aleatorios de espera . . . . .	214
6.3.1. Distribución de Poisson . . . . .	214
6.3.2. Distribución Geométrica o de Pascal . . . . .	216
6.3.3. Distribución Exponencial . . . . .	217
6.4. Distribuciones normales . . . . .	218
6.4.1. Distribución Normal o de Laplace-Gauss . . . . .	218
6.4.2. Distribución normal bidimensional . . . . .	220
6.4.3. Teorema central del límite . . . . .	220
6.5. Distribuciones derivadas de la normal . . . . .	221
6.5.1. Distribución $\chi^2$ de Pearson . . . . .	221
6.5.2. Distribución $t$ de Student . . . . .	223
6.5.3. Distribución $F$ de Fisher-Snedecor . . . . .	224
6.6. Simulación y Método de Montecarlo . . . . .	225
6.7. Relación de problemas . . . . .	227
6.8. Relación de problemas II – Temas 4, 5 y 6 . . . . .	231
6.9. Anexo I: Justificación de algunos resultados . . . . .	235

6.9.1. Distribución Binomial . . . . .	235
6.9.2. Propiedades de la función Gamma . . . . .	235
<b>7. Inferencia estadística</b>	<b>239</b>
7.1. Inferencia estadística . . . . .	239
7.1.1. Teoría de muestras . . . . .	240
7.2. Estimación paramétrica . . . . .	241
7.2.1. Estimación puntual . . . . .	241
7.2.2. Estimación por intervalos . . . . .	244
7.3. Contraste de Hipótesis . . . . .	245
7.4. Inferencia no paramétrica . . . . .	249
7.4.1. Bondad de ajuste. Tabla de contingencia . . . . .	250
7.4.2. Contraste de homogeneidad de varias muestras . . . . .	252
7.4.3. Contraste de dependencia o independencia de caracteres. Tablas de contingencia $K \times M$ . . . . .	253
7.5. Relación de problemas . . . . .	255
<b>A. Tablas de intervalos de confianza</b>	<b>265</b>
<b>B. Tablas de contrastes de hipótesis (regiones de rechazo)</b>	<b>269</b>
<b>C. Tablas de las distribuciones de probabilidad</b>	<b>275</b>



# Apuntes de ESTADÍSTICA

## Estadística descriptiva



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento a los profesores Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.*


## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

# Capítulo 1

## Estadística descriptiva

La estadística descriptiva es la rama de la estadística que trata la *descripción y análisis* de los datos de una población, sin pretender extender o generalizar sus resultados y conclusiones a otras poblaciones distintas o más amplias.

La descripción consiste en enumerar los elementos y rasgos que configuran una realidad mediante la observación o la medida. El análisis de la población está constituido por los procedimientos existentes para la determinación de los distintos aspectos, propiedades y relaciones de los conjuntos de datos.

La estadística descriptiva implica la colección, clasificación, análisis e interpretación de los datos en un proceso de organización y síntesis de la información. Estos sencillos trabajos de ordenar, contar, clasificar, registrar informáticamente, etc. requieren mucho tiempo (que se traduce en costes) y una especial atención para evitar posibles errores iniciales.

En este capítulo se tratan distintos métodos de clasificación y representación de los datos y se detallan los parámetros más importantes para el análisis, la interpretación y la obtención de resultados.

Entre los ejemplos que ilustran los conceptos, se han seleccionado dos de ellos que hacen referencia a un estudio del tráfico (ejemplo 1.5 de la página 16) y a las calificaciones de un grupo de alumnos (ejemplo 1.7 de la página 17). El recorrido de estos dos ejemplos a lo largo de todas las secciones, ilustra un estudio estadístico completo.

Por último, algunas cuestiones interesantes se tratan a modo de ejercicios autocontenidos en la relación de problemas propuestos al final del capítulo. Su interés queda justificado por el uso conjunto de las técnicas estudiadas en el capítulo y por sus numerosas aplicaciones prácticas.

### 1.1. Conceptos elementales

Como cualquier otra ciencia, la estadística utiliza su propia terminología y para acceder al conocimiento resulta imprescindible dominar su lenguaje. Conviene familiarizarse con los conceptos que se introducen en este capítulo y ser capaz de identificarlos.

A continuación se presentan las definiciones de los elementos básicos que intervienen en cualquier estudio estadístico.

**Población.** Se denomina *universo*, *colectivo*, *población estadística* o simplemente *población* al conjunto de elementos que son objeto de estudio. Las poblaciones podrán ser consideradas finitas o infinitas según la naturaleza o el número de elementos que la compongan, y en cualquier caso, estos elementos deben estar perfectamente delimitados y bien definidos.

**Individuo.** Se denomina *unidad estadística* o *individuo* a cada uno de los elementos de la población descritos mediante una serie de características a las que se refiere el estudio estadístico.

**Muestra.** Una *muestra* es un subconjunto no vacío de individuos de la población. La muestra, debidamente elegida, se somete a observación científica, en representación del conjunto total, con el propósito de obtener resultados válidos para toda la población.

El número de elementos que componen la muestra se denomina *tamaño muestral* y si coincide con el tamaño de la población, la muestra se denomina *censo*. Por tanto, realizar un censo implica el estudio de toda la población. Las dificultades para realizar un censo (población infinita, dificultad de acceso a todos los individuos, coste económico, capacidad de trabajo, tiempo necesario, etc.) hacen que sea preferible usar una muestra. En este caso, las técnicas de inferencia estadística permitirán obtener resultados de toda la población a partir de los obtenidos en la muestra.

**Encuesta.** La *encuesta* es un procedimiento de observación que consiste en la obtención de datos mediante la interrogación a los miembros de una población o la medida de los mismos.

**Caracteres.** Los *caracteres* son las cualidades o magnitudes de los individuos de la población que son objeto de estudio. Los caracteres pueden ser cualitativos (por ejemplo, nacionalidad o color del pelo) o cuantitativos (por ejemplo, número de hijos o metros cuadrados de vivienda).

Los caracteres cualitativos reciben el nombre de *atributos* y los designaremos utilizando preferentemente las primeras letras del alfabeto en mayúsculas (A,B,C,...). Los caracteres cuantitativos se denominan *variables estadísticas* y los designaremos utilizando preferiblemente las últimas letras del alfabeto en mayúsculas (...X,Y,Z).

A su vez, las variables pueden ser *discretas* (por ejemplo, número de acciones vendidas un día en la Bolsa de Valores, número de estudiantes matriculados en una Universidad, ...) o *continuas* (por ej. vida media de los tubos de televisión producidos por una fábrica, longitud de 1000 tornillos producidos por una empresa, temperaturas medidas en un observatorio cada media hora) según la naturaleza de los valores numéricos.

$$\text{Caracteres} \left\{ \begin{array}{l} \text{Cualitativos (atributos)} \\ \text{Cuantitativos} \\ \text{(variable estadística)} \end{array} \right. \left\{ \begin{array}{l} \text{Discretos} \\ \text{Continuos} \end{array} \right.$$

**Modalidades.** Las diferentes situaciones posibles del carácter se denominan *modalidades*. Éstas deben estar bien definidas de tal manera que cada individuo pertenezca a una y sólo una única modalidad. Las denotaremos haciendo uso de una letra minúscula, correspondiente al nombre del carácter, con un subíndice de orden. Por ejemplo,  $x_1, x_2, \dots, x_k$  denotan las distintas modalidades de la variable estadística  $X$ .

**Ejemplo 1.1** Se realiza un estudio sobre el tipo de software (libre o propietario) utilizado en los sistemas de gestión de bases de datos de las empresas malagueñas. Para ello, se consultó telefónicamente a 10 empresas elegidas al azar. Determinar los conceptos estadísticos elementales.

En este caso, la *población* está constituida por todas las empresas malagueñas que usan software para la gestión de bases de datos. La *encuesta* se realiza mediante llamada telefónica y el resultado es una *muestra* de 10 valores del *carácter* “tipo de software para la gestión de bases de datos” que resulta ser un *atributo* cuyas dos *modalidades* son “libre” y “propietario”.  $\square$

En el caso de las variables cuantitativas se pueden definir funciones que permiten obtener medidas descriptivas a partir de las observaciones. El objetivo de estas medidas es proporcionar información sobre las características de la distribución de los datos.

**Parámetro.** Un *parámetro* es una función que permite obtener una medida descriptiva numérica a partir de los valores de un carácter medible de la población. Por ejemplo, la media de una población se calcula dividiendo la suma de los valores de la variable entre el número total de individuos. Estas medidas suelen ser desconocidas pues para calcularlas se necesita efectuar un censo.

**Estadístico.** Un *estadístico* es una función definida sobre los valores numéricos de una muestra. Esta función permite obtener una medida descriptiva que se utiliza para obtener información sobre alguno de los parámetros desconocidos de la población. Por ejemplo, el estadístico “media aritmética de los datos de una muestra” se usa para estimar el parámetro “media de la población”.

**Ejemplo 1.2** *Estimar la compresión media del motor instalado en los automóviles de un cierto modelo producidos por una fábrica a partir del estudio efectuado en 100 vehículos.*

Se considera la *población* formada por todos los automóviles de ese modelo producidos por la fábrica. El conjunto de 100 automóviles extraídos de dicha población constituye una *muestra* de tamaño 100. Se realiza una *encuesta* que consiste en medir la compresión del motor en cada uno de ellos. El resultado es una muestra de 100 valores del *carácter* “compresión del motor” que es una *variable continua* cuyas *modalidades* corresponden a todas las posibles relaciones volumétricas. Si se calcula la media de los 100 datos de compresión se obtiene un valor del *estadístico* que proporciona información sobre el *parámetro* media de la población total.  $\square$

## 1.2. Distribuciones de frecuencias de un carácter

Uno de los conceptos sobre el que se basarán muchas definiciones posteriores y que simplifica la presentación de los datos es el de *frecuencia* o número de veces que aparece una determinada modalidad de un carácter o su proporción sobre el total. Las distintas modalidades junto a su frecuencia correspondiente constituye la *distribución de frecuencias* de un carácter.

### 1.2.1. Frecuencias

En adelante se considerará una población o muestra de tamaño  $N$  en la que se observará el carácter  $X$  que presenta las modalidades  $x_1, x_2, \dots, x_k$  (ordenadas de menor a mayor, si el carácter es cuantitativo).

**Frecuencia Absoluta.** Se llama frecuencia absoluta de un valor  $x_i$  del carácter  $X$ , y se denota por  $n_i$ , al número de individuos observados que presentan esta modalidad.

**Frecuencia Relativa.** Se llama frecuencia relativa de un valor  $x_i$  del carácter  $X$ , y se denota por  $f_i$ , al cociente entre la frecuencia absoluta y el total de individuos.

$$f_i = \frac{n_i}{N} \quad i = 1, 2, \dots, k$$

La frecuencia relativa representa la proporción de individuos que presentan una determinada modalidad y se puede expresar en tantos por cien sin más que multiplicar por cien el cociente de la fórmula anterior.

**Ejemplo 1.3** *De la siguiente frase: “La representación gráfica no es más que un medio auxiliar de la investigación estadística, pues ésta es fundamentalmente numérica”, obtener las distribuciones de frecuencias de las vocales.*

Las frecuencias absolutas de las modalidades “a”, “e”, “i”, “o” y “u” del atributo “vocales” son 15, 16, 11, 4 y 6 respectivamente y suman un total de 52 observaciones. Por tanto, la frecuencia relativa de cada una de las modalidades es  $15/52$ ,  $16/52$ ,  $11/52$ ,  $4/52$  y  $6/52$  que expresadas en tantos por cien son 29 %, 31 %, 21 %, 8 % y 11 % aproximada y respectivamente.

El significado de estas frecuencias está claro. Por ejemplo, la frecuencia absoluta de la vocal “a” es 15, es decir, de las 52 vocales contenidas en la frase, 15 de ellas son la vocal “a”, lo que corresponde al 29 % del total.  $\square$

Cuando el carácter es cuantitativo, tiene sentido definir también las siguientes frecuencias acumuladas:

**Frecuencias Acumuladas Absolutas y Relativas.** Se llama frecuencia acumulada de un valor  $x_i$  de la variable  $X$  a la suma de las frecuencias de los valores que son menores o iguales a él. Las frecuencias acumuladas se definen, tanto para las frecuencias absolutas, que se denotan por  $N_i$ , como para las relativas, que se denotan por  $F_i$ .

Si los valores  $x_i$  están ordenados de forma creciente entonces

$$N_i = \sum_{j=1}^i n_j \quad y \quad F_i = \sum_{j=1}^i f_j = \frac{N_i}{N} \quad i = 1, 2, \dots, k$$

Dualmente, se podrían haber definido estas frecuencias con los datos ordenados de forma decreciente. Según la definición utilizada se denominan frecuencias absolutas/relativas acumuladas crecientes o decrecientes.

De las definiciones anteriores se destacan las siguientes propiedades elementales:

$$\begin{array}{lll} 1) & 0 \leq n_i \leq N & 2) \quad \sum_{i=1}^k n_i = N & 3) \quad n_i = N_i - N_{i-1} \\ 4) & 0 \leq f_i \leq 1 & 5) \quad \sum_{i=1}^k f_i = 1 & 6) \quad f_i = F_i - F_{i-1} \end{array}$$

que pueden usarse a modo de prueba para detectar posibles errores iniciales en el cálculo de la distribución de frecuencias.

**Ejemplo 1.4** Como estudio preliminar a una encuesta de tráfico, fue necesario recabar cierta información acerca del número de ocupantes en los automóviles que entraban a una población el domingo por la tarde; para ello se contó el número de ocupantes en 40 de esos automóviles, y se obtuvieron los siguientes datos:

1 3 2 2 3 1 1 2 2 1 1 4 3 1 3 2 3 2 2 2  
1 2 5 1 3 1 2 1 3 1 4 1 1 3 4 2 2 1 1 4

Obtener la distribución de frecuencias acumuladas de la variable  $X$  que representa el “número de ocupantes en los automóviles”.

Si ordenamos de 1 a 5 las modalidades de la variable  $X$  y contamos el número de observaciones correspondientes a cada modalidad, obtenemos las frecuencias absolutas 15, 12, 8, 4 y 1, de cada una de las modalidades. Por lo tanto, las frecuencias acumuladas absolutas para las modalidades 1 a 5 son 15, 27, 35, 39 y 40 respectivamente. Las correspondientes frecuencias acumuladas relativas se obtienen dividiendo las absolutas por 40 que es el tamaño de la muestra, y obtenemos 0'375, 0'625, 0'875, 0'975 y 1.  $\square$

Generalmente, las distribuciones de frecuencias se presentan en forma de tabla, donde los datos se agrupan por modalidades. A cada modalidad se le asigna su frecuencia (absoluta, relativa o acumulada) para constituir la denominada *tabla estadística o de frecuencias*. Esta forma de representación permite tener organizada y resumida la información contenida en el conjunto de datos y presentada de forma más comprensible y significativa.

Las distribuciones de frecuencias de una sola variable son básicamente de dos tipos: discretas y continuas. Esta clasificación no corresponde exactamente con los tipos de caracteres sino más bien en consideración al número de observaciones y al número de valores distintos que toma la variable.

### 1.2.2. Distribuciones discretas

Se considera que la distribución de los datos es discreta si el carácter es cualitativo, o si el carácter es cuantitativo, pero el número de modalidades es “pequeño” en relación con el número de observaciones. Este tipo de distribuciones también se conoce como distribuciones de tipo II.

Para construir la tabla estadística correspondiente basta con disponer en columnas los pocos valores distintos de la variable, ordenados de menor a mayor, y sus correspondientes frecuencias, como se muestra en la figura 1.1.

$x_i$	$n_i$	$f_i$	$N_i$	$F_i$
$x_1$	$n_1$	$f_1$	$N_1$	$F_1$
$x_2$	$n_2$	$f_2$	$N_2$	$F_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_k$	$n_k$	$f_k$	$N_k$	$F_k$

Figura 1.1: Tabla de frecuencias de una distribución discreta

Para realizar los cálculos de algunos parámetros, que estudiaremos más adelante (media, varianza, momentos, etc.), se pueden añadir columnas que contienen operaciones para los valores

de cada modalidad. Además, este tipo de tablas se completan añadiendo una fila que contiene algunas de las sumas por columnas, de los datos correspondientes (véase el ejercicio 24 de la página 46, en la relación de problemas).

**Ejemplo 1.5** *Representar, en una tabla estadística, la distribución de frecuencias de los datos del ejemplo 1.4 de la página 15.*

Se observa que la variable  $X$  que determina el “número de ocupantes en los automóviles” presenta un reducido número de modalidades (1, 2, 3, 4 y 5), de tal manera que, aunque haya un elevado número de observaciones, éstas se pueden agrupar haciendo uso de la frecuencia, tal y como se recoge en la tabla de la figura 1.2.

$x_i$	$n_i$	$f_i$	$N_i$	$F_i$
1	15	0'375	15	0'375
2	12	0'300	27	0'675
3	8	0'200	35	0'875
4	4	0'100	39	0'975
5	1	0'025	40	1
Suma	40	1		

Figura 1.2: Tabla de frecuencias para los datos del ejemplo 1.5

□

Existen distribuciones que constan de un reducido número de observaciones y, en consecuencia, la variable toma un reducido número de valores distintos. Estas distribuciones también se conoce como distribuciones de tipo I, y para construir la tabla estadística basta simplemente con anotar ordenadamente las observaciones en fila o en columna, generalmente de menor a mayor.

$$x_1, x_2, x_3, \dots, x_N$$

**Ejemplo 1.6** *Para realizar un estudio sobre la venta semanal de ordenadores en una determinada empresa de informática, se observa, durante 5 semanas, el número de ordenadores vendidos, obteniéndose los siguientes resultados: 10, 12, 20, 6 y 10. Representar su distribución de frecuencias.*

La distribución de frecuencias se representa ordenando los datos: 6, 10, 10, 12, 20.

□

### 1.2.3. Distribuciones continuas

Algunas variables discretas y, en general, las variables de naturaleza continua dan lugar a conjuntos de datos en los que el número de modalidades es muy variado. Consideraremos que una distribución es continua cuando presenta un elevado número de observaciones y de modalidades distintas. En estos casos no resulta apropiado escribir todas las modalidades en una columna, como se hizo en el caso discreto. Para tabular estos datos conviene *agruparlos* en *intervalos* que constituyen una partición, y determinar el número de individuos que pertenecen a cada uno de ellos. Este tipo de distribuciones también se conoce como distribuciones de tipo III.



Tomar el intervalo como unidad de estudio, en lugar de cada valor de la variable, supone una simplificación pero resulta una pérdida de información. Por lo tanto, es importante elegir un número adecuado de intervalos que equilibre estos dos aspectos y que constituyan una partición del mismo. Según las características del conjunto de datos, en la bibliografía se proponen distintas formas de establecer el número de intervalos en función del tamaño ( $N$ ) de la muestra. Un criterio sencillo usado frecuentemente es considerar un número de intervalos aproximadamente igual a la raíz cuadrada del número de datos, es decir,  $\sqrt{N}$ .

Cada intervalo se denomina *clase* y a la diferencia entre el extremo superior ( $L_i$ ) e inferior ( $L_{i-1}$ ) se le llama *amplitud de la clase o del intervalo* y se denota por  $a_i$  que puede ser variable o constante para todos los intervalos. Al ser una partición, la unión de todos los intervalos ha de recubrir a todos los valores de la variable (exhaustivo) pero sin solaparse (excluyente). La elección del número de intervalos y su amplitud es importante si se quiere identificar el tipo de distribución y sus características.

Se llama *marca de clase* del intervalo  $i$ -ésimo y se denota por  $x_i$  al punto medio del intervalo y será el valor que representará la información del intervalo al que pertenece como si fuera un valor de la variable.

Para construir ahora la tabla estadística se colocan ordenadamente y por columnas los intervalos, las marcas de clase y las frecuencias correspondientes, como se muestra en la tabla de la figura 1.3.

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

Figura 1.3: Tabla de frecuencias de una distribución continua

**Ejemplo 1.7** Las calificaciones finales en Matemáticas de 100 estudiantes fueron:

11	46	58	25	48	18	41	35	59	28	35	2	37	68	70	31	44	84	64	82
26	42	51	29	59	92	56	5	52	8	1	12	21	6	32	15	67	47	61	47
43	33	48	47	43	69	49	21	9	15	11	22	29	14	31	46	19	49	51	71
52	32	51	44	57	60	43	65	73	62	3	17	39	22	40	65	30	31	16	80
41	59	60	41	51	10	63	41	74	81	20	36	59	38	40	43	18	60	71	44

Representar, en una tabla estadística, la distribución de frecuencias de las notas de Matemáticas.

Se define la variable  $X$  que representa la “nota final en Matemáticas”. Se observa un gran número de observaciones correspondientes a un elevado número de modalidades distintas, lo que sugiere agruparlas en clases. Veamos dos agrupamientos distintos:

1. Intervalos de la misma amplitud: Si consideramos 10 intervalos ( $\sqrt{N}$ ) de igual amplitud, podemos representar la distribución de las notas como se muestra en la tabla de la figura 1.4.

$L_{i-1}, L_i$	$x_i$	$n_i$	$f_i$	$N_i$	$F_i$
[0, 10]	5	8	0'08	8	0'08
(10, 20]	15	12	0'12	20	0'20
(20, 30]	25	10	0'10	30	0'30
(30, 40]	35	14	0'14	44	0'44
(40, 50]	45	21	0'21	65	0'65
(50, 60]	55	16	0'16	81	0'81
(60, 70]	65	10	0'10	91	0'91
(70, 80]	75	5	0'05	96	0'96
(80, 90]	85	3	0'03	99	0'99
(90, 100]	95	1	0'01	100	1
		100	1		

Figura 1.4: Tabla de frecuencias para los datos del ejemplo 1.7

2. Intervalos de diferente amplitud: Si atendemos a la calificación correspondiente a cada nota y consideramos 4 clases de distinta amplitud (suspense, aprobado, notable y sobresaliente), podemos representar la distribución de las notas como se muestra en la tabla de la figura 1.5.

$L_{i-1}, L_i$	$x_i$	$n_i$	$f_i$	$N_i$	$F_i$
[0, 50)	25	65	0'65	65	0'65
[50, 70)	60	25	0'25	90	0'90
[70, 90)	80	9	0'09	99	0'99
[90, 100]	95	1	0'01	100	1
		100	1		

Figura 1.5: Tabla de frecuencias para los datos del ejemplo 1.7

□

### 1.3. Representaciones gráficas

Estamos acostumbrados a recibir información a través de imágenes. En este sentido, la estadística utiliza la representación gráfica para presentar visualmente la distribución de los datos de la muestra. Al igual que las tablas estadísticas, las representaciones gráficas muestran la distribución de frecuencias y deben ser capaces de transmitir información de la muestra permitiendo observar algunas características de los datos.

Para conseguir estos objetivos, conviene cuidar la presentación de un gráfico (colores, formas,...) y utilizar adecuadamente los elementos que lo componen: título, ejes, leyenda, etc. Cuando se observa una representación gráfica hay que prestar especial atención al significado de los ejes y a las marcas de graduación que determinan la escala. Una visión rápida y descuidada puede inducir a conclusiones erróneas.

Los distintos tipos de gráficas representan las frecuencias absolutas, relativas o acumuladas. El tipo de carácter, según sea cualitativo o cuantitativo, establece una clasificación de las representaciones gráficas. Aunque algunas de ellas se pueden utilizar indistintamente, conviene conocer sus características para elegir la representación gráfica que resulta más apropiado a cada caso.

A continuación se relacionan los tipos de representación más utilizados y se detallan las características principales y la interpretación de los elementos que lo constituyen. La creatividad y la originalidad pueden dar lugar a otros tipos de gráficas, siempre y cuando cumplan con el objetivo de garantizar una imagen sencilla y real de los datos.

### 1.3.1. Caracteres cualitativos

Las distintas modalidades de los caracteres cualitativos no contemplan ningún orden numérico. Por tanto, estas representaciones gráficas suelen ser más icónicas y hacen uso del etiquetado de las clases o de la leyenda.

**Diagrama de rectángulos o barras.** Para cada modalidad, se representa un rectángulo o barra cuya altura (o longitud) coincide con la frecuencia absoluta (o relativa). En la figura 1.6 se representa la distribución de frecuencia de las vocales del ejemplo 1.3 de la página 14, utilizando distintos diagramas de columnas en vertical u horizontal.

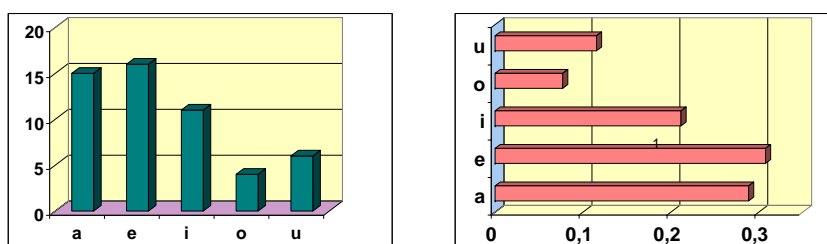


Figura 1.6: Diagrama de rectángulos

**Diagrama de Pareto.** Diagrama de barras de frecuencias relativas donde las modalidades se representan por orden decreciente en altura. Además, se superpone una curva con la frecuencia relativa acumulada cuya escala se representa a la derecha. Con este diagrama es fácil identificar las modalidades con mayor frecuencia. En la figura 1.7 se representa la distribución de frecuencias de las vocales del ejemplo 1.3 de la página 14, utilizando un diagrama de Pareto.

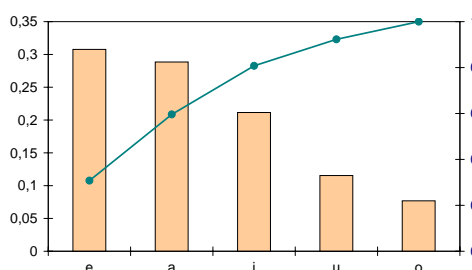


Figura 1.7: Diagrama de Pareto

**Diagrama de sectores.** Se descompone un círculo en sectores de área proporcional a la frecuencia de la modalidad correspondiente. El ángulo (en grados) del sector circular correspondiente a la modalidad  $i$ -ésima es  $\alpha_i = 360 \cdot f_i$ . En la figura 1.8 se representa la distribución de frecuencia de las vocales del ejemplo 1.3 de la página 14, utilizando distintas variedades de diagramas de sectores.

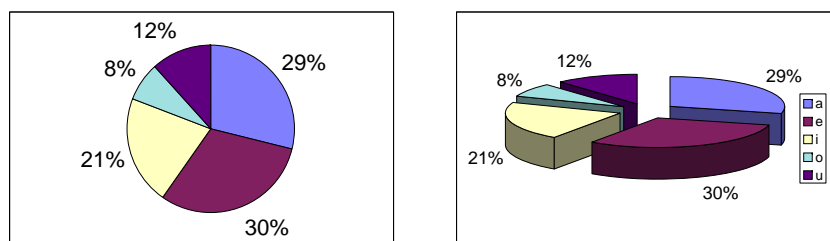


Figura 1.8: Diagrama de sectores

**Pictograma y cartogramas.** Representación icónica del fenómeno que utiliza dibujos simbólicos o mapas donde aparecen los iconos. El pictograma de la figura 1.9 representa la distribución de frecuencias de las vocales del ejemplo 1.3 de la página 14.

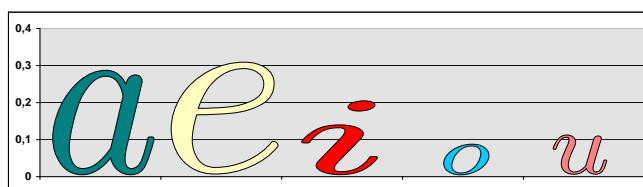


Figura 1.9: Pictograma

### 1.3.2. Caracteres cuantitativos

Este tipo de representaciones gráficas se realizan sobre los ejes de coordenadas. Para que sean más significativas, puede ser interesante un cambio de origen o escala en los ejes, si bien esto debe indicarse convenientemente para no inducir a engaño. Por ejemplo, un cambio de origen suele indicarse mediante una línea en zigzag en el eje correspondiente.

**Diagrama de barras o puntos.** Se utiliza en el caso discreto y es similar al de rectángulos pero con barras verticales o puntos en los extremos. La frecuencia absoluta (o relativa) determina la longitud de la barra y el valor de la variable determina el lugar del eje horizontal donde se apoya. En la figura 1.10 se representa la distribución de frecuencias (absolutas) del ejemplo 1.5 de la página 16, haciendo uso de un diagrama de puntos (izquierda) y de barras (derecha).

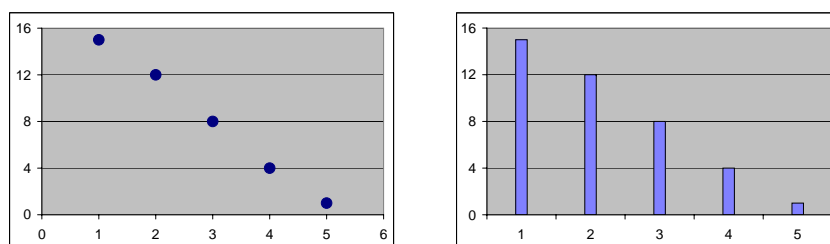


Figura 1.10: Diagrama de puntos – diagrama de barras

**Histograma.** Se utiliza para representar los datos agrupados en intervalos. Para cada clase, se dibuja un rectángulo sobre el eje X cuya base sea el intervalo y cuya área sea proporcional a la frecuencia a representar. Por lo tanto, la altura ( $h_i$ ) queda determinada por el cociente entre la frecuencia ( $n_i$ ) y la amplitud ( $a_i$ ) del intervalo. En la figura 1.11 se representa la distribución de frecuencias del ejemplo 1.7 de la página 17 cuando los intervalos tienen la misma amplitud (izquierda) y cuando la tienen distinta (derecha).

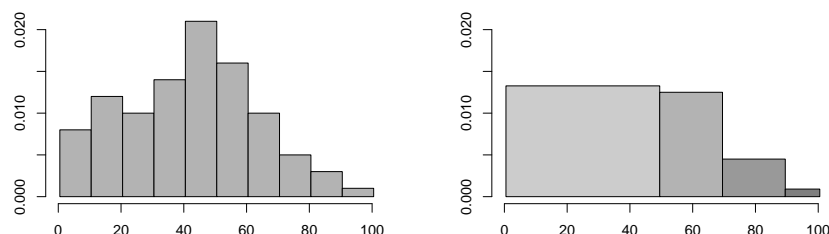


Figura 1.11: Histogramas

**Polígono de frecuencias.** Se construye uniendo los extremos de las barras en el diagrama de barras o los puntos medios superiores de los rectángulos en el histograma. En la figura 1.12 se representan las distribuciones de frecuencias absolutas del ejemplo 1.5 de la página 16 (izquierda), y las de frecuencias relativas del ejemplo 1.7 de la página 17 (derecha).

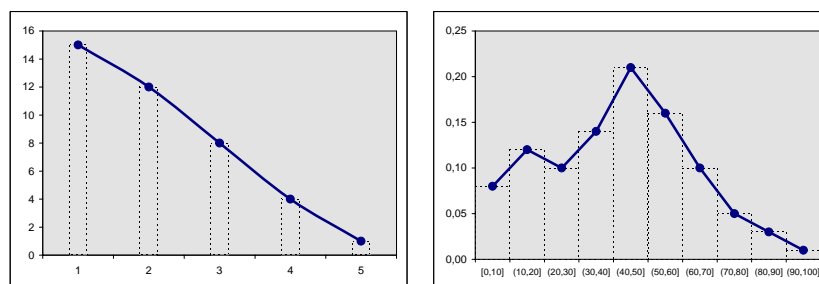


Figura 1.12: Polígonos de frecuencias

**Diagrama de frecuencias acumuladas.** Similar al polígono de frecuencias pero utilizando las frecuencias acumuladas (absolutas o relativas). En la figura 1.13 se representa la distribución de frecuencias del ejemplo 1.5 de la página 16 (izquierda) y del ejemplo 1.7 de la página 17 (derecha), utilizando diagramas de frecuencias acumuladas absolutas, para el primero, y relativas, para el segundo.

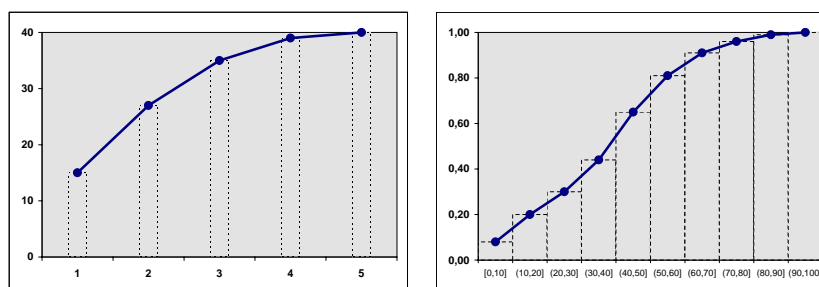


Figura 1.13: Diagrama de frecuencias (absolutas/relativas) acumuladas

Las tablas estadísticas y las representaciones gráficas constituyen distintas formas de presentar los datos de manera clara y ordenada. Ambas proporcionan información sobre la distribución de las observaciones. A veces conviene resumir toda esta información en uno o varios valores cuantitativos que sean más o menos representativos y que permitan comparar distintas muestras. Por este motivo, vamos a introducir las medidas de posición, de dispersión, de asimetría y de apuntamiento.

## 1.4. Medidas de posición

Las medidas de posición son valores numéricos descriptivos calculados a partir de los datos de la muestra. Estos valores ayudan a encontrar el “centro” de la distribución, en torno al cual se agrupan las observaciones, o la posición relativa de una observación, dentro del conjunto de datos.

Dentro de las medidas de posición destacan las medidas de tendencia central como la media, la mediana o la moda. También se definirán los cuantiles que no son propiamente medidas de tendencia central pero que se utilizan para situar los valores de la variable.

En la definición de las medidas de posición se considera una muestra de una variable  $X$  que toma los valores  $x_1, x_2, \dots, x_k$  con las frecuencias absolutas  $n_1, n_2, \dots, n_k$  respectivamente, haciendo un total de  $N$  datos.

### 1.4.1. Media aritmética

La *media aritmética* o simplemente *media* es una medida de tendencia central aplicable específicamente en el caso de variables cuantitativas. Se calcula dividiendo la suma de los valores de todos los datos entre el número total de datos, es decir

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

donde  $\bar{x}$  denota la media muestral. La media aritmética poblacional se obtiene aplicando la misma fórmula a todos los datos de la población (censo), y se suele denotar por  $\mu$ .

La media es una medida que se encuentra siempre entre los valores extremos de la variable y se considera el centro de gravedad de las observaciones, en el sentido de que la suma de las diferencias (desviaciones) de las observaciones respecto de la media es cero. Es decir, es el único valor que verifica  $\sum (x_i - \bar{x}) f_i = 0$ .

**Ejemplo 1.8** Calcular la media aritmética en los ejemplos 1.5 de la página 16, 1.6 de la página 16 y 1.7 de la página 17.

La media de la variable discreta del ejemplo 1.5 se calcula aplicando directamente la fórmula:

$$\bar{x} = \frac{1 \cdot 15 + 2 \cdot 12 + 3 \cdot 8 + 4 \cdot 4 + 5 \cdot 1}{40} = \frac{84}{40} = 2.1$$

En el ejemplo 1.6, donde la frecuencia para cada uno de sus valores es 1, la media se calcula como sigue

$$\bar{x} = \frac{6 + 10 + 10 + 12 + 20}{5} = \frac{58}{5} = 11'6$$

Si las observaciones están agrupadas por intervalos, como en el ejemplo 1.7, se consideran las marcas de clase como valores de la variable. En el caso de que los intervalos tienen la misma amplitud, obtenemos:

$$\bar{x} = \frac{5 \cdot 8 + 15 \cdot 12 + 25 \cdot 10 + \cdots + 95 \cdot 1}{100} = \frac{4160}{100} = 41'6$$

Para calcular la media aritmética también podemos utilizar la tabla estadística. El procedimiento consiste en añadir una nueva columna ( $x_i f_i$ ) en la que, para cada modalidad de la variable, aparece el producto de su valor por su frecuencia relativa. Finalmente, la suma de los números obtenidos en esta columna corresponde a la media aritmética.

Consideremos el ejemplo 1.7 donde las observaciones se agrupan en intervalos de distinta amplitud. En este caso, añadimos una nueva columna a la tabla estadística donde anotamos los productos de cada uno de los valores de la variable (las marcas de clase) por su correspondientes frecuencia relativa. Al final, en la fila de sumas, aparecerá, en esta columna, el valor de la media aritmética, calculada como  $\sum x_i f_i$ .

$L_{i-1}, L_i$	$x_i$	$n_i$	$f_i$	$x_i f_i$
[0, 50)	25	65	0'65	16'25
[50, 70)	60	25	0'25	15
[70, 90)	80	9	0'09	7'2
[90, 100]	95	1	0'01	0'95
Suma		100	1	$\bar{x}=39'4$

Obsérvese que el valor obtenido para la media (39'4) no coincide con el obtenido antes (41'6), cuando consideramos intervalos de la misma amplitud, para este mismo conjuntos de datos. La razón es que los dos valores son aproximaciones del verdadero valor de la media, que es 41'67, y que se obtendría utilizando los valores originales de las 100 observaciones, sin hacer agrupaciones.

Cuando los datos se agrupan en intervalos, perdemos el valor individual de cada observación. Por eso, al utilizar la marca de clase, como representante de todos los datos de un intervalo, estamos haciendo una aproximación. Las distintas formas de agrupar las observaciones en intervalos, dan lugar a distintas aproximaciones de las medidas resultantes calculadas.  $\square$

En muchos casos y con el fin de simplificar los cálculos (hacer que la media sea 0 o trabajar con números más pequeños) se ve la conveniencia de aplicar una transformación a la variable. En este caso, será necesario estudiar cómo se ve modificada la media de la nueva variable. En las transformaciones afines, que son las más usuales, si  $\bar{x}$  es la media de la variable  $X$ , entonces  $a\bar{x} + b$  es la media aritmética de la variable  $aX + b$ .

**Ejemplo 1.9** *Los salarios de los 6 obreros de una empresa son 800, 1.100, 1.200, 1.400, 1.600 y 1.700 euros. Calcular la media aritmética de los mismos.*

Sea  $X$  la variable estadística que representa los salarios de los obreros. Se considera la variable  $Y = 1/100 \cdot X - 13$  que toma los valores -5, -2, -1, 1, 3, 4. Ahora, la media de la variable  $Y$  es 0 y aplicando la transformación afín se obtiene la media de la variable  $X$ .

$$\text{Si } \bar{y} = \frac{\bar{x}}{100} - 13 \text{ entonces } \bar{x} = 100(\bar{y} + 13) = 100(0 + 13) = 1.300$$

También podíamos haber considerado la variable  $Z = \frac{X - 1300}{100}$  que toma los valores -5, -2, -1, 1, 3, 4 y cuya media vale 0 y en este caso

$$\text{como } \bar{z} = \frac{\bar{x} - 1300}{100} \text{ entonces } \bar{x} = 100\bar{z} + 1300 = 100 \cdot 0 + 1300 = 1.300$$

Obsérvese que en ambos casos, hemos aplicado, en distinto orden, dos transformaciones: una de ellas es, dividir por 100 para cambiar la escala y obtener números más pequeños; y la otra es, restar la media (13, en el primer caso, y 1300 en el segundo) para que la media de la nueva variable sea cero. Como podremos comprobar en algunas de las fórmulas que aparecen en este, y en otros temas, el hecho de que la media sea cero, simplifica notablemente los cálculos.  $\square$

Por último, hay que tener en cuenta que la media aritmética tiene dos graves inconvenientes. Por un lado, este promedio calculado puede no corresponder con ningún valor de la variable, por ejemplo, decir que el número medio de hijos de las familias españolas es 1'2. Por otro lado, la media aritmética es muy sensible a valores extremos de la variable (valores inusuales de la población), por ejemplo, si uno de los datos es “muy distinto” del resto, el valor de la media no es representativo de la muestra. Estos dos problemas se resuelven con el uso de la *moda*, para el primer caso, y de la *mediana*, para el segundo.

### 1.4.2. Moda

La *moda* de un conjunto de datos, que denotaremos por “Mo”, es el valor de la variable que presenta mayor frecuencia. La moda puede no ser única o incluso no existir porque todos los valores tengan la misma frecuencia. Puede usarse incluso con variables cualitativas y viene a solucionar el problema que tiene la media cuando no coincide con ningún valor de la variable o cuando interesa destacar la frecuencia de los valores de la misma.

**Ejemplo 1.10** *Determinar la moda de los datos del ejemplo 1.3 de la página 14.*

Para determinar la moda, se busca la modalidad del atributo “vocales” que tenga mayor frecuencia, que resulta ser la vocal “e”. Por lo tanto, la moda de las vocales de nuestro ejemplo es “e”.  $\square$

Este parámetro es muy fácil de calcular pero tiene el problema de que dos muestras con datos muy parecidos puedan tener modas muy distintas lo que dificulta la comparación. Además aunque se enmarca como medida de tendencia central puede ocurrir que el valor con mayor frecuencia no esté cerca del centro de los datos.



**Ejemplo 1.11** Calcular la moda de las muestras:  $M_1 = \{2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18\}$ ,  $M_2 = \{3, 5, 8, 10, 12, 15, 16\}$  y  $M_3 = \{2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9\}$ .

Buscamos, en cada conjunto de datos, el valor o valores que más se repiten: En  $M_1$  la moda es 9 que corresponde al valor con mayor frecuencia; en  $M_2$  no hay moda porque todos los valores tienen la misma frecuencia; y en  $M_3$  hay dos modas (distribución bimodal) que corresponden a los valores 4 y 7.  $\square$

Si se dispone de una tabla de frecuencias, la moda es sencilla de calcular sin más que buscar el valor de la variable que mayor frecuencia absoluta o relativa presenta.

**Ejemplo 1.12** Calcular la moda de los datos del ejemplo 1.5 de la página 16.

Para calcular la moda, se busca en la columna de la frecuencia absoluta (o relativa) el mayor valor, que resulta ser 15 (o 0'375) y que corresponde al valor 1 de la variable, que es la moda (ver la figura 1.2 de la página 16).  $\square$

En el caso de variables continuas, cuando los datos están agrupados en intervalos, se toma como *intervalo modal*  $(L_{i-1}, L_i]$  el que resulta con mayor altura<sup>1</sup> en el histograma, e interpolando<sup>2</sup>, como se muestra en la figura 1.14, se obtiene la siguiente fórmula para el cálculo de la moda:

$$Mo = L_{i-1} + \frac{\Delta_1}{\Delta_1 + \Delta_2} a_i \quad \text{donde} \quad \Delta_1 = h_i - h_{i-1} \quad \text{y} \quad \Delta_2 = h_i - h_{i+1}$$

siendo  $h_i = n_i/a_i$ , la altura del intervalo  $(L_{i-1}, L_i]$ , teniendo en cuenta que el área del rectángulo es igual a la frecuencia de dicho intervalo.

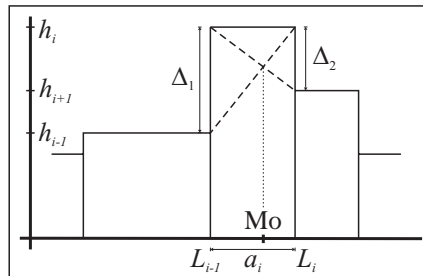


Figura 1.14: Cálculo de la moda en el histograma

Si todos los intervalos tienen la misma amplitud, es igual considerar la altura correspondiente a cada intervalo, o su frecuencia, pues son proporcionales. En tal caso, podemos considerar  $\Delta_1 = n_i - n_{i-1}$  y  $\Delta_2 = n_i - n_{i+1}$ , si consideramos las frecuencias absolutas ( $n_i$ ), o bien  $\Delta_1 = f_i - f_{i-1}$  y  $\Delta_2 = f_i - f_{i+1}$ , si consideramos las frecuencias relativas ( $f_i$ ).

Además, si el intervalo modal fuese el primero de los intervalos, entonces, para aplicar la fórmula de interpolación, se consideraría que la frecuencia del intervalo anterior es cero, es decir,  $n_{i-1} = f_{i-1} = 0$ . De igual manera, consideraremos  $n_{i+1} = f_{i+1} = 0$ , si el intervalo modal es el último de los intervalos considerados.

<sup>1</sup>Hay que tener especial cuidado cuando los intervalos no tienen la misma amplitud pues una mayor frecuencia no está relacionada con una mayor altura del intervalo sino con una área mayor.

<sup>2</sup>La interpolación utilizada para calcular la moda en un intervalo es de tipo cuadrática.

**Ejemplo 1.13** *Calcular la moda de las calificaciones finales en Matemáticas del ejemplo 1.7 de la página 17.*

Si consideramos el caso donde todos los intervalos tienen la misma amplitud (ver la figura 1.4 de la página 18), podemos utilizar la columna de la frecuencia para determinar el intervalo con mayor frecuencia que es el intervalo modal (40,50]. Aplicando la fórmula de interpolación obtenemos:

$$Mo = 40 + \frac{7}{7+5} 10 \approx 45'833$$

Pero si consideramos el caso donde los intervalos no tienen la misma amplitud, entonces tenemos que calcular, necesariamente, la altura correspondiente a cada intervalo. Para ello, utilizamos la tabla de frecuencias donde incluimos dos nuevas columnas correspondientes a la amplitud ( $a_i$ ) y a la altura ( $h_i$ ) de cada intervalo.

$L_{i-1}, L_i$	$x_i$	$n_i$	$f_i$	$a_i$	$h_i = n_i/a_i$
[0, 50)	25	65	0'65	50	1'3
[50, 70)	60	25	0'25	20	1'25
[70, 90)	80	9	0'09	20	0'45
[90, 100]	95	1	0'01	10	0'1
		100	1		

Figura 1.15: Tabla de frecuencias (ejemplo 1.7) con amplitudes y alturas

En la tabla de la figura 1.15 observamos que el intervalo modal es [0,50), pues es el intervalo con mayor altura. Aplicando la fórmula de interpolación obtenemos:

$$Mo = 0 + \frac{1'3}{1'3 + 0'05} 50 \approx 48'148$$

□

### 1.4.3. Mediana

Uno de los inconvenientes de la media aritmética es su sensibilidad a los valores extremos de la variable (valores inusuales de la población), por ejemplo, si uno de los datos difiere bastante del resto, el valor de la media no es representativo de la muestra como vemos en el siguiente ejemplo.

**Ejemplo 1.14** *Consideramos las medidas de los diámetros de diez cilindros, anotadas por un científico: 3'88, 4'09, 3'92, 3'97, 4'02, 3'95, 4'03, 3'92, 3'98, 40'6 cm. Calcular la media aritmética y determinar si es significativo su valor.*

La media aritmética de tales medidas es 7'636 que no es significativa ya que la mayoría de los datos están en torno a 4. Es posible que el último dato provenga de otra población o incluso que esté equivocado (se puede pensar que la coma decimal está mal puesta y el verdadero valor sería 4'06) y sin embargo la media se ha visto muy afectada. □

A la vista del resultado obtenido en el ejemplo anterior, se hace necesario definir una medida central más robusta frente a los datos extremos de la muestra, para que sea más representativa en estos casos.

La *mediana*, o valor mediano, que denotaremos por “Me”, es aquel valor que divide a la población en dos partes de igual tamaño, la mitad son mayores que él y la otra mitad inferiores a él. Si  $N$  es impar, existirá dicho valor y coincidirá con uno de los valores observados, mientras que si es par, se tomarán los dos valores centrales y se calculará la media. Veámoslo en el siguiente ejemplo.

**Ejemplo 1.15** *Calcular la mediana de los conjuntos de datos*

$$C_1 = \{3, 6, 4, 4, 8, 8, 8, 5, 10\} \quad y \quad C_2 = \{15, 5, 7, 18, 11, 12, 5, 9\}.$$

Para calcular la mediana es conveniente ordenar previamente los conjuntos de datos y localizar el valor, o valores, que ocupan la posición central:

$$C_1 = \{3, 4, 4, 5, \underline{6}, 8, 8, 8, 10\} \quad y \quad C_2 = \{5, 5, 7, \underline{9}, \underline{11}, 12, 15, 18\}$$

En  $C_1$  hay 9 datos, y la mediana corresponde al valor de la variable situado en la posición 5, que es el número 6. En  $C_2$  se tienen 8 datos y, por tanto, la mediana es 10 que se calcula como la media aritmética de los valores que ocupan las posiciones 4 (el 9) y 5 (el 11).  $\square$

**Ejemplo 1.16** *Calcular la mediana para los datos del ejemplo 1.14.*

Para calcular la mediana es conveniente ordenar los 10 datos de la muestra y localizar el valor, o valores, que ocupan la posición central:

$$\{3'88, 3'92, 3'92, 3'95, \underline{3'97}, \underline{3'98}, 4'02, 4'03, 4'09, 4'06\}$$

De esta manera, la mediana es 3'975 que se obtiene calculando la media aritmética de los valores de la variable que ocupan las posiciones 5 (el 3'97) y 6 (el 3'98). Obsérvese que este número (3'975) es más representativo que el valor de la media aritmética (7'636) que habíamos calculado en el ejemplo 1.14.  $\square$

Si se dispone de una tabla de frecuencias donde los valores de la variable están ordenados, la mediana corresponde al primer valor de la variable cuya frecuencia relativa acumulada sea mayor o igual que  $1/2$ . Si esta frecuencia es exactamente  $1/2$ , entonces el número de valores de la variable es par y la mediana se obtiene calculando la media aritmética de este valor de la variable y del siguiente.

**Ejemplo 1.17** *Calcular la mediana de los datos del ejercicio 1.5 de la página 16.*

La mediana es 2, pues corresponde al primer valor de la variable que verifica que  $F_i \geq 0'5$ , en concreto,  $F_i = 0'675$  (ver la tabla de la figura 1.2 de la página 16). Si  $F_i$  hubiese valido exactamente 0'5 entonces la mediana hubiese sido 2'5 que es la media aritmética de 2 y 3.  $\square$

En el caso en que los datos vengan agrupados por intervalos se calculará el intervalo que contenga la mediana (intervalo mediano), es decir, el intervalo  $(L_{i-1}, L_i]$  donde  $F_i \geq 1/2$ , o lo

que es lo mismo,  $N_i \geq N/2$ . Si se da la igualdad, entonces la mediana es  $L_i$ . En otro caso, es necesario interpolar en el intervalo mediana, mediante la fórmula

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

que se obtiene, suponiendo que las observaciones están distribuidas uniformemente en el intervalo mediana.

**Ejemplo 1.18** *Calcular la mediana de las calificaciones finales en Matemáticas en el ejemplo 1.7 de la página 17.*

Primero consideramos el caso donde los intervalos tiene la misma amplitud. En la tabla de frecuencias (figura 1.4 de la página 18) se busca el intervalo mediano, que resulta ser  $(40,50]$ , pues corresponde al primer intervalo cuya frecuencia relativa acumulada supera el valor 0'5. En este intervalo se aplica la fórmula de interpolación para obtener el valor de la mediana:

$$\text{Me} = 40 + \frac{50 - 44}{21} 10 \approx 42'857$$

Si consideramos el caso donde los intervalos tiene distinta amplitud (figura 1.5 de la página 18), entonces el intervalo mediana es  $[0,50)$  e interpolando se obtiene el valor de la mediana:

$$\text{Me} = 0 + \frac{50 - 0}{65} 50 \approx 38'462$$

□

#### 1.4.4. Cuantiles

Los *cuantiles* no se clasifican dentro del grupo de medidas de tendencia central, pero sí que son medidas de posición o de orden. Los cuantiles son parámetros que dividen en partes a los datos ordenados de la población determinando así la posición de cada uno de ellos. Por ejemplo, la mediana que hemos definido antes, divide al conjunto de las observaciones en dos partes iguales, es decir, la mitad de las observaciones es menor que la mediana, y la otra mitad son mayores que ella.

En general, un *cuantil de orden*  $k$ , que denotaremos por  $C(k)$ , divide a la población en dos partes de tal manera que una proporción  $k$  de la población es menor que dicho valor y el resto mayor. Se distinguen cuatro tipos de cuantiles que dividen a la población en 4, 5, 10 o 100 partes iguales.

**Cuartiles:** Son 3 y dividen a la población en 4 partes iguales. El primer cuartil, que denotamos por  $Q_1$ , deja a su izquierda a la cuarta parte de la población ( $k = 1/4$ ) que es menor que él. El segundo cuartil, que denotamos por  $Q_2$ , coincide con la mediana, y el tercer cuartil, que denotamos por  $Q_3$ , deja a su izquierda las tres cuartas partes de la población que son menores que él ( $k = 3/4$ ).

**Quintiles:** Son 4 y dividen a la población en 5 partes iguales. El primer quintil deja a su izquierda el 20 % de la población ( $k = 1/5$ ) que es menor que él, el segundo quintil deja al 40 % ( $k = 2/5$ ), el tercer quintil deja al 60 % ( $k = 3/5$ ) y el cuarto quintil deja al 80 % ( $k = 4/5$ ).

**Deciles:** Son 9 y dividen a la población en 10 partes iguales. Se llama decil de orden  $d$  al valor que divide a la población en dos partes, de tal forma que la proporción  $k = d/10$  de la población sea menor que él y el resto mayor.

**Percentiles o Centiles:** Son 99 y dividen a la población en 100 partes iguales. Se llama centil de orden  $c$ , que denotaremos por  $P_c$ , al valor que divide a la población en dos partes de tal forma que la proporción  $k = c/100$  de la población sea menor que él y el resto mayor.

Para calcular el cuantil de orden  $k$  en una distribución discreta, se procede de manera similar al cálculo de la mediana, buscando en la columna de la frecuencia relativa acumulado, cuál es el primer valor mayor o igual que  $k$ .

**Ejemplo 1.19** Calcular los cuartiles  $Q_1$  y  $Q_3$ , los quintiles de orden 1 y 4, los deciles de orden 1 y 9, y los percentiles  $P_1$  y  $P_{99}$  para los datos del ejemplo 1.5 de la página 16.

Para encontrar los cuartiles  $Q_1$  y  $Q_3$  se busca en la columna de las frecuencias relativas acumuladas cuál es el primer valor mayor o igual que 0'25 y 0'75 respectivamente. En este caso, los valores de la variable correspondientes determinan los cuartiles  $Q_1 = 1$  y  $Q_3 = 3$ .

Para calcular los quintiles se procede de la misma manera pero con los valores de  $k$  igual a 1/5 y 4/5 y se obtiene 1 y 3. Análogamente, para los valores de  $k$  igual a 1/10 y 9/10 y se obtiene los deciles de orden 1 y 10 que son respectivamente 1 y 4; y para los valores de  $k$  igual a 1/100 y 99/100 se determinan los percentiles  $P_1=1$  y  $P_{99}=5$ .  $\square$

En el caso de datos agrupados en intervalos, el cuantil de orden  $k$  se calcula interpolando en el intervalo  $(L_{i-1}, L_i]$  donde  $F_i \geq k$  o lo que es lo mismo  $N_i \geq Nk$ . Si se da la igualdad, entonces el cuantil  $C(k)$  es  $L_i$ , y en otro caso, aplicamos la fórmula:

$$C(k) = L_{i-1} + \frac{N \cdot k - N_{i-1}}{n_i} a_i$$

que se obtiene, suponiendo que las observaciones del intervalo están distribuidas uniformemente.

**Ejemplo 1.20** Calcular los cuantiles  $Q_1$ ,  $Q_3$  y  $P_{99}$  para el ejemplo 1.7 de la página 17.

Primero consideramos el caso donde los intervalos tiene la misma amplitud. Para calcular  $Q_1$  se busca el primer intervalo cuya frecuencia relativa acumulada es mayor o igual que 0'25 (ver figura 1.4 de la página 18), que resulta ser (20,30], y después se interpola para obtener el cuartil

$$Q_1 = 20 + \frac{25 - 20}{10} 10 = 25$$

Análogamente, se interpola en el intervalo (50,60] para obtener  $Q_3 = 50 + \frac{75 - 65}{16} 10 = 56'25$ . Sin embargo, cuando se busca el intervalo correspondiente al percentil  $P_{99}$ , se observa que la frecuencia relativa acumulada correspondiente al intervalo (80,90] es igual a 0'99 y por tanto el valor de este percentil es 90.

Si consideramos el caso donde los intervalos tiene distinta amplitud (figura 1.5 de la página 18), entonces  $Q_1 \in [0, 50)$  y  $Q_3 \in [50, 70)$ , y se calculan interpolando así:

$$Q_1 = 0 + \frac{25 - 0}{65} 50 \approx 19'2 \quad , \quad Q_3 = 50 + \frac{75 - 65}{25} 20 = 58$$

Mientras que  $P_{99} = 90$ , sin necesidad de interpolar, pues la frecuencia relativa acumulada correspondiente al intervalo [70,90) es exactamente 0'99.  $\square$

## 1.5. Medidas de dispersión

Las medidas de dispersión constituyen otro importante tipo de medidas descriptivas numéricas que ayudan a determinar la variación de los datos. Estas medidas se usan para determinar lo agrupada o dispersa que está una población y por tanto si la medida de tendencia central calculada, es representativa. Es tan importante buscar un valor central como saber la distribución de los datos en torno a ese valor central. Por ello, las medidas de tendencia central junto a las medidas de dispersión aportan una valiosa información sobre la distribución de los datos.

**Ejemplo 1.21** Para las siguientes muestras, estudiar la representatividad que tiene el valor de la media, en función de la distribución de los datos:

$$M_1 = \{2'2, 2'6, 2'9, 3'4, 3'9\} \quad , \quad M_2 = \{0'5, 1'2, 1'9, 5'2, 6'2\}$$

La media aritmética de las observaciones en cada una de las muestras es la misma, y vale 3. Si embargo, como se observa en la figura 1.16, en  $M_1$  (a la izquierda), las observaciones se agrupan en torno a ese valor, mientras que en  $M_2$  (a la derecha), no ocurre lo mismo. Por lo tanto, el valor 3 de la media es “más representativo” en el conjunto  $M_1$  que en el conjunto  $M_2$ . Es decir, aporta más información puesto que da una mejor imagen del conjunto de datos.



Figura 1.16: Muestras con igual media y distinta dispersión

□

Como se observa en el ejercicio anterior, se hace necesaria la definición de medidas descriptivas de la dispersión de los datos de una muestra. Estas medidas también servirán para determinar la representatividad de las medidas de tendencia central en esas muestras.

En la definición de las medidas de dispersión se considera una muestra de una variable  $X$  que toma los valores  $x_1, x_2, \dots, x_k$  con las frecuencias absolutas  $n_1, n_2, \dots, n_k$  respectivamente, haciendo un total de  $N$  datos.

### 1.5.1. Rango

La medida de dispersión más simple es el *rango*, *recorrido* o *intervalo*, que denotaremos por  $R$ , y que se define como la diferencia entre el mayor valor observado de la variable y el menor.

**Ejemplo 1.22** Calcular los rangos de los conjuntos de datos del ejemplo 1.21.

Si en cada conjunto se busca el mayor y el menor valor de la variable, restando ambos valores se obtiene:

$$R_{C_1} = 3'9 - 2'2 = 1'7 \quad \text{y} \quad R_{C_2} = 6,5 - 0'5 = 6$$

lo que nos indica que los datos de  $C_2$  están más dispersos que los de  $C_1$ , pues el rango es mayor. Más adelante veremos que hay una medida que se utiliza específicamente para comparar la dispersión de dos muestras: el coeficiente de variación. □

**Ejemplo 1.23** Calcular el rango en los ejemplos 1.5 de la página 16, 1.6 de la página 16 y 1.7 de la página 17.

Si en cada ejemplo se busca el mayor y el menor valor de la variable, restando se obtiene:

$$R_{\text{ej:1.5}} = 5 - 1 = 4 \quad , \quad R_{\text{ej:1.6}} = 20 - 6 = 14 \quad \text{y} \quad R_{\text{ej:1.7}} = 92 - 1 = 91$$

□

En algunas ocasiones, para determinar la dispersión de un conjunto de datos, evitando la influencia de los valores extremos, se utilizan otras definiciones de rango que hacen uso de los distintos cuantiles. Los más comunes son:

**Rango intercuartílico**, que se denotaremos por  $R_Q$ , es la diferencia entre el cuartil de orden 3 y el de orden 1

$$R_Q = Q_3 - Q_1$$

**Rango intercentílico**, que se denotaremos por  $R_C$ , es la diferencia entre el percentil de orden 99 y el de orden 1

$$R_C = P_{99} - P_1$$

**Ejemplo 1.24** Calcular los rangos intercuartílico e intercentílico para los datos del ejemplo 1.5 de la página 16.

La única dificultad que tiene el cálculo de rangos es la obtención de los diferentes cuantiles tal y como se explicaba en la sección 1.4.4

$$R_Q = 3 - 1 = 2 \quad \text{y} \quad R_C = 5 - 1 = 4$$

□

Estas medidas de dispersión, además de ser sencillas de calcular, su importancia radica en la capacidad que tienen de detectar posibles datos anómalos (los que están fuera del rango). En la relación de problemas, el ejercicio 29 de la página 48 explica una de estas técnicas de detección.

El rango se utiliza como medida de dispersión en muestras pequeñas porque es una medida relativamente insensible de la variación de los datos. Es decir, es posible que dos conjuntos de datos distintos tengan el mismo rango pero difieran considerablemente en el grado de variación de los datos y esta medida no serviría para detectar esa diferencia.

### 1.5.2. Desviación media

Otra medida de la dispersión de los datos de la muestra se puede obtener calculando la media de las distancias desde cada uno de los valores hasta un punto elegido previamente.

En primer lugar, definimos la *desviación del valor  $x_i$  de la variable respecto del parámetro  $p$*  como la distancia entre estos dos valores, es decir,  $|x_i - p|$ . Normalmente se toma una medida de tendencia central (media o mediana) como valor del parámetro. Después, se calcula la media aritmética de estas desviaciones respecto del promedio, para obtener una medida de la dispersión de la muestra.

La *desviación media respecto a un promedio  $p$*  es la media de las desviaciones de los valores de la variable respecto a una determinada medida de tendencia central  $p$ .

$$DM(p) = \frac{\sum_{i=1}^k |x_i - p| \cdot n_i}{N} = \sum_{i=1}^k |x_i - p| \cdot f_i$$

**Ejemplo 1.25** Calcular la desviación media respecto a la mediana para los datos del ejemplo 1.5 de la página 16.

Aplicando la fórmula se obtiene

$$DM(\text{Me}) = \frac{|1 - 2| \cdot 15 + |2 - 2| \cdot 12 + |3 - 2| \cdot 8 + |4 - 2| \cdot 4 + |5 - 2| \cdot 1}{40} = \frac{34}{40} = 0'85$$

□

Los problemas de cálculo que presenta la utilización de los valores absolutos, sugiere la definición de una nueva medida de dispersión. En cualquier caso, no se perderá de vista la idea de medir desviaciones respecto de un promedio, como procedimiento para medir la dispersión.

### 1.5.3. Varianzas y desviación típica

Al igual que la media aritmética es el promedio más utilizado, la varianza es la medida de dispersión por excelencia. Ambos parámetros suelen presentarse conjuntamente y forman parte de muchas definiciones.

**Varianza poblacional.** Se define la *varianza poblacional* o simplemente *varianza* de un conjunto de datos, que denotaremos por  $\sigma^2$ , como la media aritmética de los cuadrados de las desviaciones con respecto a la propia media de las observaciones, es decir

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

De la definición de varianza se puede deducir una fórmula más simple para su cálculo que consiste en calcular la media de los cuadrados y restarle el cuadrado de la media:

$$\sigma^2 = \sum_{i=1}^k x_i^2 \cdot f_i - \bar{x}^2$$

Para “compensar de algún modo” el cuadrado de las desviaciones y mantener la misma unidad de medida de las observaciones, se define la *desviación típica* o *estándar* de una conjunto de datos como la raíz cuadrada positiva de la varianza:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot f_i}$$



**Ejemplo 1.26** Calcular la varianza y la desviación típica poblacional de los datos del ejemplo 1.5 de la página 16.

La varianza es

$$\sigma^2 = \frac{(1-2'1)^2 \cdot 15 + (2-2'1)^2 \cdot 12 + (3-2'1)^2 \cdot 8 + (4-2'1)^2 \cdot 4 + (5-2'1)^2 \cdot 1}{40} = \frac{47'6}{40} = 1'19$$

y la desviación típica

$$\sigma = \sqrt{1'19} \approx 1'091$$

Otra forma más sencilla de calcular la varianza (con menos operaciones) es

$$\sigma^2 = \frac{1^2 \cdot 15 + 2^2 \cdot 12 + 3^2 \cdot 8 + 4^2 \cdot 4 + 5^2 \cdot 1}{40} - 2'1^2 = \frac{224}{40} - 4'41 = 1'19$$

□

Para aplicar la fórmula y calcular la varianza poblacional podemos utilizar la tabla estadística. Para ello, se añade una nueva columna ( $x_i^2 f_i$ ) en la que, para cada modalidad de la variable, aparece el producto del cuadrado de su valor por su frecuencia relativa. La suma de los números obtenidos en esta columna menos el cuadrado de la media corresponde a la varianza. También podíamos haber añadido una columna para calcular los valores  $(x_i - \bar{x})^2 f_i$  y, en este caso, la varianza sería simplemente la suma de los valores de esta columna.

Como resulta de su definición, la varianza y la desviación típica son números positivos. Ambos parámetros son independientes del cambio de origen, pero no de escala, es decir, si  $\sigma^2$  es la varianza de la variable  $X$ , entonces  $a^2 \sigma^2$  es la varianza de la variable  $aX + b$ .

**Ejemplo 1.27** Calcular la varianza y la desviación típica poblacional para los datos del ejemplo 1.9 de la página 23.

Sea  $X$  la variable estadística que representa los salarios de los obreros. Se considera la variable  $Y = 1/100 \cdot X - 13$  que toma los valores -5, -2, -1, 1, 3, 4. Ahora, la varianza de la variable  $Y$  es 56/6 y aplicando la transformación lineal se obtiene la varianza de la variable  $X$

$$\sigma_x^2 = 100^2 \cdot \sigma_y^2 = 100^2 \cdot \frac{56}{6} \approx 93.333$$

□

A continuación vamos a introducir dos conceptos que están muy relacionados con la media y la varianza poblacional: la variable tipificada y la varianza muestral.

### La variable tipificada.

Haciendo uso de la media y de la desviación típica de la variable  $X$ , se puede considerar una nueva variable que viene dada por:

$$Z = \frac{X - \bar{x}}{\sigma} \quad \text{que toma los valores} \quad z_i = \frac{x_i - \bar{x}}{\sigma} \quad i = 1, 2, \dots, k$$

y que se denomina *variable tipificada*. El proceso de restar la media y dividir por la desviación típica, se conoce como *tipificar*.

**Ejemplo 1.28** *Tipificar los datos del ejemplo 1.5 de la página 16.*

La variable  $X$  definida en el ejemplo 1.5 toma los valores 1 al 5 con frecuencia 15, 12, 8, 4 y 1; su media es 2'1, y su desviación típica es 1'091. Por lo tanto, para calcular los valores ( $z_i$ ) que toma la variable tipificada correspondiente, restaremos la media aritmética ( $\bar{x}$ ), a cada valor original ( $x_i$ ) de la muestra, y el resultado, lo dividiremos por la desviación típica ( $\sigma$ ), y obtenemos:

$$\frac{1 - 2'1}{1'091} \approx -1'008, \quad \frac{2 - 2'1}{1'091} \approx -0,092, \quad \frac{3 - 2'1}{1'091} \approx 0'825, \quad \frac{4 - 2'1}{1'091} \approx 1'742, \quad \frac{5 - 2'1}{1'091} \approx 2'658$$

Esto cinco números son los valores que toma la variable tipificada, y la frecuencias de cada uno de ellos es la misma que la correspondiente frecuencia del valor original.  $\square$

La variable tipificada es adimensional (independiente de las unidades usadas) y mide la desviación de la variable  $X$  respecto de su media en términos de la desviación típica, por lo que resulta de gran valor para comparar valores aislados de distintas distribuciones.

**Ejemplo 1.29** *Un estudiante obtuvo 84 puntos en el examen final de matemáticas, en el que la nota media fue 76 y la desviación típica 10. En el examen final de física obtuvo 90 puntos, siendo la media 82 y la desviación típica 16. Aunque en las dos asignaturas estuvo muy por encima de la media, ¿en cuál sobresalió más?*

Tipificando las variables para poder compararlas se obtiene

$$M = \frac{84 - 76}{10} = 0'8 \qquad F = \frac{90 - 82}{16} = 0'5$$

y se observa que la nota tipificada ( $M$ ) de matemáticas es mejor que la de física ( $F$ ) debido a que se encuentra más alejada de la media en términos de desviación típica. Es decir, la nota de matemáticas se encuentra a 0'8 desviaciones típicas por encima de la nota media y por tanto es superior a la nota de física que sólo supera a la nota media en 0'5 desviaciones típicas.  $\square$

### La cuasivarianza.

Se define la *varianza muestral* o *cuasi-varianza* como

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

siendo  $s = \sqrt{s^2}$  la *cuasidesviación típica* o *desviación típica muestral*.

Este parámetro tendrá una gran importancia en la inferencia estadística donde se trabaja con muestras. Como veremos, el estadístico  $s^2$ , calculado a partir de los datos de la muestra, será el mejor estimador del valor del parámetro  $\sigma^2$  de la población. Obsérvese que cuando el tamaño muestral es muy grande, la muestra resulta ser muy significativa, y el valor de  $s^2$  es muy próximo a  $\sigma^2$  ya que  $N - 1 \approx N$ .

Conviene no confundir la varianza de la muestra, que se calcula aplicando la fórmula de  $\sigma^2$  a los valores de la muestra, con la varianza muestral que corresponde a  $s^2$ . Sin embargo, de la expresión de sus fórmulas se deducen las siguientes relaciones entre ellas:

$$s^2 = \frac{N}{N - 1} \sigma^2 \qquad \text{o bien} \qquad \sigma^2 = \frac{N - 1}{N} s^2$$

### 1.5.4. Coeficiente de variación

Las medidas de dispersión que se han visto hasta ahora, vienen expresadas en las unidades de la variable, y por tanto, no resultan útiles para establecer una comparación entre las dispersiones de dos muestras distintas, o que simplemente, que vengan expresadas en unidades distintas.

Para solucionar este problema se define el *coeficiente de variación de Pearson* que es el cociente entre la desviación típica y el valor absoluto de la media:

$$CV = \frac{\sigma}{|\bar{x}|}$$

si bien, para su mejor interpretación, es bastante común expresarlo como porcentaje (multiplicado por 100).

El principal problema que tiene este coeficiente es que pierde representatividad cuando la media se acerca a cero.

**Ejemplo 1.30** *Calcular el coeficiente de variación de Pearson del ejemplo 1.5 de la página 16.*

En los ejemplos anteriores se había calculado la media (2'1) y la varianza (1'19). Ahora sólo habrá que aplicarla la fórmula para obtener

$$CV = \frac{\sqrt{1'19}}{2'1} \approx 0'52 \quad (52 \%)$$

□

Este coeficiente mide la dispersión relativa de la muestra y su ventaja es que resulta independiente de la unidad de medida o cambio de escala; por tanto, permite establecer una comparación entre las dispersiones de dos muestras que vengan expresadas en distintas unidades.

**Ejemplo 1.31** *Un fabricante de tubos de televisión produce dos tipos de tubos, A y B, que tienen vidas medias respectivas  $\bar{x}_A=1495$  horas y  $\bar{x}_B=1875$  horas, y desviación típica  $\sigma_A=280$  horas y  $\sigma_B=310$ . Comparar las dispersiones de las dos poblaciones en términos absolutos y relativos.*

Los coeficientes de variación para cada tipo de tubos

$$CV_A = \frac{280}{1495} \cdot 100 \approx 18'73 \% \qquad CV_B = \frac{310}{1875} \cdot 100 \approx 16'53 \%$$

indican que, en términos relativos, la dispersión es mayor en la población A; a pesar de que las desviaciones típicas sugieran lo contrario. □

En general, también se define el *coeficiente de variación media* respecto al promedio  $p$  de la forma:

$$CVM(p) = \frac{DM(p)}{|p|}$$

Como en el caso de la desviación media, el parámetro  $p$  puede ser cualquier valor pero suele utilizarse la media o la mediana.

**OBSERVACIÓN:** Es importante no confundir la variable tipificada con el coeficiente de variación. Ambos son adimensionales y permiten hacer comparaciones. Sin embargo, utilizaremos el coeficiente de variación para comparar las dispersiones de dos muestras o poblaciones, mientras que, utilizaremos la variable tipificada para comparar dos valores concretos de dos muestras o poblaciones distintas.

### 1.5.5. Momentos

Los momentos son medidas descriptivas que resultan muy útiles para calcular determinados parámetros. Estas medidas generalizan las definiciones de media aritmética y varianza, y como veremos, forman parte de la definición de algunos coeficientes.

En general, se define el *momento de orden  $r$  respecto al punto  $c$*  de la forma:

$$M_r(c) = \sum_{i=1}^k (x_i - c)^r \cdot f_i$$

aunque resultan de especial interés los siguientes dos casos particulares:

**Momentos ordinarios:** Si  $c = 0$  entonces el momento de orden  $r$  recibe el nombre de momento ordinario, se denota por  $m_r$ , se calcula así

$$m_r = \sum_{i=1}^k x_i^r \cdot f_i$$

y se observa que si  $r = 1$  se tiene la definición de media aritmética.

**Momentos centrales:** Si  $c = \bar{x}$  entonces el momento de orden  $r$  recibe el nombre de momento central, se denota por  $\mu_r$ , se calcula así

$$\mu_r = \sum_{i=1}^k (x_i - \bar{x})^r \cdot f_i$$

y se observa que si  $r = 2$  se tiene la definición de varianza.

Para aplicar la fórmula y calcular los momentos podemos utilizar la tabla estadística, tal y como se ha explicado en el cálculo de la media o la varianza. El procedimiento consiste en añadir una nueva columna con las operaciones correspondientes para cada modalidad de la variable  $((x_i - c)^r \cdot f_i)$  y sumar los números obtenidos.

**Ejemplo 1.32** Calcular los momentos ordinario y central de orden 4 de los datos del ejemplo 1.5 de la página 16.

Aplicamos directamente la fórmula para calcular el momento ordinario

$$m_4 = \frac{1^4 \cdot 15 + 2^4 \cdot 12 + 3^4 \cdot 8 + 4^4 \cdot 4 + 5^4 \cdot 1}{40} = \frac{2504}{40} = 62'6$$

y sabiendo que la media es 2'1 calculamos el momento central

$$\mu_4 = \frac{(1-2'1)^4 15 + (2-2'1)^4 12 + (3-2'1)^4 8 + (4-2'1)^4 4 + (5-2'1)^4 1}{40} = \frac{150'068}{40} = 3'7517$$

□

Se destacan las siguientes propiedades relativas a los momentos:

$$\begin{array}{lll} 1) & m_0 = 1 & 2) & m_1 = \bar{x} & 3) & m_2 = \sigma^2 + \bar{x}^2 \\ 4) & \mu_0 = 1 & 5) & \mu_1 = 0 & 6) & \mu_2 = \sigma^2 = m_2 - \bar{x}^2 \end{array}$$

y las relaciones entre los momentos centrales y ordinarios, como por ejemplo,

$$\mu_2 = m_2 - m_1^2 \quad \mu_3 = m_3 - 3m_1m_2 + 2m_1^3 \quad \mu_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4$$

que nos permiten calcular los momentos centrales, en términos de los momentos ordinarios, que son más simples de calcular.

**Ejemplo 1.33** Calcular el momento central de orden 3 de los datos del ejemplo 1.5 de la página 16 a partir de los momentos ordinarios.

Primero se calculan los momentos ordinarios de orden 1, 2 y 3 que son  $m_1 = 2'1$ ,  $m_2 = 5'6$  y  $m_3 = 17'7$  y se aplica la relación correspondiente para obtener

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3 = 17'7 - 3 \cdot 2'1 \cdot 5'6 + 2 \cdot (2'1)^3 = 0'942$$

□

## 1.6. Medidas de forma

La forma que presenta su representación gráfica permite clasificar una distribución de frecuencias. En esta sección nos fijaremos en dos características: la simetría y el apuntamiento, y proporcionaremos coeficientes que nos permitan comparar dos distribuciones.

### 1.6.1. Medidas de asimetría

Se dice que una distribución de frecuencias es simétrica cuando los valores de la variable que equidistan de un valor central tienen las mismas frecuencias. Esta situación ideal viene representada por una gráfica simétrica y en tal caso se verifica que  $\bar{x} = Me = Mo$ .

Se dice que una distribución de frecuencias es *asimétrica* si no es simétrica y esta asimetría puede presentarse a la derecha o a la izquierda (ver figura 1.17):

- Una *distribución asimétrica a la derecha o positiva* se caracteriza porque la gráfica de frecuencias presenta cola a la derecha, es decir, éstas descienden más lentamente por la derecha que por la izquierda. En este caso se verifica que  $Mo \leq Me \leq \bar{x}$ .
- Una *distribución asimétrica a la izquierda o negativa* se caracteriza porque la gráfica de frecuencias presenta cola a la izquierda, es decir, éstas descienden más lentamente por la izquierda que por la derecha. En este caso se verifica que  $\bar{x} \leq Me \leq Mo$ .

A continuación, se presentan dos coeficientes que permiten estudiar el grado de asimetría o sesgo de una distribución, sin necesidad de representarla.

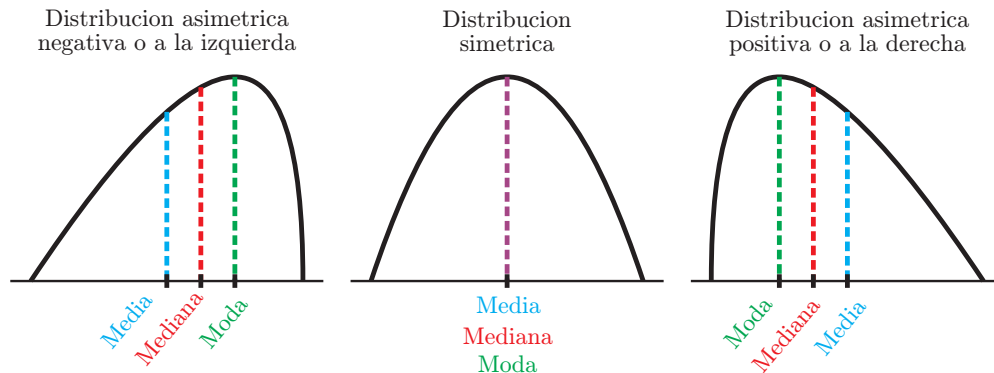


Figura 1.17: Formas de la distribución de frecuencias

**Coefficiente de asimetría de Pearson.** De acuerdo a las relaciones entre media, mediana y moda, establecidas para las distintas asimetrías, se define y se interpreta el coeficiente de sesgo de Pearson como sigue

$$A_P = \frac{\bar{x} - Mo}{\sigma} \quad \text{donde} \quad \begin{cases} A_P > 0 & \text{Asimetría a la derecha o positiva} \\ A_P = 0 & \text{Simetría} \\ A_P < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

**Ejemplo 1.34** Utilizar el coeficiente de Pearson para determinar el sesgo en el ejemplo 1.5 de la página 16.

Utilizando los datos obtenidos en los ejemplos anteriores y aplicando la fórmula se obtiene

$$A_P = \frac{2'1 - 1}{\sqrt{1'19}} \approx 1 > 0$$

lo que indica que la distribución es asimétrica a la derecha (ver figura 1.18). □

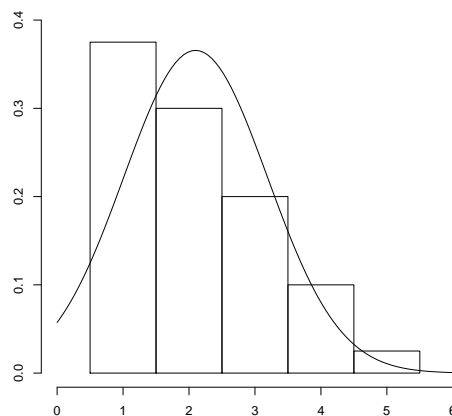


Figura 1.18: Formas de la distribución de frecuencias del ejemplo 1.5

**Coefficiente de asimetría de Fisher o 2º de Pearson.** Otro coeficiente adimensional que mide el sesgo, haciendo uso del momento central de orden 3, es el coeficiente de asimetría de Fisher que se define y se interpreta como sigue

$$g_1 = \frac{\mu_3}{\sigma^3} \quad \text{donde} \quad \begin{cases} g_1 > 0 & \text{Asimetría a la derecha o positiva} \\ g_1 = 0 & \text{Simetría} \\ g_1 < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

y que tiene su explicación en la comparación con la distribución normal que es simétrica y cuyo coeficiente de asimetría de Fisher toma el valor 0 para cualquier media y varianza.

**Ejemplo 1.35** Utilizar el coeficiente de Fisher para determinar el sesgo en el ejemplo 1.5 de la página 16.

Utilizando los datos obtenidos en los ejemplos anteriores y aplicando la fórmula se obtiene

$$g_1 = \frac{0'942}{(\sqrt{1'19})^3} \approx 0'726 > 0$$

lo que confirma que la distribución es asimétrica a la derecha (ver figura 1.18).  $\square$

### 1.6.2. Medidas de apuntamiento

El *apuntamiento* o la *curtosis* determina si la distribución de frecuencias es más o menos afilada o aplastada que la función de densidad de la distribución normal<sup>3</sup> con igual media y varianza, que se toma como referencia.

En la figura 1.19 se representan tres distribuciones de frecuencias que, de izquierda a derecha, son platicúrtica (más aplastada que la distribución normal), mesocúrtica (similar a la distribución normal) y leptocúrtica (más apuntada que la distribución normal). En cada una de ellas se ha representado la respectiva distribución normal con igual media y varianza.

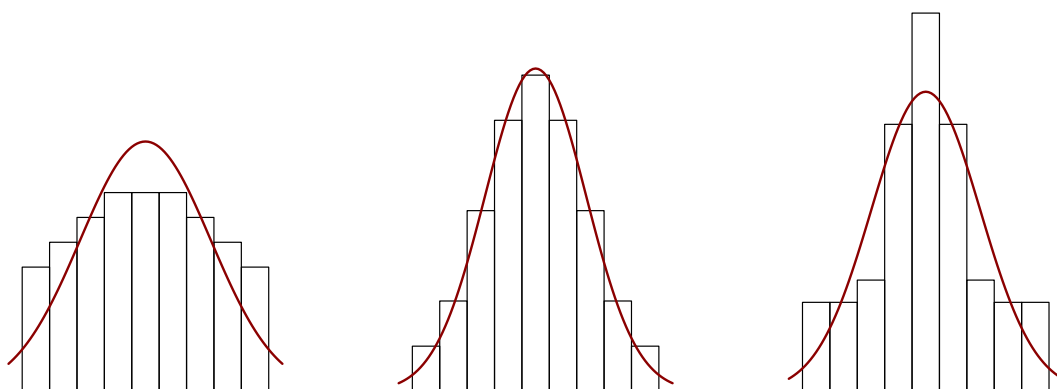


Figura 1.19: Formas de la distribución de frecuencias

<sup>3</sup>La función de densidad de la distribución normal de media  $\mu$  y desviación  $\sigma$  es la función definida por  $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ , y cuya gráfica se conoce como “campana de Gauss”.

Para determinar el grado de apuntamiento se define la siguiente medida:

**Coefficiente de aplastamiento de Fisher.** Un coeficiente adimensional que mide la curtosis de una muestra, haciendo uso del momento central de orden 4, es el coeficiente de aplastamiento de Fisher que se define y se interpreta como sigue

$$g_2 = \frac{\mu_4}{\sigma^4} - 3 \quad \text{donde} \quad \begin{cases} g_2 < 0 & \text{Menos apuntamiento que la normal.} \\ g_2 = 0 & \text{Igual apuntamiento que la normal.} \\ g_2 > 0 & \text{Más apuntamiento que la normal.} \end{cases}$$

Esta fórmula tiene su explicación en la comparación con la distribución normal. Se puede comprobar que el cociente  $\mu_4/\sigma^4$  siempre toma el valor 3 en la distribución normal de cualquier media y varianza. Por lo tanto, el coeficiente de aplastamiento de Fisher de la distribución normal toma siempre el valor 0.

**Ejemplo 1.36** *Determinar el apuntamiento de la distribución de los datos del ejemplo 1.5 de la página 16.*

Utilizando los datos obtenidos en los ejemplos anteriores y aplicando la fórmula del coeficiente de apuntamiento se obtiene

$$g_2 = \frac{3'7517}{(\sqrt{1'19})^4} - 3 \approx -0,35 < 0$$

lo que indica que la distribución es menos apuntada (más aplastada) que la normal de igual media y varianza.  $\square$



## 1.7. Relación de problemas

1. La fiabilidad de un ordenador se mide en términos de la vida de un componente de hardware específico (por ejemplo, la unidad de disco). Con objeto de estimar la fiabilidad de un sistema en particular, se prueban 100 componentes de un ordenador hasta que fallan, y se registra su vida.
  - a) Determinar la población de interés, los individuos y la muestra.
  - b) Determinar el carácter, su tipo y las posibles modalidades.
  - c) ¿Cómo podría utilizarse la información de la muestra para estimar la fiabilidad del sistema?
2. Cada cinco años, la División de Mecánica de la American Society of Engineering Education realiza una encuesta a nivel nacional sobre la educación en Mecánica, en el nivel de licenciatura, en las Universidades. En la encuesta más reciente, 66 de las 100 universidades muestreadas cubrían la estática de fluidos en su programa de ingeniería en el nivel de licenciatura.
  - a) Determinar la población de interés, los individuos y la muestra.
  - b) Determinar el carácter, su tipo y las modalidades del estudio.
  - c) Utilice la información de la muestra para inferir resultados de la población.
3. Para cada uno de los siguientes conjuntos de datos, indique si son cualitativos o cuantitativos y describir las distintas modalidades.
  - a) Tiempos de llegada de 16 ondas sísmicas reflejadas.
  - b) Marcas de calculadoras empleadas por 100 estudiantes de Ingeniería.
  - c) Velocidad máxima alcanzada por 12 automóviles impulsados con energía solar.
  - d) Número de caracteres impresos por línea de salida de computadora en 20 impresoras de línea.
  - e) Número de miembros de las familias malagueñas.
  - f) Estado civil del personal de una empresa.
  - g) Horas de vuelo de los pilotos de una compañía aérea.
4. En cada caso, determinar el tipo de distribución, organizar los datos en una tabla de frecuencias y representar gráficamente la distribución. También se pide, calcular algunas medidas de tendencia central, medidas de dispersión, de simetría y de apuntamiento.
  - a) Resistencia a la tensión ( $\text{Kg/mm}^2$ ) de láminas de acero.
 

44	43	41	41	44	44	43	44	42	45	43	43	44	45	46
42	45	41	44	44	43	44	46	41	43	45	45	42	44	44
  - b) Tiempo de espera (redondeado en minutos) de un conmutador, para cierto tren subterráneo.
 

3	4	1	0	2	2
---	---	---	---	---	---

- c) En ciertos entornos, los aceros inoxidable son especialmente susceptibles al agrietamiento. A continuación se relacionan las causas asignables y el número de casos detectados correspondientes a estas causas, en un estudio realizado entre 200 aceros observados.

Entorno húmedo	144
Entorno seco	45
Defectos de materiales	4
Defectos de soldadura	7

- d) Contenido de carbono (%) del carbón mineral.

87	86	85	87	86	87	86	81	77	85
86	84	83	83	82	84	83	79	82	73

- e) Consumo de combustible (litros/100km a 90km/h) de seis automóviles de la misma marca.

6'7	6'3	6'5	6'5	6'4	6'6
-----	-----	-----	-----	-----	-----

- f) Número de hojas de papel, por encima y por debajo del número deseado de 100 por paquete, en un proceso de empaquetado.

0	-1	0	0	1	1	2	0	1	0
---	----	---	---	---	---	---	---	---	---

- g) Resultados obtenidos en las pruebas de durabilidad de 80 lámparas eléctricas con filamento de tungsteno. La vida de cada lámpara se da en horas, aproximando las cifras a la hora más cercana.

854	1284	1001	911	1168	963	1279	1494	798	1599	1357	1090	1082
1494	1684	1281	590	960	1310	1571	1355	1502	1251	1666	778	1200
849	1454	919	1484	1550	628	1325	1073	1273	1710	1734	1928	1416
1465	1608	1367	1152	1393	1339	1026	1299	1242	1508	705	1199	1155
822	1448	1623	1084	1220	1650	1091	210	1058	1930	1365	1291	683
1399	1198	518	1199	2074	811	1137	1185	892	937	945	1215	905
1810	1265											

- h) Los clientes de una empresa necesitan contactar telefónicamente con el departamento de mantenimiento para realizar consultas y aclarar dudas. La gerencia ha recibido quejas de los clientes que suelen encontrar la línea ocupada. Para determinar el número de líneas nuevas que necesita incorporar a la centralita se realizó una encuesta entre algunos de los clientes. La siguiente tabla recoge el número de reintentos que necesitaron realizar esos clientes en su última llamada telefónica a la empresa.

3	4	3	3	1	4	1	3	2	3
1	1	4	2	3	3	2	6	1	1
3	3	2	2	2	2	1	3	2	1
6	3	1	2	2	3	2	2	4	2

5. Calcular los valores que se piden en función de los datos:

- a) Si  $N = 2$ ,  $\bar{x} = 2'6$  y  $\sigma = 1'1$ , ¿cuáles son los datos de la muestra?
- b) Si  $CV = 0'5$ ,  $\bar{x} = 2$  y  $m_3 = 14$ , ¿cuánto vale  $\mu_3$ ?

6. Se considera la siguiente tabla de frecuencias donde las distintas modalidades están ordenadas de menor a mayor

$x_i$	$n_i$	$N_i$	$f_i$	$F_i$
	10			
0		15		0'3
3				
5			0'08	
20				0'8
25		46		
50				1

- a) Completar la tabla estadística, utilizando los datos que ya contiene, y los valores de las siguientes medidas:  $N=50$ ,  $\bar{x}=10$ ,  $Me=4$ ,  $Mo=10$ , Rango=51 y  $\sigma^2=201$ .
- b) Determinar qué datos y medidas resultan irrelevantes para completar la tabla.
7. Se atribuye a George Bernard Shaw (el célebre dramaturgo y polemista irlandés) la siguiente observación: Si dos amigos encuentran un pollo y se lo come uno de ellos, la estadística afirma que en promedio cada amigo se ha comido medio pollo. Utilícese la metodología estadística para precisar el contenido de esta proposición.
8. El tamaño de la muestra A es 10, y la media y la mediana son respectivamente 16'5 y 13. El tamaño de la muestra B es 20, y la media y la mediana son respectivamente 11'4 y 10. Consideremos la unión de las dos muestras, que denotaremos por C, cuyo tamaño es 30. Si es posible, calcule la media y la mediana de la muestra C, y en otro caso, determine la posición aproximada de la medida desconocida.
9. El sueldo medio de los obreros de una fábrica es 1.500 euros. En las negociaciones del nuevo convenio colectivo se presentan dos alternativas: un aumento de 150 euros euros a cada obrero o un aumento del 10 % del sueldo de cada uno. Estudiar qué modalidad es más social en el sentido de que iguala más los salarios.
10. Busque un ejemplo donde la diferencia entre la mediana y la moda sea mayor que el rango intercuartílico.
11. Sea  $k$  un número entero positivo. Determine la media, la varianza y el sesgo en cada una de las siguientes muestras:
- a)  $M_1 = \{1, 2, 3, \dots, k\}$
- b)  $M_2 = \{p, p + c, p + 2c, p + 3c, \dots, p + kc\}$ , con  $p \in \mathbb{R}$ .
12. En un examen final de Estadística, la puntuación media de 150 estudiantes fue de 7'8, y la desviación típica de 0'8. En Cálculo, la media fue 7'3 y la desviación típica 0'76. ¿En qué materia fue mayor la dispersión en términos absolutos? ¿y en términos relativos? Explicar la respuesta. Si un alumno obtuvo 7'5 en Estadística y 7'1 en Cálculo, ¿en qué examen sobresalió más?
13. En una muestra se obtienen los valores 2, 4, 6 y 8 de la variable  $X$ . Se pide:
- a) Calcular la media y la varianza de los valores de la muestra.

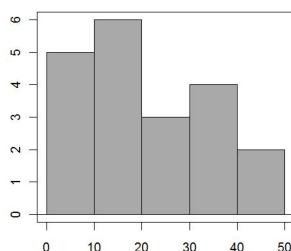
- b) Hallar los valores tipificados de la variable  $X$  y comprobar que la media de estos nuevos valores es 0 y la varianza es 1.
- c) Demostrar que el resultado del apartado anterior constituye una propiedad de cualquier variable tipificada.

14. Las distribuciones de frecuencias de las variables  $X$  e  $Y$  son campaniformes y simétricas. Además, se sabe conocen los siguientes datos:

Variable $X$	Me=10	$\sigma_x^2=4$	N=2	$\sum x_i^4 f_i=12416$
Variable $Y$	Mo=8	$\sigma_y^2=4$	N=82	$\sum y_i^4 f_i= 5648$

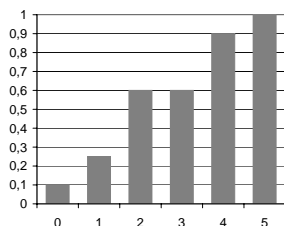
Determinar los dos valores de  $X$ , y comparar la dispersión y la curtosis de ambas variables.

15. Demostrar la igualdad  $\sum_{i=1}^k (x_i - \bar{x})^2 f_i = \sum_{i=1}^k x_i^2 f_i - \bar{x}^2$  que define a la varianza.
16. Encontrar una fórmula general que relacione el momento central de orden  $r$  con los momentos ordinarios de un orden menor o igual. Utilizar esta fórmula para comprobar las relaciones entre los momentos que aparecen en la sección 1.5.5 y calcular el momento central de orden 5 a partir de los momentos ordinarios.
17. Consideremos el siguiente histograma que representa la frecuencia absoluta de los valores de una muestra. Se pide:



- a) Calcular la media, mediana y moda.
- b) Calcular el rango intercuartílico.
- c) Calcular la varianza.

18. Consideremos el siguiente diagrama de frecuencias relativas acumuladas. Se pide:



- a) Calcular la media, mediana y moda de la variable  $X$ .
- b) Calcular el rango intercuartílico.
- c) Calcular la varianza.

19. **Sentido crítico.** Antes de extraer conclusiones de unos resultados estadísticos, conviene examinar detenidamente los valores numéricos obtenidos. El gran número de operaciones realizadas y el volumen de datos manejados son fuentes de error que inciden en los resultados. Un poco de sentido crítico puede ayudar a determinar si unos resultados son consistentes con los datos del problema. En este ejercicio se propone una serie de casos donde el resultado numérico no es correcto. Se trata de explicar razonadamente la inconsistencia del resultado en función de los datos.

- a) El número medio de accesos a una página web es -3.
- b) La mediana del número de hijos de las familias españolas es 2'1.
- c) La moda del número de hijos es 1'5.

- d) El cuartil  $C_3$  es 28 y el cuartil  $C_1$  es 32.
- e) El centil  $P_1$  es 32 y el decil  $D_1$  es 28.
- f) La varianza es -100.
- g) La media es 10, la mediana 12 y la desviación típica es 0.
- h) La expresión  $g_2 + 3$  toma un valor negativo.

20. **Modificar los datos de una muestra** En este ejercicio se va a estudiar el comportamiento de la media y la varianza cuando se pierde, se gana o se modifica algún dato de la variable. Se consideran los valores  $\{2, 4, 6, 8\}$  obtenidos en una muestra. Se pide:

- a) Calcular la media y la varianza.
- b) En cada caso, obtener el nuevo valor de la media y la varianza sin tener que aplicar nuevamente las fórmulas a todos los datos:

Caso1: Se descubre que el valor 8 observado es erróneo y se elimina.

Caso2: Se cuenta con un nuevo valor, el 5, para la muestra.

Caso3: Se descubre que el valor 8 observado es erróneo y se cambia por el verdadero valor que es el 9.

21. Estudiamos el tiempo de duración de un proceso donde, en algunos casos, el proceso ni siquiera comienza y, por tanto, el tiempo de duración es cero. Realizamos 200 pruebas y obtenemos un tiempo medio de 3'5 segundos con una varianza de 7.

- a) Si el 23 % de las pruebas fueron consideradas de tiempo 0. ¿Cuál es la media y la varianza de las restantes.
- b) Si en las 200 pruebas se obtuvieron tiempos positivos y consideramos 50 nuevas pruebas de tiempo 0, ¿cuál es la nueva media y varianza para las 250 observaciones?
- c) Obtener una fórmula que permita obtener la nueva media y varianza de una muestra cuando añadimos o eliminamos un número arbitrario de observaciones de valor 0.

22. En ocasiones, determinar si los resultados de un problema son coherentes con los datos, no es tan directo como en los apartados del ejercicio 19. Por ejemplo, supongamos que en una muestra de 200 observaciones, se obtiene que la media es 35 y la varianza es 7. ¿Son coherentes estos resultados, si sabemos que el 23 % de las observaciones toma el valor 0? Intenta razonar la respuesta y después, calcula el valor de la varianza de la muestra, sin considerar los valores nulos, pues el resultado indica la incoherencia de los datos del problema.

23. **Datos agrupados.** Se consideran los datos del ejemplo 1.7 de la página 17 y los resultados obtenidos a lo largo del capítulo. Se estudia cómo afecta la partición en intervalos a los parámetros calculados. Para ello, se pide:

- a) Dividir el rango en intervalos de amplitud 20 y calcular los distintos parámetros: Media, mediana, moda, rango intercuartílico, varianza, coeficiente de variación, coeficiente de asimetría de Fisher y coeficiente de apuntamiento.
- b) Repetir el ejercicio anterior dividiendo el rango en intervalos regulares de amplitud 5, 25 y 50. Considerar también la partición irregular por calificaciones:  $[0,20)$ ,  $[20,50)$ ,  $[50,60)$ ,  $[60,70)$ ,  $[70,90)$  y  $[90,100]$ .

- c) Comparar los datos obtenidos en las distintas particiones y determinar cómo afecta al resultado numérico de cada parámetro.
- d) Comparar los valores numéricos obtenidos para los distintos parámetros con los que se obtienen si no se consideran los datos agrupados.

24. **Tablas de frecuencias.** En el tema se comenta que las tablas de frecuencias pueden resultar muy útiles para realizar los cálculos de determinados parámetros y son fácilmente implementables en una hoja de cálculo. Para ello, basta con añadir columnas (a la derecha) que contengan operaciones entre los valores calculados en la columnas anteriores y una fila (al final de la tabla) que representa la suma de los valores de la columna correspondiente. En la siguiente tabla se incluyen algunas de estas columnas:

$x_i$	$n_i$	$f_i$	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$ x_i - \bar{x}  \cdot f_i$
$x_1$	$n_1$	$f_1$	$x_1 \cdot f_1$	$x_1^2 \cdot f_1$	$ x_1 - \bar{x}  \cdot f_1$
$x_2$	$n_2$	$f_2$	$x_2 \cdot f_2$	$x_2^2 \cdot f_2$	$ x_2 - \bar{x}  \cdot f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_k$	$n_k$	$f_k$	$x_k \cdot f_k$	$x_k^2 \cdot f_k$	$ x_k - \bar{x}  \cdot f_k$
	$N$	1			

Se pide

- a) Determinar la utilidad de las columnas introducidas en la tabla de frecuencias.
  - b) Utilizar este método para calcular la media, la varianza y los momentos ordinario y central de orden 3 en el ejemplo 1.7 de la página 17
25. **Media ponderada.** Una generalización del concepto de media aritmética es la *media ponderada*. Se utiliza cuando se asocian ciertos valores  $(w_1, w_2, \dots, w_k)$ , denominados pesos, a los valores  $(x_1, x_2, \dots, x_k)$  de la variable con el fin de dar más relevancia a unos que a otros.

$$MP = \frac{\sum_{i=1}^k w_i x_i}{\sum_{i=1}^k w_i}$$

El conjunto de los pesos  $\{w_1, w_2, \dots, w_k\}$  se denomina *ponderación*, y diremos que una ponderación es *propia* si todos los pesos son distintos de cero, es decir,  $w_i \neq 0$  para todo  $i = 1, \dots, n$ .

Ahora, veamos un ejemplo: Si la nota final de una asignatura se obtiene mediante la realización de tres pruebas parciales con pesos 1, 2 y 2, indica que la prueba segunda y tercera tiene el doble de importancia que la primera. En este caso, un alumno cuyas notas hubiesen sido 7'5, 3'0 y 5'5, su nota final sería:

$$\frac{1 \cdot 7'5 + 2 \cdot 3'0 + 2 \cdot 5'5}{1 + 2 + 2} = \frac{24'5}{5} = 4'9$$

Se pide:

- a) ¿Qué nota tendría que haber sacado en la tercera prueba para aprobar la asignatura?  
 b) ¿Cuál habría sido su nota final si los pesos hubiesen sido 2, 1, y 1?

26. Tenemos dos muestras  $A$  y  $B$

$$\begin{array}{lcl} A & \rightarrow & 1 \quad 2 \quad 3 \quad 4 \quad 5 \\ B & \rightarrow & 2 \quad 4 \quad 5 \quad 6 \quad 8 \end{array}$$

y observamos que por pares, los datos de la muestra  $A$  son menores que los valores de la muestra  $B$ . En este caso, si calculamos las medias aritméticas, obviamente, obtenemos un valor menor para  $B$ . Pero, ¿qué sucede con la media ponderada?

- a) Calcular las medias aritméticas de las muestras  $A$  y  $B$ .  
 b) Encontrar una ponderación para cada una de las variables, de manera que la media resultante de la muestra  $A$  sea mayor que la de la muestra  $B$ .  
 c) ¿Existe alguna ponderación propia de los datos de la muestra  $A$  que permita obtener una media mayor o igual de 5 o menor o igual de 1?  
 d) Obtener una ponderación propia para los datos de la muestra  $A$  de tal forma que la media sea 4. Y análogamente para la muestra  $B$ .
27. **Otras medias.** Aunque la media aritmética es la más utilizada, existen otras medidas de tendencia central que pueden resultar interesantes para determinados casos. Otro tipo de medias lo constituye un grupo denominado  $\varphi$ -medias que se obtienen aplicando la fórmula

$$\varphi^{-1} \left( \sum_{i=1}^k \varphi(x_i) f_i \right)$$

para alguna función  $\varphi$  que sea continua y monótona en el intervalo de valores posibles de la variable. Las más usuales son la media cuadrática, armónica y geométrica que utilizan la función que se indica:

Media cuadrática	$MQ = \sqrt{\frac{x_1^2 n_1 + x_2^2 n_2 + \dots + x_k^2 n_k}{N}}$	$\varphi(x) = x^2$
Media armónica	$H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_k}{x_k}}$	$\varphi(x) = \frac{1}{x}$
Media geométrica	$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \dots x_k^{n_k}}$	$\varphi(x) = \ln(x)$

Entre ellas se establece la siguiente relación:

$$H \leq G \leq \bar{x} \leq MQ$$

Se pide

- a) Comprobar que se verifica la relación anterior haciendo uso de los datos del ejemplo 1.5 de la página 16  
 b) Calcular, si es posible, el valor de las cuatro medias anteriores para los valores 2, 6 y 10, y analiza los distintos resultados pensando que esos valores corresponden a las notas de los tres exámenes de una asignatura.

- c) Repetir el apartado anterior con los valores 0, 5 y 10.
- d) Buscar, en la bibliografía, las características de cada una de estas medias y sus aplicaciones.
- e) Definir una nueva  $\varphi$ -media utilizando la función exponencial y alguna función trigonométrica. Observación: Las funciones utilizadas han de ser monótonas en el rango de valores de la variable.
28. Un manera estándar, para determinar el tiempo que se tarda en realizar un proceso, es calcular el tiempo medio empleado en cada ejecución, al realizar un número elevado de simulaciones. Puede ocurrir que determinadas ejecuciones del procesos caigan en bucles o tarden un tiempo indeterminado que obliguen a parar el proceso. En estos casos, asignamos un tiempo infinito a esas ejecuciones del proceso.
- a) Indicar los inconvenientes que presentan los posibles indicadores del tiempo empleado: tiempo medio, mediano, moda, media armónica, cuadrática o geométrica.
- b) Elegir el indicador(es) más adecuado(s) y aplicarlo(s) a los siguientes tiempos de ejecución de un proceso: 23, 56, 12, 25,  $\infty$ , 22, 23, 26, 23, 39.
29. **Datos anómalos.** En ocasiones, hay muestras que contienen “observaciones anómalas”, es decir, observaciones que están muy alejadas del cuerpo central de los datos. Este tipo de observaciones se pueden atribuir a varias causas: el dato se observa, se registra o se introduce incorrectamente; el dato proviene de una población distinta; el dato es correcto pero representa un suceso poco común, etc. Veamos un método para detectar posibles datos anómalos en una muestra utilizando el rango intercuartílico.

Primero se calculan  $Q_1$  y  $Q_3$  que determinan el rango intercuartílico  $R_Q$ . A partir de ellos se obtienen los valores  $I_I = Q_1 - 1'5 \cdot R_Q$  e  $I_S = Q_3 + 1'5 \cdot R_Q$  denominados cotas interiores inferior y superior. Estas cotas se localizan a una distancia de  $1'5 \cdot R_Q$  por debajo de  $Q_1$  en el caso de  $I_I$  y por encima de  $Q_3$  en el caso de  $I_S$ . Por último, se calculan los valores  $E_I = Q_1 - 3 \cdot R_Q$  y  $E_S = Q_3 + 3 \cdot R_Q$  denominados cotas exteriores inferior y superior. Estas cotas se localizan a una distancia de  $3 \cdot R_Q$  por debajo de  $Q_1$  en el caso de  $E_I$  y por encima de  $Q_3$  en el caso de  $E_S$ . Todo esto queda representado en la figura 1.20.

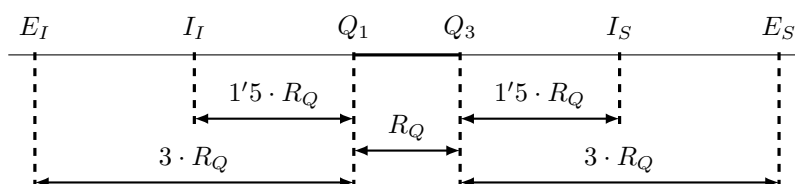


Figura 1.20: Intervalo para detectar datos anómalos

Ahora, si los datos caen entre las cotas interiores y exteriores se denominan “posibles valores fuera de intervalo”. Si los datos caen fuera de las cotas exteriores se denominan “valores fuera del intervalo muy probables”.

Detectar los posibles datos anómalos de la siguiente muestra del tiempo (en segundos) de ejecución de 25 trabajos, en un ordenador.

1'17	1'61	1'16	1'38	3'53	1'23	3'76	1'94	0'96	4'75	0'15	2'41	0'71	0'02	1'59
0'19	0'82	0'47	2'16	2'01	0'92	0'75	2'59	3'07	1'40					



## 1.8. Anexo I: Comandos de R

Comando	Descripción
<code>x=c(1,2,3,4,5);x</code>	Introduce y muestra los datos en forma de vector
<code>n=c(2,4,6,8,9);n</code>	Introduce y muestra los datos en forma de vector
<code>x=edit(x)</code>	Edita una variable ya definida
<code>length(x)</code>	Tamaño del vector de datos
<code>ls.str()</code>	Listar objetos
<code>ls()</code>	Listar objetos
<code>rm(x)</code>	Borra el objeto x
<b>Medidas de posición</b>	
<code>mean(x)</code>	Media aritmética
<code>median(x)</code>	Mediana (Me)
<code>max(x);min(x)</code>	Máximo y Mínimo
<code>quantile(x,0.25)</code>	Cuantiles
<code>summary(x)</code>	Min,Q1,Me,Media,Q3,Max
<b>Medidas de dispersión</b>	
<code>range(x)</code>	Rango = Min,Max
<code>IQR(x)</code>	Rango intercuartílico
<code>var(x)</code>	Varianza Muestral o Cuasivarianza ( $s^2$ )
<code>sd(x)</code>	Desviación estándar muestral o CuasiDesviación típica ( $s$ )
<b>Medidas de la forma</b>	
<code>library(fBasics)</code>	Cargar el paquete fBasics
<code>skewness(x)</code>	Coefficiente de asimetría
<code>kurtosis(x)</code>	Coefficiente de apuntamiento $g_2$
<b>Representaciones gráficas</b>	
<code>barplot(table(x))</code>	Diagrama de rectángulos de frecuencias absolutas
<code>barplot(table(x)/length(x))</code>	Diagrama de rectángulos de frecuencias relativas
<code>plot(table(x))</code>	Diagrama de barras de frecuencias absolutas
<code>pie(table(x))</code>	Diagrama de sectores
<code>hist(x)</code>	Histograma de frec. absolutas
<code>hist(x,freq=F)</code>	Histograma de frec. relativas
<code>hist(x,breaks=10)</code>	Histograma con 10 puntos de ruptura
<code>hist(x,10)</code>	Histograma con 10 puntos de ruptura
<code>hist(x,breaks=c(1,1.5,3,max(x)))</code>	Histograma con los puntos de ruptura
<code>boxplot(x)</code>	Diagrama de caja:
<code>boxplot(x,horizontal=TRUE)</code>	min,(Q1-1.5*IQR),Q1,Med,Q3,(Q3+1.5*IQR),max
<code>plot(x,n)</code>	Gráfico de dispersión
<b>Datos categóricos</b>	
<code>y=c("Si","No","Si","NS/NC","No","Si")</code>	Introduce los datos
<code>table(y)</code>	Genera la tabla de frecuencias absolutas
<code>barplot(table(y))</code>	Diagrama de rectángulos de frecuencias absolutas
<code>barplot(table(y)/length(y))</code>	Diagrama de rectángulos de frecuencias relativas
<code>plot(table(y))</code>	Diagrama de barras de frecuencias absolutas
<code>pie(table(y))</code>	Diagrama de sectores

### Definir y calcular otras medidas

---

Rango

```
rango = function(x) max(x)-min(x)
rango(x)
```

Varianza poblacional ( $\sigma^2$ )

```
varp = function(x) var(x)*(length(x)-1)/length(x)
varp = function(x) sum((x-mean(x))^2)/length(x)
varp(x)
```

Desviación Típica Poblacional ( $\sigma$ )

```
sdp = function(x) sqrt(var(x)*(length(x)-1)/length(x))
sdp = function(x) sqrt(sum((x-mean(x))^2)/length(x))
sdp = function(x) sqrt(varp(x))
sdp(x)
```

Variable tipificada

```
tipifica = function(x) (x-mean(x))/sqrt(var(x)*(length(x)-1)/length(x))
tipifica = function(x) (x-mean(x))/sdp(x)
tipifica(x)
```

Coefficiente de variación

```
CV = function(x) sqrt(var(x)*(length(x)-1)/length(x))/abs(mean(x))
CV(x)
```

Momentos generales, centrales y ordinarios

```
momento = function(x,c,r) sum((x-c)^r)/length(x)
momento(x,mean(x),2)
cmomento = function(x,r) sum((x-mean(x))^r)/length(x)
cmomento(x,2)
omomento = function(x,r) sum((x)^r)/length(x)
omomento(x,1)
```

### Tratamiento de datos tabulados

---

Tabla de frecuencias absolutas y relativas

```
table(x)
table(x)/length(x)
```

Media ponderada

```
weighted.mean(x,n)
```

Momentos generales, centrales y ordinarios

```
fmomento = function(x,n,c,r) sum((x-c)^r*n)/sum(n)
fcmomento = function(x,n,r) sum((x-weighted.mean(x,n))^r*n)/sum(n)
fcmomento(x,n,2)
fomomento = function(x,n,r) sum((x^r*n))/sum(n)
fomomento(x,n,1)
```

Varianza poblaciones ( $\sigma^2$ )

```
fvarp = function(x,n) sum((x-weighted.mean(x,n))^2*n)/sum(f)
fvarp = function(x,n) fcmomento(x,n,2)
fvarp(x,n)
```

# Apuntes de ESTADÍSTICA

## Regresión y correlación



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento a los profesores Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.*

## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Capítulo 2

# Regresión y correlación

En el capítulo anterior se proporcionan las herramientas para describir una población en función de los datos de una variable obtenidos en una muestra. En este capítulo se considera la observación conjunta de dos caracteres en el individuo. Los pares de datos obtenidos constituyen muestras de una variable estadística bidimensional. El objetivo del tema será describir la población a partir de las variables estudiadas, establecer la posible relación entre ellas, determinar un modelo matemático que represente dicha relación y poder cuantificar la bondad de dicho modelo.

### 2.1. Distribuciones bidimensionales

Para el estudio conjunto de dos caracteres de la población, consideraremos la variable  $X$  que presenta las modalidades  $x_1, x_2, \dots$  y la variable  $Y$  con modalidades  $y_1, y_2, \dots$ . Los distintos valores que podemos obtener al observar conjuntamente las dos variables constituyen una muestra de la variable bidimensional  $(X, Y)$ . La distribución de frecuencias de esta nueva variable viene determinada por las parejas  $(x_i, y_j)$  de valores observados junto a sus correspondientes frecuencias absolutas  $(n_{ij})$ , que indican el número de veces que se repiten dichas parejas. Análogamente al caso unidimensional, se pueden definir las frecuencias relativas  $(f_{ij})$  que indican la proporción de veces que se repite la pareja de valores  $(x_i, y_j)$  sobre el total de datos de la muestra. Si  $N$  es el tamaño de la muestra, entonces  $f_{ij}$  se calcula mediante el cociente  $n_{ij}/N$ .

Ahora mostramos distintas formas de representar la distribución de frecuencias haciendo uso de las tablas y las gráficas. La naturaleza de las variables y el tamaño o la variabilidad de los datos de la muestra determinará el procedimiento más adecuado para su representación.

#### 2.1.1. Representación tabular

La distribución de frecuencias de una variable bidimensional se puede mostrar en forma de tabla que contiene los distintos pares de valores la variable junto a sus frecuencias. Independientemente de la naturaleza discreta o continua de las variables, consideramos tres casos en función de la cantidad y variedad de datos de la muestra.

Cuando el número de observaciones es pequeño, los valores de las variables se pueden presen-

tar en forma de *tabla simple* con dos filas (o dos columnas) conteniendo las parejas de valores. Por ejemplo, la tabla

variable $X$	$x_1$	$x_2$	$\dots$	$x_N$
variable $Y$	$y_1$	$y_2$	$\dots$	$y_N$

representa los datos de la muestra  $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$  de la variable  $(X, Y)$ .

**Ejemplo 2.1** Se prueban cinco trozos experimentales de un material aislante bajo diferentes presiones. A continuación se presentan los valores ( $P$ ) de presión (en  $\text{Kg/cm}^2$ ) y las magnitudes ( $C$ ) de compresión resultantes (en  $\text{mm}$ ):  $(1,1)$ ,  $(2,1)$ ,  $(3,2)$ ,  $(4,2)$  y  $(5,4)$ . Representar la distribución de frecuencias.

Se construye una tabla simple de valores

$P$	1	2	3	4	5
$C$	1	1	2	2	4

con los pares de datos de la muestra. □

Cuando el número de observaciones es grande, pero corresponden a pocas parejas (modalidades) distintas, los valores de las variables se pueden presentar en forma de *tabla simple* con tres filas o columnas conteniendo las parejas de valores y sus frecuencias correspondientes. Por ejemplo, la tabla de la figura 2.1 representa la distribución de frecuencias de los datos de una muestra de tamaño  $N$  que contiene  $k$  tipos de pares de datos  $(x_i, y_i)$  observados  $n_i$  veces cada uno, con  $i = 1, 2, \dots, k$ .

variable $X$	variable $Y$	frecuencia absoluta	frecuencia relativa
$x_1$	$y_1$	$n_1$	$f_1$
$x_2$	$y_2$	$n_2$	$f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i$	$y_i$	$n_i$	$f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_k$	$y_k$	$n_k$	$f_k$
		$N$	1

Figura 2.1: Tabla estadística de frecuencias

**Ejemplo 2.2** Una empresa de software somete a sus programas a determinados controles para depurar errores durante su desarrollo. El número de controles efectuados disminuye los posibles errores finales pero incrementa los costes de producción. Para determinar la influencia de estas variables se observan conjuntamente el número de controles  $C$  efectuados a un software y el número de errores graves detectados  $D$  al finalizar su desarrollo obteniéndose la muestra:  $(0,0)$ ,  $(0,1)$ ,  $(1,1)$ ,  $(0,1)$ ,  $(1,1)$ ,  $(0,1)$ ,  $(0,1)$ ,  $(1,1)$ ,  $(1,0)$ ,  $(1,0)$ ,  $(1,1)$ ,  $(1,1)$ ,  $(1,1)$ ,  $(0,0)$ ,  $(1,0)$ ,  $(1,0)$ ,  $(2,1)$ ,  $(1,1)$ ,  $(1,1)$ ,  $(2,1)$ . Utilizar una tabla estadística para representar la distribución de frecuencias.

Se ordenan los valores de la muestra y se agrupan los que corresponden al mismo par de modalidades. Después, se construye una tabla donde se representan los distintos pares de valores junto a su frecuencia absoluta y relativa.

$C$	$D$	$n_i$	$f_i$
0	0	2	0'1
0	1	4	0'2
1	0	4	0'2
1	1	8	0'4
2	1	2	0'1
		20	

Por ejemplo, la fila 4 indica que hemos observado 8 veces (frec. absoluta) que con 1 control ( $C$ ) se detecta 1 error ( $D$ ), y esto supone el 40 % (frec. relativa) de los casos observados.  $\square$

Cuando hay un gran número de observaciones y de modalidades distintas, los valores de las variables se disponen en una *tabla de doble entrada*, donde los valores de cruce de cada fila y columna representan la frecuencia de la correspondiente pareja de valores. En la tabla de la figura 2.2, consideramos la variable  $X$  con  $k$  modalidades  $x_1, x_2, \dots, x_k$  y la variable  $Y$  con  $p$  modalidades  $y_1, y_2, \dots, y_p$ .

$X \backslash Y$	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_p$	
$x_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1j}$	$\dots$	$n_{1p}$	$n_{1\cdot}$
$x_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2j}$	$\dots$	$n_{2p}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_i$	$n_{i1}$	$n_{i2}$	$\dots$	$n_{ij}$	$\dots$	$n_{ip}$	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_k$	$n_{k1}$	$n_{k2}$	$\dots$	$n_{kj}$	$\dots$	$n_{kp}$	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot p}$	$N$

Figura 2.2: Tabla de doble entrada

Las distintas modalidades de las variables  $X$  e  $Y$  se ordenan en los márgenes izquierdo y superior respectivamente. La frecuencia absoluta del par  $(x_i, y_j)$  se denomina  $n_{ij}$  y se sitúa en la intersección de la fila y columna correspondiente. También se puede construir otra tabla estadística a partir de las frecuencias relativas, sin más que dividir por  $N$  las frecuencias absolutas de tal manera que

$$f_{ij} = \frac{n_{ij}}{N} \quad \text{siendo} \quad N = \sum_{i=1}^k \sum_{j=1}^p n_{ij}$$

En el margen derecho de la tabla se sitúan las frecuencias  $(n_{i\cdot})$  de los valores de la variable  $X$ , que se calculan sumando por filas. En el margen inferior se localizan las frecuencias  $(n_{\cdot j})$  de los valores de la variable  $Y$ , que se calculan sumando por columnas. Como veremos, los valores de las variables y sus frecuencias, representadas al margen, determinan las *distribuciones marginales*; mientras que los valores en interior de la tabla constituyen la denominada *distribución conjunta*.

**Ejemplo 2.3** Representar en tablas de doble entrada las distribuciones de frecuencias absolutas y relativas para los datos del ejemplo 2.2 de la página 54.

A partir de los datos de la muestra, o de la tabla de frecuencias, construimos la tabla de doble entrada, situando en el margen izquierdo y superior las distintas modalidades de las variables  $X$  e  $Y$  respectivamente, y en el interior de la tabla, se escribe las frecuencias para cada par de valores.

$C$	$D$	$n_i$	$f_i$		$n_{ij}$	0	1	$D$		$f_{ij}$	0	1	$D$
0	0	2	0'1		0	2	4	6		0	0'1	0'2	0'3
0	1	4	0'2		1	4	8	12		1	0'2	0'4	0'6
1	0	4	0'2	$\Rightarrow$	2	0	2	2		2	0	0'1	0'1
1	1	8	0'4		$C$	6	14	20		$C$	0'3	0'7	1
2	1	2	0'1										
		20											

Observe la tabla estadística (izquierda), de la que se derivan las dos tablas de doble entrada, una para las frecuencias absolutas (centro) y otra para las frecuencias relativas (derecha).  $\square$

Este tipo de representación en forma de tabla de doble entrada también se utiliza si estamos interesados en agrupar los datos en intervalos. En este caso, recuperamos los conceptos de clase, amplitud y marca, introducidos en el tema anterior.

**Ejemplo 2.4** Organizar los siguientes datos de la variable  $(X, Y)$  en una tabla de doble entrada:  $(1'72, 63)$ ,  $(1'70, 75)$ ,  $(1'70, 68)$ ,  $(1'68, 70)$ ,  $(1'75, 74)$ ,  $(1'69, 72)$ ,  $(1'71, 67)$ ,  $(1'69, 69)$ ,  $(1'67, 70)$ ,  $(1'74, 74)$ ,  $(1'76, 71)$ ,  $(1'70, 70)$ ,  $(1'69, 66)$ ,  $(1'66, 60)$ ,  $(1'78, 74)$ ,  $(1'74, 69)$ ,  $(1'70, 65)$ ,  $(1'69, 71)$ ,  $(1'71, 73)$ ,  $(1'78, 69)$

Agrupando los valores de las variables  $X$  e  $Y$  en intervalos de amplitud 5 construimos la tabla de doble entrada

$X \backslash Y$	[60, 65]	(65, 70]	(70, 75]	
$(1'65, 1'70]$	2	6	3	11
$(1'70, 1'75]$	1	2	3	6
$(1'75, 1'80]$	0	1	2	3
	3	9	8	20

que contiene las frecuencias absolutas de los intervalos correspondientes.  $\square$

Como hemos comentado, las tablas simples se utilizan para representar distribuciones de frecuencias con muchos datos de pocas modalidades distintas. Por el contrario, las tablas de doble entrada resultan más apropiadas para representar distribuciones de frecuencias con muchos datos pertenecientes a un gran número de modalidades distintas. Sin embargo, en cualquier caso podemos utilizar indistintamente un tipo u otro de representación tabular. Así, en los ejemplos 2.2 de la página 54 y 2.3 de la página 55 hemos representado la misma distribución de frecuencias utilizando los dos tipos de tablas. Es importante saberlas utilizar indistintamente y construir una de ellas a partir de la otra.



### 2.1.2. Representaciones gráficas

La representación gráfica constituye una forma ordenada de presentar la distribución de frecuencias. Las representaciones gráficas más importantes para las distribuciones bidimensionales de caracteres cuantitativos son el diagrama de dispersión, el diagrama de frecuencias y el estereograma.

**Diagrama de Dispersión.** Consiste en la representación de los distintos pares de valores sobre unos ejes cartesianos. De esta forma, cada par viene representado por un punto del plano  $XY$  que forman la llamada *nube de puntos*. La frecuencia de cada par de puntos puede representarse utilizando distintos tamaños de puntos.

En la figura 2.3 se muestran dos diagramas de dispersión. El primero representa la nube de puntos correspondiente a los datos (sin agrupar) del ejemplo 2.4 de la página 56 y el segundo representa los datos del ejemplo 2.2 de la página 54 donde el tamaño de los puntos es proporcional a su frecuencia.

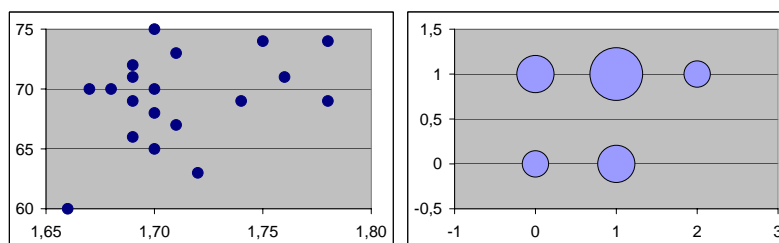


Figura 2.3: Diagramas de dispersión

**Diagrama de frecuencias.** Este tipo de representación está indicado para el caso discreto y es análogo a los diagramas de barras o puntos en el caso unidimensional. Consiste en una representación en tres dimensiones donde el plano base corresponde a los valores de las variables y la altura representa las frecuencias. El resultado es una serie de barras verticales apoyadas en los puntos del plano  $XY$  correspondientes a los valores  $(x_i, y_j)$  y cuya altura representa la frecuencia absoluta  $(n_{ij})$  o relativa  $(f_{ij})$  del par.

Este tipo de representación también se puede utilizar para representar distribuciones cuando las variables son cualitativas. En la figura 2.4 se representa mediante un diagrama de frecuencias los datos del ejemplo 2.2 de la página 54.

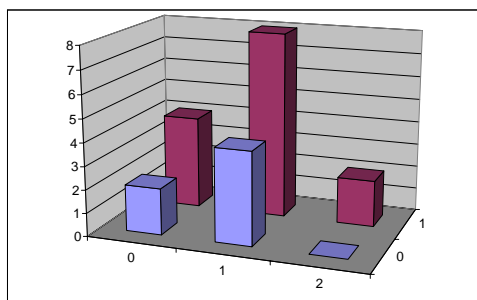


Figura 2.4: Diagrama de frecuencias

**Estereograma.** Se utiliza para representar aquellas distribuciones donde los datos se agrupan en intervalos y equivale al histograma para una variable. Se realiza análogamente al diagrama de frecuencias utilizando paralelepípedos, en vez de barras o puntos, cuya base son las regiones del plano correspondientes a los intervalos. En este caso, el volumen representa la frecuencia absoluta o relativa.

En la gráfica de la izquierda de la figura 2.3 de la página 57 se representaban los datos del ejemplo 2.4 de la página 56, en forma de nube de puntos. Ahora, en la figura 2.5 se muestran esos mismos datos, pero agrupados en intervalos.

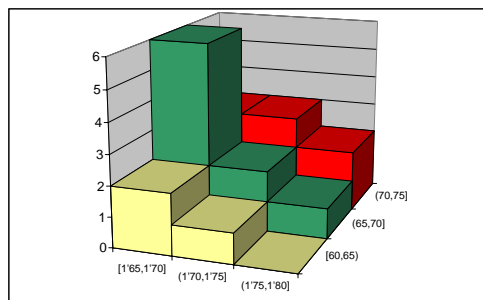


Figura 2.5: Estereograma

### 2.1.3. Distribuciones Marginales

La distribución de frecuencias bidimensional contiene la información conjunta de dos variables. Sin embargo, podemos estar interesados en estudiar una variable de manera aislada, sin considerar su relación con la otra. En este caso, debemos “separar” la información relativa a cada variable.

A partir de las distribuciones bidimensionales definimos las distribuciones marginales que son las distribuciones unidimensionales correspondientes a uno de los caracteres sin considerar el otro. Para obtenerlas basta con prescindir de la información de una de las variables eliminando los datos correspondientes.

**Ejemplo 2.5** Calcular la distribución marginal de la variable  $C$  (número de controles efectuados a un software) del ejemplo 2.2 de la página 54 a partir de su tabla estadística de frecuencias.

Si eliminamos la columna correspondiente a la variable  $D$  y agrupamos las modalidades que sean iguales,

$C$	$D$	$n_i$	$f_i$		$C$	$n_i$	$f_i$		$C$	$n_i$	$f_i$
0	0	2	0'1		0	2	0'1		0	6	0'3
0	1	4	0'2		0	4	0'2		1	12	0'6
1	0	4	0'2	$\Rightarrow$	1	4	0'2	$\Rightarrow$	2	2	0'1
1	1	8	0'4		1	8	0'4			20	
2	1	2	0'1		2	2	0'1				
		20				20					

el resultado es la distribución de frecuencias unidimensional correspondiente a la variable  $C$   $\square$

Cuando la distribución se representa en una tabla de doble entrada, las distribuciones marginales aparecen “en el margen” de la tabla que contiene la suma por filas (o columnas) de los valores conjuntos de las variables. Para obtener estas distribuciones marginales sólo hay que prescindir de los valores de la variable en el interior de la tabla.

En la tabla de doble entrada de la figura 2.2 de la página 55, la distribución marginal de la variable  $X$  aparecen en el margen derecho de la tabla y cuenta con las modalidades  $x_1, x_2, \dots, x_k$  cuyas frecuencias absolutas  $(n_{1\cdot}, n_{2\cdot}, \dots, n_{k\cdot})$  y relativas  $(f_{1\cdot}, f_{2\cdot}, \dots, f_{k\cdot})$  correspondientes a cada modalidad se calculan sumando por filas:

$$n_{i\cdot} = \sum_{j=1}^p n_{ij} \quad \text{y} \quad f_{i\cdot} = \frac{n_{i\cdot}}{N} \quad \text{son las frecuencias marginales del valor } x_i \text{ de la variable } X.$$

Análogamente, en el margen inferior se observa la marginal de la variable  $Y$  que toma los valores  $y_1, y_2, \dots, y_p$  cuyas frecuencias absolutas  $(n_{\cdot 1}, n_{\cdot 2}, \dots, n_{\cdot p})$  y relativas  $(f_{\cdot 1}, f_{\cdot 2}, \dots, f_{\cdot p})$  correspondientes a cada modalidad se calculan sumando por columnas:

$$n_{\cdot j} = \sum_{i=1}^k n_{ij} \quad \text{y} \quad f_{\cdot j} = \frac{n_{\cdot j}}{N} \quad \text{son las frecuencias marginales del valor } y_j \text{ de la variable } Y.$$

**Ejemplo 2.6** Calcular las distribuciones marginales de las variables  $C$  (número de controles efectuados a un software) y  $D$  (número de errores graves detectados), a partir de la tabla de doble entrada del ejemplo 2.3 de la página 55.

Para calcular la distribución marginal de la variable  $C$  se eliminan las dos columnas interiores de la tabla y permanece la columna de la derecha que contiene la suma por filas de los valores de las columnas eliminadas.

$$\begin{array}{c|cc|c} n_{ij} & 0 & 1 & D \\ \hline 0 & 2 & 4 & 6 \\ 1 & 4 & 8 & 12 \\ 2 & 0 & 2 & 2 \\ \hline C & 6 & 14 & 20 \end{array} \quad \Rightarrow \quad \begin{array}{c|c|c} C & n_i & f_i \\ \hline 0 & 6 & 0'3 \\ 1 & 12 & 0'6 \\ 2 & 2 & 0'1 \\ \hline & 20 & \end{array}$$

y para calcular la distribución marginal de la variable  $D$  se eliminan las tres filas interiores de la tabla y permanece la fila inferior que contiene la suma por columnas de los valores de las filas eliminadas.

$$\begin{array}{c|cc|c} n_{ij} & 0 & 1 & D \\ \hline 0 & 2 & 4 & 6 \\ 1 & 4 & 8 & 12 \\ 2 & 0 & 2 & 2 \\ \hline C & 6 & 14 & 20 \end{array} \quad \Rightarrow \quad \begin{array}{c|c|c} D & n_i & f_i \\ \hline 0 & 6 & 0'3 \\ 1 & 14 & 0'7 \\ \hline & 20 & \end{array}$$

En ambos casos, hemos añadido una columna correspondiente a las frecuencias relativas.  $\square$

#### 2.1.4. Distribuciones Condicionadas

Al igual que las marginales, las distribuciones condicionadas son también distribuciones unidimensionales. Surgen al considerar sólo aquellos valores de la muestra que presentan una determinada modalidad en una de las variables.

Se llama distribución condicionada del carácter  $X$ , respecto a la clase  $j$  del carácter  $Y$ , y se denota  $X/y_j$ , a la distribución unidimensional de la variable  $X$ , cuando sólo se consideran los individuos de la clase  $j$  de  $Y$ .

En la tabla de doble entrada de la figura 2.2 de la página 55, la distribución condicionada del carácter  $X$ , respecto a la clase  $j$  del carácter  $Y$  corresponde a la columna  $j$ -ésima y cuenta con las modalidades  $x_1, x_2, \dots, x_k$  cuyas frecuencias absolutas  $(n_1^j, n_2^j, \dots, n_k^j)$  aparecen directamente en la columna  $j$ -ésima  $(n_{1j}, n_{2j}, \dots, n_{kj})$  de la tabla. Las frecuencias relativas  $(f_1^j, f_2^j, \dots, f_k^j)$  correspondientes a cada modalidad se calculan dividiendo las absolutas entre el total de valores de  $X$  con la modalidad  $j$ , es decir,  $n_{.j}$ . Por tanto

$$n_i^j = n_{ij} \quad \text{y} \quad f_i^j = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} \quad i = 1, 2, \dots, k$$

Análogamente se puede definir la distribución condicionada del carácter  $Y$ , respecto a la modalidad  $i$  de  $X$ . Esta distribución considera los valores  $y_j$  con frecuencias:

$$n_j^i = n_{ij} \quad \text{y} \quad f_j^i = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}} \quad j = 1, 2, \dots, p$$

**Ejemplo 2.7** Determinar la distribución condicionada del carácter  $C$  respecto de la modalidad 1 del carácter  $D$ , a partir de la tabla de doble entrada del ejemplo 2.3 de la página 55.

Para determinar esta distribución condicionada, seleccionamos la segunda columna correspondiente a todos los valores de la variable  $C$  que corresponden al valor 1 de la variable  $D$ .

$n_{ij}$	0	1	$D$		$C$	$n_i$	$f_i$
0	2	4	6	$\Rightarrow$	0	4	4/14
1	4	8	12		1	8	8/14
2	0	2	2		2	2	2/14
$C$	6	14	20			14	

Las modalidades de  $C$ , junto a sus frecuencias correspondientes, en la columna seleccionada, constituyen la distribución de frecuencias del carácter  $C$ , respecto a la modalidad 1 del carácter  $D$ .  $\square$

### 2.1.5. Distribuciones conjuntas: Momentos mixtos

En este apartado vamos a presentar algunas características de las distribuciones conjuntas y su relación con las distribuciones marginales y condicionadas.

En las tablas de doble entrada, la distribución conjunta de frecuencias se puede obtener a partir de las distribuciones de frecuencias marginales y condicionadas según las relaciones

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{i.}} \cdot \frac{n_{i.}}{N} = f_{ij}^i \cdot f_{i.} \quad \text{o bien} \quad f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{.j}} \cdot \frac{n_{.j}}{N} = f_{ij}^j \cdot f_{.j}$$

A continuación vamos a definir los momentos de una distribución conjunta que se utilizan para determinar medidas de relación entre las variables. Como veremos, algunos casos particulares corresponden a las medias y varianzas de las distribuciones marginales.

**Momentos mixtos.** Se define el momento de orden  $(r, s)$  respecto al punto  $(a, b)$  como

$$M_{rs}(a, b) = \sum_{i=1}^N (x_i - a)^r \cdot (y_i - b)^s \cdot f_i \quad \text{o bien} \quad M_{rs}(a, b) = \sum_{i=1}^k \sum_{j=1}^p (x_i - a)^r \cdot (y_j - b)^s \cdot f_{ij}$$

según consideremos la distribución de frecuencias correspondiente a una tabla simple (figura 2.1 de la página 54) o a una tabla de doble entrada (figura 2.2 de la página 55).

**Ejemplo 2.8** Calcular el momento de orden  $(2, 3)$  respecto al punto  $(0, 1)$  para la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$\begin{aligned} M_{23}(0, 1) &= \frac{2(0-0)^2(0-1)^3 + 4(0-0)^2(1-1)^3 + 4(1-0)^2(0-1)^3 + \dots}{20} = \\ &= \frac{0 + 0 - 4 + 0 + 0}{20} = -\frac{4}{20} = -0'2 \end{aligned}$$

□

Resultan de especial interés los siguientes dos casos particulares:

**Momentos mixtos ordinarios.** Si  $a = b = 0$  entonces el momento de orden  $(r, s)$  recibe el nombre de momento ordinario y se denota por

$$m_{rs} = \sum_{i=1}^N x_i^r \cdot y_i^s \cdot f_i \quad \text{o bien} \quad m_{rs} = \sum_{i=1}^k \sum_{j=1}^p x_i^r \cdot y_j^s \cdot f_{ij}$$

**Ejemplo 2.9** Calcular el momento ordinario de orden  $(2, 3)$  para la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$m_{23} = \frac{2 \cdot 0^2 \cdot 0^3 + 4 \cdot 0^2 \cdot 1^3 + 4 \cdot 1^2 \cdot 0^3 + 8 \cdot 1^2 \cdot 1^3 + 2 \cdot 2^2 \cdot 1^3}{20} = \frac{0 + 0 + 0 + 8 + 8}{20} = \frac{16}{20} = 0'8$$

□

**Momentos mixtos centrales.** Si  $a = \bar{x}$  y  $b = \bar{y}$  entonces el momento de orden  $(r, s)$  recibe el nombre de momento central y se denota por

$$\mu_{rs} = \sum_{i=1}^N (x_i - \bar{x})^r \cdot (y_i - \bar{y})^s \cdot f_i \quad \text{o bien} \quad \mu_{rs} = \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s \cdot f_{ij}$$

**Ejemplo 2.10** Calcular el momento central de orden  $(2, 3)$  para la distribución de frecuencias del ejemplo 2.2 de la página 54.

Para calcular el momento central es necesario disponer de la media de las distribuciones marginales:

$$\bar{c} = \frac{0 \cdot 6 + 1 \cdot 12 + 2 \cdot 2}{20} = \frac{16}{20} = 0'8 \quad \text{y} \quad \bar{d} = \frac{0 \cdot 6 + 1 \cdot 14}{20} = \frac{14}{20} = 0'7$$

Después aplicamos la fórmula del momento central

$$\begin{aligned}\mu_{23} &= \frac{2(0-0'8)^2(0-0'7)^3 + 4(0-0'8)^2(1-0'7)^3 + 4(1-0'8)^2(0-0'7)^3 + \dots}{20} = \\ &= \frac{-0'43907 + 0'06912 - 0'01372 + 0'00864 + 0'07776}{20} = -\frac{0'29727}{20} = -0'0148635\end{aligned}$$

□

Para los momentos centrales y ordinarios, destacamos las siguientes propiedades que muestran su relación con algunas medidas de posición (media) y dispersión (varianza y desviación típica) para las distribuciones marginales:

$$\begin{array}{lll}m_{00} = 1 & m_{10} = \bar{x} & m_{01} = \bar{y} \\ \mu_{00} = 1 & \mu_{10} = 0 & \mu_{01} = 0\end{array}$$

Como en el caso unidimensional, se puede establecer una relación entre los momentos centrales y ordinarios. Destacamos las siguientes propiedades que establecen fórmulas alternativas para calcular determinadas medidas:

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

**Medias marginales.** La media marginal de la variable  $X$  corresponde a la medida de tendencia central *media aritmética* de la distribución marginal de la variable  $X$ . Análogamente se define la media marginal de la variable  $Y$  y ambas se calculan a partir de los momentos ordinarios:

$$\bar{x} = m_{10} = \sum_{i=1}^k x_i \cdot f_i \qquad \bar{y} = m_{01} = \sum_{j=1}^p y_j \cdot f_j$$

El punto  $(\bar{x}, \bar{y})$  es el *punto medio* o *centro de gravedad* de la distribución.

**Ejemplo 2.11** Calcular el centro de gravedad para la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$\bar{c} = \frac{0 \cdot 6 + 1 \cdot 12 + 2 \cdot 2}{20} = \frac{16}{20} = 0'8 \qquad \text{y} \qquad \bar{d} = \frac{0 \cdot 6 + 1 \cdot 14}{20} = \frac{14}{20} = 0'7$$

Por lo tanto, el centro de gravedad de la distribución es el punto  $(0'8, 0'7)$ . □

**Varianzas marginales.** La varianza marginal de la variable  $X$  corresponde a la medida de dispersión *varianza* de la distribución marginal de la variable  $X$ . Análogamente se define la varianza marginal de la variable  $Y$  y ambas se calculan a partir de los momentos centrales

$$\sigma_x^2 = V(X) = \mu_{20} = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot f_i \qquad \sigma_y^2 = V(Y) = \mu_{02} = \sum_{j=1}^p (y_j - \bar{y})^2 \cdot f_j$$

o de los momentos ordinarios, aplicando la propiedad que los relaciona.

Las desviaciones típicas marginales se definen como la raíz cuadrada positiva de las varianzas marginales correspondientes.

**Ejemplo 2.12** Calcular la desviación típica marginal de la variable  $C$  en la distribución de frecuencias del ejemplo 2.2 de la página 54.

$$\sigma_c^2 = \frac{6 \cdot 0^2 + 12 \cdot 1^2 + 2 \cdot 2^2}{20} - 0'8^2 = 1 - 0'64 = 0'36$$

□

**Covarianza.** La covarianza o varianza conjunta es el momento central de orden (1,1)

$$\mu_{11} = \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x}) \cdot (y_j - \bar{y}) \cdot f_{ij} \quad \text{o bien} \quad \mu_{11} = \sum_{i=1}^k (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot f_i$$

y se denota por  $Cov(X, Y)$ , o bien por  $\sigma_{xy}$ . Las propiedades que relaciona los momentos centrales y ordinarios nos permiten obtener una nueva fórmula para calcular la covarianza: *la media de los productos menos el producto de las medias*.

$$Cov(X, Y) = \sum_{i=1}^k \sum_{j=1}^p x_i y_j f_{ij} - \bar{x} \bar{y}$$

La covarianza es una medida de la variación conjunta de las variables y forma parte en la definición de los coeficientes que miden la relación entre esas variables. La covarianza se basa en las unidades de medida originales de las dos variables  $X$  e  $Y$ . Por lo tanto, no es posible comparar la covarianza de distintas distribuciones conjuntas. Si dividimos su fórmula por el producto de las desviaciones típicas de las variables  $X$  y  $Y$  obtenemos el coeficiente de correlación lineal de Pearson

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y}$$

que es una medida adimensional que permite comparar covarianzas de distintas distribuciones conjuntas.

**Ejemplo 2.13** Calcular la covarianza y el coeficiente de correlación lineal de Pearson para la distribución de frecuencias del ejemplo 2.2 de la página 54.

En primer lugar, calculamos la covarianza:

$$Cov(C, D) = \frac{2 \cdot 0 \cdot 0 + 4 \cdot 0 \cdot 1 + 4 \cdot 1 \cdot 0 + 8 \cdot 1 \cdot 1 + 2 \cdot 2 \cdot 1}{20} - 0'8 \cdot 0'7 = 0'6 - 0'56 = 0'04$$

Después calculamos las desviaciones típicas marginales

$$\sigma_c^2 = 0'36 \quad \text{y} \quad \sigma_d^2 = 0'21$$

y finalmente aplicamos la fórmula para obtener el coeficiente de correlación

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{0'04}{\sqrt{0'36} \cdot \sqrt{0'21}} \approx 0'145$$

□

## 2.2. Regresión y correlación

En esta sección se introducen algunas técnicas estadísticas que nos permitirán estudiar la relación entre dos variables de una misma población o muestra. El interés se centrará en aquellos casos donde intuimos que existe una relación entre las variables, pero no somos capaces de encontrar una función matemática que describa esta relación. Por ejemplo, intuimos que el peso y la altura de un individuo están relacionados, sin embargo, no existe ninguna fórmula matemática que nos permita determinar el peso exacto de una persona en función de su altura.

El objetivo es encontrar un modelo o función matemática que recoja, de la manera más acertada, la relación entre dos variables de este tipo. Además, cuando hayamos determinado el modelo, será necesario proporcionar alguna medida de la bondad de dicho modelo. Por tanto, hay que resolver dos problemas:

1. Encontrar un modelo que permita relacionar dos variables
2. Determinar el grado de relación entre esas dos variables.

**La regresión** estudia la naturaleza estadística de la relación entre dos variables y nos proporciona un modelo de dicha relación. El modelo consiste en una función matemática cuya gráfica se aproxima a los datos observados. La función encontrada permitirá obtener los valores aproximados de una de las variables a partir de los valores prefijados de la otra variable.

**La correlación** se encarga de solucionar el segundo problema estableciendo la correspondencia en las pautas de variación de dos variables. La correlación cuantifica esta dependencia entre las variables mediante el cálculo de los coeficientes de correlación.

Veamos, en primer lugar, qué tipos de relaciones pueden existir entre las variables. En segundo lugar, presentaremos algunos métodos para obtener modelos que determinan la relación entre las variables. En tercer lugar, introduciremos medidas que permitan estudiar la bondad de esos modelos. Y, por último, presentaremos, a modo de ejemplo, algunos modelos importantes, como el modelo lineal.

### 2.2.1. Relación entre variables

El objetivo de analizar conjuntamente dos variables diferentes en una misma población o muestra es estudiar el tipo de relación que hay entre ellas. Según el grado extremo de relación existente distinguimos tres casos: Si no hay relación alguna decimos que las variables son independientes; si, por el contrario, hay una relación total decimos que las variables dependen funcionalmente; y en los casos intermedios decimos que las variables mantienen una dependencia estadística. Desde el punto de vista estadístico, este último caso es el más interesante pues permite estudiar el grado de dependencia entre las variables, proporcionando un modelo matemático que explique la relación entre ellas.

#### Independencia

Cuando no existe relación alguna entre las variables, es decir, ninguna de ellas proporciona información sobre la otra, decimos que existe una independencia entre las variables. En este



caso, se dice que las variables son *independientes* una de la otra. Por ejemplo: la velocidad de un ordenador y el grosor del papel utilizado en la impresora.

Formalmente la independencia se define así:

1. Se dice que el carácter  $X$  es independiente de  $Y$ , si todas las condicionadas de  $X$  respecto a cualquier clase de  $Y$  coinciden con la marginal de  $X$ , es decir  $f_i^j = f_i$  para todo  $j$ .
2. Análogamente se define la independencia de  $Y$  respecto a  $X$  si  $f_j^i = f_j$  para todo  $i$ .

Se deduce que si  $X$  es independiente de  $Y$ , entonces  $Y$  es independiente de  $X$  y esto ocurre si y sólo si  $f_{ij} = f_i \cdot f_j$ . En este caso, es fácil determinar, a la vista de la tabla de frecuencias, la independencia de caracteres porque las columnas son proporcionales entre sí, al igual que ocurre con las filas.

**Ejemplo 2.14** Comprobar que la siguiente tabla de frecuencias corresponde a dos variables independientes.

	$y_1$	$y_2$	$y_3$	$y_4$
$x_1$	1	3	2	4
$x_2$	3	9	6	12
$x_3$	2	6	4	8

Se puede observar que las columnas de la tabla son proporcionales: la segunda columna es tres veces la primera, la tercera es dos veces la primera y la cuarta es cuatro veces la primera.

Si representamos la distribución de frecuencias relativas en forma de tabla de doble entrada

	$y_1$	$y_2$	$y_3$	$y_4$	
$x_1$	1/60	3/60	2/60	4/60	1/6
$x_2$	3/60	9/60	6/60	12/60	3/6
$x_3$	2/60	6/60	4/60	8/60	2/6
	1/10	3/10	2/10	4/10	1

observamos que el producto de las frecuencias de las distribuciones marginales coincide con la frecuencia correspondiente de la distribución conjunta. Por ejemplo,  $f_2 \cdot f_{\cdot 3} = f_{23}$ , es decir,  $3/6 \cdot 2/10 = 6/60$ .

También se puede comprobar que las distribuciones de frecuencias condicionadas y marginal son iguales. Por ejemplo, en las siguientes tablas calculamos las distribución de frecuencias de la variable  $X$  condicionada a cualquier modalidad de la variable  $Y$  (izquierda) y comprobamos que todas coinciden y son iguales a la distribución de frecuencias marginal de la variable  $X$  (derecha).

$x_i$	$f_i^j$	$f_i^1$	$f_i^2$	$f_i^3$	$f_i^4$	$x_i$	$f_i$
$x_1$	$\frac{1}{6}$	$= \frac{1}{1+3+2}$	$= \frac{3}{3+9+6}$	$= \frac{2}{2+6+4}$	$= \frac{4}{4+12+8}$	$x_1$	$\frac{1}{6}$
$x_2$	$\frac{3}{6}$	$= \frac{3}{1+3+2}$	$= \frac{9}{3+9+6}$	$= \frac{6}{2+6+4}$	$= \frac{12}{4+12+8}$	$x_2$	$\frac{3}{6}$
$x_3$	$\frac{2}{6}$	$= \frac{2}{1+3+2}$	$= \frac{6}{3+9+6}$	$= \frac{4}{2+6+4}$	$= \frac{8}{4+12+8}$	$x_3$	$\frac{2}{6}$

Análogamente podíamos comprobar que se verifica para la variable  $Y$  calculando sus frecuencias condicionadas y marginal.  $\square$

### Dependencia funcional

En el estudio conjunto de dos variables puede ocurrir que la aparición de un determinado valor de una de las variables esté perfectamente determinado conociendo el valor de la otra para esa misma observación. En este caso, decimos que existe una dependencia funcional entre las variables y podemos establecer un modelo matemático que relaciona ambas variables.

Por ejemplo, si tomamos varias muestras de las longitudes de las circunferencias ( $L$ ) y sus radios ( $R$ ) observamos que los valores de las variables están relacionados por la fórmula:  $L = 2\pi R$ . Es decir, existe un modelo matemático que me permite calcular el valor que toma la variable  $L$  sin necesidad de observarlo, conociendo el valor correspondiente de la variable  $R$ .

A la vista de la tabla de frecuencias es fácil determinar la dependencia funcional. Si para cada modalidad  $x_i$  de  $X$  existe una única modalidad  $y_j$  de  $Y$  tal que  $n_{ij} \neq 0$ , decimos que la variable  $Y$  depende funcionalmente de la variable  $X$ . Esta relación de dependencia funcional no es recíproca, es decir, si  $X$  depende funcionalmente de  $Y$  no implica que  $Y$  dependa funcionalmente de  $X$ . Por ejemplo:  $Y = a \cdot X^2$  donde  $Y$  depende de  $X$  y no al revés.

**Ejemplo 2.15** *Comprobar que la siguiente tabla de frecuencias corresponde a dos variables que dependen funcionalmente. Determinar la dependencia y establecer el modelo matemático.*

	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$
$x_1$	0	0	3	0	0
$x_2$	0	0	0	0	1
$x_3$	0	0	2	0	0
$x_4$	4	0	0	0	0

Como se observa en la tabla, para cada modalidad  $x_i$  de la variable  $X$  existe una única modalidad  $y_j$  de la variable  $Y$  cuya frecuencia conjunta es distinta de 0. En este caso, decimos que la variable  $Y$  depende funcionalmente de la variable  $X$  y se establece el siguiente modelo matemático en forma de tabla

$X$	$x_1$	$x_2$	$x_3$	$x_4$
$Y$	$y_3$	$y_5$	$y_3$	$y_1$

que permite determinar los valores de  $Y$  en función de la observación del valor de  $X$ . □

### Dependencia estadística

La independencia y la dependencia funcional son dos casos extremos de la relación entre las variables cuando ésta no existe o es total. Generalmente, cuando se estudian conjuntamente dos variables para establecer la relación entre ambas surgen los casos intermedios.

Cuando una variable puede dar información sobre otra, pero la relación entre ambas no es determinista y por tanto no existe o no se conoce una expresión matemática que las relacione, se dice que existe una *dependencia aleatoria o estadística*. Por ejemplo, sabemos que el peso y la estatura de una persona son dos variables relacionadas y sin embargo no se puede establecer una fórmula matemática que determine, en todos los casos, el peso de una persona en función de su altura.

La dependencia estadística también suele considerarse en aquellos procesos o variables cuya relación es determinista pero resulta muy complejo su estudio. Por ejemplo, en el comportamiento atmosférico sólo intervienen fenómenos físicos perfectamente estudiables y sin embargo, su estudio es intratable cuando pretendemos establecer una predicción meteorológica. Igual ocurre con las placas tectónicas terrestres, aunque su movimiento se rige por leyes físicas, su complejidad impide la predicción exacta de un terremoto. En estos casos, se considera que las variables presentan una dependencia estadística y se estudia su relación a partir de muestras.

### 2.2.2. Regresión: Método de los mínimos cuadrados

Cuando existe una dependencia estadística entre variables, el objetivo es encontrar un modelo o función matemática que determine, de manera aproximada, la relación entre las variables

La representación de los datos obtenidos en la muestra de una variable estadística bidimensional  $(X, Y)$  sobre el plano (diagrama de dispersión) constituye una nube de puntos. Se llama *línea o curva de regresión* a la función que mejor se ajusta a esa nube de puntos.

Si todos los valores de la variable satisfacen la ecuación calculada, se dice que las variables están perfectamente correlacionadas o que hay *correlación perfecta* entre ellas. En general, como se observa en la figura 2.6, se trata de una línea ideal en torno a la cual se distribuyen los puntos de la nube.

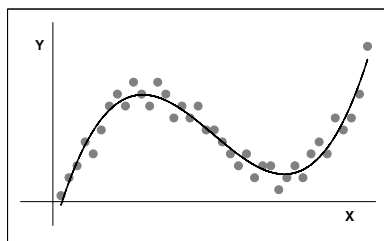


Figura 2.6: Nube de puntos y curva de regresión

En la práctica, la obtención de esta línea no es sencilla y, en general, no tiene que tener una expresión matemática en forma de ecuación. Por esta razón, la estadística se limita a calcular líneas “ideales” con expresiones matemáticas conocidas con formas rectas, parabólicas, exponenciales, logarítmicas, hiperbólicas, etc.

Cuando dispongamos de la ecuación de esta curva de regresión podemos utilizarla para estudiar las características de la relación entre las variables y predecir valores desconocidos.

El problema general de la regresión es ajustar una función o curva de ecuación conocida a la nube de puntos que representa las observaciones de una variable bidimensional  $(X, Y)$ . En primer lugar, hay que determinar qué variable es la dependiente, y cual es la independiente. Después, y a la vista de la nube de puntos, hay que elegir un tipo de modelo o función  $y = f(x)$ , que puede ser lineal, cuadrático, exponencial, etc., que determina la relación entre las variables.

El tipo de modelo de regresión  $y = f(x)$ , elegido para ajustar la nube de puntos, dependerá de una serie de coeficientes o parámetros. Los métodos de regresión nos permiten calcular los coeficientes o parámetros que determinan el modelo que mejor se ajusta a la nube de puntos.

Por ejemplo, si hemos elegido un modelo lineal de regresión del tipo  $y = a + bx$ , el método de regresión nos ayudará a calcular los valores de  $a$  y  $b$  que determinan la recta  $y = a + bx$  que mejor se ajusta a la nube de puntos.

Para poder determinar los coeficientes de un modelo de regresión es necesario disponer de un mínimo número de puntos. En general, será necesario que haya tantos puntos como coeficientes haya que determinar en el modelo. Por ejemplo, si consideramos el modelo lineal  $y = a + bx$ , entonces será necesario que la nube de puntos tenga, al menos, dos puntos, pues, con un sólo punto habría infinitud de rectas que ajustasen (perfectamente) el modelo y ninguna de ellas sería mejor que las otras. O, por ejemplo, pensemos en un modelo parabólico  $y = a + bx + cx^2$ . En este caso, será necesario que la nube de puntos tenga más de tres elementos, pues con un número menor, por ejemplo dos, hay infinitud de parábolas que pasan por esos dos puntos, y todas ellas, se ajustan perfectamente a la nube de puntos.

Por lo tanto, y para evitar trivialidades, consideraremos que el número de observaciones de una variable bidimensional  $(X, Y)$  es mayor o igual al número de coeficientes del modelo de regresión que deseamos ajustar. Además, como veremos en esta sección, será necesario que esos puntos tengan valores distintos de la variable independientes, es decir, que el número de coeficientes del modelo debe ser menor o igual al número de observaciones con valores distintos de la variable independiente.

### Método de los mínimos cuadrados

El método de los mínimos cuadrados permite ajustar modelos de regresión, y consiste en minimizar las distancias entre el modelo y los puntos correspondientes a los valores observados en la muestra. Estas distancias reciben el nombre de *errores* o *residuos*.

Consideramos una muestra de tamaño  $N$  de una variable bidimensional  $(X, Y)$  que toma los valores  $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)$  con frecuencias absolutas  $n_1, n_2, \dots, n_k$ , y supongamos que hemos determinado que la variable  $Y$  depende de la variable  $X$ . Primero, elegimos el modelo o función  $y = f(x)$  que depende de ciertos parámetros  $a_1, \dots, a_m$ . Después, a cada valor  $x_i$  de la variable  $X$  le asignamos un valor teórico  $y_i^* = f(x_i)$  calculado a partir del modelo.

Como se ve en la figura 2.7, las diferencias entre los verdaderos valores  $y_i$  y los valores  $y_i^*$  estimados por el modelo, a partir de los correspondientes valores  $x_i$ , determinan los errores cometidos al utilizar el modelo, que se denotan por  $e_i = (y_i - y_i^*)$ .

El objetivo es minimizar los errores, pero hay que tener en cuenta que los valores  $e_i$  pueden ser positivos o negativos en función de la posición relativa del punto  $(x_i, y_i)$  respecto de la función  $y = f(x)$ . Por lo tanto, la simple suma de estos errores puede dar una visión equivocada del ajuste del modelo a la nube de puntos. Por ejemplo, si la suma de los errores es 0, puede ser que la función pase efectivamente por todos los puntos de la nube indicando un ajuste perfecto; o puede ser también que los errores de signo positivo se hayan compensado con los negativos y el ajuste no sean tan bueno como creíamos.

Utilizar los valores absolutos de los errores puede dificultar notablemente los cálculos, de manera que, para evitar estos problemas, utilizaremos los cuadrados de los errores. Y ya estamos en disposición de construir una función objetivo  $F$ , definida como la suma de los cuadrados de los errores  $e_i$ . Esta función sólo depende de los parámetros de la función  $f(x)$  que hay que

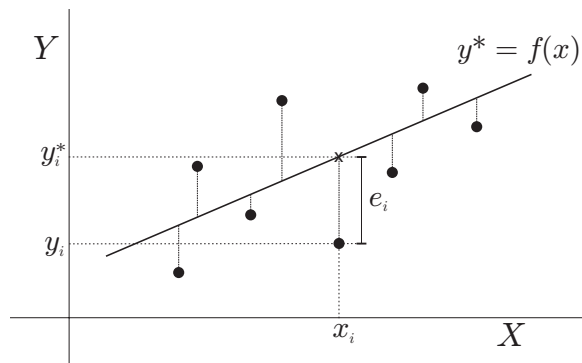


Figura 2.7: Método de los mínimos cuadrados Y/X

determinar:

$$F(a_1, \dots, a_m) = \sum_{i=1}^k e_i^2 \cdot n_i = \sum_{i=1}^k (y_i - y_i^*)^2 \cdot n_i = \sum_{i=1}^k (y_i - f(x_i))^2 \cdot n_i$$

Para calcular el valor de los parámetros que minimizan la función basta con resolver el sistema obtenido al igualar a cero las derivadas parciales de  $F$  respecto de los parámetros de los que depende  $f(x)$ , es decir, resolver el sistema:  $\nabla F = 0$  donde  $\nabla$  es el operador gradiente. En definitiva, el método consiste en minimizar la suma de los cuadrados de los errores y de ahí su nombre.

Para explicar el método de los mínimos cuadrados hemos considerado que la variable independiente era  $X$ . En este caso, los errores se definían como las diferencias entre los valores observados de la variable  $Y$  y los valores estimados según el modelo  $y = f(x)$ . Si consideramos que  $Y$  es la variable independiente entonces el modelo es de la forma  $x = g(y)$  y los errores se determinan como diferencias de los valores observado y los estimados para la variable  $X$ , es decir,  $e_i = (x_i - x_i^*)$  donde  $x_i^* = g(y_i)$  (ver figura 2.8).

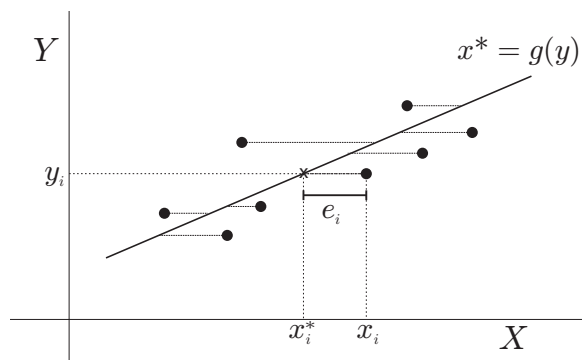


Figura 2.8: Método de los mínimos cuadrados X/Y

### Curva general de regresión

La curva general de regresión es un conjunto de puntos que representa a la nube de puntos. Como veremos, ajustar un modelo de regresión a la nube de puntos, equivale a ajustarlo a la curva de regresión. Este resultado simplificará notablemente los cálculos en aquellos ejemplos cuyos datos se presentan en forma de tabla de doble entrada.

Consideramos la distribución de frecuencias de la variable  $(X, Y)$  que presenta las modalidades  $(x_i, y_j)$  con frecuencias relativas  $f_{ij}$  con  $i = 1, \dots, k$  y  $j = 1, \dots, p$ . Se define la **curva general de regresión de  $Y$  sobre  $X$**  como la función que asigna, a cada valor  $x_i$  de la variable de  $X$ , la media  $\bar{y}_i$  de la distribución de la variable  $Y$  condicionada al valor  $x_i$  de la variable  $X$ .

Con esta definición, podemos decir que la curva de regresión está formada por los valores  $(x_i, \bar{y}_i)$  con frecuencia relativa  $f_i$  con  $i = 1, \dots, k$ , siendo  $\bar{y}_i = \sum_{j=1}^p y_j f_{ij}$ . Obsérvese que se podría definir, de manera análoga, la curva general de regresión de  $X$  sobre  $Y$  como la función que asigna, a cada valor  $y_j$  de la variable de  $Y$ , la media  $\bar{x}_j = \sum_{i=1}^k x_i f_{ij}$  de la distribución de la variable  $X$  condicionada al valor  $y_j$  de la variable  $Y$ .

La importancia de estas curvas radica en la siguiente propiedad de la curva general de regresión: *El problema de ajustar un modelo de regresión  $Y$  sobre  $X$  a la nube de puntos, por el método de los mínimos cuadrados, es equivalente a ajustar dicho modelo a la curva general de regresión, por el método de los mínimos cuadrados.*

Esta propiedad tiene dos implicaciones inmediatas en el ajuste por mínimos cuadrados. Por un lado, cuando tengamos un conjunto de observaciones donde algunos puntos tienen el mismo valor de la variable independiente, podemos simplificar el conjunto de datos. En particular, cuando tengamos un problema donde la distribución de frecuencias viene expresada con una tabla de doble entrada, podemos transformarla en una tabla estadística de frecuencias. Para ello, sustituiremos los valores de la variable dependiente por las medias de las distribuciones condicionadas correspondientes.

En la figura 2.9 se muestra como se transforma la tabla de doble entrada de la distribución de frecuencias de la variable  $(X, Y)$ , en una tabla de frecuencias donde cada modalidad  $(x_i, y_j)$  ha sido sustituida por la modalidad  $(x_i, \bar{y}_i)$ , siendo  $\bar{y}_i$  la media de la variable  $Y/X = x_i$ , para todo  $i = 1, \dots, k$ .

$X \setminus Y$	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_p$			$x_i$	$y_i$	$n_i$
$x_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1j}$	$\dots$	$n_{1p}$	$n_{1\cdot}$		$x_1$	$\bar{y}_1$	$n_{1\cdot}$
$x_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2j}$	$\dots$	$n_{2p}$	$n_{2\cdot}$		$x_2$	$\bar{y}_2$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$x_i$	$n_{i1}$	$n_{i2}$	$\dots$	$n_{ij}$	$\dots$	$n_{ip}$	$n_{i\cdot}$	$\longrightarrow$	$x_i$	$\bar{y}_i$	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$x_k$	$n_{k1}$	$n_{k2}$	$\dots$	$n_{kj}$	$\dots$	$n_{kp}$	$n_{k\cdot}$		$x_k$	$\bar{y}_k$	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot p}$	$N$				$N$

Figura 2.9: Aplicación de la propiedad de la curva general de regresión

Esta simplificación del conjunto de observaciones no tiene sentido realizarse cuando todos los puntos tiene distinto valor de la variable independiente. En tal caso, la curva de regresión coincide con la nube de puntos.

Por otro lado, veamos que la propiedad de la curva general de regresión que hemos presentado, tiene otra consecuencia inmediata, para evitar trivialidades, imponiendo una restricción al modelo de regresión.

Sabemos que el número de observaciones de una variable bidimensional  $(X, Y)$  debe ser mayor o igual que el número de coeficientes del modelo de regresión que deseamos ajustar. Pero según la propiedad de las curvas de regresión, ajustar un modelo a la nube de puntos es igual que ajustarlo a la curva de regresión. Por lo tanto, el número de coeficientes del modelo debe ser menor o igual al número de puntos de la curva de regresión, es decir, al número de observaciones con valores distintos de la variable independiente.

### 2.2.3. Correlación

La correlación mide el grado de relación entre las variables, a partir del modelo de regresión. Para ello, se definen medidas que determinan la bondad de dicho modelo.

La aproximación de la curva de regresión a la nube de puntos viene determinada por los residuos. Las medidas de correlación deben cuantificar la dispersión de los datos en torno al modelo, es decir, lo cerca o lejos de la curva que están los puntos. Para ello, será necesario hacer un estudio de las varianzas y de los residuos.

En las fórmulas que vamos a obtener para estas medidas, consideramos una muestra de tamaño  $N$  de una variable  $(X, Y)$  que toma los valores  $(x_i, y_i)$ , con frecuencias absolutas  $n_i$ , y relativas  $f_i$ , respectivamente para todo  $i = 1, \dots, k$ .

#### Varianzas del modelo

En el estudio del modelo general de regresión  $y = f(x)$  para las variables  $X$  e  $Y$ , hemos considerado dos nuevas variables: los valores  $(E)$  de los errores o residuos y los valores  $(Y^*)$  estimados por el modelo. Para cada pareja de valores  $(x_i, y_i)$  de la variable  $(X, Y)$  hemos considerado un valor  $y_i^* = f(x_i)$  de la variable  $Y^*$  y un valor  $e_i = y_i - y_i^*$  de la variable  $E$ .

Vamos a considerar las varianzas de estas variables  $Y^*$  y  $E$ , cuyos valores se obtienen a partir del modelo ajustado. Ambas medidas se utilizan para determinar la bondad del ajuste y, junto a la varianza de  $Y$ , forman parte en la definición de algunos coeficientes de correlación.

Se llama **varianza explicada** a la varianza de los valores estimados  $y_i^*$  de la variable  $Y^*$

$$\sigma_{y^*}^2 = \sum_{i=1}^k (y_i^* - \bar{y}^*)^2 \cdot f_i$$

Se llama **varianza residual** o **varianza no explicada** a la varianza de los errores  $e_i$  de la variable  $E$

$$\sigma_e^2 = \sum_{i=1}^k (e_i - \bar{e})^2 \cdot f_i \quad \text{siendo} \quad \bar{e} = \sum_{i=1}^k e_i \cdot f_i \quad \text{y} \quad e_i = y_i - y_i^*$$

Y, se llama **varianza total** a la varianza de la variable dependiente  $Y$ .

### Coefficiente de determinación

En el estudio del modelo general de regresión  $y = f(x)$  para las variables  $X$  e  $Y$ , la variable  $E$  (errores o residuos) mide las diferencias entre los valores de la variable  $Y$  y los valores de la variable ( $Y^*$ ) estimados por el modelo. Por lo tanto, se espera que  $E$  sea una variable cuya media debe ser 0, y cuya varianza debe ser pequeña (en comparación con la de  $Y$ ).

Por esta razón, se define el *coeficiente de determinación* como 1 menos el cociente entre la varianza residual y la varianza de la variable  $Y$ , y se denota por  $R^2$

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

Si el ajuste, mediante la curva de regresión, es bueno, cabe esperar que este coeficiente tome un valor próximo a 1. De esta manera, el coeficiente de determinación mide el grado de bondad del ajuste.

### Residuos

Los residuos indican la discrepancia entre el modelo y los datos. Una comparación entre estos valores para distintos modelos permite elegir el más adecuado.

En el método de los mínimos cuadrados, se definía la función suma de los cuadrados de los residuos. Esta función dependía de los coeficientes del modelo que se obtenían al ajustar la curva a la nube de puntos.

Por tanto, a partir del modelo ajustado  $y = f(x)$  y de los puntos  $(x_i, y_i)$  con frecuencias absolutas  $n_i$  podemos obtener los residuos  $e_i = y_i - f(x_i)$  que son los valores de la variable  $E$  correspondientes al modelo. Si calculamos la suma de los cuadrados de los residuos (sin promediarlos)

$$\text{SSE} = \sum_{i=1}^k e_i^2 \cdot n_i = \sum_{i=1}^k (y_i - y_i^*)^2 \cdot n_i$$

obtenemos un coeficiente de correlación que denotamos por SSE (Sum of Squared Errors).

Este coeficiente sirve para comparar la bondad de dos modelos que se ajustan a una misma nube de puntos. SSE determina los errores cometidos cuando se utilizan los valores estimados por el modelo en lugar de los verdaderos valores de la variable. Por tanto, el modelo que presente un menor valor de SEE corresponde al modelo que mejor se aproxima a la nube de puntos.

Veamos que la curva general de regresión que hemos presentado en la página 70 tiene una interesante propiedad de correlación que determina una cota inferior del error que se comete cuando se ajusta cualquier modelo de regresión.

Consideramos la distribución de frecuencias de la variable  $(X, Y)$  que presenta las modalidades  $(x_i, y_j)$  con frecuencias absolutas  $n_{ij}$  con  $i = 1, \dots, k$  y  $j = 1, \dots, p$ , y sean  $(x_i, \bar{y}_i)$  los puntos que definen la curva general de regresión de  $Y$  sobre  $X$ . Entonces se verifica que el valor del coeficiente SSE de cualquier modelo de regresión  $y = f(x)$  es mayor o igual que el valor del



coeficiente SSE de la curva general de regresión  $Y/X$ , es decir,

$$\sum_{i=1}^k \sum_{j=1}^p (y_j - \bar{y}_i)^2 n_{ij} \leq \sum_{i=1}^k \sum_{j=1}^p (y_j - f(x_i))^2 n_{ij}$$

Por lo tanto, la expresión del primer miembro de la ecuación, determina una cota inferior del error que se comete cuando se ajusta cualquier modelo de regresión. Sin embargo, si todos los valores de la variable independiente son distintos, entonces la cota que determina la curva general de regresión es trivial, pues vale 0.

## 2.3. El modelo lineal

Sea  $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$  una muestra de la variable estadística bidimensional  $(X, Y)$ . Para simplificar las fórmulas, hemos considerado que todas las modalidades presentan frecuencia absoluta igual a uno; en otro caso, los distintos valores aparecerían multiplicados por su frecuencia absoluta correspondiente.

Nuestro objetivo será encontrar un modelo lineal que se ajuste a la nube de puntos y un coeficiente que determine el grado de aproximación del modelo a los datos.

### 2.3.1. Regresión lineal

El modelo lineal que mejor se aproxima a la nube de puntos recibe el nombre de *recta de regresión de  $Y$  sobre  $X$* . Este modelo de ecuación  $Y = a + b \cdot X$  queda determinado conociendo los valores de los parámetros  $a$  y  $b$ . Aplicando el método de los mínimos cuadrados se obtienen fórmulas que permitan calcular estos parámetros en función de los datos de la muestra.

A cada valor  $x_i$  de la variable  $X$  le corresponde un valor  $y_i$  de la variable  $Y$ . Sin embargo, la recta de regresión le asigna a  $x_i$  el valor estimado  $y_i^* = f(x_i) = a + bx_i$ . Por tanto, la diferencia (también llamada error o residuo) entre el valor “teórico ajustado” y el valor “real” es

$$e_i = y_i - y_i^* = y_i - a - bx_i$$

Aplicando el método de los mínimos cuadrados, imponemos la condición de que la suma de los errores al cuadrado sea mínima. Para ello, minimizamos la función

$$F(a, b) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - a - bx_i)^2$$

donde  $x_i$  e  $y_i$  son datos del problema.

Ahora, los puntos críticos de la función  $F$ , que resultan ser mínimos<sup>1</sup>, se obtienen resolviendo la ecuación  $\nabla F(a, b) = 0$ .

<sup>1</sup>En el ejercicio 35 de la página 96 se propone la demostración de este resultado

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{array} \right\} \leftrightarrow \left\{ \begin{array}{l} -2 \sum_{i=1}^N (y_i - a - bx_i) = 0 \\ -2 \sum_{i=1}^N x_i (y_i - a - bx_i) = 0 \end{array} \right\} \leftrightarrow \left\{ \begin{array}{l} \sum_{i=1}^N y_i = a \cdot N + b \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i y_i = a \sum_{i=1}^N x_i + b \sum_{i=1}^N x_i^2 \end{array} \right\}$$

El sistema anterior recibe el nombre de *sistema de ecuaciones normales* que expresado en forma matricial resulta:

$$\begin{pmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{pmatrix}$$

Resolviendo el sistema se obtienen los siguientes resultados:

$$b = \frac{N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \cdot \sum_{i=1}^N y_i}{N \sum_{i=1}^N x_i^2 - \left( \sum_{i=1}^N x_i \right)^2} \quad y \quad a = \frac{1}{N} \left( \sum_{i=1}^N y_i - b \sum_{i=1}^N x_i \right)$$

Si se divide el numerador y el denominador de la expresión de  $b$  por  $N^2$  y se observa la expresión obtenida para  $a$  tenemos

$$b = \frac{Cov(X, Y)}{\sigma_x^2} \quad y \quad a = \bar{y} - b\bar{x}$$

Por tanto, la ecuación de la *recta de regresión de Y sobre X* ( $Y/X$ ) es

$$(Y - \bar{y}) = \frac{Cov(X, Y)}{\sigma_x^2} (X - \bar{x})$$

Si consideramos que  $Y$  es la variable independiente y  $X$  la dependiente, entonces la ecuación del modelo lineal es  $X = a + bY$ . Para ajustar el modelo a la nube de puntos, aplicamos el método de los mínimos cuadrados y obtenemos la recta de regresión de  $X$  sobre  $Y$  ( $X/Y$ ) que es

$$(X - \bar{x}) = \frac{Cov(X, Y)}{\sigma_y^2} (Y - \bar{y})$$

Como podemos observar, las dos rectas de regresión obtenidas pasan y se cortan en el punto del plano correspondiente al centro de gravedad  $(\bar{x}, \bar{y})$ .

En este punto, hay que hacer una observación importante sobre las rectas de regresión  $Y/X$  y  $X/Y$ . Desde el punto de vista matemático, las dos rectas son distintas pues, en general, si en

la recta de regresión  $Y/X$  despejamos la variable  $X$  en función de la variable  $Y$ , no se obtiene la recta de regresión de  $X/Y$ , y viceversa.

Los modelos de regresión permiten “predecir” los valores de la variable dependiente en función de los valores de la variable independiente. Así, la recta de regresión de  $Y/X$  determina los valores de  $Y$  en función de los valores de  $X$ , y por lo tanto, si deseamos utilizar un modelo lineal para calcular un valor de  $X$  en función de uno de  $Y$ , no podemos utilizar el modelo lineal  $Y/X$ . En este caso será necesario calcular la recta de regresión  $X/Y$ .

**Ejemplo 2.16** En el ejemplo 2.2 de la página 54 se consideran las variables “número de controles efectuados” ( $C$ ) y “número de errores detectados” ( $D$ ) en programas de software. Determinar la variable dependiente y calcular la recta de regresión para los datos de la muestra.

Evidentemente, el número de errores detectados (variable  $D$ ) depende del número de controles efectuados (variable  $C$ ), y por lo tanto, consideramos el modelo lineal  $D = a + bC$ . Veamos tres formas de calcular los valores de  $a$  y  $b$ , que determinan el modelo.

1. Resolviendo el sistema de ecuaciones normales:

$$\begin{aligned}\sum d_i n_i &= a \cdot N + b \sum c_i n_i \\ \sum c_i d_i n_i &= a \sum c_i n_i + b \sum c_i^2 n_i\end{aligned}$$

y, para ello, resulta útil considerar la siguiente tabla estadística

$c_i$	$d_i$	$n_i$	$c_i n_i$	$c_i^2 n_i$	$d_i n_i$	$c_i d_i n_i$
0	0	2	0	0	0	0
0	1	4	0	0	4	0
1	0	4	4	4	0	0
1	1	8	8	8	8	8
2	1	2	4	8	2	4
		20	16	20	14	12

que determina el sistema de ecuaciones

$$14 = 20a + 16b$$

$$12 = 16a + 20b$$

cuya solución es  $a = \frac{11}{18} \approx 0'611$  y  $b = \frac{1}{9} \approx 0'111$ .

2. Aplicando las fórmulas

$$b = \frac{\text{Cov}(X, Y)}{\sigma_x^2} \quad \text{y} \quad a = \bar{y} - b\bar{x}$$

a las variables y datos de nuestro ejemplo, siendo  $X = C$  e  $Y = D$ ,

$$b = \frac{\text{Cov}(C, D)}{\sigma_C^2} = \frac{0'04}{0'36} = \frac{1}{9} \approx 0'111$$

$$a = \bar{D} - b\bar{C} = 0'7 - \frac{1}{9} \cdot 0'8 = \frac{11}{18} \approx 0'611$$

también obtenemos esos mismos valores para los coeficientes  $a$  y  $b$  del modelo.

3. Calculamos la curva general de regresión:

$n_{ij}$	0	1	$D$		$c_i$	$d_i$	$n_i$
0	2	4	6	$\longrightarrow$	0	$2/3$	6
1	4	8	12		1	$2/3$	12
2	0	2	2		2	1	2
$C$	6	14	20				20

Aplicando la propiedad de la curva general de regresión, si ajustamos el modelo lineal de regresión a esta distribución de frecuencias, obtenemos la misma recta que en los casos anteriores.

Independientemente del método usado, la recta de regresión de  $D/C$  es el siguiente modelo lineal que relaciona ambas variables:

$$D = \frac{11}{18} + \frac{1}{9}C$$

En la figura 2.10 se representa la nube de puntos (puntos azules) y la curva general de regresión (cruces en rojo), cada una de ellas con un número que indica su frecuencia absoluta. Además, se representa la recta de regresión (línea discontinua) que se ajusta a estos datos.

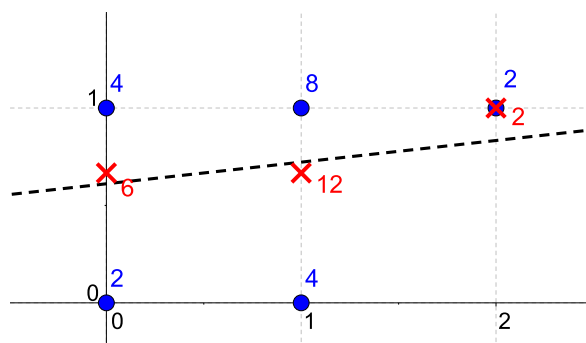


Figura 2.10: Ajuste lineal

Obsérvese que este modelo me permite “predecir” el valor de  $D$  en función del valor de  $C$ . Por ejemplo, cuando se realicen  $C = 5$  controles, se espera que se detecten  $D = 7/6 \approx 1'167$  errores.

Sin embargo, este modelo, no se debe utilizar para predecir el valor de  $C$ , en función de  $D$ , aunque, matemáticamente, si sea posible. En ese caso, habría que utilizar la recta de regresión  $C/D$ .  $\square$

### 2.3.2. Correlación lineal

Una vez visto el problema de *regresión* o *estimación* de una variable, se verá ahora el problema de la *correlación*, o grado de interconexión entre variables. Se pretende determinar con qué precisión se describe o explica la relación entre variables en una ecuación lineal.

### Coeficientes de regresión y correlación lineal

Dada una variable bidimensional  $(X, Y)$ , podemos obtener dos rectas de regresión: la de  $Y$  sobre  $X$  y la de  $X$  sobre  $Y$ . Para cada una de estas rectas definimos el *coeficiente de regresión lineal* como

$$b_{y/x} = \frac{Cov(X, Y)}{\sigma_x^2} \quad \text{y} \quad b_{x/y} = \frac{Cov(X, Y)}{\sigma_y^2}$$

siendo  $b_{y/x}$  el coeficiente de regresión de la recta de regresión  $Y/X$  y  $b_{x/y}$  de la recta de regresión de  $X/Y$ .

Estos coeficientes tienen el mismo signo y están estrechamente relacionados con las pendientes de las rectas. Por ello, los valores que toman determinan el crecimiento, decrecimiento, horizontalidad o verticalidad de las rectas de regresión. Por ejemplo, si  $b_{y/x}$  es un número positivo, entonces la recta de regresión de  $Y$  sobre  $X$  es creciente e indica que aumenta la variable  $Y$  al aumentar la  $X$ .

Las pendientes de las rectas son:

$$m_{y/x} = b_{y/x} = \frac{Cov(X, Y)}{\sigma_x^2} \quad \text{y} \quad m_{x/y} = \frac{1}{b_{x/y}} = \frac{\sigma_y^2}{Cov(X, Y)}$$

siendo  $m_{y/x}$  la pendiente de la recta de regresión  $Y/X$  y  $m_{x/y}$  de la recta de regresión  $X/Y$ .

**Ejemplo 2.17** Calcular los coeficientes de regresión lineal para las variables del ejemplo 2.2 de la página 54.

Sabiendo que  $Cov(C, D) = 0'04$ ,  $\sigma_c^2 = 0'36$  y  $\sigma_d^2 = 0'21$ , aplicamos la fórmula que se deduce de la definición:

$$b_{d/c} = \frac{Cov(C, D)}{\sigma_c^2} = \frac{0'04}{0'36} = \frac{1}{9} \approx 0'111 \quad , \quad b_{c/d} = \frac{Cov(C, D)}{\sigma_d^2} = \frac{0'04}{0'21} = \frac{4}{21} \approx 0'1905$$

□

### Coeficiente de correlación lineal

Siempre que los datos tiendan a agruparse en torno a una línea recta se puede afirmar que existe *correlación lineal* o dependencia de tipo lineal. Además, distinguimos dos tipos:

- Si la recta tiene pendiente positiva, la correlación o dependencia lineal es *directa*, es decir, incrementos positivos de una variable implican aumentos en la otra.
- Si la recta tiene pendiente negativa, la correlación o dependencia lineal es *inversa*, es decir, al aumentar una variable disminuye la otra.

El *análisis de correlación* consiste en determinar un número que permita conocer cuál es el grado de asociación entre las variables y en qué sentido (directa o inversamente). Por esta razón se introduce el concepto de coeficiente de correlación lineal de Pearson.

El **coeficiente de correlación lineal de Pearson** es una medida que se sólo se define para el modelo lineal, y que determina el grado de ajuste entre una nube de puntos y la recta de

regresión correspondiente. Este coeficiente es adimensional, se denota por  $r$  o  $\rho$  y viene definido por la media geométrica de los coeficientes de regresión lineal:

$$r = \rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} \quad -1 \leq r \leq 1$$

Obsérvese que este coeficiente no puede calcularse si alguna de las variables es degenerada, es decir, toma un único valor. En ese caso, como la desviación típica de la variable degenerada es siempre 0, la definición anterior carece de sentido.

El coeficiente de correlación lineal resulta ser siempre un número en el intervalo  $[-1, 1]$  con las siguientes interpretaciones<sup>2</sup> en función a su valor y su signo:

- El signo de este indicador va a coincidir con el de la covarianza pues las desviaciones de las variables son siempre positivas. De esta manera
  - Si  $r > 0$  entonces la relación entre las variables es directa.
  - Si  $r < 0$  entonces la relación entre las variables es inversa.
- El valor del coeficiente determina el grado de ajuste de la recta. De esta manera
  - Si  $r = -1$  ó  $r = 1$  entonces la correlación es perfecta e indica que existe una dependencia funcional entre las variables. En este caso, los datos representados en la nube de puntos están situados sobre una recta, que resulta ser la recta de regresión  $Y/X$  y que coincide con la de  $X/Y$ .
  - Si  $r = 0$  entonces las rectas de regresión son paralelas a los ejes ( $y = \bar{y}$  y  $x = \bar{x}$ ), y se dice que las variables están linealmente incorreladas.
  - Los valores intermedios determinan los grados intermedios de ajustes. Cuanto más cerca de 1 ó -1 esté el valor de  $r$  la correlación será más *fuerte*, mientras que valores próximos a 0 indican una correlación *débil*.

El coeficiente  $r$  que hemos definido, resulta ser una medida objetiva de correlación lineal entre dos variables, en el sentido de que no depende de la escala de medición utilizada, es decir, es adimensional. Sin embargo, es importante tener en cuenta que sólo tiene sentido definir este coeficiente en el caso lineal.

**Ejemplo 2.18** *Determinar e interpretar el valor del coeficiente de correlación lineal de Pearson para los datos del ejemplo 2.2 de la página 54.*

Sabiendo que  $\text{Cov}(C, D) = 0'04$ ,  $\sigma_c^2 = 0'36$  y  $\sigma_d^2 = 0'21$ , aplicamos la definición y obtenemos:

$$r = \frac{\text{Cov}(C, D)}{\sigma_c \cdot \sigma_d} = \frac{0'04}{\sqrt{0'36} \cdot \sqrt{0'21}} \approx 0'1455$$

Que el valor de  $r$  sea positivo, indica que la relación entre las variables es directa, es decir, que  $D$  aumenta, cuando aumenta  $C$ . Además, el hecho de que el valor de  $|r|$  esté más próximo a 0 que a 1, indica que la correlación entre las variables es débil, es decir, que no hay mucha relación lineal entre ellas.  $\square$

<sup>2</sup>Que  $r \in [-1, 1]$  es consecuencia de la expresión  $\sigma_e^2 = \sigma_y^2(1 - r^2) \geq 0$  que se deduce en el caso lineal; y la interpretación de los posibles valores de  $r$  es consecuencia de la fórmula  $r^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$  que se deriva de la expresión anterior.

El coeficiente de correlación lineal permite establecer una relación entre las pendientes de la recta de regresión  $Y/X$  y la recta de regresión  $X/Y$  para un mismo conjunto de datos.

Si multiplicamos el numerador y denominador de  $m_{y/x}$  por  $\sigma_y$  y el de  $m_{x/y}$  por  $\sigma_x$  obtenemos

$$m_{y/x} = \frac{Cov(X, Y)}{\sigma_x^2} \cdot \frac{\sigma_y}{\sigma_y} = r \cdot \frac{\sigma_y}{\sigma_x} \quad \text{y} \quad m_{x/y} = \frac{\sigma_y^2}{Cov(X, Y)} \cdot \frac{\sigma_x}{\sigma_x} = \frac{1}{r} \cdot \frac{\sigma_y}{\sigma_x}$$

siendo  $r$  el coeficiente de correlación lineal. Como el valor de  $r$  es siempre un número comprendido entre -1 y 1, se puede establecer la siguiente relación entre las pendientes de las rectas de regresión

$$|m_{y/x}| \leq |m_{x/y}|$$

Esta relación permite determinar cuál de las dos rectas de regresión es la de  $Y$  sobre  $X$  y cuál es la de  $X$  sobre  $Y$  comparando, simplemente, sus pendientes.

**Ejemplo 2.19** Sean  $y = 4x - 7$  e  $y = x - 1$  las rectas de regresión de las variables  $X$  e  $Y$  cuya covarianza es 9. Calcular las medias y las varianzas de las variables  $X$  e  $Y$  y determinar el coeficiente de correlación lineal de Pearson.

Como las dos rectas de regresión pasan por el punto  $(\bar{x}, \bar{y})$ , basta resolver el sistema de ecuaciones formado por las dos rectas para obtener el valor de  $\bar{x}$  y  $\bar{y}$  que resulta ser 2 y 1, respectivamente.

Para obtener el resto de las medidas es necesario determinar cual de las dos rectas es la  $Y/X$ , y cual es la  $X/Y$ . Para ello, utilizamos la relación que se establece entre sus pendientes:  $|m_{y/x}| \leq |m_{x/y}|$ . Por lo tanto,  $y = x - 1$  es la recta de regresión  $Y/X$ , e  $y = 4x - 7$  es la recta de regresión  $X/Y$ .

Como la covarianza es 9, y sabemos que  $m_{y/x} = Cov(X, Y)/\sigma_x^2 = 1$ , entonces se deduce que  $\sigma_x^2 = 9$ . Y, análogamente, como sabemos que  $m_{x/y} = \sigma_y^2/Cov(X, Y) = 4$ , entonces se deduce que  $\sigma_y^2 = 36$ .

Finalmente, para calcular el coeficiente de correlación lineal, aplicamos su definición:

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{9}{\sqrt{9} \cdot \sqrt{36}} = \frac{9}{18} = 0.5$$

□

## Descomposición de la varianza

Las características del modelo lineal permiten expresar la varianza de la variable  $Y$  (varianza total), como suma de las varianzas residual y explicada. Esta fórmula se conoce como *descomposición de la varianza*.

Consideramos una muestra de tamaño  $N$  de una variable  $(X, Y)$  que toma los valores  $(x_i, y_i)$  con frecuencias relativas  $f_i$  para todo  $i = 1, \dots, k$ ; y consideramos el modelo lineal de regresión  $y = a + bx$  que determina las variables  $Y^*$ , que toma los valores  $y_i^* = a + bx_i$  con frecuencias relativas  $f_i$  para todo  $i = 1, \dots, k$ , y la variable  $E$ , que toma los valores  $e_i = y_i - y_i^*$  con frecuencias relativas  $f_i$  para todo  $i = 1, \dots, k$ .

Para el modelo lineal, como consecuencia de la primera ecuación normal, se verifican las siguientes propiedades:

(p1) Las medias de las variables  $Y$  e  $Y^*$  son iguales, es decir,  $\bar{y} = \bar{y}^*$ .

(p2) La suma de los residuos es cero y, por lo tanto, la media es cero, es decir,  $\bar{e} = \sum_{i=1}^n e_i \cdot f_i = 0$ .

Con estas propiedades podemos simplificar la fórmula de la varianza residual

$$\sigma_e^2 = \sum_{i=1}^k (e_i - \bar{e})^2 \cdot f_i \stackrel{(p1)}{=} \sum_{i=1}^k e_i^2 \cdot f_i = \sum_{i=1}^k (y_i - y_i^*)^2 \cdot f_i$$

y obtener una fórmula que relaciona las varianzas del modelo con la varianza total.

$$\sigma_y^2 = \sigma_{y^*}^2 + \sigma_e^2$$

**Ejemplo 2.20** Calcular las varianzas residual y explicada para el modelo lineal calculado en el ejemplo 2.16 de la página 75 y comprobar que la varianza marginal de la variable  $D$  es la suma de las varianzas del modelo.

Para calcular las varianzas del modelo, necesitamos obtener las distribuciones de frecuencias de las variables  $D^*$ , que representa los valores estimados por el modelo, y  $E$ , que representa los residuos. Las distribuciones de ambas variables se obtienen a partir del modelo ajustado:

$$D = \frac{11}{18} + \frac{1}{9}C$$

- La varianza explicada es la varianza de la variable  $D^*$  que toma los valores  $d_i^* = f(c_i)$  con las frecuencias absolutas de las modalidades  $c_i$  de la variable  $C$ :

$C$	$n_i$	$f_i$	$D^*$
0	6	0'3	11/18
1	12	0'6	13/18
2	2	0'1	15/18

y, por lo tanto, la varianza explicada toma el valor  $\sigma_{d^*}^2 = \frac{4}{900} \approx 0'0044$ .

- La varianza residual es la varianza de la variable  $E$  que toma los valores  $e_{ij} = d_j - f(c_i)$  con las frecuencias absolutas  $n_{ij}$  de las modalidades  $(c_i, d_j)$  de la variable  $(C, D)$ :

$e_{ij}$	$d_1=0$	$d_2=1$	con frecuencias	$n_{ij}$	$d_1=0$	$d_2=1$
$c_1=0$	-11/18	7/18		$c_1=0$	2	4
$c_2=1$	-13/18	5/18		$c_2=1$	4	8
$c_3=2$	-15/18	3/18		$c_3=2$	0	2

y, por lo tanto, la varianza residual toma el valor  $\sigma_e^2 = \frac{185}{900} \approx 0'2056$ .

Si comparamos estas dos varianzas, con la varianza de la variable  $E$ , obtenemos la siguiente relación:

$$\sigma_{e^*}^2 + \sigma_e^2 = \frac{4}{900} + \frac{185}{900} = \frac{189}{900} = \frac{21}{100} = \sigma_d^2$$

que es una característica del modelo lineal de regresión. □



La descomposición de la varianza permite, en el caso lineal, definir el coeficiente de determinación ( $R^2$ ) como el cociente entre la varianza explicada y la varianza total

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} = \frac{\sigma_{y*}^2}{\sigma_y^2} \quad \text{con} \quad 0 \leq R^2 \leq 1$$

Además, sólo en el caso lineal, donde tiene sentido calcular el coeficiente de correlación lineal ( $r$ ), se verifica la siguiente relación entre los coeficientes de correlación y determinación

$$R^2 = r^2$$

Obsérvese que el 2 que hay sobre  $r$  indica una potencia (elevar al cuadrado), mientras que el 2 de la expresión  $R^2$  es simplemente un símbolo (notación), pues no se define lo que significa  $R$ .

El valor de  $R^2$ , en el caso lineal, siempre es un número en el intervalo  $[0, 1]$ , de manera que si  $R^2$  está próximo a 1 significa que el ajuste es “bueno” mientras que un valor de  $R^2$  próximo a 0 indica que el modelo no es el adecuado.

**Ejemplo 2.21** *Calcular el coeficiente de determinación para el modelo lineal calculado en el ejemplo 2.16 de la página 75 y determinar la bondad del ajuste.*

Sabiendo los valores de las varianzas,  $\sigma_d^2 = \frac{189}{900} = 0'21$  de la variable  $D$ , y  $\sigma_e^2 = \frac{185}{900}$  de la variable  $E$  (residuos), aplicamos la fórmula

$$R^2 = 1 - \frac{185/900}{189/900} = 1 - \frac{185}{189} = \frac{4}{189} \approx 0'0212$$

y obtenemos el valor del coeficiente de determinación que, al ser próximo a 0, indica que la correlación entre las variables  $C$  y  $D$  es débil, es decir, que no hay mucha relación entre ellas.

Otra forma más sencilla de calcular este coeficiente es aplicando la fórmula que lo relaciona con el coeficiente de correlación lineal

$$R^2 = r^2 = \frac{Cov(C, D)^2}{\sigma_c^2 \cdot \sigma_d^2} = \frac{0'04^2}{0'36 \cdot 0'21} = \frac{0'0016}{0'0756} \approx 0'0212$$

□

## 2.4. Modelos de regresión no lineal

El modelo de regresión lineal que hemos estudiado es el más utilizado habitualmente. Sin embargo, la forma de la nube de puntos puede sugerir la consideración de otros modelos de regresión. Como veremos, en general, recurriremos al método de los mínimos cuadrados para ajustar el modelo y determinar el valor de los coeficientes. Sin embargo, hay algunos modelos que se pueden reducir al caso lineal, aplicando alguna transformación algebraica, y utilizar las fórmulas obtenidas antes.

### 2.4.1. Linealización de modelos

En muchos casos, los modelos de regresión utilizados pueden reducirse al caso lineal que hemos estudiado. Para ello, se realizan algunas transformaciones algebraicas y se determina un cambio de variables. Para obtener el nuevo modelo se aplican los cambios de las variables, transformando todas las modalidades.

Por ejemplo, a partir del modelo  $y = a \cdot b^x$  y aplicando logaritmos neperianos

$$y = a \cdot b^x \quad \Longrightarrow \quad \ln(y) = \ln(a) + \ln(b) \cdot x \quad \Longrightarrow \quad Y = A + B \cdot X$$

y podemos considerar el modelo lineal  $Y = A + B \cdot X$  donde

$$Y = \ln(y) \quad , \quad X = x \quad , \quad A = \ln(a) \quad \text{y} \quad B = \ln(b)$$

Ahora, aplicamos a  $y$  el cambio de variable, transformando todas sus modalidades. En este caso, las modalidades de la nueva variable  $Y$  se obtiene calculando el logaritmo neperiano de las modalidades de la variable  $y$ . Por último, ajustamos la nueva nube de puntos a la recta para obtener los valores de  $A$  y  $B$  y poder calcular los coeficientes  $a$  y  $b$  del modelo original

$$a = e^A \quad \text{y} \quad b = e^B$$

Obsérvese que para aplicar esta reducción al caso lineal, es necesario que todos los valores de  $y$  sean positivos, pues estamos considerando su logaritmo.

**Ejemplo 2.22** Ajustar el modelo  $y = a \cdot e^{bx}$  a los siguientes datos:

Variable $X$	1	2	3	4	5
Variable $Y$	4'5	6'5	10'0	15'0	22'0

Si aplicamos logaritmos neperianos al modelo  $y = a \cdot e^{bx}$  obtenemos

$$y = a \cdot e^{bx} \quad \Longrightarrow \quad \ln(y) = \ln(a) + b \cdot x \quad \Longrightarrow \quad Y = A + B \cdot X$$

y podemos considerar el modelo lineal  $Y = A + B \cdot X$  donde

$$Y = \ln(y) \quad , \quad X = x \quad , \quad A = \ln(a) \quad \text{y} \quad B = b$$

Si aplicamos estas transformaciones a los valores de las variables, obtenemos la siguiente tabla:

Nueva variable $X = x$	1	2	3	4	5
Nueva variable $Y = \ln y$	1'504	1'872	2'303	2'708	3'091

Para estos datos, calculamos la recta de regresión  $Y/X$  que resulta ser  $y = 1'0925 + 0'401x$ . Y deshaciendo los cambios de variable, obtenemos que  $a = e^A = e^{1'0925} = 0'272$  y  $b = B = 0'401$ . Por lo tanto, el modelo ajustado es

$$y = 0'272 \cdot e^{0'401x}$$

□

### 2.4.2. Ajuste parabólico

Consideramos una muestra  $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$  de una variable bidimensional  $(X, Y)$ . Nuestro objetivo es ajustar una función del tipo

$$y = a + bx + cx^2$$

Aplicando el método de los mínimos cuadrados, obtenemos la función

$$F(a, b, c) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - a - bx_i - cx_i^2)^2$$

La solución del problema pasa por minimizar la función  $F(a, b, c)$  para determinar los valores de  $a$ ,  $b$  y  $c$ . Para ellos, se resuelve el siguiente sistema de ecuaciones normales

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \\ \frac{\partial F}{\partial c} = 0 \end{array} \right\} \iff \left\{ \begin{array}{l} \sum_{i=1}^N y_i = aN + b \sum_{i=1}^N x_i + c \sum_{i=1}^N x_i^2 \\ \sum_{i=1}^N x_i y_i = a \sum_{i=1}^N x_i + b \sum_{i=1}^N x_i^2 + c \sum_{i=1}^N x_i^3 \\ \sum_{i=1}^N x_i^2 y_i = a \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i^3 + c \sum_{i=1}^N x_i^4 \end{array} \right\}$$

que escrito en forma matricial resulta

$$\begin{pmatrix} N & \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^3 \\ \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i^4 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N x_i^2 y_i \end{pmatrix}$$

Este resultado se puede generalizar (observar la estructura y disposición de los elementos de las matrices en el caso polinómico) para ajustar un modelo polinómico de cualquier grado. De manera que el sistema de ecuaciones normales, para el modelo polinómico general

$$y = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$$

expresado en forma matricial es

$$\begin{pmatrix} N & \sum x_i & \sum x_i^2 & \dots & \sum x_i^n \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{n+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \dots & \sum x_i^{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_i^n & \sum x_i^{n+1} & \sum x_i^{n+2} & \dots & \sum x_i^{2n} \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \\ \vdots \\ \sum x_i^n y_i \end{pmatrix}$$

cuya solución nos permite obtener los valores de los  $n + 1$  parámetros  $a_0, a_1, \dots, a_n$ .

**Ejemplo 2.23** Ajustar el modelo parabólico  $D = a + b \cdot C + c \cdot C^2$  a los datos del ejemplo 2.2 de la página 54.

Consideremos la tabla estadística de la distribución de frecuencias de las variables  $C$  y  $D$ , a la que hemos añadido una serie de columnas que nos resultarán útiles.

$c_i$	$d_i$	$n_i$	$c_i n_i$	$c_i^2 n_i$	$c_i^3 n_i$	$c_i^4 n_i$	$d_i n_i$	$c_i d_i n_i$	$c_i^2 d_i n_i$
0	0	2	0	0	0	0	0	0	0
0	1	4	0	0	0	0	4	0	0
1	0	4	4	4	4	4	0	0	0
1	1	8	8	8	8	8	8	8	8
2	1	2	4	8	16	32	2	4	8
		20	16	20	28	44	14	12	16

Con los valores de la tabla, podemos construir el siguiente sistema de ecuaciones lineales:

$$\left\{ \begin{array}{l} \sum d_i = aN + b \sum c_i + c \sum c_i^2 \\ \sum c_i d_i = a \sum c_i + b \sum c_i^2 + c \sum c_i^3 \\ \sum c_i^2 d_i = c \sum c_i^2 + b \sum c_i^3 + c \sum c_i^4 \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} 14 = 20a + 16b + 20c \\ 12 = 16a + 20b + 28c \\ 16 = 20a + 28b + 44c \end{array} \right\}$$

cuya solución, determina los valores de los parámetros  $a = \frac{2}{3}$ ,  $b = -\frac{1}{6}$  y  $c = \frac{1}{6}$ , y por tanto, el modelo que buscamos es:

$$D = \frac{2}{3} - \frac{1}{6} \cdot C + \frac{1}{6} \cdot C^2$$

En la figura 2.11 se representa la nube de puntos (puntos azules) y la curva general de regresión (cruces en rojo), cada una de ellas con un número que indica su frecuencia absoluta. Además, se representa la parábola de regresión (línea discontinua) que se ajusta a estos datos.

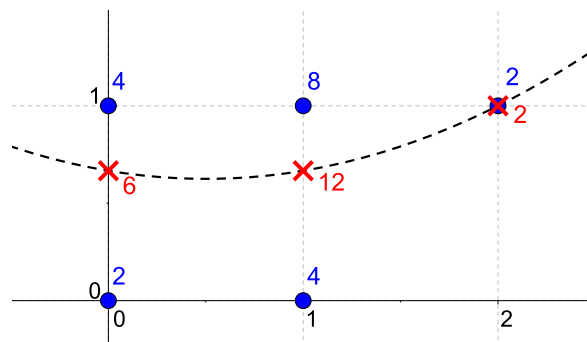


Figura 2.11: Ajuste parabólico

En este ejemplo, podíamos haber utilizado la curva general de regresión para ajustar el modelo parabólico y haber obtenido el mismo resultado. Como se observa en la figura, el modelo de regresión pasa exactamente por los puntos de la curva general de regresión, lo que supone que el ajuste es perfecto en el sentido de que los errores cometidos con cualquier otro modelo será siempre mayores.

Y todo esto ocurre porque el número de puntos distintos de la curva general de regresión (que es tres) coincide con el número de coeficientes del modelo (que también es tres por ser un modelo parabólico completo), en virtud de las propiedades de la curva general de regresión.

Si aplicamos el método de los mínimos cuadrados para ajustar cualquier otro modelo de regresión con más de tres coeficientes (por ejemplo un modelo cúbico completo) daría lugar a un sistema de ecuaciones normales compatible indeterminado, pues existirían infinitud de modelos (por ejemplo, infinitud de polinomios de grado 3) que se ajustan perfectamente a la nube de puntos, en el sentido de que, todo ellos, pasan por todos los puntos de la curva general de regresión.  $\square$

### 2.4.3. Otros ajustes

En general, para ajustar un modelo de regresión, utilizaremos el método de los mínimos cuadrados descrito en la sección 2.2.2. Este método que ya hemos usado para determinar el modelo lineal y el polinómico, se resume en los siguientes pasos:

1. Consideramos el conjunto de datos  $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ .
2. Representamos la nube de puntos para determinar qué modelo resulta más adecuado.
3. Si el modelo es  $f(x)$  y depende de los parámetros  $a_1, a_2, \dots, a_n$ , entonces consideramos la función

$$F(a_1, a_2, \dots, a_n) = \sum_{i=1}^N (y_i - f(x_i))^2$$

4. Calculamos todas las derivadas parciales de la función  $F$  y las igualamos a 0 para obtener el sistema de ecuaciones normales.
5. Al resolver este sistema obtenemos el valor de los parámetros que determinan el modelo de regresión ajustado.

**Ejemplo 2.24** *Obtener una fórmula que permita determinar el modelo de regresión  $y = bx$  para el conjunto de datos  $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ .*

Para aplicar el método de los mínimos cuadrados, debemos minimizar la función

$$F(b) = \sum_{i=1}^N (y_i - bx_i)^2$$

que sólo depende de un parámetro. En este caso, la derivada de  $F$ , igualada a 0, determina la ecuación normal:

$$\frac{dF}{dx}(b) = 0 \implies \sum_{i=1}^N x_i y_i - b \sum_{i=1}^N x_i^2 = 0$$

La solución de esta ecuación determina el mínimo<sup>3</sup> de la función  $F$  que corresponde al valor del coeficiente  $b$ , calculado a partir del conjunto de puntos.

$$b = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}$$

---

<sup>3</sup>El punto crítico obtenido es el mínimo de la función  $F$  pues  $\frac{d^2}{dx^2} F(b) = 2 \sum x_i^2 > 0$ .

OBSERVACIÓN: Aunque el modelo  $y = bx$  que hemos ajustado es lineal, no debemos confundirlo con el modelo lineal general  $y = a + bx$ . Un error muy común, que debemos evitar, es aplicar la fórmula  $Cov(X, Y)/\sigma_x^2$  del modelo lineal general para calcular el valor del parámetro  $b$ , considerando que el término independiente ( $a$ ) es igual a 0.  $\square$

En ocasiones, es posible aplicar los dos métodos (linealizar o aplicar, directamente, mínimos cuadrados) a un mismo modelo de regresión. Veamos un ejemplo.

**Ejemplo 2.25** *Ajustar el modelo  $y = ax + bx^3$  de dos maneras distintas (linealización del modelo y método de los mínimos cuadrados) para ajustarlo al siguiente conjunto de datos de la variable  $(X, Y)$ :*

$$\{(1, 5), (2, 8), (3, 9), (4, 8), (5, 0)\}$$

En primer lugar ajustamos el modelo reduciéndolo a un modelo lineal:

- Si dividimos por  $x$  la expresión del modelo obtenemos

$$y = ax + bx^3 \implies \frac{y}{x} = a + bx^2 \implies Y = A + B \cdot X$$

y podemos considerar el modelo lineal  $Y = A + B \cdot X$  donde

$$Y = \frac{y}{x}, \quad X = x^2, \quad A = a \quad \text{y} \quad B = b$$

Si aplicamos estas transformaciones a los valores de las variables, obtenemos el siguiente conjunto de datos de las nuevas variables  $(X, Y)$ :

$$\{(1, 5), (4, 4), (9, 3), (16, 2), (25, 0)\}$$

Para estos valores, calculamos la recta de regresión  $Y/X$  que resulta ser  $Y = 4'9765 - 0'1979 \cdot X$ . Deshaciendo los cambios de variable, obtenemos que  $a = A = 4'9765$  y que  $b = B = -0'1979$ . Por tanto, el modelo ajustado es

$$y = 4'9765x - 0'1979x^3$$

Obsérvese que este método no podría haberse utilizado si el valor de  $x$  de alguna de las observaciones hubiese sido 0, pues no hubiese sido posible aplicar la transformación.

Y ahora, utilizamos el método de los mínimos cuadrados para ajustar directamente el mismo modelo:

- En primer lugar, consideramos la función

$$F(a, b) = \sum (y_i - ax_i - bx_i^3)^2$$

Después, calculamos todas las derivadas parciales de la función  $F$ , y las igualamos a 0 para obtener el sistema de ecuaciones normales.

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{array} \right\} \iff \left\{ \begin{array}{l} \sum x_i y_i = a \sum x_i^2 + b \sum x_i^4 \\ \sum x_i^3 y_i = a \sum x_i^4 + b \sum x_i^6 \end{array} \right\}$$

que aplicado a los valores de nuestras variables, resulta ser el sistema de ecuaciones:

$$\begin{aligned} 80 &= 55a + 979b \\ 824 &= 979a + 20515b \end{aligned}$$

cuya solución, determina el valor de los parámetros  $a \approx 4'912$  y  $b \approx -0'194$  que determinan el modelo de regresión ajustado:

$$y = 4'912x - 0'194x^3$$

En este ejemplo, para hacer el ajuste, no tendría sentido simplificar el conjunto de observaciones utilizando la curva general de regresión, pues todos los valores de la variable independiente son distintos, y por lo tanto, la curva general de regresión coincide con la propia nube de puntos.  $\square$

Obsérvese, en el ejemplo anterior, que aunque son muy parecidos, los coeficientes de los modelos obtenidos por cada uno de los métodos son distintos. El objetivo de los dos métodos es minimizar los errores, sin embargo, la transformación aplicada en la linealización del modelo distorsiona estos errores. Por lo tanto, el uso directo del método de los mínimos cuadrados proporciona un modelo más ajustado que el método de linealización, si bien, en muchos casos puede resultar más sencillo aplicar este último.

#### 2.4.4. Bondad del ajuste

En los modelos que se reducen al caso lineal, se suele calcular el coeficiente de correlación lineal para el modelo transformado que es tipo lineal. Este coeficiente se puede utilizar como indicativo de la bondad del propio ajuste. Sin embargo, no debemos utilizarlo para comparar dos ajustes distintos.

En los modelos polinómicos completos (con todos sus términos) se verifica la fórmula de la descomposición de la varianza que, en general, no es cierta para cualquier modelo. Por lo tanto, en el caso polinómico resulta apropiado utilizar el coeficiente de determinación ( $R^2$ ) como medida de correlación. Además, este coeficiente toma valores en el intervalo  $[0, 1]$  con la misma interpretación que se le daba en el caso lineal.

**Ejemplo 2.26** Utilizar el coeficiente de determinación para estudiar la bondad de los modelos lineal y parabólico ajustados a los datos del ejemplo 2.16 de la página 75.

Para el modelo lineal  $D = \frac{11}{18} + \frac{1}{9}C$  podemos determinar los residuos ( $E$ )

$e_{ij}$	$d_1=0$	$d_2=1$	con frecuencias	$n_{ij}$	$d_1=0$	$d_2=1$
$c_1=0$	$-11/18$	$7/18$		$c_1=0$	2	4
$c_2=1$	$-13/18$	$5/18$		$c_2=1$	4	8
$c_3=2$	$-15/18$	$3/18$		$c_3=2$	0	2

y calcular la varianza residual  $\sigma_e^2 = 0'2056$  que nos permite calcular el coeficiente de determinación para el modelo lineal

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_d^2} = 1 - \frac{0'2056}{0'21} = 0'0212$$

Para el modelo parabólico  $y = \frac{2}{3} - \frac{1}{6}C + \frac{1}{6}C^2$  podemos determinar los residuos ( $E$ )

$e_{ij}$	$d_1=0$	$d_2=1$	con frecuencias	$n_{ij}$	$d_1=0$	$d_2=1$
$c_1=0$	-2/3	1/3		$c_1=0$	2	4
$c_2=1$	-2/3	1/3		$c_2=1$	4	8
$c_3=2$	-1	0		$c_3=2$	0	2

y calcular la varianza residual  $\sigma_e^2 = 0'2$  que nos permite calcular el coeficiente de determinación para el modelo lineal

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_d^2} = 1 - \frac{0'2}{0'21} = 0'0476$$

En ambos casos, el coeficiente de determinación es muy próximo a cero, lo que indica que los ajustes no son apropiados. Además, de los resultados se deduce que la parábola es un modelo mejor que la recta para ajustar los datos de la muestra, pues el valor de  $R^2$  es mayor. Esta conclusión es siempre cierta para estos dos modelos en cualquier conjunto de datos pues la expresión de la parábola  $y = a + bx + cx^2$  generaliza a la de la recta  $y = a + bx$ , que es un caso particular que se obtiene cuando  $c = 0$ .  $\square$

En general, para determinar el grado de bondad de un modelo cualquiera, se suele utilizar el coeficiente de determinación. Sin embargo, hay que tener en cuenta que, sólo en el caso polinómico completo, incluyendo el caso lineal, este coeficiente toma un valor entre 0 y 1. Por esa razón, para comparar la bondad de dos ajustes cualesquiera, a una misma nube de puntos, es preferible utilizar el coeficiente SSE que determina la suma de los cuadrados de los residuos

$$\text{SSE} = \sum_{i=1}^N e_i^2 n_i = \sum_{i=1}^N (y_i - f(x_i))^2 n_i \quad \text{o bien} \quad \text{SSE} = \sum_{i=1}^k \sum_{j=1}^p e_{ij}^2 n_{ij} = \sum_{i=1}^k \sum_{j=1}^p (y_j - f(x_i))^2 n_{ij}$$

**Ejemplo 2.27** Determinar qué modelo, el lineal o el parabólico, se ajusta mejor a los datos del ejemplo 2.16 de la página 75.

Para el modelo lineal podemos determinar los residuos (ver ejemplo anterior) y calcular el valor de  $\text{SSE} = 37/9 \approx 4'111$ . De la misma manera, para el modelo parabólico podemos determinar los residuos (ver ejemplo anterior) y calcular el valor de  $\text{SSE} = 4$ .

La curva general de regresión establece una cota inferior del valor de SSE para cualquier modelo que se ajuste a este conjunto de datos.

$n_{ij}$	0	1	$D$
0	2	4	6
1	4	8	12
2	0	2	2
$C$	6	14	20

$c_i$	$\bar{d}_i$	$n_i$
0	2/3	6
1	2/3	12
2	1	2
		20

 $\xrightarrow{(*)}$ 

$$\sum_{i=1}^3 \sum_{j=1}^2 (d_j - \bar{d}_i)^2 n_{ij} = 4$$

$$(*) \sum_{i=1}^3 \sum_{j=1}^2 (d_j - \bar{d}_i)^2 n_{ij} = 2(0 - \frac{2}{3})^2 + 4(1 - \frac{2}{3})^2 + 4(0 - \frac{2}{3})^2 + 8(1 - \frac{2}{3})^2 + 2(1 - 1)^2 = 4$$

Lo que significa que cualquier modelo que se ajuste a los datos del ejemplo, por el método de los mínimos cuadrados, debe tener un valor de SSE mayor o igual a 4. El hecho de que el modelo parabólico haya sido exactamente 4, indica que este ajuste parabólico es perfecto, en el sentido de que ningún otro ajuste puede disminuir la suma de los cuadrados de los residuos.  $\square$



## 2.5. Relación de problemas

1. En la elaboración de la siguiente tabla de frecuencias de la variable  $(X, Y)$  se ha cometido un error.

$Y \setminus X$	0	1	2	3	4	5	
$[0, 4]$	3	3	1	0	0	0	7
$(4, 6]$	3	4	2	0	0	0	9
$(6, 8]$	1	3	2	1	0	0	7
$(8, 12]$	0	0	1	2	3	2	9
	7	11	6	3	3	2	32

Se pide:

- Detectar y corregir la errata.
  - Representar la distribución condicionada  $(Y/X = 2)$  y calcular el sesgo y la curtosis.
  - Calcular las rectas de regresión de  $X$  sobre  $Y$  y de  $Y$  sobre  $X$ .
  - Calcular el coeficiente de correlación lineal y la varianza residual del modelo lineal  $Y/X$ .
2. Demostrar la igualdad de las dos siguientes fórmulas que permiten calcular la covarianza:

$$\text{Cov}(X, Y) = \sum_{i=1}^k \sum_{j=1}^p (x_i - \bar{x}) \cdot (y_j - \bar{y}) \cdot f_{ij} = \sum_{i=1}^k \sum_{j=1}^p x_i \cdot y_j \cdot f_{ij} - \bar{x}\bar{y}$$

3. Demostrar las siguientes propiedades de los momentos ordinarios y centrales que se establecen en la sección 2.1.5 de la página 60:

$$\begin{array}{lll} m_{00} = 1 & m_{10} = \bar{x} & m_{01} = \bar{y} \\ \mu_{00} = 1 & \mu_{10} = 0 & \mu_{01} = 0 \end{array}$$

4. La siguiente tabla recoge los valores de fuerza ( $F$ ) y elongación ( $E$ ), registrados en 6 pruebas de tensión de acero.

$F$	1	2	3	4	5	6
$E$	15	35	41	63	77	84

Estimar el modelo lineal de regresión  $E/F$  y obtener una medida de la bondad del ajuste.

5. Representar gráficamente los datos de las muestras de las variables  $(X, Y_i)$  con  $i = 1, 2, \dots, 7$  que se proporciona en la siguiente tabla:

$X$	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	$Y_6$	$Y_7$
0	0	13	4	11	0	10	2
4	2	11	3	13	2	7	5
6	3	10	8	8	4	12	3
8	4	9	6	4	3	4	8
12	6	7	7	7	7	5	4
14	7	6	13	6	6	2	4
16	8	5	2	3	8	8	10
22	11	2	11	2	11	4	12
26	13	0	0	1	13	5	6

- a) A la vista de las gráficas, elegir, en la siguiente lista, un valor para el coeficiente de correlación lineal de cada una de las muestras anteriores y justificar la elección.

$$-1 \quad , \quad -0'875 \quad , \quad -0'543 \quad , \quad 0 \quad , \quad 0'606 \quad , \quad 0'986 \quad , \quad 1$$

- b) Calcular los coeficientes de correlación lineal y comprobar que se ha elegido correctamente.

6. Representar gráficamente, calcular la recta de regresión y determinar el grado de correlación lineal de la variable  $X$  con cada una de las variables  $Y$  que se presentan en la siguiente tabla:

$X$	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	$Y_6$	$Y_7$
1	4	1	6	1	7	6	1
2	2	3	5	1	5	4	4
3	3	5	4	3	4	3	2
4	2	7	3	5	2	1	1
5	4	9	2	6	2	5	5

7. Los siguientes datos están tomados de un estudio sobre el flujo de tráfico a través de un túnel para vehículos. Las cifras son los valores promedio basados en las observaciones que se hicieron en 10 intervalos de 5 minutos.

Densidad(veh/km)	43	55	40	52	39	33	50	33	44	21
Velocidad(km/h)	27'0	23'8	30'7	24'0	34'8	41'4	27'0	40'4	31'7	51'2

Se pide:

- a) Representar el diagrama de dispersión.
- b) A la vista del diagrama, elegir el valor correcto de  $r$  entre estos tres valores: 0'968, -0'968, -0'198.
- c) Verificar la respuesta calculando  $r$ .
- d) ¿Hay alguna evidencia real de que exista asociación entre la velocidad de los vehículos y la densidad?

8. Recordando que dos variables son linealmente incorreladas si  $r = 0$ . Se pide

- a) Justificar que 2 variables aleatorias son linealmente incorreladas si y solo si su covarianza es 0.
- b) Dados los puntos (1,0), (2,1), (4,1) y (5,a), hallar el valor de  $a$  sabiendo que las variables  $X$  e  $Y$  son incorreladas. Determinar las rectas de regresión.

9. Veamos la importancia de la representación gráfica de los datos. Las siguientes tablas presentan tres conjuntos de datos que tienen la misma correlación y la misma recta de

regresión:

$X_1$	$Y_1$	$X_2$	$Y_2$	$X_3$	$Y_3$
10	8'04	10	9'14	10	7'70
8	6'95	8	8'14	8	6'60
13	7'58	13	8'74	13	9'60
9	8'81	9	8'77	9	7'80
11	8'33	11	9'26	11	8'70
14	9'96	14	8'10	14	9'90
6	7'24	6	6'13	6	7'96
4	4'26	4	3'10	4	5'92
12	10'84	12	9'13	12	8'80
7	4'82	7	7'26	7	6'90
5	5'68	5	4'74	5	2'62

- Calcular la recta de regresión y el coeficiente de correlación lineal de cada conjunto y comprobar que son iguales.
  - Utilizar el diagrama de dispersión para representar los conjuntos de datos junto a la recta de regresión calculada.
  - ¿En qué conjunto de datos utilizarías la recta de regresión para predecir el valor de la variable  $Y$  cuando  $X = 16$ ?
10. Sea  $(X, Y)$  una variable estadística bidimensional. La variable  $X$  presenta las modalidades  $a$  y 0 mientras que la variable  $Y$  toma los valores  $a - 1$  y 1. Además, se conoce que la proporción de datos muestrales que presentan la modalidad 0 en la variable  $X$  es 0'75 y la proporción de datos muestrales que presentan la modalidad  $a - 1$  en la variable  $Y$  es 0'5. Sabiendo que la recta de regresión mínimo cuadrática de  $X$  sobre  $Y$  es  $X + Y = 1$ . Calcular:
- El coeficiente de correlación.
  - Estimar el valor de  $X$  para  $Y = 0$  y el de  $Y$  para  $X = 1$ .
11. Las rectas  $x - 2y = 4$  y  $2x - 9y = 8$  son las rectas de regresión de una variable estadística bidimensional  $(X, Y)$ , con  $N = 10$  y  $\sigma_x^2 = 9$ .
- Hallar el coeficiente de correlación lineal, la varianza de  $Y$  y la covarianza.
  - Si se descubre que uno de los puntos considerados, el  $(2, -1)$ , no debería haberse utilizado, hallar las nuevas rectas de regresión.
12. A partir de 30 observaciones de una variable estadística bidimensional  $(X, Y)$  se obtuvieron las rectas de regresión:  $X = (Y - 1)/2$  e  $X = Y - 1$ , sabiéndose que la varianza de  $X$  es 1. Más adelante se obtuvo una nueva observación que resultó ser el punto  $(0, 1)$ .
- Obtener las nuevas rectas de regresión.
  - Las varianzas residuales de ambos ajustes. ¿Han aumentado o disminuido?
  - ¿Mejoran los ajustes al tomar una nueva observación?
13. Sea una regresión lineal mínimo cuadrática del tipo  $Y/X$  obtenida a partir de  $N$  observaciones de una variable estadística bidimensional  $(X, Y)$ , con centro de gravedad el origen de coordenadas.

Con objeto de obtener más información, se realiza una nueva observación, que resulta ser de nuevo el centro de gravedad.

Ante la duda de que esta información adicional que parece reiterativa, no aporte nada nuevo, se decide realizar una nueva regresión lineal del tipo  $Y/X$  con las  $N+1$  observaciones.

- a) Estudiar si esta información es de utilidad, pues hace disminuir la varianza residual.
- b) Comprueba si aumenta o no, el coeficiente de correlación lineal.
- c) ¿ En qué tanto por ciento como máximo, disminuye la varianza residual con respecto a la inicial ?

14. Consideremos los siguientes modelos de regresión:

$$y = a \cdot e^{bx} \quad , \quad y = a \cdot x^b \quad , \quad y = \frac{1}{a + b \cdot x}$$

Para cada uno de ellos, se pide:

- a) Determinar los cambios de variable necesarios para reducir los siguientes modelos al caso lineal.
- b) Determinar las ecuaciones que permiten calcular los coeficientes del modelo original, a partir de los coeficientes del modelo lineal
- c) Determinar las restricciones que debe verificar el conjunto de datos para poder aplicar la reducción.

15. Ajustar el modelo  $y = a \cdot b^x$  (reduciéndolo al caso lineal) a los siguientes datos:

Variable X	1	2	3	4	5
Variable Y	3'0	4'5	7'0	10'0	15'0

16. Ajustar el modelo  $y = a \cdot x^b$  (reduciéndolo al caso lineal) a los siguientes datos:

Variable X	1	2	3	4	5
Variable Y	0'5	2'0	4'5	8'0	12'5

17. Ajustar el modelo  $y = \frac{1}{a + b \cdot x}$  (reduciéndolo al caso lineal) a los siguientes datos:

Variable X	1	2	3	4	5
Variable Y	1'00	0'50	0'33	0'25	0'20

18. Consideramos la muestra (1,0), (2,1), (1,2), (-1,0), (2,2) de la variable (X,Y). Se pide:

- a) Ajustar un modelo del tipo  $Y = a + b(1/X)$ .
- b) Ajustar la recta  $Y/X$ .
- c) ¿Qué modelo resulta más apropiado?

19. Dados los puntos: (1,1) , (2,1) , (3,2) , (4,4) y (5,8), se pide:

- a) Estudiar si resultaría conveniente realizar un ajuste lineal.
- b) Ajustar una función del tipo  $y = a \cdot b^x$ .

- c) Utilizar los modelos para predecir y comparar los valores  $y$  para  $x = 6$  y  $x = 10$ . A la vista de los resultados, elegir el modelo más adecuado para la predicción y justificar la respuesta.
- d) Comparar los dos modelos utilizando el coeficiente de correlación lineal y SSE.
20. Dados los puntos  $(0,0'9)$ ,  $(2,1/3)$ ,  $(3,1/7)$ ,  $(4,1/10)$  y  $(6,1/82)$  obtener los coeficientes del ajuste por transformación al modelo lineal, para una relación entre ambas variables del tipo  $y = 1/(ab^x + 1)$ .
21. Se probó el desgaste ( $d$  en  $mm.$ ) de seis moldes, probando cada uno de ellos bajo una diferente temperatura ( $t$  en unidades de  $100^\circ C$ ) de operación controlada en un baño de aceite. Los resultados de la prueba fueron:

$t$	1	1,5	2	3	3,5	4
$d$	3'3	5'0	5'5	9'4	11'4	12'8

Puede suponerse que los valores de la temperatura no tienen error y hay bases para suponer que el desgaste y la temperatura están relacionados por una función lineal. Se pide:

- a) Obtener la ecuación del modelo lineal de regresión.
- b) Estimar el desgaste cuando la temperatura de operación es  $250^\circ C$ .
- c) Elegir otro modelo de regresión que resulte más apropiado y que no contemple desgaste cuando la temperatura es de 0 grados.
22. Vamos a estudiar el movimiento uniformemente acelerado de un objeto a partir de los datos del espacio ( $e$ ) y del tiempo ( $t$ ) recogidos en la siguiente tabla:

tiempo	1	2	3	4	5	6	7
espacio	13	41	67	119	176	245	333

- a) Ajustar mediante mínimos cuadrados la expresión del espacio en función del tiempo.
- b) Estimar el espacio inicial, la velocidad inicial y la aceleración.
- c) Predecir el espacio recorrido cuando  $t = 10$ .
- d) Hallar la nueva ecuación considerando el nuevo dato  $e = 6$  para  $t = 0$ . (Observación: utilizar los cálculos anteriores).
23. El tiempo total necesario para detener un automóvil después de percibir un peligro se compone del tiempo de reacción más el tiempo de frenado. Por tanto, la velocidad del vehículo no es suficiente para calcular este tiempo total aplicando las leyes de la mecánica. Para estudiar este fenómeno se considera la siguiente tabla que contiene las distancias ( $d$  en metros) de frenada de un automóvil que marcha a la velocidad ( $v$  en  $Km/h$ ) desde el instante en que se observa el peligro.

$v$	30	45	60	75	90	105
$d$	1'30	2'25	3'50	5'20	7'50	10'50

Representar gráficamente los datos, determinar un modelo que se ajuste a la nube de puntos y utilizarlo para estimar  $d$  cuando  $v$  es  $80 Km/h$ . (interpolación) y  $120 km/h$ . (extrapolación). Estudiar las limitaciones del modelo.

24. Ajustar una recta y una parábola de regresión  $Y/X$  al conjunto de puntos  $\{(2, 3), (3, 4), (8, 9), (9, 8)\}$ . Comprobar que la parábola se ajusta mejor a los datos y justificar por qué ocurre siempre esto, independientemente del conjunto de puntos.

25. Dada la tabla:

$X \backslash Y$	0	1	2
20	2	0	0
30	1	3	2
40	1	3	2
50	2	0	0

se pide

- Ajustar un modelo lineal de regresión.
  - Calcular el coeficiente de correlación lineal y la covarianza.
  - Estudiar la dependencia e independencia de las distribuciones.
  - Ajustar una parábola de regresión y comparar la bondad del modelo con el caso lineal.
26. **Cambio de variable.** Al analizar los datos, a veces conviene aplicar una transformación que simplifique su aspecto general. La siguiente tabla muestra el contenido de oxígeno  $Y$  a los  $X$  metros de profundidad de un lago:

$X$	10	20	30	40	50	60	70
$Y$	6'5	5'6	5'4	6'0	4'6	1'4	0'1

Dar respuesta a las siguientes cuestiones:

- Aplicar el cambio de variable  $X' = (X - 40)/10$  y calcular la media de la nueva variable  $X'$ .
  - Ajustar la recta de regresión  $Y/X'$ .
  - Estudiar la correlación lineal.
  - Utilizar el modelo para predecir el contenido de oxígeno a los 65 metros.
  - Ajustar una parábola y comparar la bondad del ajuste con el modelo lineal.
27. Dada la siguiente tabla de frecuencias:

$Y \backslash X$	1	2	3	4	5
3 - 4	5	3	1		
4 - 5	1	2	1	2	
5 - 6			4	3	1
6 - 7			1	2	2
7 - 8					2

- Aplicar el cambio de variable  $W = X - 3$  y  $Z = Y - 5'5$ .
- Calcular la recta de regresión  $Z/W$ .
- Ajustar el modelo parabólico de regresión  $z = a + bw^2$ .
- Ajustar el modelo de regresión  $z = a + bw + cw^3$ .
- Comparar la bondad de los modelos utilizando una medida de correlación apropiada.

28. El número de agricultores españoles, en millones viene dado por los puntos (1973, 9'47), (1974, 9'26), (1975, 8'86), (1976, 8'25), (1977, 7'81), (1978, 8'01), (1979, 7'55), (1980, 7'24), (1981, 7'01), (1982, 6'88) y (1983, 7'03).
- Aplicar una traslación a los años para obtener una nueva variable con media 0.
  - Predecir el número de agricultores en el año 1970 suponiendo una dependencia lineal entre las variables.
  - Hallar el coeficiente de correlación lineal.
  - Ajustar una curva del tipo  $y = a \cdot b^x$  y comparar la bondad del ajuste con el modelo lineal.
29. Estudiar en qué medida le afectan los cambios de origen y de escala al coeficiente de correlación lineal.
30. Los datos que muestra el siguiente ejemplo provienen del registro del número de automóviles que salen de una población grande por la carretera principal hacia la costa en cada uno de los 10 domingos seleccionados al azar. Las observaciones se hicieron en un punto de observación sobre la carretera durante un intervalo de tiempo fijo, y para mantener los números sencillos, se expresan redondeándolos al 1000 más cercano. También se muestra la temperatura (en grados centígrados) que se registró en la población al principio del día.

$t$	13	16	9	10	18	23	19	27	15	10
$v$	18	19	9	12	21	25	26	30	24	14

Se pide:

- Representar gráficamente los datos.
  - Elegir y ajustar un modelo que permita establecer la relación que existe entre la temperatura ( $t$ ) y el número de vehículos ( $v$ ).
  - Justificar la elección del modelo del apartado anterior.
31. Algunas veces se requiere que la curva de regresión pase por el origen. En estos casos, elegimos modelos que no tengan término independiente, como en el siguiente ejercicio. Ajustar el modelo  $E = aC$  a los siguientes datos obtenidos en un experimento para determinar la rigidez de un resorte. Se midió la extensión ( $E$ ) del resorte (a partir de su longitud natural) bajo la acción de diferentes cargas ( $C$ ):

Carga (Newtons)	2	4	6	8	10	12
Extensión (mm)	10	19	29	40	48	56

32. **Regresión múltiple.** En la tabla,  $z$  representa una propiedad física particular de las barras de acero forjado, y  $x$  e  $y$  son los porcentajes de elementos  $a$  y  $b$  que se encuentran presentes en la aleación. Se escogieron cuatro niveles para  $x$  y cuatro para  $y$ , lo que da 16 posibles combinaciones, y se registró experimentalmente un valor de  $z$  para barra de cada tipo. (Este es un ejemplo de lo que se conoce como diseño factorial completo).

$x$	5	5	5	5	10	10	10	10	15	15	15	15	20	20	20	20
$y$	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
$z$	28	30	48	74	29	50	57	42	20	24	31	47	9	18	22	31

- a) Demostrar que las ecuaciones normales para el modelo lineal de regresión múltiple  $z = a + b \cdot x + c \cdot y$  en forma matricial son

$$\begin{pmatrix} n & \sum x & \sum y \\ \sum x & \sum x^2 & \sum xy \\ \sum y & \sum xy & \sum y^2 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum z \\ \sum xz \\ \sum yz \end{pmatrix}$$

- b) Resolver el sistema para deducir las fórmulas que determinan los coeficientes  $a$ ,  $b$  y  $c$  en función de los momentos:

$$a = \bar{z} - b\bar{x} - c\bar{y} \quad , \quad b = \frac{\sigma_{xz}\sigma_y^2 - \sigma_{yz}\sigma_{xy}}{\sigma_x^2\sigma_y^2 - (\sigma_{xy})^2} \quad \text{y} \quad c = \frac{\sigma_x^2\sigma_{yz} - \sigma_{xy}\sigma_{xz}}{\sigma_x^2\sigma_y^2 - (\sigma_{xy})^2}$$

- c) Utilizar las fórmulas anteriores para calcular los valores de  $a$ ,  $b$  y  $c$  a partir de los datos del experimento.
- d) La linealidad del modelo significa que hay una relación lineal entre  $z$  y  $x$  cuando  $y$  está fija, y entre  $z$  e  $y$  cuando  $x$  está fija. Sobre el diagrama de dispersión, representar las distintas rectas obtenidas al fijar los valores de  $y$  que se investigan (1, 2, 3 y 4). Hacer lo mismo para los valores de  $x$  que se investigan (5, 10, 15 y 20).
- e) Calcular el mayor valor de  $z$  estimado por el modelo dentro del intervalo de valores de  $x$  e  $y$  que se investigan. Justificar la respuesta.

33. Fórmulas de codificación.

- a) Obtener la fórmula lineal de codificación que transforme respectivamente los valores 5, 10, 15 y 20 de la variable  $x$  en los valores en -3, -1, 1 y 3 de la variable  $u$ .
- b) Análogamente, obtener la fórmula lineal de codificación que transforme los valores 1, 2, 3 y 4 de la variable  $y$  en los valores en -3, -1, 1 y 3 de la variable  $v$ .
- c) Aplicar estas fórmulas de codificación a los datos del ejercicio 32 de la página 95 y comprobar que el nuevo modelo de regresión lineal múltiple  $z = a + bu + cv$  coincide con la ecuación que relacionaba  $x$  e  $y$  con  $z$ .

34. Consideramos los datos (1,2,1), (1,4,3), (2,2,4), (2,2,5), (2,4,3), (1,4,3) y (2,4,5) de una muestra de la variable  $(x, y, z)$ . Se pide:

- a) Ajustar un plano de regresión a la nube de puntos.
- b) Ajustar un modelo del tipo  $z = a + b \cdot \ln(xy)$ .
- c) Determinar el modelo de regresión más apropiado.

35. Para determinar el modelo  $y = a + bx$ , aplicando el método de los mínimos cuadrados, tenemos que minimizar la función  $F(a, b) = \sum (y_i - a - bx_i)$ . Los valores de  $a$  y  $b$  obtenidos, resolviendo la ecuación  $\nabla F(a, b) = 0$ , son puntos críticos de la función  $F$ , pero ¿son mínimos de la función? Para ello, es necesario aplicar algún criterio de clasificación de extremos de un campo escalar. Un criterio sencillo, que podemos aplicar aquí, consiste en calcular la matriz Hessiana de  $F$  ( $\nabla^2 F(a, b)$ ) y comprobar que, tanto el elemento que ocupa la posición (1,1), como el determinante de la matriz, son números positivos. Se pide:

- a) Calcule la matriz Hessiana de  $F$  y úsela para determinar que el punto crítico obtenido, aplicando el método de los mínimos cuadrados, es un mínimo de la función  $F$ .
- b) Realice esta misma comprobación para el modelo parabólico de regresión.



## 2.6. Anexo I: Justificación de algunos resultados

En esta sección vamos a presentar la justificación de algunos de los resultados que hemos visto en este tema. Incluiremos aquellas demostraciones que utilizan resultados básicos de matemáticas o aquellas que se apoyan en los conocimientos aprendidos en otras asignaturas de matemáticas de la titulación.

Consideramos una muestra de tamaño  $N$  de una variable bidimensional  $(X, Y)$  que toma los valores  $(x_i, y_i)$  con frecuencias absolutas  $n_i$ , siendo  $N = \sum n_i$ , y frecuencias relativas  $f_i$  para todo  $i = 1, \dots, k$ .

Sea  $y = a + bx$  la recta de regresión de  $Y$  sobre  $X$  con  $b = \sigma_{xy}/\sigma_x^2$  y  $a = \bar{y} - b\bar{x}$ , y consideremos las variables  $Y^*$  (de los valores estimados) que toma los valores  $y_i^* = a + bx_i$  con frecuencias  $f_i$ , y la variable  $E$  (de los residuos) que toma los valores  $e_i = y_i - y_i^*$  con frecuencias  $f_i$ .

### 2.6.1. Descomposición de las varianzas para el modelo lineal de regresión

Vamos a demostrar que, en el caso lineal, se verifica la propiedad:  $\sigma_y^2 = \sigma_{y^*}^2 + \sigma_e^2$ . Para ello, vamos a demostrar el resultado equivalente  $\sigma_e^2 = \sigma_y^2 - \sigma_{y^*}^2$ , mediante la siguiente cadena de igualdades:

$$\begin{aligned}\sigma_e^2 &= \sum_{i=1}^N (y_i - a - bx_i)^2 f_i \stackrel{(1)}{=} \sum_{i=1}^N (y_i - \bar{y} + b\bar{x} - bx_i)^2 f_i = \sum_{i=1}^N [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 f_i = \\ &= \sum_{i=1}^N [(y_i - \bar{y})^2 - 2b(y_i - \bar{y})(x_i - \bar{x}) + b^2(x_i - \bar{x})^2] f_i = \sigma_y^2 - 2b\sigma_{xy} + b^2\sigma_x^2 \stackrel{(2)}{=} \\ &\stackrel{(2)}{=} \sigma_y^2 - 2\frac{\sigma_{xy}^2}{\sigma_x^2} + \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} \stackrel{(3)}{=} \sigma_y^2 - \sigma_{y^*}^2\end{aligned}$$

donde

$$(1) \text{ Sustitución: } a = \bar{y} - b\bar{x}$$

$$(2) \text{ Sustitución: } b = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$(3) \text{ Sustitución: } \sigma_{y^*}^2 = \frac{\sigma_{xy}^2}{\sigma_x^2} \text{ pues si } y^* = a + bx \text{ entonces } \sigma_{y^*}^2 = b^2\sigma_x^2 = \frac{\sigma_{xy}^2}{\sigma_x^4}\sigma_x^2 = \frac{\sigma_{xy}^2}{\sigma_x^2}$$

### 2.6.2. El coeficiente de correlación lineal de Pearson ( $r$ ) es un número comprendido entre -1 y 1

Veamos que se verifica la propiedad:  $\sigma_{y^*}^2 = r^2\sigma_y^2$

$$\sigma_{y^*}^2 \stackrel{(1)}{=} b^2\sigma_x^2 \stackrel{(2)}{=} \left(\frac{\text{Cov}(X, Y)}{\sigma_x^2}\right)^2 \sigma_x^2 \stackrel{(3)}{=} \left(\frac{\text{Cov}(X, Y)}{\sigma_x}\right)^2 \frac{\sigma_y^2}{\sigma_y^2} \stackrel{(2)}{=} \left(\frac{\text{Cov}(X, Y)}{\sigma_x\sigma_y}\right)^2 \sigma_y^2 \stackrel{(4)}{=} r^2\sigma_y^2$$

donde

- (1) Propiedad de la varianza frente a la transformación afín  $Y^* = a + bX$
- (2) Operar y simplificar
- (3) Multiplicar y dividir por  $\sigma_y^2$
- (4) Definición del coeficiente  $r$

Considerando el resultado anterior y la fórmula de la descomposición de la varianza obtenemos:

$$\left. \begin{array}{l} \sigma_y^2 = \sigma_{y^*}^2 + \sigma_e^2 \\ \sigma_{y^*}^2 = r^2 \sigma_y^2 \end{array} \right\} \implies \sigma_y^2 = r^2 \sigma_y^2 + \sigma_e^2 \implies \sigma_e^2 = (1 - r^2) \sigma_y^2$$

Ahora bien, como  $\sigma_e^2 \geq 0$  y  $\sigma_y^2 \geq 0$  (por definición de la varianza) entonces  $(1 - r^2) \geq 0$  y, por lo tanto,  $r^2 \leq 1$ , es decir,  $-1 \leq r \leq 1$ .

## 2.7. Anexo II: Comandos de R

```

> x=c(1,2,3,4,5)
> y=c(2,4,6,8,9)
> table(x,y)           # Tabla de doble entrada
> cov(x,y)             # Covarianza muestral (dividido por N-1)
> cor(x,y)             # Coef. Correlación lineal de Pearson

###  MODELOS DE REGRESIÓN LINEAL  ###

> reg1<-lm(y ~ x)           # Regresión lineal:  $Y = a_0 + a_1 * X$ 
> reg2<-lm(y ~ x+I(x^2)+I(x^3)) # Regresión:  $Y = a_0+a_1*X+a_2*X^2+a_3*X^3$ 

###  MODELOS DE REGRESION NO LINEAL  ###

> reg3<-nls(y ~ a*exp(b*x)   # Regresión:  $Y = a * e^{bX}$ 
> reg4<-nls(y ~ a*b^x)       # Regresión:  $Y = a * b^X$ 
> reg5<-nls(y ~ a+b*x)       # Regresión:  $Y = a + b * X$ 

###  DATOS DEL MODELO: Regresión y Correlación  ###

> reg=lm(y ~ x)
> plot(x,y);abline(reg) # Representa la nube de puntos y el modelo ajustado
> summary(reg)          # Resumen datos del modelo
> names(reg)            # Datos de la Regresión lineal almacenados en "reg"
> reg$fitted.values     # Valores estimados de "y" por el modelo
> reg$residuals         # Residuos estimados
> coef(reg)             # Coeficientes del Modelo
> resid(reg)            # Residuos del Modelo
> fitted(reg)           # Valores ajustados por el modelo

# Fórmulas para definir  $R^2$  y SSE

> 1-var(resid(reg))/var(y) # Coeficiente de determinación ( $R^2$ )
> sum(resid(reg)^2)        # Suma de los cuadrados de los residuos (SSE)

#####

> reg1<-lm(y~x)

> reg1

Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
          0.4          1.8

```

```

> summary(reg1)

Call:
lm(formula = y ~ x)
Residuals:
    1      2      3      4      5 
-2.000e-01  3.193e-16  2.000e-01  4.000e-01 -4.000e-01 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4000     0.3830   1.044 0.373021
x              1.8000     0.1155  15.588 0.000574 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3651 on 3 degrees of freedom
Multiple R-squared:  0.9878, Adjusted R-squared:  0.9837 
F-statistic: 243 on 1 and 3 DF,  p-value: 0.0005737

#####

> reg2<-nls(y~a+b*x)

> reg2

Nonlinear regression model
  model: y ~ a + b * x
 data:  parent.frame()
  a      b
0.4 1.8
residual sum-of-squares: 0.4
Number of iterations to convergence: 1
Achieved convergence tolerance: 1.251e-07

> summary(reg2)

Formula: y ~ a + b * x
Parameters:
      Estimate Std. Error t value Pr(>|t|)
a    0.4000     0.3830   1.044 0.373021
b    1.8000     0.1155  15.588 0.000574 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3651 on 3 degrees of freedom
Number of iterations to convergence: 1
Achieved convergence tolerance: 1.251e-07

```

# Apuntes de ESTADÍSTICA

## Series estadísticas



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento al profesor Carlos Cerezo Casermeiro, por sus correcciones y sugerencias en la elaboración de estos apuntes.*



## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Capítulo 3

# Series estadísticas

En un estudio estadístico, los datos de una muestra proceden de las observaciones de una variable estadística. Si estas observaciones están ordenadas y estamos interesados en estudiar su evolución entonces la muestra constituye una *serie estadística* de datos.

En este capítulo estudiaremos dos tipos de series de datos: los números índice y las series temporales. Para cada una de ellas, veremos sus características y determinaremos algunos métodos que permitan extraer la información que proporcionan las series estadísticas.

### 3.1. Números índice

Normalmente, cuando se quiere estudiar la evolución de determinados fenómenos complejos donde intervienen varias variables, uno de los mayores problemas es la forma de medir algunos agregados (sumas) que son heterogéneas (no se parecen). Dichos problemas se presentan sobre todo en el análisis de variables económicas como listas de precios, cantidades, etc. El problema de dicha medición consiste, en obtener un único número que sea descriptivo del volumen total del agregado que se quiera estudiar, o en obtener un único número que nos posibilite estudiar la evolución en el tiempo de dicho agregado. La solución a este problema se tiene mediante el uso de una técnica estadística llamada *número índice*.

Llamamos *número índice* o simplemente *índice* a una medida estadística diseñada para poner de relieve cambios en una variable o en un grupo de variables relacionadas con respecto al tiempo, situación geográfica o cualquier otra característica. Una colección de números índice para diferentes años, lugares, etc., recibe el nombre de *serie de índices*.

En el caso más sencillo, los números índice sirven para conocer la variación porcentual de una determinada magnitud en el tiempo o en el espacio. En este caso, los números índice no son otra cosa que el porcentaje de variación de cada valor de la variable con respecto a un valor de referencia llamado *periodo base* o *periodo de referencia*.

Por ejemplo, sean  $x_a$  y  $x_b$  dos valores de una variable  $X$  en dos instantes de tiempo  $a$  y  $b$ . Entonces, el cociente entre  $x_b$  y  $x_a$

$$x_{b/a} = \frac{x_b}{x_a}$$

determina un número índice, que denotaremos por  $x_{b/a}$ , y que representa la relación entre los

valores de la variable en esos dos instantes. Este número se suele multiplicar por 100 para expresarlo en tantos por ciento. Además, el instante  $a$ , que determina el denominador del cociente, se denomina *periodo base* o *de referencia*.

**Ejemplo 3.1** *Calcular e interpretar el número índice que determina la relación entre el precio del gasóleo A en febrero de 2011, que era de 1'23 euros, respecto al precio en marzo del 2009, que era de 84 céntimos de euro.*

El número índice ( $p$ ) que determina la relación entre los precios del gasóleo A en esos dos instantes de tiempo es

$$p_{2011/2009} = \frac{123}{84} \approx 1'464(146'4\%)$$

Lo que significa que el precio del gasóleo A se incrementó más de un 46 % en esos dos años.  $\square$

Con los números índice podemos comparar los costes de alimentación o de otros servicios en una ciudad durante un año con los del año anterior, o la producción de acero en un año en una zona del país con la de otra zona. Aunque se usan principalmente en economía e industria, los números índice son aplicables en muchos otros campos. En educación, por ejemplo, se pueden usar los números índices para comparar la inteligencia relativa de estudiantes en sitios diferentes o en años diferentes.

### 3.1.1. Clasificación de números índice

En función del número de variables que queramos relacionar, podemos hablar de dos tipos de números índices: números índices simples y números índices complejos. Los índices simples se refieren a una sola variable mientras que los índices complejos hacen intervenir a más de una variable.

$$\text{Índices} \left\{ \begin{array}{l} \text{Simples} \left\{ \begin{array}{l} \text{Elementales} \\ \text{En cadena} \end{array} \right. \\ \text{Complejos} \left\{ \begin{array}{l} \text{Sin ponderar} \\ \text{Ponderados} \end{array} \right. \end{array} \right.$$

Atendiendo al periodo base considerado, los índices simples pueden ser elementales o en cadena. Los índices simples elementales están referidos a un mismo periodo base, mientras que los índices simples en cadena están referidos al periodo inmediatamente anterior en la serie y, por tanto, no es fijo. En cuanto a los índices complejos podemos establecer otra clasificación atendiendo a la ponderación o no de las variables que intervienen.

### 3.1.2. Propiedades de los números índice

A continuación vamos a relacionar las propiedades más importantes que deseamos que cumpla un número índice. Para todas ellas, consideramos los índices  $x_a, x_b, x_c \dots$  expresados en tantos por 1 y correspondientes a los periodos de tiempo  $a, b, c, \dots$  respectivamente de la variable  $X$ .



1. Propiedad identidad: El índice de un periodo respecto al mismo periodo es 1, es decir,  $x_{a/a} = 1$ .
2. Propiedad de inversión temporal: Establece una relación entre los índices correspondientes a dos periodos de tiempo.

$$x_{a/b} \cdot x_{b/a} = 1 \quad \Longleftrightarrow \quad x_{a/b} = \frac{1}{x_{b/a}}$$

3. Propiedad cíclica o circular: Establece una relación entre los índices de varios periodos de tiempo encadenados.

$$x_{a/b} \cdot x_{b/c} \cdot x_{c/a} = 1 \quad , \quad x_{a/b} \cdot x_{b/c} \cdot x_{c/d} \cdot x_{d/a} = 1 \quad , \quad \dots$$

4. Propiedad cíclica o circular modificada: Establece otra relación entre los índices de varios periodos de tiempo encadenados.

$$x_{a/b} \cdot x_{b/c} = x_{a/c} \quad , \quad x_{a/b} \cdot x_{b/c} \cdot x_{c/d} = x_{a/d} \quad , \quad \dots$$

Desde un punto de vista teórico, sería deseable que los números índice verificasen estas propiedades. Si bien, los índices simples que vamos a definir cumplen todas ellas, no se conoce ningún índice complejo que verifique todas las propiedades.

## 3.2. Índices simples

Llamamos *índices simples* a los que hacen referencia a una variable concreta, es decir, a los que dan a conocer la evolución de una única variable comparándola con ella misma al tomar un periodo de tiempo como referencia o base.

Los números índices simples se calculan dividiendo el valor actual de la variable entre el valor de la variable en el tiempo utilizado como base. En función de que el tiempo considerado como base sea fijo para todos los valores o vaya cambiando, se distinguen dos tipos de números índice: *elementales* o *en cadena*. Ambos tipos verifican todas las propiedades definidas anteriormente.

En esta sección vamos a definir, calcular e interpretar estos dos tipos de índices y veremos tres ejemplos de índices elementales: las relaciones de precios, de cantidad y de valor.

### 3.2.1. Índices simples elementales (ISE)

Los índices elementales son un tipo de índices simples que responderían estrictamente a la definición como cociente de valores de la variable. En este caso se toma un único valor como periodo base o periodo de referencia y es fijo para todos los valores de la variable.

Consideramos una serie de valores  $x_0, x_1, \dots, x_k$  observados de la variable  $X$  en los instantes o periodos de tiempo  $t = 0, 1, \dots, k$ . Los números índice simples elementales se obtienen dividiendo cada uno de los valores de la variable  $X$  por el valor fijo de la variable que corresponde con el momento que se toma como base. El índice obtenido con ese cociente se multiplica por 100 para expresar el resultado en tantos por cien.

En la siguiente tabla se calculan estos índices para los distintos valores de la variable  $X$  en los periodos de tiempo ( $t$ ) correspondientes y tomando como base el instante  $t = 0$ .

tiempo $t$	0	1	2	...	$k$
variable $X$	$x_0$	$x_1$	$x_2$	...	$x_k$
ISE	1	$\frac{x_1}{x_0}$	$\frac{x_2}{x_0}$	...	$\frac{x_k}{x_0}$
ISE en %	100	$\frac{x_1}{x_0} \cdot 100$	$\frac{x_2}{x_0} \cdot 100$	...	$\frac{x_k}{x_0} \cdot 100$

El índice elemental para un periodo dado con respecto al mismo periodo es siempre 100. En particular, el número índice correspondiente al periodo base es siempre 100. Esto da cuenta de la notación (frecuente en la literatura estadística) de escribir, por ejemplo, “1969=100”, para indicar que se ha tomado 1969 como periodo base.

Si en un periodo, el número índice es mayor de 100, significa que existe un incremento del valor de la variable en ese periodo con respecto al valor de dicha variable en el periodo tomado como base. Por ejemplo, un índice de 134 %, significa que existe un incremento del 34 % respecto del periodo base. Si el número índice es menor de 100, significa que existe una disminución del valor de la variable en ese periodo con respecto al valor de dicha variable en el periodo base. Así, si este índice es 98 %, significa que existe una disminución del 2 %, siempre con respecto al periodo base.

**Ejemplo 3.2** La siguiente tabla contiene las cifras de ventas (en miles de millones) de una empresa durante los últimos cinco años de existencia de la peseta como moneda de curso legal.

Año	1997	1998	1999	2000	2001
ventas	1'5	2'4	2'4	1'8	2'7

Calcular la serie de números índice simples elementales tomando como referencia el año 1997 e interpretar los resultados.

Para calcular los índices simples elementales, dividimos las cifras de ventas de cada año entre las ventas registradas durante el año 1997 y multiplicamos por 100.

Año	1997	1998	1999	2000	2001
ISE	100	160	160	120	180

Por ejemplo, el número índice 180 correspondiente al año 2001 se ha calculado dividiendo la cifra de ventas de este año entre las ventas registradas durante el año 1997. Esto significa que en el año 2001 las ventas se han visto incrementadas en un 80 % respecto al año base 1997. Además, podemos observar que el índice para el año 1997 es 100 por ser el periodo tomado como base.

Estos índices elementales permiten conocer fácilmente que durante el lustro se ha vendido por encima de los registros obtenidos en el año 1997. Para ello, sólo necesitamos comprobar que todos los índices elementales obtenidos son mayores que 100.  $\square$

### 3.2.2. Índices simples en cadena (ISC)

Los índices en cadena son un tipo de índices simples donde el periodo base va a ir cambiando de un valor de la variable a otro. Para calcular el índice de un periodo tomaremos como base el valor de la variable en el periodo inmediatamente anterior.

Consideramos una serie de valores  $x_0, x_1, \dots, x_k$  observados de la variable  $X$  en los instantes o periodos de tiempo  $t = 0, 1, \dots, k$ . Para cada periodo  $t$ , el índice simple en cadena se obtiene dividiendo el valor de la variable en ese periodo ( $x_t$ ) por el valor de la variable en el periodo anterior ( $x_{t-1}$ ). El índice obtenido con ese cociente se multiplica por 100 para expresar el resultado en tantos por cien.

En la siguiente tabla se calculan estos índices para los distintos valores de la variable  $X$  en los periodos de tiempo ( $t$ ) correspondientes.

tiempo $t$	0	1	2	...	$k$
variable $X$	$x_0$	$x_1$	$x_2$	...	$x_k$
ISC	—	$\frac{x_1}{x_0}$	$\frac{x_2}{x_1}$	...	$\frac{x_k}{x_{k-1}}$
ISC en %	—	$\frac{x_1}{x_0} \cdot 100$	$\frac{x_2}{x_1} \cdot 100$	...	$\frac{x_k}{x_{k-1}} \cdot 100$

Obsérvese que la definición no tiene sentido para el primer valor de la serie. Además, si en un periodo, el número índice es mayor de 100, significa que existe un incremento del valor de la variable en ese periodo con respecto al valor de dicha variable en el periodo anterior. Por ejemplo, un índice de 134 %, significa que existe un incremento del 34 % respecto del periodo anterior. Si el número índice es menor de 100, significa que existe una disminución del valor de la variable en ese periodo con respecto al valor de dicha variable en el periodo anterior. Así, si este índice es 98 %, significa que existe una disminución del 2 %, siempre con respecto al periodo anterior.

**Ejemplo 3.3** *Calcular los índices simples en cadena para la serie de datos de ventas del ejemplo 3.2 de la página 106, e interpretar los resultados.*

Para calcular los índices simples en cadena, dividimos las cifras de ventas de cada año entre las ventas registradas durante el año anterior y multiplicamos por 100 %.

Año	1997	1998	1999	2000	2001
ventas	1'5	2'4	2'4	1'8	2'7
ISC	—	160	100	75	150

Por ejemplo, el número índice 150 correspondiente al año 2001 se ha calculado dividiendo la cifra de ventas de este año entre las ventas registradas durante el año 2000. Esto significa que en el año 2001 las ventas se han visto incrementadas en un 50 % respecto al año anterior. Sin embargo, el número 75 del año 2000 significa que en este año se redujeron las ventas en un 25 % respecto al año anterior. Por otro lado, el número 100 del año 1999 indica que las cifras de ventas de este año coinciden con las ventas registradas el año anterior.

Estos índices en cadena permiten estudiar la evolución de las ventas año a año. En nuestro ejemplo, resulta fácil determinar el año donde no se ha producido una evolución favorable (incremento) de las ventas que corresponde al año 2000 cuyo índice es inferior a 100.  $\square$

**Ejemplo 3.4** *La siguiente tabla contiene el consumo de petróleo en España durante la década de los 90, medido en miles de toneladas.*

Año	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Consumo	47'741	49'367	50'464	49'709	51'894	54'610	55'433	57'396	61'670	63'04

*Calcular los índices simples elementales y en cadena tomando como referencia el año 1990 e interpretar los resultados.*

Para calcular los índices simples elementales, dividimos el consumo de cada año entre el consumo del año 1990 y multiplicamos por 100 %. En la casilla correspondiente al año base colocamos 100 %.

Año	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
ISE	100'00	103'41	105'70	104'12	108'70	114'39	116'11	120'22	129'18	132'05

El número 120'22 del año 1997 significa que en este año se produjo un incremento del 20'22 % del consumo respecto al año 1990.

Para calcular los índices simples en cadena, dividimos el consumo de cada año entre el consumo del año anterior y multiplicamos por 100 %.

Año	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
ISC	—	103'41	102'22	98'50	104'40	105'23	101'51	103'54	107'45	102'22

El número 107'45 del año 1998 significa que en este año se produjo un incremento del 7'45 % del consumo respecto al año anterior. El número 98'50 del año 1993 significa que en este año se produjo un descenso del consumo de energía equivalente al 1'50 % del consumo respecto al año anterior.

Podemos observar que los índices elementales son todos mayores que 100 lo que indica que el consumo durante toda la década fue superior al consumo producido en el año 1990. Con este tipo de índice es más difícil apreciar si durante todos los años se ha producido este incremento. Sin embargo, al observar los índices en cadena se aprecia más fácilmente este fenómeno. Los años con índice inferior o superior a 100 indican una disminución o aumento respectivamente del consumo de este combustible.  $\square$

### 3.2.3. Relación de precios, cantidades y valores

Uno de los ejemplos más usuales de índice simple es lo que se conoce como *relación de precios*, que no es más que el cociente entre el precio de un artículo en un periodo dado y su precio en otro periodo (base). Además, por sencillez se supone que los precios en cada periodo son constantes, ya que en caso de no ser así, podemos tomar un promedio adecuado para el periodo de modo que la suposición sea válida.

Si  $p_a$  y  $p_b$  son los precios de un artículo durante los periodos  $a$  y  $b$  respectivamente, entonces la *relación de precios en el periodo  $b$  con respecto al periodo  $a$*  se denota por  $p_{b/a}$  y viene definida por la fórmula

$$p_{b/a} = \frac{p_b}{p_a}$$

En vez de comparar los precios de un artículo, podemos estar interesados en comparar las cantidades (o volúmenes) de producción, consumo, exportación, etc. En este caso, el número índice simple se conoce como *relación de cantidad* o *relación de volumen*.

Si  $q_a$  y  $q_b$  representan las cantidades durante los periodos  $a$  y  $b$  respectivamente, entonces la *relación de cantidad en el periodo  $b$  con respecto al periodo  $a$*  se denota por  $q_{b/a}$  y se calcula de manera análoga a los precios

$$q_{b/a} = \frac{q_b}{q_a}$$

Si  $p$  es el precio de un artículo durante un periodo y  $q$  es la cantidad (o volumen) producida, vendida, etc., durante ese mismo periodo, entonces el producto  $p \cdot q$  recibe el nombre de *valor total*. Por ejemplo, si 10 artículos se venden a 2'15 euros, el valor total es  $p \cdot q = 2'15 \cdot 10 = 21'5$  euros.

La *relación de valor* es un índice simple que permite comparar el valor total en dos periodos de tiempo. Sean  $p_a$  y  $q_a$  el precio y la cantidad de artículos registrados durante el periodo  $a$ , y  $p_b$  y  $q_b$  durante el periodo  $b$ . Ahora los valores totales durante estos periodos vienen dados por  $v_a = p_a \cdot q_a$  y  $v_b = p_b \cdot q_b$ . Definimos la *relación de valor del periodo  $b$  respecto del periodo  $a$*  como el cociente entre los valores totales en esos periodos

$$v_{b/a} = \frac{v_b}{v_a} = \frac{p_b \cdot q_b}{p_a \cdot q_a} = \left( \frac{p_b}{p_a} \right) \cdot \left( \frac{q_b}{q_a} \right) = p_{b/a} \cdot q_{b/a}$$

donde  $p_{b/a}$  y  $q_{b/a}$  son los índices de precios y cantidades expresados en tantos por 1. Es decir, la relación de valor es el producto de la relación de precios por la relación de cantidad.

**Ejemplo 3.5** La siguiente tabla contiene los precios en euros y las cantidades (en miles) producidas, de un mismo artículo, por una factoría durante la década de los 90.

Año	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Precio	1'0	1'2	1'5	1'8	2'1	2'8	3'2	2'4	2'7	3'0
Cantidad	12	14	18	18	20	15	12	16	20	24

Calcular la relación de precios, de cantidad y de valor tomando como referencia el año 1990 e interpretar los resultados.

Primero calculamos las relaciones de precios ( $p_t$ ) y de cantidad ( $q_t$ ) como índices elementales.

$t$	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
$p_t$	100	120	150	180	210	280	320	240	270	300
$q_t$	100	116'67	150	150	166'67	125	100	133'33	166'67	200

Ahora, para calcular la relación de valor ( $v_t$ ), necesitamos primero obtener las cifras del valor multiplicando el precio y la cantidad en cada periodo. Después, calculamos el índice elemental

de la serie de datos obtenida y obtenemos el resultado.

$t$	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
valor	12	16'8	27	32,4	42	42	38'4	38'4	54	72
$v_{t/0}$	100	140	225	270	350	350	320	320	450	600

El número 270 del año 1993 significa que en este año se produjo un incremento del 170 % del valor de la producción respecto al año 1990. Obsérvese que durante los años 1994 y 1995 la relación de valor fue la misma y se debió a que una disminución de la producción se compensó con un incremento de los precios. Igual ocurrió durante los años 1996 y 1997. Por último, el número 600 del año 1999 indica que el valor de la producción se ha visto multiplicado por 6 durante la década de los noventa condicionado por un incremento tanto de las cantidades como de los precios.  $\square$

### 3.3. Índices complejos

Llamamos índices complejos a los que hacen referencia a dos o más variables, es decir, a los que dan a conocer la evolución de varias variables a lo largo del tiempo comparándolas con respecto a ellas mismas, tomando un periodo de tiempo como referencia o base. Además, las variables tienen que estar relacionadas entre sí de alguna forma, ya que no podemos mezclar variables diferentes.

Existen dos tipos de índices complejos:

- **Índice complejo sin ponderar:** Se trata de construir un índice complejo a partir de índices simples, dándole a todos la misma importancia.
- **Índice complejo ponderado:** Se trata de construir un índice complejo a partir de índices simples, dándole distinta importancia o peso a cada uno de ellos.

En lo que sigue se considerarán  $n$  variables  $X_1, X_2, \dots, X_n$  que toman valores en  $k$  instantes o periodos de tiempo  $t$  como se recoge en la siguiente tabla

$t$	$X_1$	$X_2$	$\dots$	$X_n$
0	$x_{10}$	$x_{20}$	$\dots$	$x_{n0}$
1	$x_{11}$	$x_{21}$	$\dots$	$x_{n1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$k$	$x_{1k}$	$x_{2k}$	$\dots$	$x_{nk}$

donde  $x_{i,t}$  representa el valor de la variable  $X_i$  en el tiempo  $t$ .

Como hemos de llegar a un solo número índice resumiendo una gran cantidad de información, es fácil comprender que los promedios (media aritmética, media geométrica, mediana, etc.) juegan un papel importante en el cálculo de números índices. Así como existen muchos métodos para calcular promedios, también hay muchos para calcular los números índices, cada uno con sus ventajas y desventajas propias.

### 3.3.1. Índices complejos sin ponderar

Veamos dos métodos para calcular índices complejos sin ponderar.

#### Método de agregación simple

En este método de calcular un índice, expresamos el valor total de la variable en el tiempo dado como porcentaje del valor total de las variables en el tiempo base. Es decir, para cada tiempo  $t$  definimos el índice:

$$I_t = \frac{\sum_{i=1}^n x_{it}}{\sum_{i=1}^n x_{i0}} \cdot 100 \quad t = 0, 1, 2, \dots, k$$

que recibe el nombre de *índice de Bradstreet y Dutot*

Aunque este método es fácil de aplicar, tiene dos grandes desventajas que lo convierten en insatisfactorio. Por un lado, no tiene en cuenta la importancia relativa de las distintas variables (no es ponderado). Así pues, por ejemplo asignaría igual peso a la leche que a la crema de afeitar a la hora de calcular el IPC. Por otro lado, las unidades escogidas al anotar los valores de la variable afectan al índice.

#### Método del promedio simple

El índice producido por este método depende del procedimiento utilizado para promediar las relaciones de precios; los procedimientos incluyen la media aritmética, la geométrica, la armónica y la mediana. Por ejemplo, si consideramos la media aritmética, el índice correspondiente al tiempo  $t$  respecto al base  $t = 0$  es:

$$I_t = \frac{\sum_{i=1}^n \frac{x_{it}}{x_{i0}}}{n} \cdot 100 \quad t = 0, 1, 2, \dots, k$$

y recibe el nombre de *índice de Sanerbeck*.

Si bien este método no se ve afectado por la unidad de medida elegida, conserva aún la desventaja citada de dar la misma importancia a todas las variables.

**Ejemplo 3.6** La siguiente tabla recoge los valores de las variables  $X_1, X_2, \dots, X_6$  en 5 instantes de tiempo ( $t$ ).

$t$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
0	2	22	7	8	7	8
1	3	25	9	13	8	9
2	3	27	10	15	9	9
3	4	28	11	18	11	10
4	4	30	11	22	12	11

Calcular los índices complejos sin ponderar por el método de agregación simple y por el método del promedio simple.

Para aplicar el método de agregación simple, calculamos la suma o agregado para cada periodo de tiempo y, a partir de ella, calculamos los números índice tomando como base el periodo 0.

$t$	Agregado	Índice
0	54	100
1	67	124'1
2	73	135'2
3	82	151'9
4	90	166'7

Para aplicar el método del promedio simple utilizando la media aritmética, calculamos las series de índices para cada variable.

$t$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	Índice
0	100	100	100	100	100	100	100
1	150	113'6	128'6	162'5	114'3	112'5	130'2
2	150	122'7	142'9	187'5	128'6	112'5	140'7
3	200	127'3	157'1	225	157'1	125	165'2
4	200	136'4	157'1	275	171'4	137'5	179'6

La última columna contiene, para cada periodo de tiempo, la media aritmética de los índices correspondientes de las variables.  $\square$

### 3.3.2. Índices complejos ponderados

Con el fin de evitar las desventajas del método de agregación simple, asignamos un peso  $w_i$  a cada variable  $X_i$ . Tales pesos indican la importancia de la variable en cuestión.

#### Método de agregación ponderada

Este método generaliza al método de agregación simple y se utiliza si las variables son homogéneas. Para cada tiempo  $t$  definimos el índice así:

$$I_t = \frac{\sum_{i=1}^n x_{it} w_i}{\sum_{i=1}^n x_{i0} w_i} \cdot 100 \quad t = 0, 1, 2, \dots, k$$

#### Método del promedio ponderado

Este método generaliza al método del promedio simple utilizado en los índices complejos sin ponderar y se utiliza si las variables no son homogéneas. El promedio ponderado más utilizado



es la media aritmética ponderada, aunque también se utilizan otros, como la media geométrica ponderada. Para cada tiempo  $t$  definimos el índice así:

$$I_t = \frac{\sum_{i=1}^n \frac{x_{it}}{x_{i0}} \cdot w_i}{\sum_{i=1}^n w_i} \cdot 100 \quad t = 0, 1, 2, \dots, k$$

**Ejemplo 3.7** La siguiente tabla recoge los valores de  $X_1$ ,  $X_2$  y  $X_3$  en 5 instantes de tiempo ( $t$ ):

$t$	$X_1$	$X_2$	$X_3$
0	8	7	8
1	13	8	9
2	15	9	9
3	18	11	10
4	22	12	11

Calcule los índices complejos por el método de agregación ponderada y por el método del promedio ponderado, sabiendo que a la variable  $X_1$  le asignamos el doble de importancia que al resto de variables.

La ponderación asignada es de 2, 1 y 1 respectivamente para las variables  $X_1$ ,  $X_2$  y  $X_3$ .

Para aplicar el método de agregación ponderada, calculamos la suma o agregado ponderado para cada periodo de tiempo y, a partir de ella, calculamos los números índice tomando como base el periodo 0. Por ejemplo, para el periodo 2, el agregado  $48 = 2 \cdot 15 + 1 \cdot 9 + 1 \cdot 9$  y el índice  $154'8 = (48/31) \cdot 100$ .

Para aplicar el método del promedio ponderado utilizando la media aritmética ponderada, calculamos las series de índices para cada variable. La última columna contiene, para cada periodo de tiempo, la media aritmética ponderada de los índices correspondientes de las variables. Por ejemplo, para el periodo 2, el índice  $154 = (2 \cdot 187'5 + 1 \cdot 128'6 + 1 \cdot 112'5)/4$ .

Agregación ponderado		
$t$	Agregado	Índice
0	31	100
1	43	138'7
2	48	154'8
3	57	183'9
4	67	216'1

Promedio ponderado				
$t$	$I_1$	$I_2$	$I_3$	Índice
0	100	100	100	100
1	162'5	114'3	112'5	138
2	187'5	128'6	112'5	154
3	225	157'1	125	183
4	275	171'4	137'5	214'7

□

### 3.3.3. Índices de precios

Los índices de precios son los tipos de índices complejos ponderados más empleados en las actividades económicas e industriales. Consideran que las variables  $X_i$ , con  $i = 1, \dots, n$  son los

precios de los artículos cuyos valores en el periodo  $t$  se denotan por  $p_{it}$ . A cada relación de precios asignamos un peso dado por el valor total del artículo en términos de alguna unidad monetaria. Como el valor de un artículo se obtiene multiplicando su precio  $p$  por la cantidad  $q$ , los pesos vienen dados por  $w = p \cdot q$ .

En las fórmulas que aparecen a continuación, las sumas se extienden a los valores de todas las variables para un tiempo  $t$ , es decir, la expresión  $\sum$  equivale a  $\sum_{i=1}^n$ .

Según el valor del artículo considerado se distinguen tres índices:

- El índice de Laspeyres (método del año base) es un índice complejo ponderado que utiliza como ponderación, el valor, a precios del periodo base, de la cantidad en dicho periodo, es decir,  $w_i = p_{i0} \cdot q_{i0}$

$$L_t = \frac{\sum \frac{p_{it}}{p_{i0}} \cdot p_{i0} q_{i0}}{\sum p_{i0} q_{i0}} = \frac{\sum p_{it} \cdot q_{i0}}{\sum p_{i0} \cdot q_{i0}}$$

- El índice de Paasche (método del año dado) es un índice complejo ponderado que utiliza como ponderación, el valor, a precios del periodo base, de la cantidad del periodo actual, es decir,  $w_i = p_{i0} \cdot q_{it}$

$$P_t = \frac{\sum \frac{p_{it}}{p_{i0}} p_{i0} q_{it}}{\sum p_{i0} q_{it}} = \frac{\sum p_{it} \cdot q_{it}}{\sum p_{i0} \cdot q_{it}}$$

- El índice de Marshall-Edgeworth es un índice complejo ponderado que, a diferencia de los anteriores, utiliza como ponderación la media aritmética de las cantidades del año base y del año dado, es decir,  $w_i = p_{i0} \cdot (q_{i0} + q_{it})/2$

$$M_t = \frac{\sum p_{it} \cdot (q_{i0} + q_{it})}{\sum p_{i0} \cdot (q_{i0} + q_{it})}$$

Si utilizamos los índices de Laspeyres y de Paasche dados anteriormente obtenemos que

- El índice ideal de Fisher es la media geométrica de los índices de Laspeyres y de Paasche

$$F_t = \sqrt{L_t \cdot P_t} = \sqrt{\left( \frac{\sum p_{it} \cdot q_{i0}}{\sum p_{i0} \cdot q_{i0}} \right) \cdot \left( \frac{\sum p_{it} \cdot q_{it}}{\sum p_{i0} \cdot q_{it}} \right)}$$

El índice ideal de Fisher, que en particular verifica el criterio de inversión temporal y el de inversión de factores, es mejor que cualquier otro número índice útil en cuanto a satisfacer las propiedades consideradas importantes (de ahí el apelativo de “ideal”). No obstante, desde una perspectiva práctica, también sirven y se utilizan con mucha frecuencia los otros índices que hemos definido.

**Ejemplo 3.8** Calcule los índices de precios correspondientes a los datos que aparecen en la siguientes tabla, tomando 2000 como año base:

$t$	$A$		$B$		$C$	
	<i>Precios</i>	<i>Cantidades</i>	<i>Precios</i>	<i>Cantidades</i>	<i>Precios</i>	<i>Cantidades</i>
2000	2	8	3	5	1	3
2001	3	7	4	6	2	3
2002	3	10	5	6	2	5
2003	3	12	7	7	4	8
2004	4	11	8	8	5	10

Aplicamos las fórmulas como se indica en las siguientes tablas:

$$\text{Índice de Laspeyres: } L_t = \frac{\sum p_{it} \cdot q_{i0}}{\sum p_{i0} \cdot q_{i0}}$$

$$\text{Índice de Paasche: } P_t = \frac{\sum p_{it} \cdot q_{it}}{\sum p_{i0} \cdot q_{it}}$$

$t$	cálculos	$L_t$
2000	$\frac{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3}{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3} \cdot 100 = 100$	100
2001	$\frac{3 \cdot 8 + 4 \cdot 5 + 2 \cdot 3}{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3} \cdot 100 = 147'1$	147'1
2002	$\frac{3 \cdot 8 + 5 \cdot 5 + 2 \cdot 3}{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3} \cdot 100 = 161'8$	161'8
2003	$\frac{3 \cdot 8 + 7 \cdot 5 + 4 \cdot 3}{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3} \cdot 100 = 208'8$	208'8
2004	$\frac{4 \cdot 8 + 8 \cdot 5 + 5 \cdot 3}{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3} \cdot 100 = 255'9$	255'9

$t$	cálculos	$P_t$
2000	$\frac{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3}{2 \cdot 8 + 3 \cdot 5 + 1 \cdot 3} \cdot 100 = 100$	100
2001	$\frac{3 \cdot 7 + 4 \cdot 6 + 2 \cdot 3}{2 \cdot 7 + 3 \cdot 6 + 1 \cdot 3} \cdot 100 = 145'7$	145'7
2002	$\frac{3 \cdot 10 + 5 \cdot 6 + 2 \cdot 5}{2 \cdot 10 + 3 \cdot 6 + 1 \cdot 5} \cdot 100 = 162'8$	162'8
2003	$\frac{3 \cdot 12 + 7 \cdot 7 + 4 \cdot 8}{2 \cdot 12 + 3 \cdot 7 + 1 \cdot 8} \cdot 100 = 220'8$	220'8
2004	$\frac{4 \cdot 11 + 8 \cdot 8 + 5 \cdot 10}{2 \cdot 11 + 3 \cdot 8 + 1 \cdot 10} \cdot 100 = 282'1$	282'1

$$\text{I. de Marshall-Edgeworth: } M_t = \frac{\sum p_{it} \cdot (q_{i0} + q_{it})}{\sum p_{i0} \cdot (q_{i0} + q_{it})}$$

Índice de Fisher:

$$F_t = \sqrt{L_t \cdot P_t}$$

$t$	cálculos	$M_t$
2000	$\frac{2 \cdot (8+8) + 3 \cdot (5+5) + 1 \cdot (3+3)}{2 \cdot (8+8) + 3 \cdot (5+5) + 1 \cdot (3+3)} \cdot 100 = 100$	100
2001	$\frac{3 \cdot (8+7) + 4 \cdot (5+6) + 2 \cdot (3+3)}{2 \cdot (8+7) + 3 \cdot (5+6) + 1 \cdot (3+3)} \cdot 100 = 146'4$	146'4
2002	$\frac{3 \cdot (8+10) + 5 \cdot (5+6) + 2 \cdot (3+5)}{2 \cdot (8+10) + 3 \cdot (5+6) + 1 \cdot (3+5)} \cdot 100 = 162'3$	162'3
2003	$\frac{3 \cdot (8+12) + 7 \cdot (5+7) + 4 \cdot (3+8)}{2 \cdot (8+12) + 3 \cdot (5+7) + 1 \cdot (3+8)} \cdot 100 = 216'1$	216'1
2004	$\frac{4 \cdot (8+11) + 8 \cdot (5+8) + 5 \cdot (3+10)}{2 \cdot (8+11) + 3 \cdot (5+8) + 1 \cdot (3+10)} \cdot 100 = 272'2$	272'2

$t$	cálculos	$F_t$
2000	$\sqrt{100 \cdot 100} = 100$	100
2001	$\sqrt{147'1 \cdot 145'7} = 146'4$	146'4
2002	$\sqrt{161'8 \cdot 162'8} = 162'3$	162'3
2003	$\sqrt{208'8 \cdot 220'8} = 214'7$	214'7
2004	$\sqrt{255'9 \cdot 282'1} = 268'7$	268'7

□

Por último, queremos hacer notar que todos los índices de precios de esta sección se pueden definir análogamente para cantidades y obtener los índices de cantidades de Laspeyres, Paasche, Marshall-Edgeworth o Fisher. Las fórmulas que hemos mostrado para calcular los índices de precios son válidas para obtener estos índices de cantidades sin más que cambiar los valores de los precios por los de las cantidades y viceversa.

### Índice de precios al consumo

Uno de los índices de Laspeyres más conocidos es el llamado *índice del coste de la vida* o *índice de precios al consumo*, más conocido como IPC. En este índice que elabora el Instituto Nacional de Estadística (INE), los precios están ponderados por las cantidades, y la ponderación son las cantidades consumidas por la población.

La importancia de este índice está en su significado y sus implicaciones sociales. Pensemos que, por ejemplo, en muchos contratos aparecen ciertas cláusulas de revisión salarial que producen aumentos anuales automáticos en correspondencia con los aumentos del índice de precios al consumo.

Con este índice, estamos interesados en comparar precios, cantidades o valores de grandes grupos de artículos. Por ejemplo, al calcular un índice de precios al consumo no sólo queremos comparar los precios de la leche en dos periodos, sino también los precios de los huevos, de la carne, del calzado, de la vivienda, etc., de modo que se consiga una visión general de la evolución de los precios. Naturalmente, podríamos simplemente hacer una lista con todos estos precios, pero eso no sería muy satisfactorio. Lo deseable es disponer de un solo número índice que compare los precios en ambos periodos en promedio.

No es difícil ver que los cálculos de números índice que afecten a un grupo de artículos conllevan muchos problemas que hay que solventar. Por ejemplo, debemos decidir qué artículos o servicios deben incluirse, así como su peso de importancia relativa; hemos de recolectar datos referentes a precios y cantidades de tales artículos; hemos de decidir que hacer con las distintas calidades dentro de un mismo artículo, o con ciertos artículos o servicios que están disponibles en un año pero no en el año base; por fin, hemos de decidir cómo reunir toda esa información y sacar un sólo número índice del coste de la vida que tenga significado práctico.

### 3.4. Series de números índice

Como vimos en la primera sección, la colección de números índice correspondientes a los valores de una variable constituyen una serie de números índice. Una utilización directa de las series de índices consiste en analizar las variaciones o fluctuaciones de una variable o de un conjunto de variables en un periodo de tiempo.

Las principales características de estas series son las variables que intervienen, sus ponderaciones y el periodo considerado como base. En esta sección vamos a estudiar cómo se obtienen nuevas series de índices cuando modificamos alguna de sus características y como se relacionan entre sí.

Además, veremos una aplicación de las series de índices para eliminar la influencia de unas variables sobre otras en un proceso que se denomina *deflación*.

#### 3.4.1. Cambio de periodo base

Una serie de números índice se calcula a partir de los valores observados temporalmente en una variable, tomando uno de ellos como periodo base. En la práctica es deseable que el período base elegido para la comparación sea un periodo de estabilidad no muy alejado en el pasado. Por tanto, de cuando en cuando puede ser necesario cambiar el periodo base.

Una posibilidad es recalcular todos los números índice en términos del nuevo periodo base aunque para ello es necesario disponer de los valores de la variable. Un método aproximado más simple consiste en dividir todos los números índice para los diversos años correspondientes al periodo base antiguo por el número índice correspondiente al nuevo periodo base, expresando los resultados como porcentajes. Estos resultados representan los nuevos números índices, siendo el

número índice para el nuevo periodo base 100.

Sea  $I_{t/0}$  el número índice correspondiente al periodo  $t$  tomando como base el periodo 0. Si queremos cambiar de base considerando un nuevo periodo  $a$ , aplicamos la fórmula:

$$I_{t/a} = \frac{I_{t/0}}{I_{a/0}}$$

para calcular el nuevo índice correspondiente al periodo  $t$ . Matemáticamente hablando, este método es estrictamente aplicable sólo si los números índice satisfacen el criterio circular. Sin embargo, para muchos tipos de índices, el método afortunadamente da resultados que en la práctica son suficientemente próximos a los que se obtendrían teóricamente.

**Ejemplo 3.9** *Consideremos los datos del ejemplo 3.8 de la página 114. Recalcular el índice de Paasche tomando 2002 como año base utilizando los dos procedimientos.*

El primer procedimiento ( $I_{2002}$ ) consiste en volver a calcular todos los índices de Paasche igual que se hizo en el ejemplo 3.8 pero tomando como base el año 2002. Para ello, será necesario disponer de los datos originales. El segundo procedimiento (última columna de la tabla) es más sencillo y se calcula aplicando una simple regla de tres a los índices  $I_{2000}$  ya conocidos.

t	$I_{2000}$	$I_{2002}$	$I_{t/2002}$
2000	100	$\frac{2.8+3.5+1.3}{3.8+5.5+2.3} \cdot 100 = 61'8$	$\frac{100}{162'8} \cdot 100 = 61'4$
2001	145'7	$\frac{3.7+4.6+2.3}{3.7+5.6+2.3} \cdot 100 = 89'5$	$\frac{145'7}{162'8} \cdot 100 = 89'5$
2002	162'8	$\frac{3.10+5.6+2.5}{3.10+5.6+2.5} \cdot 100 = 100$	$\frac{162'8}{162'8} \cdot 100 = 100$
2003	220'7	$\frac{3.12+7.7+4.8}{3.12+5.7+2.8} \cdot 100 = 134'5$	$\frac{220'7}{162'8} \cdot 100 = 135'6$
2004	282'1	$\frac{4.11+8.8+5.10}{3.11+5.8+2.10} \cdot 100 = 169'9$	$\frac{282'1}{162'8} \cdot 100 = 173'3$

□

### 3.4.2. Renovación y empalme

Las principales características de los índices complejos ponderados son el periodo base y los pesos asignados a cada variable. Con el fin de que estos indicadores sean lo más representativos posible de la realidad, conviene de vez en cuando, revisar las variables que intervienen y los pesos asignados a las mismas. A partir de ese momento, vamos a calcular una nueva serie con parámetros distintos y a relacionarla con la anterior.

El proceso de *renovación* consiste en obtener esta nueva serie de números índices, a partir de los mismos valores de la variable pero cambiando los pesos asignados a las variables. Para ello, volvemos a aplicar las mismas fórmulas utilizando las nuevas características de la serie.

El proceso de *empalme* consiste en relacionar ambas series truncadas en el periodo de renovación. Para ello, aplicamos un cambio de base a la serie antigua, tomando como periodo base el periodo de renovación.

**Ejemplo 3.10** *Utilizando los datos del ejemplo 3.8 de la página 114, renovar el índice de Paasche tomando como nuevo año base el 2002 y efectuar el empalme correspondiente.*

Primero se calculan los índices de Paasche tomando 2002 como nuevo año base. Para ello será necesario disponer de la tabla de datos del ejemplo 3.8. Después se calculan los índices de empalme aplicando una simple regla de tres a partir del índice antiguo y el nuevo para el año 2002. Por último se toman los índices de empalme para los años anteriores a 2002 y los nuevos índices de Paasche para los años posteriores a 2002.

$t$	$I_{2000}$	Renovación	Empalme	$I_{2002}$
2000	100	— — —	$\frac{100}{162'8} 100 = 61'4$	61'4
2001	145'7	— — —	$\frac{100}{162'8} 145'7 = 89'5$	89'5
2002	162'8	$\frac{3 \cdot 10 + 5 \cdot 6 + 2 \cdot 5}{3 \cdot 10 + 5 \cdot 6 + 2 \cdot 5} \cdot 100 = 100$	$\frac{100}{162'8} 162'8 = 100$	100
2003	— — —	$\frac{3 \cdot 12 + 7 \cdot 7 + 4 \cdot 8}{3 \cdot 12 + 5 \cdot 7 + 2 \cdot 8} \cdot 100 = 134'5$	— — —	134'5
2004	— — —	$\frac{4 \cdot 11 + 8 \cdot 8 + 5 \cdot 10}{3 \cdot 11 + 5 \cdot 8 + 2 \cdot 10} \cdot 100 = 169'9$	— — —	169'9

□

### 3.4.3. Deflación de series estadísticas

Como vimos, el producto del precio de un artículo por su cantidad da lugar a una cifra que tiene carácter de valor. Por lo tanto, el valor  $v_t$  de un conjunto de  $n$  artículos distintos en un periodo  $t$  viene determinado

$$v_t = \sum_{i=1}^n p_{it} \cdot q_{it}$$

siendo  $p_{it}$  y  $q_{it}$  el precio y la cantidad del artículo  $i$  en el periodo  $t$ .

Los índices simples elementales para los valores  $v_t$  se denominan *números índice de valor* y determinan una serie de índices conocida como serie de valor. Podemos comprobar que el índice de precios de Laspeyres ( $L^P$ ) por el índice de cantidades de Paasche ( $P^Q$ ) da lugar al índice de valor

$$L_t^P \cdot P_t^Q = \frac{\sum p_{it} q_{i0}}{\sum p_{i0} q_{i0}} \cdot \frac{\sum q_{it} p_{it}}{\sum q_{i0} p_{it}} = \frac{\sum p_{it} q_{it}}{\sum p_{i0} q_{i0}} = \frac{v_t}{v_0} = v_{t/0}$$

y, de la misma manera, también se puede calcular este índice de valor como el producto del índice de precios de Paasche ( $P^P$ ) por el índice de cantidades de Laspeyres ( $L^Q$ )

$$P_t^P \cdot L_t^Q = \frac{\sum p_{it} q_{it}}{\sum p_{i0} q_{it}} \cdot \frac{\sum q_{it} p_{i0}}{\sum q_{i0} p_{i0}} = \frac{\sum p_{it} q_{it}}{\sum p_{i0} q_{i0}} = \frac{v_t}{v_0} = v_{t/0}$$

Estas series cronológicas de valor se refieren a las variaciones en el tiempo de cifras monetarias que están sujetas a las fluctuaciones del poder adquisitivo de la moneda. Por ejemplo, aunque los ingresos de una familia pueden estar creciendo teóricamente durante un cierto número de años, sus ingresos reales pueden en verdad estar disminuyendo debido al aumento del coste de la vida, en tanto en cuanto este aumento del coste de la vida hace que disminuya su poder adquisitivo.

Denominaremos *valor nominal*, *aparente* o *corriente* a las cifras monetarias observadas y *valor real* o *constante* a las cifras corregidas convenientemente para eliminar la influencia de la depreciación monetaria. La operación de convertir valores nominales en valores reales recibe el nombre de *deflación*. En otras palabras, la deflación consiste en eliminar el efecto de la inflación.

Para deflactar hay que tener en cuenta que lo que se persigue es obtener valoraciones en términos reales, es decir, la valoración a lo largo del tiempo en euros del periodo tomado como

base, porque cuando analizamos una serie estadística en términos de valores nominales podemos estar sobrevalorando o infravalorando las fluctuaciones que tiene la variable o conjunto de variables, puesto que en términos aparentes se está recogiendo la influencia de la inflación.

Por tanto, nuestra intención es pasar de una serie de valores nominales a la serie de valores reales, es decir, con precios constantes e iguales a los correspondientes al año que se toma como base:

$$\begin{array}{ccc}
 \text{Valores nominales} & & \text{Valores reales} \\
 \hline
 \begin{array}{c} \sum p_{i0}q_{i0} \\ \sum p_{i1}q_{i1} \\ \sum p_{i2}q_{i2} \\ \vdots \\ \sum p_{ik}q_{ik} \end{array} & \Rightarrow & \begin{array}{c} \sum p_{i0}q_{i0} \\ \sum p_{i0}q_{i1} \\ \sum p_{i0}q_{i2} \\ \vdots \\ \sum p_{i0}q_{ik} \end{array}
 \end{array}$$

Para obtener los valores reales a partir de los valores nominales, basta dividir éstos por el índice de precios de Paasche. Sin embargo, es más común utilizar el *índice del coste de la vida* o *índice de precios al consumo* (IPC), que prepara el Instituto Nacional de Estadística.

Por tanto, el IPC se utiliza para eliminar la influencia de los precios en una serie que esté valorada en términos monetarios. Lo que hacemos es calcular el valor real, dividiendo el valor nominal de cada año por el número índice del coste de la vida, IPC, usando un periodo base adecuado, es decir:

$$\text{valor real} = \frac{\text{valor nominal}}{\text{IPC}} \cdot 100$$

tomando el IPC en tanto por ciento. Con esta fórmula se obtiene el valor real de una cantidad en unidades monetarias del año base considerado en el IPC. Por ejemplo, si el IPC corresponde al instante  $t$  respecto del instante  $a$  tomado como base ( $I_{t/a}$ ), entonces el valor real representará al valor nominal en el instante  $t$  en unidades monetarias del instante  $a$ .

**Ejemplo 3.11** Si el sueldo de un obrero ha crecido un 50 % en 10 años (1970-1980) y en ese mismo periodo el IPC se ha doblado, ¿cuánto ha crecido realmente el sueldo del obrero?

Si el sueldo de un individuo en 1980 es el 150 % de su sueldo en 1970 (o sea, han crecido un 50 %), y el coste de la vida se ha doblado en ese mismo periodo de tiempo ( $\text{IPC}_{1980/1970}=200$ ), entonces su sueldo real en 1980 en pesetas de 1970 se calcula así:

$$\text{Valor real}_{1970} = \frac{\text{Valor nominal}_{1980}}{\text{IPC}_{1980/1970}} \cdot 100 = \frac{150}{200} \cdot 100 = 75 \%$$

Por lo tanto, aunque aparentemente cobra un 50 % más de sueldo, realmente cobra un 25 % menos y, es decir, cobra más pero ha perdido poder adquisitivo.  $\square$

En el ejemplo anterior hemos utilizado el IPC para calcular una cantidad (sueldo) correspondiente a un año (1980) en términos de unidades monetarias de un año anterior (1970). Veamos otro ejemplo donde realizamos esta comparación, pero respecto a un año posterior.

**Ejemplo 3.12** Un SEAT 600 en el año 1960 costaba unas 65.000 ptas. (390'66 €). Se disponen de los siguientes datos del IPC:

$$I_{60/92} = 4'956 \% \qquad I_{01/92} = 136'584 \% \qquad I_{06/01} = 118'337 \%$$

Calcule el precio real que hubiese costado comprar el coche en el año 2006.

En este ejemplo queremos calcular el precio en pesetas del año 2006 de un artículo que fue comprado en 1960. Para ello, primero calculamos el IPC correspondiente al periodo 2006-1960 a partir de los índices disponibles:

$$I_{60/06} = I_{60/92} \cdot I_{92/01} \cdot I_{01/06} = \frac{I_{60/92}}{I_{01/92} \cdot I_{01/06}} = \frac{0'04956}{1'365844 \cdot 1'18337} = 0'03066(3'066 \%)$$

Ahora, para calcular el precio real del coche en el año 2006, dividimos su valor nominal entre el IPC del periodo que hemos calculado:

$$\text{Precio real (2006)} = \frac{\text{Precio nominal (1960)}}{I_{60/06}} = \frac{65.000}{3'066} \cdot 100 = 2.120.026 \text{ ptas.}$$

Y por lo tanto, un SEAT 600 que costase 65.000 ptas. en el año 1960 hubiese costado en 2006 más de dos millones de pesetas, exactamente 2.120.026 ptas. (12.741'61€).  $\square$

Por último, utilizaremos el IPC para deflactar una serie cronológica y poder comparar cantidades.

**Ejemplo 3.13** Deflactar la serie cronológica de las indemnizaciones totales (miles de pesetas), abonadas en España por las compañías de seguros, durante el periodo 1956-1960, tomando como deflacionador el índice del coste de la vida. Utilice los resultados para comparar las cantidades.

Año	Indemnizaciones	IPC <sub>1936=100</sub>
1956	318.511	643'1
1957	523.926	712'4
1958	670.718	807'7
1959	905.661	866'7
1960	1.036.129	876'9

En primer lugar cambiamos de base la serie de índices tomando 1956 como año base y después usamos el nuevo IPC para calcular las cantidades pagadas por las compañías en pesetas de 1956 lo que nos permitirá comparar unos años con otros.

Año	IPC <sub>1956=100</sub>	Siniestros deflacionados
1956	100	$\frac{318.511}{100} \cdot 100 = 318.511$
1957	$\frac{712'4}{643'1} \cdot 100 = 110'8$	$\frac{523.926}{110'8} \cdot 100 = 472.857$
1958	$\frac{807'7}{643'1} \cdot 100 = 125'6$	$\frac{670.718}{125'6} \cdot 100 = 534.011$
1959	$\frac{866'7}{643'1} \cdot 100 = 134'8$	$\frac{905.661}{134'8} \cdot 100 = 671.855$
1960	$\frac{876'9}{643'1} \cdot 100 = 136'4$	$\frac{1.036.129}{136'4} \cdot 100 = 759.625$

Observemos cómo podemos utilizar estos resultados para comparar cantidades. Tomando los datos de la primera tabla, observamos que las cantidades pagadas por siniestros en 1956 se ven duplicadas en 1958 y triplicadas al año siguiente, en 1959. Sin embargo, una vez deflacionada la serie, observamos que la cantidad correspondiente al año 1956 no se duplica hasta 1959 y no se llega a triplicar en todo el periodo.  $\square$



### 3.5. Series Temporales o Cronológicas

Una serie temporal es un conjunto de observaciones tomadas en instantes específicos, generalmente a intervalos iguales. Es decir, es una variable estadística bidimensional donde una variable es el tiempo (variable independiente) y la otra corresponde al fenómeno cuantitativo que se quiere estudiar (variable dependiente). Por ejemplo, la cotización diaria al cierre de la sesión bursátil de ciertas acciones, producción de leche en unos años o las temperaturas cada hora por el Instituto Metereológico de una ciudad.

#### 3.5.1. Representación gráfica

La representación gráfica de una serie temporal se realiza mediante un diagrama de dispersión, donde el tiempo se representa en el eje  $X$  y la variable, objeto de estudio, se representa en el eje  $Y$ .

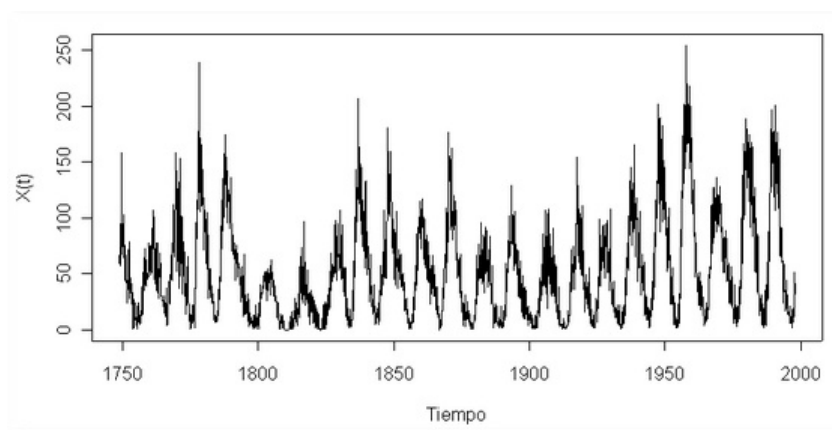


Figura 3.1: Serie Temporal

En la gráfica<sup>1</sup> de la figura 3.1 se representa una serie temporal  $(x(t))$  mediante un diagrama de dispersión que permite observar el comportamiento de dicha serie a lo largo del tiempo (años). El objetivo será poder predecir el comportamiento de esta serie temporal en un futuro no muy lejano.

#### 3.5.2. Promedios o Medias Móviles

Dado un conjunto de números  $y_1, y_2, \dots, y_N$ , llamamos *promedio o media móvil de orden  $k$* , a la siguiente sucesión de medias aritméticas:

$$\frac{y_1 + \dots + y_k}{k}, \quad \frac{y_2 + \dots + y_{k+1}}{k}, \quad \dots, \quad \frac{y_{N-k+1} + \dots + y_N}{k}$$

Si los datos se dan anual o mensualmente, se llama media móvil de  $k$  años o de  $k$  meses.

<sup>1</sup>Fuente: <http://www.seh-lelha.org/tseries.htm>

Las medias móviles tienen la propiedad de que tienden a reducir la variación presente en un conjunto de datos, es decir, originan la suavización de series en el tiempo. Si el periodo de la media móvil se hace coincidir exactamente con el periodo de cierta fluctuación sistemática, esta fluctuación queda eliminada en la serie resultante al aplicar la media móvil.

**Ejemplo 3.14** Calcular la media móvil de orden 3 ( $\hat{Y}_3$ ) para los valores 2, 6, 1, 5, 3, 7 y 2 de la variable  $Y$ .

$$\begin{array}{l} \frac{2+6+1}{3} = 3 \\ \frac{6+1+5}{3} = 4 \\ \frac{1+5+3}{3} = 3 \\ \frac{5+3+7}{3} = 5 \\ \frac{3+7+2}{3} = 4 \end{array} \quad \Rightarrow \quad \begin{array}{c|c} Y & \hat{Y}_3 \\ \hline 2 & \\ 6 & 3 \\ 1 & 4 \\ 5 & 3 \\ 3 & 5 \\ 7 & 4 \\ 2 & \end{array}$$

Como se aprecia en el ejemplo, cada media se van calculando a partir del conjunto de datos que se obtiene del anterior, eliminando el primero y añadiendo el siguiente. De ahí que reciba el nombre de media móvil.

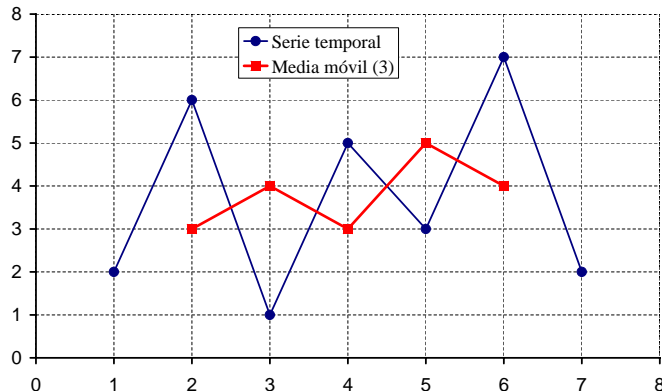


Figura 3.2: Media móvil de orden 3

Como se observa en la figura 3.2, la representación de la media móvil da lugar a una serie temporal más suave que la original. Es decir, si buscamos rectas paralelas que acoten las series, entonces, las correspondientes rectas que acotasen a las medias móviles estarían menos, entre sí, que las que acotasen a los valores originales.  $\square$

Si el orden es impar, la media móvil queda centrada pues su valor se asigna al dato que ocupa la posición central. Sin embargo, si el orden es par, la media móvil queda descentrada pues no hay ningún valor del conjunto que ocupe la posición central. En tal caso, se procede a centrar o corregir la media móvil, volviendo a calcular la media entre 2 consecutivos.

**Ejemplo 3.15** Calcular una media móvil de orden 4 con los datos del ejemplo 3.14.

$$\begin{array}{l}
 \frac{2+6+1+5}{4} = 3'5 \\
 \frac{6+1+5+3}{4} = 3'75 \\
 \frac{1+5+3+7}{4} = 4 \\
 \frac{5+3+7+2}{4} = 4'25
 \end{array}
 \Rightarrow
 \begin{array}{c|c|c}
 Y & \hat{Y}_4 & \hat{\hat{Y}}_4 \\
 \hline
 2 & & - \\
 6 & & - \\
 1 & 3'5 & 3'625 \\
 5 & 3,75 & 3'875 \\
 & 4 & \\
 3 & 4'25 & 4'125 \\
 7 & & - \\
 2 & & -
 \end{array}$$

□

### 3.6. Análisis de las series temporales

Existen una gran cantidad de componentes que conforman una serie temporal, aunque estas pueden dividirse en cuatro grandes grupos:

1. Tendencia secular (T).
2. Variaciones estacionales o periódicas (E ó S).
3. Variaciones cíclicas (C).
4. Variaciones aleatorias, irregulares o accidentales (A ó I).

La primera y la tercera son observables a largo plazo mientras que la segunda y la última se estudian en cortos periodos de tiempo. El objetivo es saber cómo se relacionan e interactúan estas componentes. Desgraciadamente esto es bastante difícil, por lo que se presentan, básicamente, dos alternativas:

1. Hipótesis Aditiva:

$$Y = T + E + C + A$$

donde  $Y$  es la conjunción de los 4 factores mediante acumulación o suma.

2. Hipótesis Multiplicativa:

$$Y = T \cdot E \cdot C \cdot A$$

donde  $Y$  es la conjunción de los 4 factores mediante el producto.

La elección de cuál de estas hipótesis es la mejor depende del grado de acierto a que conduce la aplicación de cada una. Nosotros consideraremos, principalmente, la segunda, aunque aplicar la primera se realizaría de forma análoga.

En la gráfica<sup>2</sup> de la figura 3.3 se muestra una serie temporal (arriba), junto a tres de sus componentes (T, E y A) representadas aisladamente.

<sup>2</sup>Fuente: <http://www.seh-lilha.org/tseries.htm>

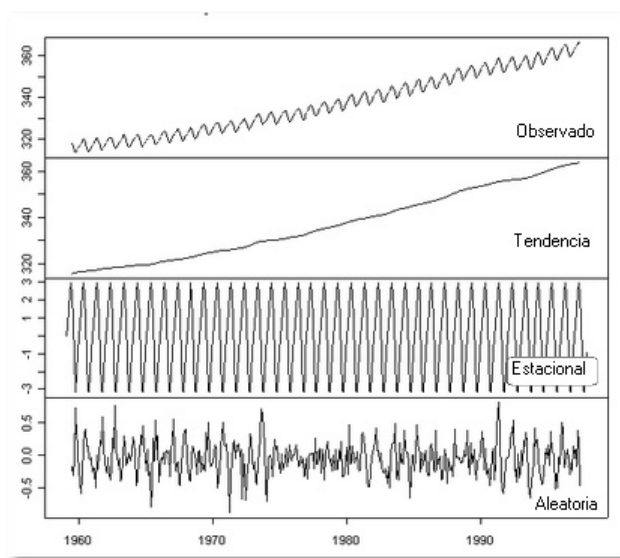


Figura 3.3: Descomposición de una serie temporal en tres componentes

### 3.6.1. Tendencia secular

La tendencia secular se refiere a la dirección general predominante de la serie observada en un espacio de tiempo suficientemente amplio. Se puede representar por una curva de tendencia (generalmente recta de tendencia).

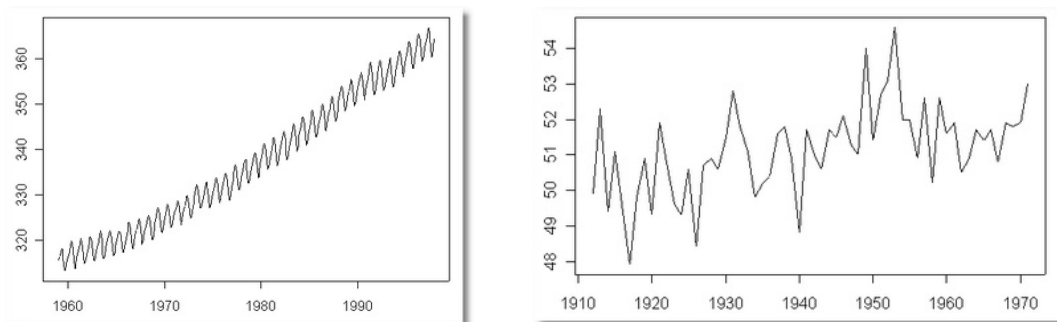


Figura 3.4: Series temporales con tendencia más (izquierda) o menos (derecha) pronunciada

En la gráfica<sup>3</sup> de la figura 3.4 se muestran dos series temporales. La tendencia es la recta imaginaria que se aproxima a la serie de datos. Y, como se observa, la serie que se representa a la izquierda, tiene una marcada tendencia creciente, mientras que la representada a la derecha, también tiene una tendencia creciente, pero es más suave.

<sup>3</sup>Fuente: <http://www.seh-lilha.org/tseries.htm>

### 3.6.2. Variaciones estacionales o periódicas

Las variaciones estacionales o periódicas son las variaciones ocurridas por los meses del año o las estaciones, y que se repiten de forma cíclica todos los años (dentro de un periodo anual). Por ejemplo, la subida de precios en Navidad, la producción de productos agrícolas, las ventas de bañadores o el número de viajeros en un autobús en las horas puntas.

Su representación gráfica viene determinada por una curva cíclica de periodo corto.

### 3.6.3. Variaciones cíclicas

Las variaciones cíclicas son aquellas variaciones que se observan a lo largo del tiempo, y que se “repiten cíclicamente”. Su representación gráfica se caracteriza por una curva de periodo largo. Por ejemplo, la recesión económica o el índice de paro.

Generalmente estas variaciones cíclicas son propias de las variables económicas y para observarse mejor es necesario que el periodo que abarca la serie temporal sea suficientemente amplio.

### 3.6.4. Variaciones aleatorias, irregulares o accidentales

Las variaciones accidentales son los movimientos esporádicos (irregulares o aleatorios) que se producen en una serie y que rompen su “tendencia”. Por ejemplo, la subida de petróleo en la guerra del Golfo, una inundación o una helada en el campo.

Además, se suele suponer que tales sucesos producen variaciones que pierden influencia tras poco tiempo.

## 3.7. Estimación de la tendencia

De entre los muchos métodos que existen para calcular la tendencia secular de una serie temporal, resaltamos los 4 siguientes:

### 3.7.1. Método gráfico

El método gráfico consiste en determinar dos curvas (poligonales), una superior y otra inferior, que acoten a nuestra serie temporal. Después, los puntos medios, localizados entre las dos curvas determinan otra curva mucho más amortiguada, que nos indica gráficamente la tendencia o dirección predominante de la serie.

Para ello, representamos gráficamente la serie temporal y procedemos de la siguiente manera:

1. Se unen, mediante segmentos, los puntos máximos de la serie, obteniéndose una línea quebrada que se denomina poligonal de cimas.
2. De la misma forma, se unen los puntos mínimos de la serie, obteniéndose la poligonal de fondos.

3. Se trazan perpendiculares al eje de abscisas que pasen por los vértices de las poligonales de cimas y fondos.
4. Se calculan los puntos medios de los segmentos determinados por los cortes de cada perpendicular con las poligonales de cimas y fondos.
5. Finalmente, la tendencia viene determinada por la curva poligonal que une los puntos medios de los segmentos.

En la gráfica de la figura 3.5 se representan los elementos necesarios para aplicar el método gráfico de estimación de la tendencia. Se ha utilizado el color azul para la serie temporal y el rojo para la tendencia. Las poligonales de cimas y fondos se representan con trazos continuos negros y las perpendiculares con trazos discontinuos.

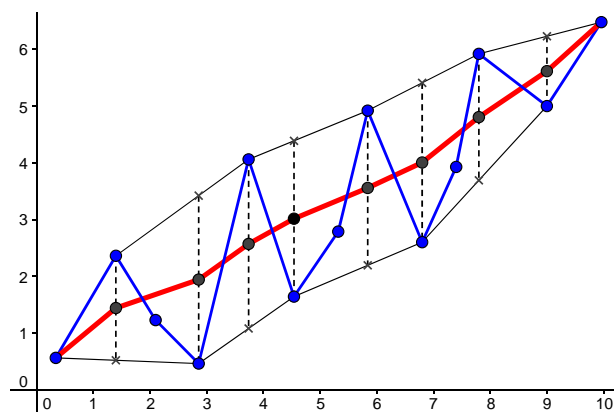


Figura 3.5: Aplicación del método gráfico para la estimación de la tendencia

### 3.7.2. Método de las medias móviles

Usando medias móviles de órdenes adecuados, podemos eliminar las variaciones estacionales, cíclicas y aleatorias, obteniendo por tanto la tendencia secular.

Las desventajas del método son la pérdida de los datos iniciales y finales de la serie y el hecho de que cuanto mayor es el orden de las medias, mayor información se pierde, por lo que habrá que mantener un cierto equilibrio.

**Ejemplo 3.16** En la siguiente tabla se muestra el cálculo de la tendencia secular utilizando medias móviles de orden 3 ( $\hat{Y}_3$ ) y 4 ( $\hat{Y}_4$ ) respectivamente:

$t$	$Y$	$\hat{Y}_3 = T$	$\hat{Y}_4$	$\hat{\hat{Y}}_4 = T$
1952	2'4	—		—
1953	3'4	3'17		—
1954	3'7	3'73	3'40	3'75
1955	4'1	4'33	4'10	4'33
1956	5'2	4'87	4'57	4'81
1957	5'3	5'37	5'05	5'28
1958	5'6	5'63	5'52	5'64
1959	6	5'93	5'77	5'74
1960	6'2	5'77	5'72	5'63
1961	5'1	5'40	5'55	5'45
1962	4'9	5'07	5'35	5'30
1963	5'2	5'30	5'25	5'48
1964	5'8	6	5'72	—
1965	7	—		—

□

### 3.7.3. Método de mínimos cuadrados

Este método se basa en el ya conocido método de los mínimos cuadrados, pues una serie temporal no es más que un caso de variable bidimensional.

Si los valores de las variables son muy grandes entonces los coeficientes con los que trabajamos son elevados. Cuando la variable tiempo,  $t$ , toma valores consecutivos formando una serie de salto constante e igual a la unidad, (lo que ocurre en la mayoría de las ocasiones), puede sustituirse por otra,  $t'$ , que se obtenga de ella mediante un sencillo cambio de origen. La técnica de simplificación a seguir es análoga en los dos casos posibles:

1. Si el número de valores de  $t$  es *impar* entonces el cambio es  $t' = t - \bar{t}$ .
2. Si el número de valores de  $t$  es *par* entonces el cambio es  $t' = 2(t - \bar{t})$ .

En cualquiera de los dos casos anteriores se verifica que:

$$\sum_{i=1}^N t'_i = 0 \quad \text{y} \quad \sum_{i=1}^N t'^3_i = 0$$

con lo que el sistema de ec. normales para obtener los parámetros  $a$  y  $b$  de una recta  $y = a + bx$  de tendencia estaría formado por dos ecuaciones con una incógnita cada una:

$$\sum_{i=1}^N t'_i y_i = b \sum_{i=1}^N t'^2_i \quad \text{y} \quad \sum_{i=1}^N y_i = aN$$

Además, se puede calcular el coeficiente de determinación, para saber si el ajuste de tendencia es representativo o no.

**Ejemplo 3.17** Consideremos la serie temporal constituida por los valores 3, 5, 8, 9, 13 y 12 de la variable  $Y$  para los años 1960 a 1965 respectivamente. Calcular los valores de tendencia por el método de los mínimos cuadrados y comprobar si son representativos.

$t_i$	$y_i$	$t'_i$	$t'_i y_i$	$t'^2_i$	$y^2_i$	$y^*_i = T$	$e_i$
1960	3	-5	-15	25	9	3'3	-0'3
1961	5	-3	-15	9	25	5'3	-0'3
1962	8	-1	-8	1	64	7'3	0'7
1963	9	1	9	1	81	9'3	-0'3
1964	13	3	39	9	169	11'3	1'7
1965	12	5	60	25	144	13'3	-1'3
	50	0	70	70	492	49'8	

$$\Rightarrow \begin{cases} 70 = 70b \\ 50 = 6a \end{cases} \Rightarrow \begin{cases} b = 1 \\ a = 8'3 \end{cases} \Rightarrow y^* = t' + 8'3$$

Como  $\sigma_e^2 = 1'16$  y  $\sigma_y^2 = 12'61$  entonces  $R^2 = 1 - \frac{1'16}{12'61} = 0'908$  (próximo a 1), lo que permite afirmar que los valores de tendencia calculados son representativos.  $\square$

Este tercer método permite realizar “predicciones” en el futuro, siempre que la línea de ajuste sea representativa ( $R^2 \approx 1$ ) y se limite a un futuro próximo.

**Ejemplo 3.18** Utilizando los datos del ejemplo 3.17, predecir el valor de  $Y$  para el año 1968:

$$t' = 2(1968 - 1962'5) = 11 \quad \Rightarrow \quad y^* = 11 + 8'3 = 19'3$$

$\square$

#### 3.7.4. Método de semipromedios

Consiste en separar los datos en 2 partes (iguales preferiblemente) y promediar los datos en cada uno de los 2 grupos, obteniendo así dos puntos  $(t_1, y_1)$  y  $(t_2, y_2)$ . La línea de tendencia se halla entonces haciendo pasar una recta por los 2 puntos calculados:

$$y - y_1 = \frac{y_2 - y_1}{t_2 - t_1} \cdot (t - t_1)$$

**Ejemplo 3.19** Utilizar el método de los semipromedios para calcular los valores de tendencia de la serie temporal del ejemplo 3.17.



Grupo	$t_i$	$y_i$	$\bar{y}$	Punto	$y_i^* = T$
1	1960	3	5'3	(1961, 5'3)	3'3
	1961	5			5'3
	1962	8			7'3
2	1963	9	11'3	(1964, 11'3)	9'3
	1964	13			11'3
	1965	12			13'3

siendo  $y^* = 5'3 + 2(t - 1961)$  la línea de tendencia obtenida a partir de los puntos (1961, 5'3) y (1964, 11'3).  $\square$

### 3.8. Estimación de la variación estacional

Existen muchos métodos para calcular la variación estacional de una serie temporal, sin embargo, la mayoría se basan en el mismo principio: aislar la variación estacional mediante la eliminación previa de las otras componentes.

Vamos a presentar dos métodos para la hipótesis multiplicativa  $Y = T \cdot E \cdot C \cdot A$  donde la eliminación de las componentes pasa por ir dividiendo la expresión anterior por las componentes aisladas. Además, presentaremos un método para la hipótesis aditiva  $Y = T + E + C + A$ , análogo a los anteriores, pero restando componentes.

En los tres casos, determinaremos unas medidas de la variación estacional *los índices de variación estacional* ( $I_E$ ) y las *diferencias de variación estacional* ( $D_E$ ), asociadas a cada estación o momento de repetición anual. Estas medidas se utilizan para desestacionalizar la serie, eliminando esta componente.

#### 3.8.1. Método de la media móvil en porcentajes

El método de la media móvil en porcentajes nos permite identificar la variación estacional de una serie temporal, procediendo de la siguiente manera:

1. Dada la serie cronológica (por meses, estaciones, trimestres, etc.) en varios años, se calcula la tendencia mediante el método de medias móviles cuyo orden coincida con el número de estaciones o periodos (orden 12 para meses, orden 4 para estaciones o trimestres, etc.). Si el orden de la media móvil es un número par, entonces la centramos.
2. El promedio móvil calculado sirve para eliminar las variaciones estacionales y las accidentales. Por lo tanto, dividiendo los datos originales ( $Y = T \cdot E \cdot C \cdot A$ ) entre los calculados en el primer paso ( $T \cdot C$ ), obtenemos conjuntamente las variaciones estacional y accidental ( $E \cdot A$ ).

$$\frac{Y}{T \cdot C} = \frac{T \cdot E \cdot C \cdot A}{T \cdot C} = E \cdot A$$

3. Por último, y para eliminar la componente accidental, basta con calcular las medias aritméticas de los valores obtenidos en el paso anterior, referidas a cada estación o periodo.

Los valores obtenidos representan la variación estacional y su media debe ser uno. Cuando no lo sea, se recomienda normalizarlos, de manera que la nueva media sea exactamente uno. Estos últimos valores obtenidos, expresados en tantos por cien, se denominan índices de variación estacional ( $I_E$ ) y representan el porcentaje sobre la media de los valores estacionales. Es decir, si  $I_E$  es mayor del 100 % entonces, en esa estación, el valor es superior a la tendencia y, en caso contrario, es inferior.

**Ejemplo 3.20** *Calcular los índices de variación estacional de la serie de datos relativos a ventas, obtenida en un estudio realizado durante 5 años, y que se recoge en la siguiente tabla:*

	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	2	2'2	2'2	2'4	2'5
Verano	3'1	3	3'5	3'6	3'6
Otoño	2'6	2'8	4'3	4'5	4'9
Invierno	1'8	2	2'1	2'2	2'3

Para calcular los índices de variación estacional, seguimos los pasos del método de la media móvil en porcentajes:

1. Calculamos  $T \cdot C$  (eliminamos  $E \cdot A$ ) utilizando medias móviles de orden 4 que al centrarlas se obtiene:

$T \cdot C$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	—	2'43	2'81	3'13	3'25
Verano	—	2'48	3'01	3'16	3'31
Otoño	2'40	2'50	3'05	3'19	—
Invierno	2'41	2'56	3'09	3'20	—

2. Dividimos los datos de la tabla original por los de la que hemos obtenido en el paso anterior para obtener  $E \cdot A$ .

$E \cdot A$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	—	0'91	0'78	0'77	0'77
Verano	—	1'21	1'16	1'14	1'09
Otoño	1'08	1'12	1'41	1'41	—
Invierno	0'75	0'78	0'68	0'69	—

3. Haciendo media aritméticas por filas eliminamos  $A$  obteniendo la variación estacional (sin normalizar). Por último, calculamos los índices de variación estacional como simples proporciones.

	$E$	$I_E$
Primavera	0'81	$\frac{100}{3'94/4} \cdot 0'81 = 81'97\%$
Verano	1'15	$\frac{100}{3'94/4} \cdot 1'15 = 116'84\%$
Otoño	1'26	$\frac{100}{3'94/4} \cdot 1'26 = 127'66\%$
Invierno	0'72	$\frac{100}{3'94/4} \cdot 0'72 = 73'53\%$
	3,94	

El valor obtenido para estos índices mide la influencia de la variación estacional sobre un nivel medio de ventas, es decir, que en primavera descienden las ventas un 18 % aproximadamente, se eleva casi un 17 % y un 28 % respectivamente en verano y en otoño, y vuelven a descender más de un 26 % en invierno.  $\square$

### 3.8.2. Método del porcentaje medio

El método del porcentaje medio nos permite calcular los índices de variación estacional de una serie temporal con el objetivo de poder desestacionalizar la serie. Para ello, procederemos de la siguiente manera:

1. Expresamos cada dato de cada periodo (mes, estación, trimestre, etc.) como porcentajes del promedio anual.
2. Se calcula la media aritmética de los porcentajes obtenidos para un mismo periodo en el paso anterior. De esta forma se obtienen los índices de variación estacional.
3. Si la media de los índices obtenidos en el paso anterior no es 100, entonces debemos ajustarlos (normalización) dividiendo cada uno de ellos, por la media. Por ejemplo habría que ajustarlos si la suma de los índices, obtenidos en el paso anterior, no corresponde al total teórico, es decir, 1200 para meses, 400 para estaciones o trimestres, etc.

**Ejemplo 3.21** Calcular los índices de variación estacional de la serie temporal del ejemplo 3.20 de la página 130 utilizando el método del porcentaje medio.

Para ello, seguimos los siguientes pasos:

1. Calculamos las medias anuales:

	Año 1	Año 2	Año 3	Año 4	Año 5
Media	2'375	2'5	3'025	3'175	3'325

2. Calculamos los porcentajes y la media de los mismos o índice de variación estacional

	Año 1	Año 2	Año 3	Año 4	Año 5	$I_E$
Primavera	84'21 %	88'00 %	72'73 %	75'59 %	75'19 %	79'144 %
Verano	130'53 %	120 %	115'70 %	113'39 %	108'27 %	117'578 %
Otoño	109'47 %	112 %	142'15 %	141'73 %	143'37 %	130'544 %
Invierno	75'79 %	80 %	69'42 %	69'29 %	69'17 %	72'734 %
						400

En este caso no es necesario ajustar los índices puesto que su suma (400) corresponde al total teórico.  $\square$

**Ejemplo 3.22** Calcular los índices de variación estacional para los dos años observados por periodos mensuales.

La siguiente tabla recoge los resultados obtenidos al aplicar los distintos pasos del método:

	Año 1	Año 2	Año 1	Año 2	Media	$I_E$
Enero	1	3	50 %	100 %	75 %	$75 \cdot (1200/1196) = 75'25 \%$
Febrero	3	4	150 %	133 %	141 %	$141 \cdot (1200/1196) = 141'47 \%$
Marzo	0	2	0 %	66 %	33 %	$33 \cdot (1200/1196) = 33'11 \%$
Abril	2	4	100 %	133 %	116 %	$116 \cdot (1200/1196) = 116'39 \%$
Mayo	1	3	50 %	100 %	75 %	$75 \cdot (1200/1196) = 75'25 \%$
Junio	4	6	200 %	200 %	200 %	$200 \cdot (1200/1196) = 200'67 \%$
Julio	1	1	50 %	33 %	41 %	$41 \cdot (1200/1196) = 41'14 \%$
Agosto	3	2	150 %	66 %	108 %	$108 \cdot (1200/1196) = 108,36 \%$
Septiembre	0	1	0 %	33 %	16 %	$16 \cdot (1200/1196) = 16,05 \%$
Octubre	2	3	100 %	100 %	100 %	$100 \cdot (1200/1196) = 100'34 \%$
Noviembre	1	2	50 %	66 %	58 %	$58 \cdot (1200/1196) = 58,19 \%$
Diciembre	6	5	300 %	166 %	233 %	$233 \cdot (1200/1196) = 233'78 \%$
Media	2	3				
Suma			1200 %	1200 %	1192 %	1200 %

Obsérvese que, en este ejemplo, para obtener los índices de variación estacional, ha sido necesario ajustar las medias de los porcentajes, puesto que no sumaban 1200 (total teórico).  $\square$

### 3.8.3. Estimación de la variación estacional para el modelo aditivo

Los índices de variación estacional, necesarios para la desestacionalización de una serie temporal, no son aplicables cuando consideramos la hipótesis aditiva ( $Y = T + E + C + A$ ). En estos casos, definimos una medida equivalente que denominamos *diferencias de variación estacional*, y que denotamos por  $D_E$ . Para ello, aplicaremos el *método de la diferencia a la tendencia*, de la siguiente manera:

1. Calculamos la tendencia ( $T$ ) por cualquiera de los métodos ya estudiados.
2. A cada dato de la serie le restamos su correspondiente valor de la tendencia:

$$Y - T = E + C + A$$

3. Para eliminar el resto de componentes ( $C + A$ ), se promedian los valores correspondientes a los mismos periodos.
4. Los valores obtenidos deben sumar 0, y si no es así, entonces hay que ajustarlos. Para ello, se calcula la media y se le resta a cada uno de los valores, obteniendo la diferencia de variación estacional ( $D_E$ ).

En este caso, la interpretación es similar a los índices de variación estacional, pero tomando el cero como centro. Por ejemplo, si  $D_E$  es positivo entonces, en ese punto, el valor de la serie es superior a la tendencia. Por el contrario, si  $D_E$  es negativo indica que el correspondiente valor de la serie es inferior a la tendencia.

**Ejemplo 3.23** *Calcular las diferencias de variación estacional de la serie temporal del ejemplo 3.20 de la página 130.*

En primer lugar, estimamos la tendencia de la serie de ventas ( $V$ ) utilizando, por ejemplo, el método de los mínimos cuadrados. Para ello, será necesario asignar un número a cada periodo de tiempo ( $t$ ). Comenzaremos asignando un 1 a la primavera (P) del año 1, y consecutivamente al resto de periodos de tiempo, hasta asignar un 20 al invierno (I) del año 5.

	Año 1				Año 2				...	Año 5			
	P	V	O	I	P	V	O	I	...	P	V	O	I
$t$	1	2	3	4	5	6	7	8	...	17	18	19	20
$V$	2'0	3'1	2'6	1'8	2'2	3'0	2'8	2'0	...	2'5	3'6	4'9	2'3

La recta de regresión que se ajusta a los datos de la tabla anterior es

$$V = 0,0617 \cdot t + 2,2326$$

y determina la tendencia ( $T$ ) que mostramos en la siguiente tabla:

$T$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	2'29	2'54	2'79	3'03	3'28
Verano	2'36	2'60	2'85	3'10	3'34
Otoño	2'42	2'66	2'91	3'16	3'40
Invierno	2'48	2'73	2'97	3'22	3'47

En segundo lugar, a cada dato de la la serie le restamos su correspondiente valor de tendencia y obtenemos los siguientes valores:

$E + C + A$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	-0'29	-0,34	-0'59	-0'63	-0'78
Verano	0'74	0'40	0'65	0'50	0'26
Otoño	0'18	0'14	1'39	1'34	1'50
Invierno	-0'68	-0'73	-0'87	-1'02	-1'17

Para eliminar el resto de componentes, se promedian los valores correspondientes a los mismos periodos.

	$D_E$
Primavera	-0,53
Verano	0,51
Otoño	0,91
Invierno	-0,89

Los valores obtenidos suman 0, de manera que no será necesario normalizarlos y, por lo tanto, corresponden a las diferencias de variación estacional. Los valores negativos obtenidos para la primavera y el invierno, indican que, en estas estaciones, el valor de la serie es inferior a la tendencia. Por el contrario, los valores positivos obtenidos para el verano y el otoño indican que los valores de la serie están por encima de la tendencia.  $\square$

#### 3.8.4. Desestacionalización de una serie temporal

La componente estacional tiene interés, por sí misma, pues nos permite conocer la evolución a corto plazo de la serie temporal. Pero, además, es interesante llegar al conocimiento de la serie temporal una vez eliminadas las variaciones estacionales, y este proceso se denomina *desestacionalización*.

Una de las aplicaciones de la desestacionalización es el cálculo de la tendencia real. La eliminación de la componente estacional se utiliza para recalcular la tendencia y obtener una mejor aproximación de la trayectoria real de la serie.

Para desestacionalizar una serie temporal, y dependiendo de la hipótesis elegida (aditiva o multiplicativa), se procede de la siguiente manera:

1. Hipótesis multiplicativa: Se divide cada dato de la serie por el índice de variación estacional:

$$T \cdot C \cdot A = \frac{Y}{I_E}$$

2. Hipótesis aditiva: A cada dato de la serie se le resta la diferencia de variación estacional:

$$T + C + A = Y - D_E$$

**Ejemplo 3.24** Desestacionalizar la serie temporal del ejemplo 3.20 de la página 130 para recalcular su tendencia, suponiendo la hipótesis multiplicativa.

En el ejemplo 3.20 se determina la estacionalidad ( $E$ ) que figura en la siguiente tabla:

$E$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	0'820	0'820	0'820	0'820	0'820
Verano	1'168	1'168	1'168	1'168	1'168
Otoño	1'277	1'277	1'277	1'277	1'277
Invierno	0'735	0'735	0'735	0'735	0'735

Para desestacionalizar la serie, dividimos los datos de la tabla original por los valores anteriores:

$T \cdot C \cdot A$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	2'44	2'68	2'68	2'93	3'05
Verano	2'65	2'57	3'00	3'08	3'08
Otoño	2'04	2'19	3'37	3'53	3'84
Invierno	2'45	2'72	2'86	2'99	3'13

Esta última tabla corresponde a las cifras de ventas obtenidas por la empresa, prescindiendo de las variaciones estacionales.

Ahora, para calcular la tendencia real, recalculamos la tendencia a los datos desestacionalizados de la tabla anterior, aplicando cualquier método, por ejemplo el de mínimos cuadrados. Para ello, será necesario asignar un número a cada periodo de tiempo ( $t$ ). Comenzaremos asignado un 1 a la primavera (P) del año 1, hasta asignar un 20 al invierno (I) del año 5.

	Año 1				Año 2				...	Año 5			
	P	V	O	I	P	V	O	I	...	P	V	O	I
$t$	1	2	3	4	5	6	7	8	...	17	18	19	20
$V$	2'44	2'65	2'04	2'45	2'68	2'57	2'19	2'72	...	3'05	3'08	3'84	3'13

La recta de regresión que se ajusta a los datos de la tabla anterior es

$$V = 0,0578 \cdot t + 2,2567$$

y determina la tendencia (T) que mostramos en la siguiente tabla:

$T$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	2'31	2'55	2'78	3'01	3'24
Verano	2'37	2'60	2'83	3'07	3'30
Otoño	2'43	2'66	2'89	3'12	3'35
Invierno	2'49	2'72	2'95	3'18	3'41

Obsérvese que estos valores de la tendencia real difieren de los valores de tendencia obtenidos en el ejemplo anterior, ya que en este caso no están afectados por la componente estacional.  $\square$

### 3.9. Estimación de las variaciones cíclicas

Una vez calculadas las variaciones Estacionales y la Tendencia, restando o dividiendo los datos originales, por estos, obtenemos:

Hipótesis multiplicativa: 
$$\frac{Y}{T \cdot E} = \frac{T \cdot E \cdot C \cdot A}{T \cdot E} = C \cdot A$$

Hipótesis aditiva: 
$$Y - (T + E) = (T + E + C + A) - (T + E) = C + A$$

Para aislar la componente cíclica, basta calcular un promedio móvil apropiado de unos pocos meses de duración (digamos 3, 5 ó 7 meses, de manera que no sea necesario el centrado). De esta forma se suavizan las variaciones accidentales para dejar sólo las variaciones cíclicas.

Si ocurre una periodicidad de ciclos, se puede construir *índices cíclicos* de manera parecida a como se han hecho los índices estacionales.



### 3.10. Estimación de las variaciones aleatorias

Para aislar las variaciones aleatorias basta con restar o dividir por el resto de las componentes ya calculadas  $T$ ,  $E$  y  $C$  según consideremos la hipótesis aditiva o multiplicativa.

$$\text{Hipótesis multiplicativa: } \frac{Y}{T \cdot E \cdot C} = \frac{T \cdot E \cdot C \cdot A}{T \cdot E \cdot C} = A$$

$$\text{Hipótesis aditiva: } Y - (T + E + C) = (T + E + C + A) - (T + E + C) = A$$

En la práctica se observa que las variaciones aleatorias tienden a tener pequeña magnitud y a seguir el esquema de una distribución normal; es decir, las pequeñas desviaciones ocurren con gran frecuencia, mientras que grandes desviaciones ocurren con pequeña frecuencia.

**Ejemplo 3.25** *Descomponer la serie temporal del ejemplo 3.20 de la página 130 en sus cuatro componentes, suponiendo la hipótesis multiplicativa.*

En el ejemplo 3.24 de la página 135 se calculan tanto la tendencia como las variaciones estacionales. Si tomamos los datos desestacionalizados ( $T \times C \times A$ ) que proporcionaba este ejemplo y los dividimos entre los datos de tendencia corregida, obtenemos una nueva tabla con las componentes  $C$  y  $A$ :

$C \cdot A$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	1'05	1'05	0'97	0'97	0'94
Verano	1'12	0'99	1'06	1'00	0'93
Otoño	0'84	0'82	1'16	1'13	1'14
Invierno	0'98	1'00	0'97	0'94	0'92

Si utilizamos medias móviles, por ejemplo, de orden 3, obtenemos la componente cíclica:

$C$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	—	1'01	1'01	0'98	0'94
Verano	1'00	0'95	1'06	1'04	1'01
Otoño	0'98	0'94	1'06	1'02	1'00
Invierno	0'96	0'93	1'04	1'00	—

Y, finalmente, para aislar la componente aleatoria ( $A$ ) basta dividir, en cada periodo de tiempo, los valores originales de la serie entre el producto de los valores calculados de las componentes ( $T \cdot E \cdot C$ ), o de manera más sencilla, dividir, simplemente, los valores de la primera

tabla de este ejemplo ( $C \cdot A$ ), por los correspondientes valores de la segunda ( $C$ ).

$A$	Año 1	Año 2	Año 3	Año 4	Año 5
Primavera	—	1'05	0'96	0'99	1'00
Verano	1'11	1'03	0'99	0'97	0'93
Otoño	0'86	0'88	1'10	1'10	1'15
Invierno	1'03	1'08	0'94	0'94	—

Obsérvese que los valores de las componentes  $C$  y  $A$  son próximos a 1, lo que indica que tienen muy poco efecto en esta serie temporal, estudiada bajo la hipótesis multiplicativa. Análogamente, cuando consideremos la hipótesis aditiva, valores de las componentes próximos a cero indicarán la poca influencia de esa componente en la serie temporal.

□

### 3.11. Relación de problemas

- Consideramos la variable  $X$  que toma los valores 22, 28, 34, 25 y 41 en cinco periodos de tiempo consecutivos. Se pide:
  - Calcular la serie de índices simples elementales con base el periodo de menor valor.
  - Calcular la serie de índices simples en cadena.
  - Calcular la serie de índices simples elementales tomando como base un periodo ficticio cuyo valor sea la media de los valores de la variable en esos 5 años.
- El porcentaje de la población mayor de 65 años sobre el total de la población en siete de los distritos de la ciudad de Málaga es: 8 , 11'3 , 9'63 , 6'78 , 7'32 , 8'96 y 6'8. Determinar el índice simple más adecuado a este caso y calcular la serie de índices correspondiente.
- Comprobar que son correctos los índices simples calculados para comparar la evolución del número de franceses y noruegos residentes en la ciudad de Málaga entre los años 1956 y 1958, e interpretar el resultado.

Año	Franceses	Noruegos	$I_F$	$I_N$
1956	1035	44	100	100
1957	1230	56	118'84	127'27
1958	1351	65	130'53	147'72

- Consideramos los valores 1, 2, 5, 8, 10, 15, 18, 20, 21, 24, 25, 28, 30, 32, 36, 45, 89, 99, 100 y 273 de una variable en 20 instantes de tiempo. Sin usar calculadora, obtener los índices simples elementales con base el instante 8 cuyo valor correspondiente es 20.
- Consideramos los valores 1, 1'2, 1'8, 2'4, 4'8, 2'4, 1'2, 0'3, 0'1, 10, 8, 8, 10, 125, 1250 de una variable en 15 instantes de tiempo consecutivos. Sin usar calculadora, obtener los índices simples en cadena.
- Sin utilizar calculadora, determinar los errores cometidos en la elaboración de esta serie de índices simples y justificar la respuesta.

variable	5	6	6	12	0	10	16	8	4	10
ISE	100	120	100	240	0	0	320	160	120	200
ISC	100	160	100	240	0	0	160	150	50	250

- Consideramos la siguiente serie de números índices simples correspondientes a los valores de una variable  $X$  en diez periodos de tiempo ( $t$ ).

$t$	1	2	3	4	5	6	7	8	9	10
IS	75	90	100	105	120	125	112	134	180	240

Obtener los valores de la variable  $X$  en cada uno de los siguientes casos:

- Los índices son elementales y el valor de la variable en el periodo  $t = 3$  es 1350.
- Los índices son elementales y el valor de la variable en el periodo  $t = 6$  es 28'1.

- c) Los índices son en cadena y el valor de la variable en el periodo  $t = 3$  es 1350.
8. Los índices simples elementales y en cadena están íntimamente relacionados. De hecho, existe una fórmula para calcular unos en función de los otros. Se pide:
- Determinar una fórmula que permita calcular los índices elementales en función de los índices en cadena.
  - Consideremos que los datos del ejercicio 7 corresponden a una serie de índices simples en cadena. A partir de ella, calcular la serie de índices elementales utilizando la fórmula obtenida en el apartado anterior.
  - Determinar una fórmula que permita calcular los índices en cadena en función de los índices elementales.
  - Consideremos que los datos del ejercicio 7 corresponden a una serie de índices simples elementales. A partir de ella, calcular la serie de índices en cadena utilizando la fórmula obtenida en el apartado anterior.
9. El precio de un kilo de azúcar entre los años 1975 y 1982 viene dado en la siguiente tabla:

Año	1975	1976	1977	1978	1979	1980	1981	1982
Precio	25	29	34	38	42	45	70	77

Se pide:

- Calcular la relación de precios tomando 1975 como año base, explicando los resultados.
  - Calcular la relación de precios tomando 1978 como año base y 1982 como año dado.
  - Calcular la relación de precios tomando como base en cada periodo de tiempo el valor que toma la variable en el periodo inmediatamente anterior (número índice simple en cadena).
10. Consideremos cuatro productos de una industria, cuyos precios de venta y producción son los siguientes:

Producto	1979		1988	
	Precio	Cantidad	Precio	Cantidad
A	225	200	314	320
B	75	15	82	21
C	68	10	75	14
D	109	34	120	50

Se pide:

- Para cada uno de los productos, determine el índice de valor para 1988 con base en 1979.
- Determine el índice de valor, de todos los productos, para 1988 con base en 1979.
- Determine los índices de precios de Laspeyres, de Paasche, de Marshall-Edgeworth y de Fisher para 1988 tomando como base el año 1979.

11. Índices complejos sin ponderar. La siguiente tabla muestra los precios y las cantidades de tres artículos para los años 1980 a 1984

t	A		B		C	
	Precios	Cantidades	Precios	Cantidades	Precios	Cantidades
1980	2	10	5	12	10	3
1981	2	12	6	10	11	2
1982	3	15	6	5	12	3
1983	4	20	7	6	12	1
1984	4	18	8	5	13	2

Se pide:

- Calcular los índices de precios, por agregación simple, de estos productos, tomando como base el año 1980.
  - Calcular los índices de cantidad, por agregación simple, de estos productos, tomando como base el año 1980.
  - Calcular los índices de precios, por la media aritmética simple, de estos productos, tomando como base el año 1980.
  - Calcular los índices de cantidad, por la media aritmética simple, de estos productos, tomando como base el año 1980.
12. Índices complejos ponderados. Con los datos de la tabla del ejemplo 11 de la página 141, calcular:
- Los índices de precios, con base 1980, por el método de agregación ponderada, y tomando como pesos las cantidades para 1980.
  - Los índices de precios, con base 1982, por el método de promedio ponderado, y tomando como pesos las cantidades para 1982.
13. Índices de precios. Con los datos de la tabla del ejemplo 11 de la página 141, calcular:
- El número índice de precios de Laspeyres para 1984 tomando como base el año 1980.
  - El número índice de precios de Paasche para 1984 tomando como base el año 1980.
  - El número índice ideal de Fisher para 1984 tomando como base el año 1980.
  - El número índice de precios de Marshall-Edgeworth para 1984 tomando como base el año 1980.
14. Cambio de periodo base. La siguiente tabla muestra los precios y las cantidades de tres artículos para los años 1990 a 1994

t	A		B		C	
	Precios	Cantidades	Precios	Cantidades	Precios	Cantidades
1990	2	10	5	12	3	10
1991	2	12	6	10	2	11
1992	3	20	7	5	3	12
1993	5	15	7	6	1	13
1994	4	18	8	4	2	14

Se pide:

- Calcular los índices de Laspeyres, con base 1990.
- Calcular los índices de Laspeyres, con base 1992, a partir de los datos obtenidos en el apartado anterior.
- Recalcular los índices de Laspeyres, con base 1992, a partir de los datos originales y compararlos con los que se han obtenido en el apartado anterior.
- Repetir el proceso con los índices de Paasche, Marshall-Edgeworth y Fisher.

15. Renovación y empalme. La siguiente tabla muestra los precios y las cantidades de tres artículos para los años 1990 a 1992

t	A		B		C	
	Precios	Cantidades	Precios	Cantidades	Precios	Cantidades
1990	2	10	5	12	3	10
1991	2	12	6	10	2	11
1992	3	20	7	5	3	12

Se pide:

- Calcular los índices de Paasche, con base 1990.
- Se consideran los siguientes nuevos datos para los años 1993 al 1995:

t	A		B		C	
	Precios	Cantidades	Precios	Cantidades	Precios	Cantidades
1993	4	25	8	6	4	13
1994	6	20	8	7	2	14
1995	5	22	9	5	3	15

Se pide:

- Renovar el índice de Paasche tomando 1993 como nuevo año base.
- Empalmar las dos series de índices

16. Consideremos que el salario medio por hora en unidades monetarias de los trabajadores de un determinado sector productivo y los índices de precios de consumo a lo largo de seis años fueron los siguientes:

Años	Salarios/hora	Índice de precios
1979	52	140
1980	58	162
1981	60	175
1982	63	190
1983	64	200
1984	84	205

Se pide:

- Estudie la evolución de los salarios/hora en términos reales.
- Cuantificar la variación en ese periodo del salario/hora en unidades monetarias corrientes y en términos reales.

17. Hallar los deflatores implícitos para el producto interior bruto a precios de mercado conociendo los datos de la siguiente tabla:

Años	Producto interior bruto	
	A precios corrientes	A precios constantes de 1980
1980	15209	15209
1981	16980	15171
1982	19567	15356
1983	22235	15633
1984	25121	15925
1985	27930	16282

18. Sabiendo que el IPC del año 1998 respecto del año 1990 es de 135 %, se pide:

- Calcule el valor real en 1990 de un producto que costase 1000 pesetas del año 1998.
- Calcule el valor real en el año 1998 de un producto que en el año 1990 costaba 1000 pesetas.

19. En el año 2006 compré un coche por valor de 24.000 euros. Suponiendo verdaderos los siguientes datos del IPC:

$$I_{90/69} = 1000 \% \qquad I_{90/06} = 75 \%$$

¿Qué le hubiese costado a mi padre (valor real) comprarlo en 1969?

20. La siguiente tabla muestra la población agricultora (en millones) en EE.UU. durante los años 1973-1983.

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Población	9'47	9'26	8'86	8'25	7'81	8'01	7'55	7'24	7'01	6'88	7'03

Se pide:

- Obtener la media móvil de orden 4 y de orden 5 y representar en una gráfica los promedios conjuntamente con los datos originales.
  - Calcular la tendencia por el método de los mínimos cuadrados, ajustando una recta y representar gráficamente el resultado junto a los valores originales.
  - Calcular la tendencia por el método de semipromedios y representar gráficamente el resultado junto a los valores originales. Hacer el ejercicio tomando primero como promedio la media aritmética y repetirlo utilizando la mediana. Sugerencia: Omitir el dato central correspondiente al año 1978 para poder dividir los datos en dos conjuntos con igual número de elementos.
  - Presentar en una tabla los valores de la tendencia obtenidos en los métodos anteriores y comparar los distintos resultados.
21. La siguiente tabla muestra la producción de energía eléctrica mensual de consumo no industrial, en miles de millones de kilovatios-hora (Kwh), en EE.UU. durante los años 1976-1981.

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
1976	178'2	156'7	164'2	153'2	157'5	172'6	185'9	185'8	165'0	163'6	169'0	183'1
1977	196'3	162'8	168'6	156'9	168'2	180'2	197'9	195'9	176'0	166'4	166'3	183'9
1978	197'3	173'7	173'2	159'7	175'2	187'4	202'6	205'6	185'6	175'6	176'3	191'7
1979	209'5	186'3	183'0	169'5	178'2	186'7	202'4	204'9	180'6	179'8	177'4	188'9
1980	200'0	188'7	187'5	168'6	175'7	189'4	216'1	215'4	191'5	178'5	178'6	195'6
1981	205'2	179'6	185'4	172'4	177'7	202'7	220'2	210'2	186'9	181'4	175'6	195'6

Se pide:

- Calcular los índices de variación estacional por el método de media móvil en porcentajes.
  - Calcular los índices de variación estacional por el método de porcentaje medio.
  - Construir una tabla de comparación para los índices estacionales hallados en los apartados anteriores.
  - Desestacionalizar los datos haciendo uso de los índices de variación estacional obtenidos por el método de la media móvil en porcentajes.
  - Representar en un mismo gráfico los datos originales y los desestacionalizados para poder comparar.
  - Calcular la tendencia por el método de los mínimos cuadrados.
  - Calcular las variaciones cíclica y accidental.
22. Las siguientes cifras corresponden a los matrimonios celebrados en España durante el periodo 1959-1962.

		1959	1960	1961	1962
1 <sup>er</sup> .	cuatrimestre	66	62	63	61
2 <sup>o</sup> .	cuatrimestre	77	77	78	78
3 <sup>er</sup> .	cuatrimestre	100	97	96	97

Se pide:

- Calcular las componentes de la serie temporal considerando la hipótesis multiplicativa.
- Calcular las componentes de la serie temporal considerando la hipótesis aditiva.



# Apuntes de ESTADÍSTICA

## Probabilidad



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento a los profesores Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.*


## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Capítulo 4

# Probabilidad

Un *experimento científico* es una acción que da lugar a resultados identificables. Este experimento puede ser *determinista* o *aleatorio* y será en este último tipo donde centraremos nuestro estudio.

Las características de un experimento aleatorio son: (1) Los posibles resultados son conocidos previamente, (2) el resultado no es predecible y (3) repeticiones en situaciones análogas puede dar resultados diferentes.

### Espacio muestral y suceso aleatorio

El *espacio muestral* de un experimento aleatorio es el conjunto formado por todos los posibles resultados del experimento. El cardinal de este conjunto puede ser finito (número obtenido al lanzar un dado) o infinito (tiempo que tarda una bombilla en fundirse).

Un *suceso aleatorio* es un subconjunto de elementos del espacio muestral. Para un experimento aleatorio, un suceso queda definido si una vez realizado el experimento, queda siempre determinado si sucedió o no.

Se llama *espacio de sucesos* al conjunto formado por todos los subconjuntos del espacio muestral, es decir, si  $E$  es el espacio muestral, entonces,  $\mathcal{P}(E)$  (conjunto de las partes de  $E$ ) es el espacio de sucesos. Por ejemplo, si jugamos a “cara o cruz” (“Head and Tail” en inglés) y lanzamos una moneda, el espacio muestral  $E = \{H, T\}$  está formado por los sucesos  $H$ =“salir cara” (Head, en inglés) y  $T$ =“salir cruz” (Tail, en inglés); y el espacio de sucesos es el conjunto  $\mathcal{P}(E) = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ .

### Suceso elemental, seguro e imposible

Un suceso se dice *elemental* si corresponde a un único resultado simple del experimento, por ejemplo,  $A$ =“salir 5 al lanzar un dado”= $\{5\}$ . Un suceso *compuesto* es la unión de varios sucesos elementales, es decir, un conjunto formado por varios resultados posibles del experimento, por ejemplo,  $B$ =“salir número impar al lanzar un dado”= $\{1, 3, 5\}$ .

Llamamos *suceso seguro* al suceso que sabemos que ocurrirá siempre al realizar el experimento y que se corresponde con el espacio muestral, por ejemplo, en el experimento del lanzamiento

de un dado, el suceso seguro es  $E = \{1, 2, 3, 4, 5, 6\}$ . El *suceso imposible* es aquel que no puede suceder nunca y se representa por  $\emptyset$ , por ejemplo, “salir un número mayor que 7” en el experimento de lanzar un dado.

Decimos que dos sucesos  $A$  y  $B$  de un espacio muestral son *incompatibles* si  $A \cap B = \emptyset$ . Por ejemplo, al lanzar un dado, los sucesos  $A$ =“salir par” y  $B$ =“salir impar” son incompatibles, pues  $A = \{2, 4, 6\}$  y  $B = \{1, 3, 5\}$  con lo cual  $A \cap B = \emptyset$ .

Si  $A$  es un suceso del espacio muestral  $E$ , llamamos *suceso contrario* o *complementario* del suceso  $A$ , y lo denotamos por  $\bar{A}$  ó bien  $A^c$  al suceso que ocurre cuando no se da el suceso  $A$ , es decir,  $\bar{A} = E - A$ . Por ejemplo, en el experimento de lanzar un dado, el complementario del suceso  $A$ =“salir par” es el suceso  $\bar{A}$ =“salir impar”, pues  $A = \{2, 4, 6\}$  y  $\bar{A} = E - A = \{1, 3, 5\}$ .

**Ejemplo 4.1** Consideramos el experimento de lanzar un dado y observar el número que aparece en la cara superior. Sean  $A$ =“salir un número par”,  $B$ =“salir impar” y  $C$ =“salir primo” tres sucesos. Describir el espacio muestral  $E$  y los sucesos  $A \cup B$ ,  $B \cup C$ ,  $A \cap B$ ,  $B \cap \bar{C}$ , determinando su tipo.

$$\begin{array}{ll} A = \{2, 4, 6\} & \\ B = \{1, 3, 5\} & \implies \\ C = \{2, 3, 5\} & \end{array} \quad \begin{array}{l} A \cup B = \{1, 2, 3, 4, 5, 6\} \text{ , suceso seguro} \\ B \cup C = \{1, 2, 3, 5\} \text{ , suceso compuesto} \\ A \cap B = \emptyset \text{ , suceso imposible} \\ B \cap \bar{C} = \{1\} \text{ , suceso elemental} \end{array}$$

□

Como podemos observar en las definiciones anteriores y en el ejemplo, existe una gran analogía entre los sucesos y la teoría de conjuntos que permite determinar la estructura del espacio de sucesos.

## 4.1. Álgebra de Boole de sucesos

Como aplicación directa de la teoría de conjuntos, el espacio de sucesos  $\mathcal{P}(E)$ , con las operaciones unión, intersección y complementario, tiene la estructura de álgebra de Boole. Pero veamos que también podemos obtener esta misma estructura en subconjuntos del espacio de sucesos.

Se llama *álgebra de sucesos* sobre el espacio muestral  $E$  a toda familia  $\mathcal{A} \subset \mathcal{P}(E)$  que verifica las siguientes condiciones:

**Ax.1)**  $E \in \mathcal{A}$

**Ax.2)** Si  $A \in \mathcal{A}$  entonces  $\bar{A} \in \mathcal{A}$

**Ax.3)** Si  $A, B \in \mathcal{A}$  entonces  $A \cup B \in \mathcal{A}$

De la condición (Ax.3) sobre la pertenencia al álgebra de la unión de dos sucesos se deduce por inducción la pertenencia al álgebra de cualquier unión finita de sucesos. Imponer que se cumpla esta condición para la unión numerable de sucesos da lugar a la definición de  $\sigma$ -álgebra.

Se llama  $\sigma$ -álgebra de sucesos sobre el espacio muestral  $E$  a toda familia  $\mathcal{A} \subset \mathcal{P}(E)$  que verifica las siguientes condiciones:

**Ax.1)**  $E \in \mathcal{A}$

**Ax.2)** Si  $A \in \mathcal{A}$  entonces  $\bar{A} \in \mathcal{A}$

**Ax.3)** Si  $A_i \in \mathcal{A}$  para todo  $i \in I$  entonces  $\cup_{i \in I} A_i \in \mathcal{A}$  (con  $I$  finito o infinito numerable).

Si un conjunto de sucesos  $\mathcal{A}$  es un álgebra o un  $\sigma$ -álgebra de sucesos sobre un espacio muestral  $E$ , diremos que  $(E, \mathcal{A})$  es un *espacio probabilizable*. Normalmente, en el caso finito, tomaremos como álgebra el conjunto  $\mathcal{P}(E)$ .

**Ejemplo 4.2** Consideramos el experimento de lanzar un dado y observar el número que aparece en la cara superior cuyo espacio muestral es  $E = \{1, 2, 3, 4, 5, 6\}$ . El conjunto de sucesos  $\mathcal{A} = \{\emptyset, \{1\}, \{2\}, \{1, 2\}, E\}$  no es un álgebra de sucesos pues no verifica el axioma (Ax.2). Sin embargo, el conjunto de sucesos  $\mathcal{B} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, E\}$  sí es un álgebra de sucesos pues verifica los tres axiomas.

Hay muchas propiedades que se deducen de la definición axiomática de álgebra de sucesos. Por ejemplo,  $\emptyset \in \mathcal{A}$  sabiendo que  $\emptyset = \bar{E}$  y aplicando las condiciones (Ax.1) y (Ax.2). Igual ocurre con la intersección de sucesos: si  $A, B \in \mathcal{A}$  entonces  $A \cap B \in \mathcal{A}$  como consecuencia de escribir  $A \cap B = \overline{\bar{A} \cup \bar{B}}$  y aplicar sucesivamente las condiciones (Ax.2) y (Ax.3). En una  $\sigma$ -álgebra, si  $A_i \in \mathcal{A}$ , entonces  $\bigcap_{i \in I} A_i \in \mathcal{A}$  como consecuencia de escribir  $\bigcap_{i \in I} A_i = \overline{\bigcup_{i \in I} \bar{A}_i}$  y aplicar sucesivamente las condiciones (Ax.2) y (Ax.3).

**Ejemplo 4.3** Si  $\mathcal{A}$  es un álgebra de sucesos sobre el espacio muestral  $E$ , demostrar la siguiente propiedad:

$$\text{Si } A, B \in \mathcal{A} \quad \text{entonces} \quad A - B \in \mathcal{A}$$

Sabemos que  $A - B = A \cap \bar{B}$ . Como  $A \in \mathcal{A}$  por hipótesis y  $B \in \mathcal{A}$  por el axioma (Ax.2), aplicamos que la intersección de dos sucesos del álgebra pertenece al álgebra y deducimos que  $A - B \in \mathcal{A}$ .  $\square$

## 4.2. Probabilidad

En primer lugar, veamos la definición axiomática de probabilidad. Después veremos su definición clásica o frecuentista que relaciona los conceptos de probabilidad y frecuencia relativa.

### 4.2.1. Definición axiomática de probabilidad

Sea  $(E, \mathcal{A})$  un espacio probabilizable, se llama *función de probabilidad*, o simplemente *probabilidad* a toda función  $P : \mathcal{A} \rightarrow [0, 1]$  que verifique las siguientes condiciones:

**Ax.1)** Para todo  $A \in \mathcal{A}$ , se verifica que  $P(A) \geq 0$

**Ax.2)**  $P(E) = 1$

**Ax.3)** Para todo  $\{A_i : A_i \in \mathcal{A}\}_{i \in I}$ , se verifica que  $P(\bigcup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$  si  $A_i \cap A_j = \emptyset$  para todo  $i \neq j$ , con  $i, j \in I$ .

Si  $P$  es una función de probabilidad sobre el espacio probabilizable  $(E, \mathcal{A})$ , se llama *espacio de probabilidad* a la terna  $(E, \mathcal{A}, P)$ .

Una función de probabilidad queda determinada conociendo el valor de la función para los sucesos elementales, pues la probabilidad de cualquier otro suceso se calcula aplicando el axioma (Ax.3).

**Ejemplo 4.4** Consideramos el experimento que consiste en lanzar un dado. Calcule la probabilidad de que salga un número par.

Los sucesos elementales del experimento son  $\{1, 2, 3, 4, 5, 6\}$  y si el dado no está trucado, los sucesos son equiprobables y la probabilidad de cada uno de ellos es  $1/6$ . El suceso que nos piden corresponde con el subconjunto  $\{2, 4, 6\}$  cuya probabilidad es la suma de las probabilidades de cada uno de los sucesos elementales que lo componen, es decir,  $1/2$ .  $\square$

**Ejemplo 4.5** Sea  $E = \{a_1, a_2, a_3\}$  el espacio muestral de un cierto experimento aleatorio. Determine si la función  $P$  definida por  $P(a_1) = \frac{1}{4}$ ,  $P(a_2) = \frac{1}{2}$  y  $P(a_3) = \frac{1}{8}$  es una probabilidad.

La unión disjunta de los sucesos elementales corresponde con el espacio muestral cuya probabilidad ha de ser 1 en virtud del axioma (Ax.2). Por lo tanto, por el axioma (Ax.3), la suma de las probabilidades de todos los sucesos elementales ha de ser 1. Pero en este caso,  $P$  no es una probabilidad pues  $P(E) = P(a_1) + P(a_2) + P(a_3) = 7/8 \neq 1$ , en contra de los axiomas (Ax.2) y (Ax.3).  $\square$

En cualquier espacio de probabilidad  $(E, \mathcal{A}, P)$  se verifican las siguientes propiedades:

1.  $P(\bar{A}) = 1 - P(A)$
2.  $P(\emptyset) = 0$
3. Si  $A \subset B \Rightarrow P(A) \leq P(B)$
4.  $0 \leq P(A) \leq 1$
5.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6.  $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
7.  $P(A - B) = P(A) - P(A \cap B)$

Además, cada una de las propiedades anteriores se demuestra a partir de la definición axiomática de probabilidad y de las propiedades que se hayan demostrado previamente, utilizando la estructura de álgebra de Boole del álgebra  $\mathcal{A}$ .

**Ejemplo 4.6** Demuestre la propiedad de la probabilidad  $P(\bar{A}) = 1 - P(A)$  y utilícela para demostrar la propiedad  $P(\emptyset) = 0$ .

Por ser  $\mathcal{A}$  un álgebra de Boole, sabemos que para cualquier suceso  $A$  se verifica que  $E = A \cup \bar{A}$ . Como esta unión es disjunta, aplicando el axioma (Ax.3), tenemos que  $P(E) = P(A) + P(\bar{A})$ . Finalmente, aplicamos el axioma (Ax.2), sustituyendo  $P(E)$  por 1, y despejamos para obtener la propiedad  $P(\bar{A}) = 1 - P(A)$ .

Para demostrar la segunda propiedad usamos la propiedad que hemos demostrado pero considerando el caso particular  $A = E$ . De esta manera tenemos que  $P(\bar{E}) = 1 - P(E)$ , es decir,  $P(\emptyset) = 1 - 1 = 0$ .  $\square$

Resulta muy práctico utilizar los diagramas de Venn para interpretar el significado de las propiedades de la probabilidad. Para ello, identificamos la probabilidad de los sucesos con su área y asignamos 1 a la probabilidad del universo donde se representan los sucesos.

**Ejemplo 4.7** Interpretar la fórmula de la probabilidad de la unión de sucesos, a partir de un diagrama de Venn

Si representamos la unión de los sucesos  $A$  y  $B$  en un diagrama de Venn como el de la figura 4.1

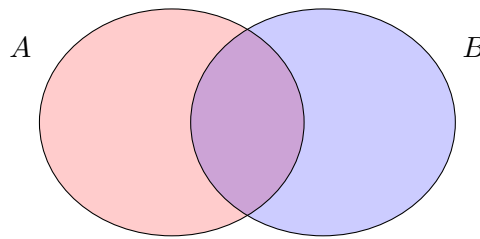


Figura 4.1: Diagrama de Venn de la unión de sucesos

observamos que al sumar las áreas de  $A$  y  $B$ , hay una región ( $A \cap B$ ) que hemos sumado dos veces y que debemos restar para calcular la probabilidad de la unión:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .  $\square$

#### 4.2.2. Relación entre frecuencias y probabilidad

Los fenómenos aleatorios son totalmente imprevisibles de manera aislada, pero presentan regularidades cuando se repiten un número elevado de veces. *Un experimento aleatorio se caracteriza porque repetido muchas veces y en idénticas condiciones el cociente entre el número de veces que aparece un resultado y el número total de veces que se realiza el experimento tiende a un número fijo.* Esta propiedad es conocida como ley de los grandes números, establecida por Jakob Bernouilli (1654-1705).

Consideremos un suceso  $A$  del espacio muestral  $E$  de un experimento aleatorio. Si se realiza  $N$  veces dicho experimento y el suceso  $A$  aparece  $n_A$  veces, se dice que la frecuencia relativa  $f_A$  del suceso  $A$  es  $n_A/N$ . La probabilidad del suceso  $A$  puede considerarse como el límite de la frecuencia relativa del suceso  $A$ , cuando el número de experiencias ( $N$ ) tiende a infinito:

$$P(A) = \lim_{N \rightarrow \infty} f_A = \lim_{N \rightarrow \infty} \frac{n_A}{N}$$

**Ejemplo 4.8** Si consideramos el experimento de lanzar una moneda, la probabilidad de salir cara es  $1/2$ . ¿Qué significado tiene este número?

Está claro que cuando lanzamos la moneda una única vez, no podemos predecir el resultado (experimento aleatorio). Sin embargo, si lanzamos la moneda muchas veces, esperamos que aproximadamente la mitad de ellas ( $1/2$ ) sean caras. La probabilidad es, por tanto, una estimación del comportamiento de un experimento cuando se realiza muchas veces.  $\square$

Hay fenómenos aleatorios, como el lanzamiento de dados, de monedas, etc., en que, por razones de simetría y regularidad, se puede suponer que todos los sucesos elementales son equiprobables, es decir, que tienen igual probabilidad de presentarse. En estos casos es útil la definición de probabilidad de Pierre-Simon Laplace (1749-1827): *La probabilidad de un suceso  $A$  es igual al cociente entre el número de casos favorables a que ocurra el suceso y el número de casos posibles, en el supuesto de que todos sean igualmente probables*

$$P(A) = \frac{\text{número de casos favorables al suceso } A}{\text{número de casos posibles}}$$

**Ejemplo 4.9** Si extraemos simultáneamente dos bolas de una urna que contiene 5 bolas blancas y 7 bolas rojas ¿cuál es la probabilidad de que ambas bolas extraídas sean del mismo color?

Como

$$P(BB) = \frac{C_{5,2}}{C_{12,2}} = \frac{\binom{5}{2}}{\binom{12}{2}} = \frac{10}{66} = \frac{5}{33} \quad \text{y} \quad P(RR) = \frac{C_{7,2}}{C_{12,2}} = \frac{\binom{7}{2}}{\binom{12}{2}} = \frac{21}{66} = \frac{7}{22}$$

se tiene

$$P(BB \cup RR) = P(BB) + P(RR) = \frac{5}{33} + \frac{7}{22} = \frac{31}{66}$$

$\square$

### 4.3. Probabilidad condicionada. Sucesos independientes

Consideremos un espacio de probabilidad  $(E, \mathcal{A}, P)$  y sea  $A \in \mathcal{A}$  un suceso con  $P(A) > 0$ . Para cualquier suceso  $B \in \mathcal{A}$  definimos la *probabilidad del suceso  $B$  condicionada al suceso  $A$* , de la siguiente manera:

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

Fijado un suceso  $A \in \mathcal{A}$ , la función  $P_A : \mathcal{A} \rightarrow [0, 1]$  es una probabilidad pues cumple la definición axiomática:

$$\text{Ax.1: } P_A(B) = \frac{P(B \cap A)}{P(A)} \geq 0 \text{ por ser cociente de números no negativos.}$$

$$\text{Ax.2: } P_A(E) = \frac{P(E \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1$$

$$\text{Ax.3: Si } B \cap C = \emptyset \Rightarrow P_A(B \cup C) = \frac{P((B \cup C) \cap A)}{P(A)} = \frac{P((B \cap A) \cup (C \cap A))}{P(A)} =$$



$$= \{(B \cap A) \cap (C \cap A) = \emptyset\} = \frac{P(B \cap A)}{P(A)} + \frac{P(C \cap A)}{P(A)} = P_A(B) + P_A(C)$$

Esta probabilidad se suele denotar por  $P(B|A)$  y su definición recoge la idea de actualización del valor de la probabilidad en función de la información que se tenga en cada momento. El valor de la probabilidad de  $B$  cambia cuando conocemos la ocurrencia de un suceso  $A$ .

**Ejemplo 4.10** *En el ejemplo del lanzamiento de un dado, calcular la probabilidad de obtener un cinco si sabemos que saldrá un número impar.*

Consideremos el espacio muestral  $E = \{1, 2, 3, 4, 5, 6\}$  correspondiente al experimento aleatorio de lanzar un dado y sea  $I = \text{“número impar”}$  un suceso del espacio de sucesos. Al principio, los sucesos elementales son equiprobables y podemos calcular  $P(5) = 1/6$  y  $P(I) = 1/2$ . Aplicando la definición de probabilidad condicionada tenemos

$$P(5|I) = \frac{P(5 \cap I)}{P(I)} = \frac{P(5)}{P(I)} = \frac{1/6}{1/2} = \frac{1}{3}$$

Obsérvese como la probabilidad original asociada al suceso elemental 5 a pasado de ser  $1/6$  a ser  $1/3$  cuando hemos conocido la información de que el número era impar.  $\square$

### Sucesos independientes

Sean  $A$  y  $B$  dos sucesos de un espacio de probabilidad  $(E, \mathcal{A}, P)$ . Decimos que el suceso  $A$  es *independiente* del  $B$  si y sólo si  $P(A|B) = P(A)$ . Esta relación de independencia es simétrica, es decir, si el suceso  $A$  es independiente del  $B$  entonces el suceso  $B$  es independiente del  $A$  y se expresa así:

$$A \text{ y } B \text{ son independientes} \iff P(A|B) = P(A) \text{ ó bien } P(B|A) = P(B)$$

Aplicando la definición de probabilidad condicionada podemos afirmar que si dos sucesos son independientes entonces se cumple:

$$P(A \cap B) = P(A) \cdot P(B)$$

**Ejemplo 4.11** *Consideremos una urna compuesta de 3 bolas blancas y 5 bolas negras. Se extrae una bola y después otra. ¿Qué probabilidad hay de que las dos bolas sean blancas?*

El enunciado no dice nada y, sin embargo, el experimento es completamente distinto si devolvemos la primera bola a la urna antes de extraer la segunda (extracciones con reemplazamiento) o no la devolvemos (extracciones sin reemplazamiento). Veamos qué ocurre en ambos casos: Sean  $B_i$  los sucesos “extraer bola blanca en la  $i$ -ésima extracción”, con  $i = 1, 2$ . La probabilidad que nos piden es:

$$P(B_1 \cap B_2) = P(B_1) \cdot P(B_2|B_1)$$

- Extracciones con reemplazamiento. Si devolvemos la bola a la urna, el resultado de la segunda extracción no depende, en absoluto, del resultado de la primera y, por lo tanto, los sucesos son independientes

$$P(B_1 \cap B_2) = P(B_1) \cdot P(B_2|B_1) = P(B_1) \cdot P(B_2) = \frac{3}{8} \cdot \frac{3}{8} = \frac{9}{64} = 0'140625$$

- Extracciones sin reemplazamiento. Sin embargo, si no devolvemos la bola a la urna, la composición de la urna (número de bolas de cada color) será distinta a la original y las probabilidades (casos favorables entre casos posibles) de los sucesos serán distintas:

$$P(B_1 \cap B_2) = P(B_1) \cdot P(B_2 | B_1) = \frac{3}{8} \cdot \frac{2}{7} = \frac{3}{28} = 0,10714$$

Obsérvese que la probabilidad condicionada  $P(B_2 | B_1)$  se ha calculado como casos favorables entre posibles, en función de la nueva composición de la urna después de la primera extracción.

□

## 4.4. Teorema de la probabilidad total. Teorema de Bayes

Consideremos un espacio de probabilidad  $(E, \mathcal{A}, P)$  y sea  $\mathcal{C} = \{C_i\}_{i \in I}$  un conjunto de sucesos. Decimos que  $\mathcal{C}$  es un *sistema completo de sucesos* de  $E$  si se verifican las siguientes condiciones:

- 1) Las intersecciones son vacías, es decir,  $C_i \cap C_j = \emptyset$  para todo  $i, j \in I$  tal que  $i \neq j$
- 2) La unión es el total, es decir,  $\bigcup_{i \in I} C_i = E$

El conjunto  $\mathcal{C}$  también se denomina partición del espacio muestral  $E$ .

**Ejemplo 4.12** El conjunto  $\mathcal{C} = \{A, B, C\}$ , con  $A = \{1, 2\}$ ,  $B = \{3, 5, 6\}$  y  $C = \{4\}$ , constituye una partición del espacio muestral  $E = \{1, 2, 3, 4, 5, 6\}$  del experimento consistente en lanzar un dado.

### 4.4.1. Teorema de la probabilidad total

Consideremos un espacio de probabilidad  $(E, \mathcal{A}, P)$  y sea  $\mathcal{C} = \{C_i\}_{i \in I}$  un sistema completo de sucesos, tal que para todo  $i \in I$ ,  $P(C_i) > 0$ . Si  $B \in \mathcal{A}$  es un suceso cualquiera, entonces:

$$P(B) = \sum_{i \in I} P(B | C_i) \cdot P(C_i)$$

**Ejemplo 4.13** Tres máquinas  $A$ ,  $B$  y  $C$  producen respectivamente el 50 %, 30 % y 20 % del número total de artículos de una fábrica. Los porcentajes de desperfectos de producción de estas máquinas son 3 %, 5 % y 10 %. Si se seleccionan al azar un lote de productos, halle la proporción de artículos defectuosos.

De los datos del problema deducimos que las probabilidades de cada uno de los sucesos elementales del espacio muestral  $E = \{A, B, C\}$  son  $P(A) = 0'5$ ,  $P(B) = 0'3$ ,  $P(C) = 0'2$ . Además, si consideramos el suceso  $D$  = “artículo defectuoso”, sabemos que  $P(D | A) = 0'03$ ,  $P(D | B) = 0'05$ ,  $P(D | C) = 0'10$ . Como  $A$ ,  $B$  y  $C$  constituyen una partición de  $E$ , podemos calcular la probabilidad que nos piden aplicando el teorema de la probabilidad total:

$$\begin{aligned} P(D) &= P(D | A) \cdot P(A) + P(D | B) \cdot P(B) + P(D | C) \cdot P(C) \\ &= 0'03 \cdot 0'5 + 0'05 \cdot 0'3 + 0'10 \cdot 0'2 = 0'05 \longrightarrow 5\% \end{aligned}$$

□

#### 4.4.2. Teorema de Bayes

El teorema de Bayes es una consecuencia directa del teorema de la probabilidad total y de la definición de probabilidad condicionada cuando tenemos un sistema completo de sucesos.

Consideremos un espacio de probabilidad  $(E, \mathcal{A}, P)$  y sea  $\mathcal{C} = \{C_i\}_{i \in I}$  un sistema completo de sucesos, tal que para todo  $i \in I$ ,  $P(C_i) > 0$ . Si  $B \in \mathcal{A}$  es un suceso cualquiera, entonces:

$$P(C_j | B) = \frac{P(B | C_j) \cdot P(C_j)}{\sum_{i \in I} P(B | C_i) \cdot P(C_i)}$$

Este teorema recoge la idea de actualización del valor de la probabilidad en función de la información que se tenga en cada momento. Al principio tenemos  $P(C_j)$  que es la *probabilidad “a priori”* y representa la “opinión inicial” sobre un asunto. Después ocurre un suceso  $B$  que representa la nueva información recibida y que determina las probabilidades  $P(B | C_i)$  para todos los  $C_i$  que se denominan *verosimilitudes*. Al final, aplicando el teorema de Bayes se obtiene  $P(C_j | B)$  que se denomina *probabilidad “a posteriori”* y que representa la “nueva opinión” sobre el asunto.

**Ejemplo 4.14** Tres máquinas  $A$ ,  $B$  y  $C$  producen respectivamente el 50 %, 30 % y 20 % del número total de artículos de una fábrica. Los porcentajes de desperfectos de producción de estas máquinas son 3 %, 5 % y 10 %. Supóngase que se selecciona al azar un artículo y resulta ser defectuoso. Calcule la probabilidad de que el artículo haya sido producido por la máquina  $A$ .

De los datos del problema deducimos que las probabilidades de cada uno de los sucesos elementales del espacio muestral  $E = \{A, B, C\}$  son  $P(A) = 0'5$ ,  $P(B) = 0'3$  y  $P(C) = 0'2$ . Además, si consideramos el suceso  $D$  = “artículo defectuoso”, sabemos que  $P(D | A) = 0'03$ ,  $P(D | B) = 0'05$  y  $P(D | C) = 0'10$ . Con todos estos datos podemos calcular la probabilidad que nos piden:

$$\begin{aligned} P(A | D) &= \frac{P(D | A) \cdot P(A)}{P(D | A) \cdot P(A) + P(D | B) \cdot P(B) + P(D | C) \cdot P(C)} \\ &= \frac{0'03 \cdot 0'5}{0'03 \cdot 0'5 + 0'05 \cdot 0'3 + 0'10 \cdot 0'2} = \frac{0'015}{0'05} = 0'3 \end{aligned}$$

Obsérvese que “a priori” la probabilidad del suceso  $A$  era 0'5 y que una vez ocurrido el suceso  $D$  se obtiene una nueva probabilidad “a posteriori” para el suceso  $A$  que es 0'3.  $\square$

## 4.5. ANEXO: Combinatoria

Para determinar la probabilidad de algunos sucesos, especialmente aquellos que se obtienen aplicando la regla de Laplace, resulta muy útil conocer la combinatoria. En las siguientes definiciones consideramos que  $\Omega$  es un conjunto de  $n$  elementos.

Llamaremos **variaciones de  $n$  elementos tomados de  $k$  en  $k$**  al número de ordenaciones distintas de  $k$  elementos de  $\Omega$ . En el primer lugar de una posible lista ordenada podemos colocar cualquier elemento de entre los  $n$  posibles. En segundo lugar, podemos colocar cualquiera de entre los  $n - 1$  restantes. En tercer lugar, cualquiera de entre los  $n - 2$  restantes. Así, hasta llegar al lugar  $k$ -ésimo en donde podemos colocar cualquier elemento de entre los  $n - k + 1$  restantes. Por tanto:

$$V_n^k = n \cdot (n - 1) \cdot (n - 2) \cdots (n - k + 1) = \frac{n!}{(n - k)!}$$

Llamaremos **permutaciones de  $n$  elementos** al número de ordenaciones posibles de todos los elementos de  $\Omega$ . Por un razonamiento similar al anterior podemos llegar a que:

$$P_n = V_n^n = n!$$

Llamaremos **combinaciones de  $n$  elementos tomados de  $k$  en  $k$**  al número de subconjuntos distintos formados por  $k$  elementos de  $\Omega$ . Aquí,  $\{a, b, c\} = \{b, a, c\} = \{c, b, a\} = \dots$  que, como variaciones, no son la misma. Es decir, de entre todas las variaciones de  $n$  elementos tomados de  $k$  en  $k$ , ahora consideraremos como iguales aquellas en las cuales sus elementos están ordenados de formas distintas. Así:

$$C_n^k = \frac{V_n^k}{P_k} = \frac{n!}{k! \cdot (n - k)!} = \binom{n}{k}$$

Llamaremos **variaciones con repetición de  $n$  elementos tomados de  $k$  en  $k$**  al número de ordenaciones distintas de elementos de  $\Omega$ , pudiendo elegirse un elemento, a lo sumo,  $k$  veces. Así, en el primer lugar de una posible lista ordenada podemos colocar cualquiera de los  $n$  elementos. En el segundo lugar podemos colocar cualquiera de los  $n$  elementos, ya que cualquier elemento lo podemos escoger, a lo sumo,  $k$  veces. Así, hasta llegar al  $k$ -ésimo lugar, en el cual podemos colocar cualquiera de los  $n$  elementos.

$$VR_n^k = n^k$$

Llamaremos **permutaciones con repetición** al número de ordenaciones posibles de todos los elementos de cuando éstos se encuentran agrupados en clases, siendo indistinguibles los elementos de cada clase. Es decir, de entre todas las permutaciones posibles de los  $n$  elementos de  $\Omega$ , cualquier permutación entre sí de los elementos de una misma clase da lugar a la misma permutación con repetición. Así:

$$PR_n^{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdots n_r!} \text{ siendo } n_1 + n_2 + \cdots + n_r = n$$

Llamaremos **combinaciones con repetición de  $n$  elementos tomados de  $k$  en  $k$**  al número de conjuntos distintos que podemos formar con  $k$  elementos de  $\Omega$ , pudiendo elegirse cualquier elemento, a lo sumo,  $k$  veces. Así, una forma de determinar el conjunto es indicando el número de veces que seleccionamos cada elemento. Para ello, tomemos  $k$  bolas en fila encerradas entre dos barras y tomemos además  $n - 1$  barras más.

- Diremos que el primer elemento de  $\Omega$  lo hemos tomado tantas veces como bolas haya entre las barras  $1^a$  y  $2^a$
- Diremos que el segundo elemento de  $\Omega$  lo hemos tomado tantas veces como bolas haya entre las barras  $2^a$  y  $3^a$
- ...

De esta forma se trata de colocar todas las barras y las bolas, es decir, se trata de colocar  $n + k - 1$  elementos ( $k$  bolas y  $n - 1$  barras) siendo las barras indistinguibles y también las bolas. Así:

$$CR_n^k = PR_{n+k-1}^{k,n-1} = \frac{(n+k-1)!}{k! \cdot (n-1)!} = C_{n+k-1}^k = C_{n+k-1}^{n-1}$$

#### 4.5.1. Identificación del problema

Para determinar si nuestro problema corresponde con variaciones, permutaciones o combinaciones puede resultar útil tener una respuesta a las siguientes preguntas:

1. ¿Cuántos elementos tengo? Esta pregunta hace referencia al total de elementos de que dispongo en el conjunto  $\Omega$  antes de plantearme las agrupaciones. La respuesta será el valor de  $n$ .
2. ¿Cuántos elementos tienen las agrupaciones? La respuesta corresponde al valor de  $k$ .
3. ¿Son distinguibles los elementos de  $\Omega$ ? Si la respuesta es no, entonces tenemos permutaciones con repetición. Si la respuesta es afirmativa, entonces nos seguimos preguntando.
4. ¿Importa el orden? Es decir, si cambiamos el orden de los elementos de una misma agrupación, ¿estamos considerando el mismo caso? Si la respuesta es afirmativa, entonces nos referimos a variaciones. Si la respuesta es negativa, entonces nos referimos a combinaciones. Finalmente nos preguntamos.
5. ¿Se pueden repetir los elementos en las agrupaciones? En cada uno de los casos anteriores, si la respuesta es negativa nos referimos a variaciones o combinaciones simples, y si la respuesta es afirmativa, entonces nos referimos a variaciones o combinaciones con repetición.

Con estas preguntas sólo nos queda identificar a las permutaciones simples que corresponden con las variaciones simples cuando  $n = k$ .

En la figura 4.2 se representa, a modo de algoritmo, el método que hemos descrito para identificar los tipos de problemas de combinatoria.

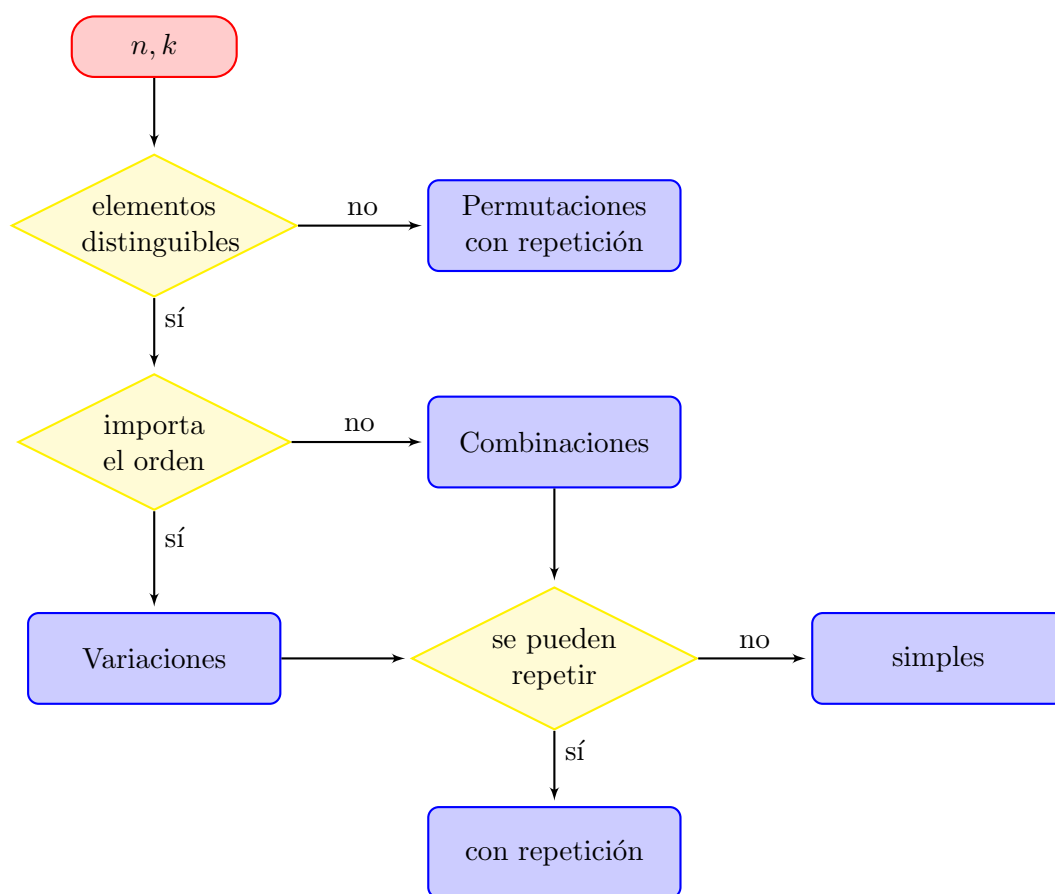


Figura 4.2: Esquema de combinatoria

**Ejemplo 4.15** ¿Cuántos grupos distintos de trabajo formados por 5 alumnos, se pueden formar con los alumnos de una clase de 25 alumnos?

Partimos de un conjunto de  $n = 25$  elementos distinguibles (alumnos de la clase) y queremos hacer agrupaciones de  $k = 5$  elementos (grupos de trabajo). Ahora bien:

- No importa el orden de los elementos en cada agrupación, y
- No es posible que haya elementos repetidos en una misma agrupación.

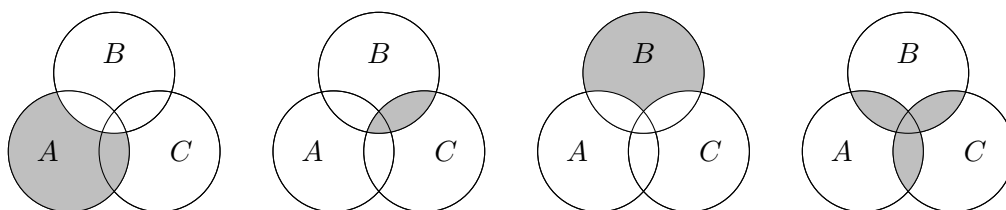
Por lo tanto, siguiendo el algoritmo, llegamos a que el número de posibles grupos de trabajo es

$$C_{25}^5 = \binom{25}{5} = \frac{25!}{5! \cdot 20!} = \frac{25 \cdot 24 \cdot 23 \cdot 22 \cdot 21}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 53\,130$$

□

## 4.6. Relación de problemas

- Sean  $A = \{1, 2, 3, 4, 5, 6\}$ ,  $B = \{3, 4, 5\}$ ,  $C = \{6, 7, 8, 9\}$  y  $D = \{4, 5, 6, 7\}$ , cuatro sucesos del espacio muestral  $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ . Calcule  $\bar{A}$ ,  $\bar{B}$ ,  $\bar{C}$ ,  $\bar{U}$ ,  $\bar{A} \cap B$ ,  $\bar{A} \cup B$ ,  $A \cap \bar{B}$ ,  $A \cup \bar{B}$ ,  $C \cup \bar{D}$ ,  $C \cap \bar{D}$ ,  $\bar{C} \cap D$ ,  $\bar{B} \cup \bar{D}$ ,  $\bar{B} \cap \bar{D}$ ,  $\overline{B \cap D}$ ,  $\overline{B \cup D}$ ,  $A - B$ ,  $B - A$ ,  $B - C$ ,  $C - B$ ,  $B - D$ ,  $D - B$ ,  $A - D$ ,  $D - A$  y  $D - C$ .
- Sean  $A$ ,  $B$  y  $C$ , tres sucesos cualesquiera de un espacio muestral  $U$ . Represente mediante un diagrama de Venn los conjuntos  $(A \cup B) - (B \cap C)$  y  $(B - A) \cap (B - C)$ .
- Determine el conjunto que representan cada uno de los siguientes diagramas de Venn



- Si en una clase de 100 alumnos, 54 han aprobado el examen de Matemáticas, 75 el de física y 40 han aprobado los dos exámenes, ¿cuántos alumnos no han aprobado ninguna de las dos asignaturas?
- En una clase  $C$  de 30 alumnos, 18 estudian Matemáticas ( $M \subset C$ ), 13 Filosofía ( $F \subset C$ ) y 5 Historia ( $H \subset C$ ). Sabiendo que sólo hay 3 alumnos que estudian simultáneamente Matemáticas y Filosofía, se pide:
  - Determine cuantos alumnos estudian Matemáticas o Filosofía.
  - ¿Cuántos elementos tiene el conjunto  $C - (M \cup F)$ ?
  - ¿Puede saberse cuantos alumnos sólo estudian Historia?
- Se sometió un grupo de personas a un cuestionario formado por tres preguntas. Sabemos que el 8% contestaron bien las tres preguntas, el 9% contestaron bien sólo a la 1ª y 2ª, el 11% contestaron bien sólo la 1ª y 3ª, el 16% contestaron bien la 2ª y 3ª, el 45% contestaron bien a la 1ª, el 32% a la 2ª y el 39% a la 3ª. ¿Qué porcentaje de personas no contestaron bien a ninguna pregunta?
- Definir tres espacios de probabilidad distintos sobre el espacio muestral  $E = \{0, 1, 2\}$ .
- Consideremos el espacio muestral de 4 elementos  $E = \{a, b, c, d\}$ . Justifique si alguno de los siguientes casos define una probabilidad:
  - $P(a) = \frac{1}{2}$   $P(b) = \frac{1}{3}$   $P(c) = \frac{1}{4}$   $P(d) = \frac{1}{5}$
  - $P(a) = \frac{1}{2}$   $P(b) = \frac{1}{4}$   $P(c) = -\frac{1}{4}$   $P(d) = \frac{1}{2}$
  - $P(a) = \frac{1}{2}$   $P(b) = \frac{1}{4}$   $P(c) = \frac{1}{8}$   $P(d) = \frac{1}{8}$
  - $P(a) = \frac{1}{2}$   $P(b) = \frac{1}{4}$   $P(c) = \frac{1}{4}$   $P(d) = 0$
- Consideremos el espacio muestral de 4 elementos  $E = \{a, b, c, d\}$ . Calcule las probabilidades que se piden

- a) Hallar  $P(a)$  si  $P(b) = \frac{1}{3}$ ,  $P(c) = \frac{1}{6}$  y  $P(d) = \frac{1}{9}$   
 b) Hallar  $P(a)$  y  $P(b)$  si  $P(c) = P(d) = \frac{1}{4}$  y  $P(a) = 2P(b)$   
 c) Hallar  $P(b, c, d)$  si  $P(b, c) = \frac{1}{3}$ ,  $P(b, d) = \frac{1}{4}$  y  $P(b) = \frac{1}{5}$   
 d) Hallar  $P(a)$  si  $P(c, d) = \frac{2}{3}$ ,  $P(b, d) = \frac{1}{2}$  y  $P(b) = \frac{1}{3}$
10. Consideremos el espacio de sucesos  $\mathcal{A} = \{A, B, C, D\}$ . Determine si la siguiente función:
- $$P(A) = 3/7; P(B) = 0; P(C) = 2/7; P(D) = 2/7$$
- define una probabilidad sobre ese espacio.
11. Demostrar que si  $P$  es una probabilidad sobre  $E$ , entonces  $P(A - B) = P(A) - P(A \cap B)$  para cualquiera dos sucesos  $A$  y  $B$  de  $E$ .
12. Sean  $A$  y  $B$  dos sucesos tales que  $P(A \cup B) = 0'8$ ,  $P(A \cap B) = 0'3$  y  $P(B - A) = 0'2$ .
- a) Represente mediante un diagrama de Venn la situación planteada.  
 b) Calcule la probabilidad de los sucesos  $A$  y  $B$ .  
 c) Calcule la probabilidad de los siguientes sucesos:  $A - B$ ,  $B - \bar{A}$ ,  $B \cap \bar{A}$  y  $\bar{B} \cap \bar{A}$ .
13. Halla la probabilidad de un suceso sabiendo que la suma de su cuadrado y del cuadrado de la probabilidad del suceso contrario es  $1/2$ .
14. Un programa informático combina al azar los colores rojo, azul, verde, amarillo y negro, para obtener una bandera de tres franjas horizontales de colores (no necesariamente distintos). ¿Qué probabilidad hay de que la bandera obtenida coincida con la alemana? ¿Qué probabilidad hay de que la bandera obtenida coincida con la española?
15. Cinco amigos que van de viaje, llegan a un hotel donde sólo quedan libres dos habitaciones, una doble y una triple. Si en la recepción del hotel asignan las habitaciones al azar, se pide:
- a) ¿Qué probabilidad hay de que Juan duerma en la misma habitación que Marta?  
 b) ¿Cómo cambiaría esa probabilidad si el hotel dispusiera de tres habitaciones, una individual y dos dobles?
16. En una pandilla de cinco amigos, ¿qué probabilidad hay de que haya, al menos, dos amigos que cumplan años el mismo día? ¿Cuántos amigos tendría que tener la pandilla para que esa probabilidad fuese  $1/2$ ?
17. Si elegimos al azar un punto en el cuadrado de vértices  $(0, 0)$ ,  $(2, 0)$ ,  $(2, 2)$ ,  $(0, 2)$ , ¿qué probabilidad hay de que pertenezca al círculo de centro  $(1, 1)$  y radio 1 inscrito en el cuadrado. Generalizar este resultado a un círculo de radio  $r$  inscrito en un cuadrado de lado  $2r$ .
18. Un ejercicio de oposición consiste en responder adecuadamente a las preguntas relativas a dos temas. Para cada opositor, se realiza un sorteo entre los 100 temas que componen el temario y se extraen, al azar, tres temas, de los cuales, el opositor elige los dos temas del ejercicio de oposición. Se pide:
- a) Si un opositor se ha estudiado 65 temas, ¿qué probabilidad tiene de realizar satisfactoriamente el ejercicio, es decir, de que coincidan, al menos, dos de los tres temas obtenidos al azar, con los que ha estudiado?



- b) Determine una fórmula general que permita calcular la probabilidad de realizar satisfactoriamente el ejercicio de oposición en función del número ( $x$ ) de temas estudiados por el opositor.
  - c) ¿Cuántos temas ( $x$ ) debe estudiar un opositor si desea tener una probabilidad de aprobar, superior al 90 %?
  - d) Determine una fórmula que permita conocer el número de temas, que debe estudiar un opositor, en función de la probabilidad de conocer, al menos, dos de los tres temas obtenidos al azar.
  - e) Fórmula general: Determine la fórmula general que relaciona el número de temas estudiados ( $x$ ) con la probabilidad de éxito ( $p$ ), en función del número total de temas del temario ( $N$ ), de los temas extraídos al azar ( $T$ ) y del número de ellos ( $t$ ) que debe conocer para aprobar.
19. Otro examen de oposición tiene un temario de 50 temas clasificados en dos bloques de 30 y 20 respectivamente. Para aprobar el examen hay que responder acertadamente a las preguntas de 3 temas elegidos al azar en un sorteo; 2 del primer bloque y 1 del segundo.
- a) Si un opositor ha estudiado 15 temas del primer bloque y 10 del segundo, ¿qué probabilidad tiene de conocer los tres temas de la oposición?
  - b) Determinar la fórmula general que permita calcular la probabilidad de aprobar en función del número total de temas del temario ( $N$ ), clasificados en dos bloques (de  $N_1$  y  $N_2$  temas respectivamente, con  $N = N_1 + N_2$ ), del número de temas estudiados del primer bloque ( $x_1$ , con  $0 \leq x_1 \leq N_1$ ) y del número de temas estudiados del segundo bloque ( $x_2$ , con  $0 \leq x_2 \leq N_1$ ).
20. Distribución Hipergeométrica: En un pueblo de 100 vecinos, 60 de ellos son mujeres y 40 son hombres. Si el ayuntamiento sortea cuatro entradas gratuitas para el concierto de la feria del pueblo, qué probabilidad hay de que haya el mismo número de hombres que de mujeres agraciadas. Generalización: Supongamos que en el pueblo viven  $N_1$  hombres y  $N_2$  mujeres que hacen un total de  $N$  habitantes. Si el ayuntamiento sortea  $n$  entradas, qué probabilidad hay de que sean agraciados  $n_1$  hombres y  $n_2$  mujeres, con  $n = n_1 + n_2$ .
21. En todas las monedas españolas de 2 euros figura, en el reverso, la inscripción “2 EUROS”, pero en el anverso puede aparecer una imagen de S.M. el Rey Juan Carlos I de Borbón, de don Quijote de la Mancha o, recientemente, de la Mezquita Catedral de Córdoba. Determine el espacio muestral y la función de probabilidad del experimento consistente en lanzar
- a) dos monedas de 2 euros con la imagen de S.M. el Rey Juan Carlos I de Borbón.
  - b) ¿dos monedas de 2 euros, pero una de ellas con la imagen de don Quijote de la Mancha y la otra con la imagen de la Mezquita Catedral de Córdoba.
  - c) una misma moneda de 2 euros, dos veces.
  - d) Repetir todo el ejercicio utilizando tres monedas de 2 euros.
22. Consideramos el experimento aleatorio consistente en lanzar dos dados iguales:
- a) Determinar el espacio muestral.

- b) Determinar los sucesos:  $A$ ="Las caras son iguales",  $B$ ="La suma de las caras es mayor que 8",  $C$ ="La suma de las caras es igual a 5" y  $D$ ="La suma de las caras es par".
- c) Calcular la probabilidad de los sucesos del apartado anterior.
- d) Determinar los sucesos:  $\bar{A}$ ,  $\bar{B}$ ,  $A \cup B$ ,  $A \cap B$ ,  $A \cup C$ ,  $A \cap C$  y  $A - D$ .
- e) Calcular la probabilidad de los sucesos del apartado anterior haciendo uso de las propiedades de la función de probabilidad.
- f) Analizar las diferencias de este experimento respecto a otro que utilizase dos dados distintos (diferente color por ejemplo).
23. Consideremos el experimento de lanzar simultáneamente una dado y una moneda.
- a) Determine el espacio muestral y la función de probabilidad.
- b) Si me dan un euro por cada punto obtenido en el dado y un euro más si sale cara o dos euros más si sale cruz, determine el nuevo espacio muestral de las ganancias que esperamos obtener y la probabilidad del suceso "ganar 6 euros o más".
24. Un dado se lanza dos veces. Halla la probabilidad de obtener 4, 5 ó 6 en el primer lanzamiento y 1, 2 ó 3 en el segundo.
25. Dos amigos salen de caza. El primero acierta un promedio de 2 piezas cada 5 disparos y el segundo una pieza cada dos disparos. Si los dos disparan al mismo tiempo a una misma pieza. ¿Cuál es la probabilidad de que la pieza haya sido alcanzada?
26. En una tienda de electrodomésticos nos informan de que, la probabilidad de que se averíe una lavadora durante su periodo de garantía es  $1/4$  y la de que se averíe un frigorífico, durante el periodo de garantía, es  $1/3$ . Supongamos que adquirimos ambos electrodomésticos. Calcule la probabilidad de los siguientes sucesos:
- a) Durante el período de garantía se averían los dos electrodomésticos.
- b) Algún electrodoméstico se avería durante su periodo de garantía.
- c) Durante el período de garantía sólo se avería la lavadora.
- d) Durante el período de garantía sólo se avería el frigorífico.
27. ¿Cuál es la probabilidad de hundir un barco, sabiendo que sólo pueden lanzarse 3 torpedos, y que la probabilidad de hundir un barco con cada torpedo es  $0'2$ ?
- ¿Cuántos torpedos habría que lanzar para que la probabilidad de hundir un barco fuera, al menos, del 90 %?
28. Un aparato consta de dos partes  $A$  y  $B$ , que se fabrican de manera independiente. Se sabe que en el proceso de fabricación la probabilidad de que la parte  $A$  salga defectuosa es  $0'1$  y la probabilidad de un defecto en  $B$  es de  $0'03$ . ¿Cuál es la probabilidad de que el aparato sea defectuoso?
29. Sean  $A$  y  $B$  dos sucesos independientes tales que la probabilidad de que ocurran los dos sucesos es  $1/3$  y de que no ocurra ninguno de los dos es  $1/6$ . Calcule el valor de  $P(A)$  y de  $P(B)$ .

30. Un sistema electrónico de dos componentes se conecta en paralelo de modo que falle sólo si sus dos componentes fallan. La probabilidad de que el primer componente falle es  $0'10$  y de que falle el segundo es  $0'05$ . Suponiendo que ambos componentes funcionan independientemente, se pide:
- ¿Qué probabilidad hay de que el sistema funcione?
  - Si el sistema dispone los componentes en serie, ¿cómo varía esa probabilidad?
  - Recalcular esta probabilidad si al sistema original (en paralelo) le añadimos un nuevo componente en serie que tiene una probabilidad de fallar de  $0'2$ .
31. Una resistencia  $R$  se quema una de cada 100 veces que se enciende un aparato durante más de 12 horas. Recientemente han salido al mercado unas nuevas resistencias  $R+$  que se queman una de cada 300 veces que el aparato está encendido durante más de 12 horas. Las resistencias  $R$  vienen en un blíster de 3 unidades y su precio es de 5 euros. Las resistencias  $R+$  se venden sueltas a 5 euros cada una. Con 5 euros, ¿qué es más eficiente: un sistema con las 3 resistencias  $R$  en paralelo (de manera que el sistema no funciona si no funciona ninguna de las tres) o con una única resistencia  $R+$ ? ¿Y si hubiera una oferta de lanzamiento de  $R+$  de  $2 \times 1$ ?
32. Distribución Binomial: Si lanzamos al aire 4 veces una moneda perfecta, ¿qué probabilidad hay de salgan dos caras y dos cruces? Generalización: ¿qué probabilidad hay de que salgan  $n_1$  caras y  $n_2$  cruces, si lanzamos al aire  $n$  veces ( $n = n_1 + n_2$ ) una moneda trucada con probabilidad  $p$  de salir cara?
33. Distribución Geométrica: Se lanza una moneda al aire tantas veces como sea necesario hasta obtener una cara. ¿Qué probabilidad hay de tener que lanzar cinco veces la moneda? Generalización: ¿Qué probabilidad hay de tener que lanzar  $x$  veces una moneda trucada con probabilidad  $p$  de salir cara?
34. Distribución Binomial negativa: Se lanza una moneda al aire tantas veces como sea necesario hasta obtener cara tres veces. ¿Qué probabilidad hay de tener que lanzar diez veces la moneda? Generalización: ¿Qué probabilidad hay de tener que lanzar  $x$  veces una moneda trucada con probabilidad  $p$  de salir cara para obtener  $n$  caras?
35. Un tirador dispara sobre una diana y sabe que la probabilidad de que acierte es  $1/3$ . Se pide:
- Calcular la probabilidad de que acierte al menos una vez si dispara 8 veces.
  - Calcular la probabilidad de no acertar en 8 disparos consecutivos.
36. Si dos dados se lanzan 20 veces, hallar:
- La probabilidad de obtener alguna vez “doble 6”.
  - No haya sumado nunca 8 puntos.
  - Alguna vez sume 8 puntos.
37. En un taller trabajan 10 obreros y la probabilidad de que uno cualquiera de ellos esté de baja es  $0'1$ . Determine la probabilidad de que un día
- vengan todos los obreros a trabajar.

- b) falte al trabajo al menos un obrero.
38. 5 profesores imparten todos los días una hora de clase a un grupo de 20 alumnos. La probabilidad de que falte un día a clase un profesor es  $0'01$  y la de que falte un alumno es  $1/20$ . Calcule la probabilidad de los siguientes sucesos:
- No venga a clase ningún profesor.
  - Falte a clase algún profesor.
  - Falte algún alumno a clase.
  - Vengan a clase todos los alumnos y todos los profesores.
  - Generalizar los resultados anteriores cuando la clase tiene  $n$  alumnos y la probabilidad de que falte uno de ellos es  $1/n$ .
39. Una moneda está trucada y la probabilidad de salir cara es tres veces mayor que la probabilidad de salir cruz. Consideremos el experimento de lanzar tres veces esta moneda y anotar el número de caras obtenido en los tres lanzamientos.
- Determine el espacio muestral y la función de probabilidad.
  - ¿Qué resultado es más probable, que salga alguna cruz en los tres lanzamientos o que el resultado de los tres lanzamientos sea el mismo?
  - Si repetimos dos veces el experimento, ¿qué probabilidad hay de que, en alguna de las dos veces, hayan salido tres cruces?
40. Si  $P(A) = 1/3$ ,  $P(B) = 1/4$  y  $P(A \cap B) = 1/5$ .
- Halle las probabilidades de los sucesos:  $A|B$ ,  $B|A$ ,  $\bar{A}|B$ ,  $\bar{B}|\bar{A}$ ,  $A \cup B$  y  $\bar{B}|A$ .
  - ¿Son  $A$  y  $B$  incompatibles? ¿Son  $A$  y  $B$  independientes?
41. Si  $P(A \cup B) = 0'8$ ,  $P(A \cap B) = 0'3$  y  $P(B - A) = 0'2$ .
- Halle las probabilidades de los sucesos:  $A|B$ ,  $A|\bar{B}$ ,  $(B - A)|(A \cup B)$ .
  - ¿Son  $A$  y  $B$  incompatibles? ¿Son  $A$  y  $B$  independientes?
42. Sean  $A$  y  $B$  dos sucesos que verifican lo siguiente:
- La probabilidad de no ocurran simultáneamente los dos sucesos es  $0'5$ ,
  - la probabilidad de que no ocurra el suceso  $B$  es  $0'9$  y
  - la probabilidad de que ocurra el suceso  $B$ , sabiendo que ha ocurrido el suceso  $A$  es  $1/3$ .

Determine la probabilidad de que ocurra el suceso  $A$  y responda a las siguientes preguntas:

- ¿Son  $A$  y  $B$  sucesos equiprobables? Justifique la respuesta.
  - ¿Son  $A$  y  $B$  sucesos independientes? Justifique la respuesta.
  - ¿Son  $A$  y  $B$  sucesos incompatibles? Justifique la respuesta.
43. En el experimento de lanzar un dado se consideran los siguientes sucesos:  $A =$  “Obtener un número mayor que 4” y  $B =$  “Obtener un múltiplo de 3”. Se pide:

- a) Utilice la definición frecuentista de probabilidad para calcular  $P(A)$ ,  $P(B)$ ,  $P(A \cap B)$ ,  $P(A \cup B)$ ,  $P(A|B)$  y  $P(B|A)$ .
  - b) Probar que el resultado obtenido en el apartado anterior es el mismo si aplicamos la definición de probabilidad condicionada para calcular  $P(A|B)$  y  $P(B|A)$ .
44. Se lanzan dos dados. Si la suma de los puntos de las caras superiores es 5, hallar la probabilidad de que en alguno de los dados salga 2. Realice este ejercicio de dos maneras distintas (utilizando o no la probabilidad condicionada).
45. Se lanzan dos dados al aire y se anota la suma de los puntos obtenidos. Se pide:
- a) Determinar el espacio muestral.
  - b) Calcular la probabilidad de anotar un 7.
  - c) ¿Son equiprobables todos los sucesos elementales?
  - d) Calcular la probabilidad de que el número obtenido sea par.
  - e) Calcular la probabilidad de que el número obtenido sea impar.
  - f) Si sabemos que uno de los dados salió 4, ¿cómo cambia esta información, el valor de la probabilidad de obtener 6 puntos?
46. La probabilidad de fallo en tres máquinas A, B y C son: 0'1, 0'05 y 0'01. Determine el espacio muestral y la función de probabilidad y calcule las siguientes probabilidades:
- a) Probabilidad de que funcione alguna.
  - b) Probabilidad de que fallen 2 máquinas a la vez.
  - c) Probabilidad de que funcionen las 3.
  - d) Probabilidad de que si existe un único fallo, éste se deba a la máquina A.
  - e) Probabilidad de que si existen dos fallos, alguno se haya producido en la máquina A.
  - f) Probabilidad de que si existe fallo (uno o varios), esté averiada la máquina A.
47. Una urna contiene tres bolas rojas y siete negras. Se extraen dos bolas al azar. Describir el espacio muestral  $E$  y la función de probabilidad  $P$  cuando:
- a) Se extraen con reemplazamiento.
  - b) Se extraen sin reemplazamiento.
48. Una caja contiene cuatro bolas blancas y dos negras. Se saca una bola y a continuación (sin devolver la primera a la caja) se extrae otra. Consideramos los sucesos:  $A$  = "la primera bola extraída es blanca" y  $B$  = "la segunda bola extraída es blanca". Se pide:
- a) Hallar  $P(A)$  y  $P(B|A)$ .
  - b) ¿Son  $A$  y  $B$  dos sucesos independientes?
  - c) ¿Cual es la probabilidad de que las dos bolas extraídas sean blancas?
  - d) ¿Cual es la probabilidad de que las dos bolas extraídas sean negras?
49. Se sacan dos bolas de una urna que se compone de una bola blanca, otra roja, otra verde y otra negra. Describa el espacio muestral  $E$  y la función de probabilidad  $P$  cuando:
- a) La primera bola se devuelve a la urna antes de sacar la segunda.

- b) La primera bola no se devuelve a la urna antes de sacar la segunda.
  - c) Se extraen simultáneamente de la urna las dos bolas.
50. Una caja contiene 12 objetos de los cuales 5 son defectuosos. Si se van tomando hasta encontrar uno defectuoso. Encontrar:
- a) Espacio muestral.
  - b) Probabilidad de que se obtenga en la sexta extracción.
51. Una caja contiene dos bolas blancas y dos bolas negras y, sin mirar, se van sacando bolas de la caja, consecutivamente y sin reemplazamiento, hasta que aparezcan las dos bolas negras. Determine el espacio muestral del experimento y calcule la probabilidad de los sucesos elementales, justificando matemáticamente los cálculos realizados.
52. Lanzamos una moneda perfecta tantas veces como sea necesario hasta que salga cara, y anotamos el número total de lanzamientos que han sido necesarios. Determine el espacio muestral y la función de probabilidad, y compruebe que la suma de las probabilidades de todos los sucesos elementales es 1.
53. Demostrar que si  $A$  y  $B$  son dos sucesos independientes entonces  $\bar{A}$  es también independiente de  $B$ . Indicación: Probar que  $P(B|\bar{A}) = P(B)$  aplicando que  $B \cap \bar{A} = B - (A \cap B)$ .
54. La probabilidad de que un hombre viva más de 75 años es  $1/4$  y la de que su mujer viva más de 75 años es  $1/3$ . Se pide:
- a) Calcular la probabilidad de que ambos vivan más de 75 años.
  - b) Calcular la probabilidad de que el hombre viva más de 75 años y la mujer no.
  - c) Calcular la probabilidad de que ambos mueran antes de los 75 años.
55. Una urna  $A$  contiene 4 bolas blancas y 6 rojas y otra  $B$  contiene 7 bolas blancas y 5 rojas. Si se extrae una bola de la urna  $B$
- a) ¿Cuál es la probabilidad de sacar una bola roja?
  - b) ¿Cómo cambia el valor de esta probabilidad si sabemos que antes de la extracción se saca una bola de la urna  $A$  y se pasa a la urna  $B$ ?
56. Se tiene una urna vacía y se lanza una moneda al aire. Si sale cara se introduce en la urna una bola blanca y si sale cruz se introduce una bola negra. El experimento se repite tres veces y a continuación se introduce la mano en la urna y se saca una bola. ¿Cuál es la probabilidad de que en la urna quede una bola de cada color?
57. En la estantería de libros de Matemáticas de una biblioteca hay un libro de álgebra, siete copias del mismo libro de cálculo y cuatro copias del mismo libro de estadística. El libro de álgebra tiene 300 páginas y 24 capítulos; el de cálculo tiene 350 páginas y 20 capítulos; y el de estadística tiene 400 páginas y 22 capítulos.
- a) Determine la probabilidad de que, elegido un libro al azar, al abrirlo obtengamos una página que encabeza un capítulo.
  - b) Sabiendo que, al abrir al azar un libro elegido también al azar, hemos obtenido una página que encabeza un capítulo, ¿de qué rama de la Matemática es más probable que sea?

- c) Sabiendo que, al abrir al azar un libro elegido también al azar, no hemos obtenido una página que encabeza un capítulo, ¿de qué rama de la Matemática es más probable que sea?
58. El 20 % de los productos fabricados por la empresa A y el 5 % de los fabricados por la empresa B tienen algún defecto.
- a) Si mis únicos suministradores son estas dos empresas, ¿qué porcentaje de productos debo adquirir en cada una si estoy dispuesto a admitir entre mis productos un total del 10 % de defectuosos.
- b) Utilizando el porcentaje anterior, ¿qué probabilidad hay de que haya sido fabricado por la empresa A un producto que elegido al azar resultó ser defectuoso?
59. El 2 % de una población padece una enfermedad E, existiendo un síntoma S, tal que el 27 % de los enfermos presentan el síntoma, mientras que un 5 % de los individuos no enfermos presentan el síntoma. Calcular los porcentajes de individuos con el síntoma y de individuos enfermos que presentan el síntoma.
60. En una operación de fabricación se utilizan dos líneas de producción para ensamblar fusibles electrónicos. Ambas líneas producen fusibles con la misma velocidad y generalmente 2'5 % de los fusibles que producen están defectuosos. Sin embargo, la línea 1 de producción experimentó recientemente problemas mecánicos y produjo 6 % de fusibles defectuosos durante un periodo de 3 semanas. Esta situación no se conoció antes de que varios lotes de fusibles electrónicos producidos en este periodo se enviaran a los clientes. Si uno de los dos fusibles probados por un cliente resultó tener defectos, ¿qué probabilidad hay de que el lote del que provino se haya producido en la línea que tuvo problemas? (Suponga que todos los fusibles del lote se produjeron en la misma línea).
61. En una planta de electrónica, se sabe por experiencia que la probabilidad de que un obrero de nuevo ingreso que haya asistido al programa de capacitación de la compañía cumpla la cuota de producción es de 0'86 y que la probabilidad correspondiente de un obrero de nuevo ingreso que no ha asistido a dicho curso de capacitación es de 0'35. Si el 80 % de la totalidad de los obreros de nuevo ingreso asisten al curso de capacitación, se pide:
- a) ¿Qué probabilidad existe de que un trabajador de nuevo ingreso cumpla la cuota de producción?
- b) ¿Qué probabilidad hay de que un obrero de nuevo ingreso que satisface la cuota de producción haya asistido al curso de capacitación de la compañía?
62. *Errores de diagnóstico.* Una cierta enfermedad la padece el  $p$  % de la población. Se sabe que la probabilidad de detectar la enfermedad, mediante un análisis no del todo fiable, en una persona enferma es la misma que la de no detectarla en una persona sana, siendo estas probabilidades la proporción de personas que no padecen la enfermedad en dicha población.
- a) Se le aplica el análisis a una persona y resulta negativo. Calcular la probabilidad de que haya habido un error en el diagnóstico.
- b) Se le aplica el análisis a una persona y resulta positivo. Calcular la probabilidad de que haya habido un error en el diagnóstico.

63. *Secuencialidad del Teorema de Bayes.* Un prisionero político en Rusia será exiliado a Siberia o a los Urales, y él no sabe a cual de los dos será enviado, pero sabe que la probabilidad de ser exiliado a Siberia es  $0'8$ . También sabe que si un residente en Siberia es seleccionado aleatoriamente, la probabilidad de que lleve un abrigo de pieles es  $0'5$ , mientras que en los Urales, ésta es de  $0'7$ . Al llegar a su lugar de exilio, la primera persona que ve no lleva abrigo de pieles. Se pide:
- ¿Cuál es la probabilidad de que esté en Siberia?
  - Teniendo en cuenta la información anterior, la siguiente persona que ve tampoco lleva abrigo de pieles. ¿Cuál es ahora la probabilidad de que esté en Siberia?
  - ¿Y si hubiese visto juntas a las dos personas en el primer momento?
64. Supóngase que una caja contiene 5 monedas, y que la probabilidad de obtener cara en un lanzamiento es distinta para cada moneda. Sea  $p_i$  la probabilidad de obtener cara al lanzar la  $i$ -ésima moneda ( $i=1,2,3,4,5$ ) y supóngase que  $p_i = (i - 1)/4$ .
- Supóngase que se selecciona una moneda de la caja al azar, y que al lanzarla una vez se obtiene cara, ¿cuál es la probabilidad de que se haya seleccionado la  $i$ -ésima moneda?
  - Si la misma moneda es lanzada otra vez, ¿cuál será la probabilidad de obtener otra cara?
  - Si se ha obtenido una cruz en el primer lanzamiento de la moneda seleccionada y se lanza otra vez la misma moneda, ¿cuál es la probabilidad de obtener una cara en el segundo lanzamiento?
  - Supongamos que, con la misma caja realizamos el siguiente experimento: seleccionamos aleatoriamente una moneda de la caja y la lanzamos repetidamente hasta que obtenemos una cara. Si se obtiene la primera cara en el cuarto lanzamiento, ¿cuál es la probabilidad de que se haya seleccionado la  $i$ -ésima moneda?
  - Si se continúa lanzando la misma moneda hasta que aparece otra cara, ¿cuál es la probabilidad de que se necesiten exactamente tres lanzamientos?
65. En una ciudad existen dos fábricas de pelotas de tenis. En la fábrica  $F_1$  el porcentaje de ellas que se fabrican de calidad  $A$  es del 80 %, de calidad  $B$  es del 5 % y de calidad  $C$  del 15 %. En la fábrica  $F_2$  los porcentajes son  $a$ ,  $b$  y  $c$  respectivamente.
- Dar una expresión general, lo más simplificada posible, de la proporción de pelotas de calidad  $A$  para toda la ciudad.
  - Sabiendo que  $a=92\%$  y que el porcentaje de pelotas de calidad  $A$  en toda la ciudad es del 89 %, ¿cuál de las dos fábricas produce más pelotas de tenis?
  - Si el porcentaje de pelotas de calidad  $B$  en toda la ciudad es del 5 %, ¿qué valores toman  $b$  y  $c$ ? y entonces ¿cuál es la proporción de pelotas fabricadas por  $F_2$  entre las de calidad  $C$ ?
66. *En contra de la intuición.* Proponemos cuatro ejemplos de la vida cotidiana donde nuestra intuición no coincide con la realidad, poniendo de manifiesto que saber un poco de matemáticas puede ayudarnos a no dejarnos engañar por las falsas apariencias.



- a) Coincidencia de cumpleaños. En ocasiones nos sorprendemos por “coincidencias” que no son extraordinarias. Por ejemplo. En una comida con 25 personas, dos cumplen años el mismo día. La probabilidad de que eso suceda puede parecernos bastante baja, ya que hay 366 fechas posibles. Pero no lo es. A partir de 23 personas ya hay un 50 % de probabilidades de que dos compartan día de nacimiento. Con 30 personas supera el 70 %. Y en una reunión de 70 pueden apostar lo que quieran con garantías de ganar: supera el 99 %.
- b) Saber y ganar. El concursante de un programa de televisión se enfrenta a la prueba final, en la que hay tres puertas. Detrás de una de ellas hay un coche, y tras las otras dos, nada. Elige una y el presentador ordena abrir alguna de las otras dos, siempre una sin premio. Entonces, tienta al concursante: “¿Desea cambiar de puerta?”. La intuición nos dice que da igual, que tendremos un 50 % de probabilidades de acertar. Pero no es así. Si nos quedamos en la misma solo tendremos una probabilidad de  $1/3$  (33 %) de conseguir el premio, igual que al principio. Pero si cambiamos, la probabilidad de obtener el coche será de  $2/3$ : seremos ganadores siempre que nuestra primera opción no fuera la correcta. Y partíamos con un 66 % de probabilidades de equivocarnos.
- c) Diagnóstico terrible. Nos hacen una prueba para averiguar si padecemos una grave enfermedad que afecta a una de cada 200 personas. El análisis tiene el 98 % de fiabilidad, esto es, falla el 2 % de las veces. Damos positivo. ¿Debemos asustarnos? Sí, pero no en exceso. La probabilidad de que padezcamos el mal es del 20 %. De cada 10.000 personas, unas 50 tendrán la enfermedad. De ellas, 49 obtendrán un resultado positivo en la prueba y una dará negativo (por el margen de error). En cuanto a la población sana (9.950 personas), 9.751 darán negativo y 199 positivo. Luego la mayoría de las personas diagnosticadas del mal en ese análisis (199 de 248) serán en realidad falsos positivos (80 %).
- d) ¿Es tan improbable? 30 personas van a una fiesta y dejan su sombrero en un perchero. A la salida, cada una toma uno sin fijarse bien si es el suyo. ¿Qué probabilidad hay de que ninguna acierte? La intuición nos señala que es muy difícil que suceda, pero no lo es tanto. La probabilidad de que ninguno de los asistentes se lleve su sombrero es de alrededor del 37 %. Aproximadamente la misma, por cierto, que la de que acierte solo uno.

Utilice los conocimientos adquiridos en este tema de probabilidad para justificar los razonamientos y comprobar los cálculos que se proporcionan en los cuatro ejemplos anteriores.



# Apuntes de ESTADÍSTICA

## Variable aleatoria



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento a los profesores Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.*

## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Capítulo 5

# Variable aleatoria

Los posibles resultados de un experimento son todos los sucesos que constituyen el espacio muestral. A menudo, nos interesa que estos resultados sean numéricos. En este caso, utilizamos una función que permita clasificar a los sucesos, asignando valores numéricos a cada uno de ellos.

Por ejemplo, si el experimento aleatorio consiste en lanzar tres veces una moneda, entonces el espacio muestral se puede representar por  $\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$  donde  $H$  al suceso “salir cara” y  $T$  es el suceso “salir cruz”. Pero si estamos interesados en determinar el número de caras obtenidas en los tres lanzamientos de la moneda, entonces podemos definir una función  $X$  que asigna un valor numérico (número de caras) a cada resultado del experimento. De esta manera, tenemos, por ejemplo, que  $X(HTH) = 2$  o que  $X(TTT) = 0$ . Este tipo de funciones, cuyos valores dependen de los resultados de un experimento aleatorio, se llaman variables aleatorias.

Las variables aleatorias y sus distribuciones de probabilidad, pueden considerarse una generalización del concepto frecuentista de probabilidad. Se introducen como el modelo matemático ideal al que se aproximan las distribuciones de frecuencias que se obtendrían en una repetición indefinida de pruebas de este experimento. Por ello, nos recuerdan a las variables estadísticas y a sus distribuciones de frecuencia que ya hemos estudiado en estadística descriptiva.

Las variables aleatorias se clasifican conforme al rango de valores que pueden asumir, y llamaremos *soporte* a ese conjunto de posibles valores (números reales) que puede tomar una variable aleatoria.

En este capítulo estudiaremos principalmente las variables aleatorias discretas, cuyo soporte está formado por un número finito, o infinito numerable de valores (p.e. número de defectos en una inspección de productos, número de elementos en espera en una cola, etc.) y las variables aleatorias continuas cuyo soporte es un intervalo o conjunto de intervalos de números reales (p.e. durabilidad de un dispositivo, velocidad de un automóvil, resistencia a la tensión de una nueva aleación, etc.).

Además, al final del tema, estudiaremos las variables aleatorias bidimensionales y algunos aspectos asociados relativos a las distribuciones, medidas y regresión. Su analogía con ellas, nos recordará lo estudiado en el tema de regresión y correlación, sin más que cambiar la frecuencia por la probabilidad, en la mayoría de las fórmulas.

## 5.1. Variable aleatoria unidimensional

Sea  $(\Omega, \mathcal{A}, P)$  un espacio probabilístico asociado a un experimento aleatorio. Una variable aleatoria  $X$  es una función definida sobre el espacio muestral  $\Omega$  (conjunto de resultados de un experimento aleatorio) que toma valores en un conjunto de números reales, llamado *soporte*, y que denotaremos por  $S_x$ . Se suelen utilizar las abreviaturas “v.a.u.” o simplemente “v.a.” para referirse a las variables aleatorias unidimensionales.

En términos matemáticos precisos, una variable aleatoria unidimensional es una aplicación  $X : \Omega \rightarrow \mathbb{R}$  que verifica la siguiente propiedad:

$$\text{para todo } x \in \mathbb{R} \text{ el conjunto } \{\omega \in \Omega / X(\omega) \leq x\} \in \mathcal{A}.$$

**Ejemplo 5.1** *Utilice una variable aleatoria para modelizar el experimento que consiste en lanzar dos veces un dado y anotar la suma de las puntuaciones obtenidas.*

El espacio muestral del experimento que consiste en lanzar dos veces un dado se puede representar así:

$$\Omega = \{(11), (12), (13), (14), (15), (16), (21), (22), \dots, (66)\}$$

y nos permite considerar la variable aleatoria  $X$  que suma el valor de las puntuaciones obtenidas en los dos dados:

$$X : \Omega \rightarrow \mathbb{R} \quad \text{tal que} \quad X(ij) = i + j \quad \text{siendo} \quad (ij) \in \Omega$$

Así, por ejemplo,  $X(11) = 2$ ,  $X(36) = 9$  ó  $X(66) = 12$ , de manera que el soporte de esta variable es el conjunto

$$S_x = \{2, 3, 4, 5, \dots, 12\}$$

y, por lo tanto, la variable aleatoria  $X$  es discreta. □

## 5.2. Función de distribución

Sea  $(\Omega, \mathcal{A}, P)$  un espacio probabilístico asociado a un experimento aleatorio y sea  $X$  una variable aleatoria. Definimos la función de distribución  $F : \mathbb{R} \rightarrow [0, 1]$ , asociada a la variable aleatoria  $X$ , de la siguiente manera:

$$F(x) = P(\{\omega \in \Omega / X(\omega) \leq x\}) = P(X \leq x) \quad \text{para todo } x \in \mathbb{R}$$

La función de distribución es única para cada variable aleatoria a la que caracteriza, resulta especialmente útil para calcular probabilidades ya que:

- $P(X \leq x) = F(x)$
- $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$
- $P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$

y sus principales propiedades son:

1.  $0 \leq F(x) \leq 1$  para todo  $x \in \mathbb{R}$
2.  $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$
3.  $F(\infty) = \lim_{x \rightarrow \infty} F(x) = F(\infty) = 1$
4.  $F$  es monótona no decreciente, es decir, si  $x_1 < x_2$  entonces  $F(x_1) \leq F(x_2)$ .
5.  $F$  es continua por la derecha, es decir,  $\lim_{h \rightarrow 0^+} F(x+h) = F(x)$

Veamos ahora que forma tiene esta función de distribución en cada uno de los tipos de variables (discretas y continuas) que vamos a estudiar.

### 5.3. Variable aleatoria discreta

Una variable aleatoria  $X$  se dice que es *discreta* si el soporte  $S_x$  es un conjunto discreto, es decir, cuando la variable  $X$  toma un número finito o infinito numerable de valores reales. Por ejemplo, el número de defectos observados en un control de calidad o el número de elementos que esperan en una cola son variables aleatorias discretas. Se suelen utilizar la abreviatura “v.a.d.” para referirse a las variables aleatorias discretas.

A continuación, haremos corresponder una probabilidad a cada valor de la variable aleatoria, lo cual constituye la distribución de probabilidad de la variable aleatoria, que nos recuerda a las distribuciones de frecuencias asociadas a las variables estadísticas.

#### 5.3.1. Distribución de probabilidad

Sea  $(\Omega, \mathcal{A}, P)$  un espacio probabilístico asociado a un experimento aleatorio y sea  $X$  una variable aleatoria discreta que toma los valores en el conjunto  $S_x = \{x_1, x_2, x_3, \dots\}$ . Definimos la probabilidad  $p(x_i)$  para cada uno de los elementos del soporte, de la siguiente manera:

$$p(x_i) = P(\{\omega \in \Omega / X(\omega) = x_i\}) = P(X = x_i)$$

La distribución de probabilidad de la variable  $X$  está constituida por los elementos del soporte  $S_x$  junto a sus correspondientes valores de probabilidad. Normalmente, se representa en forma de tabla, de la siguiente manera:

$x$	$p(x)$
$x_1$	$p(x_1)$
$x_2$	$p(x_2)$
$\vdots$	$\vdots$
$x_n$	$p(x_n)$
$\vdots$	$\vdots$

La representación gráfica de la distribución de probabilidad se realiza en un diagrama de barras. En el eje OX se representan los distintos elementos del soporte, y en el eje OY se representa la probabilidad correspondiente a cada uno de ellos.

**Ejemplo 5.2** Consideramos el experimento consistente en lanzar una moneda tres veces al aire. Definimos la variable aleatoria  $X$  que determina el número de caras ( $H$ ) que aparecen en cada serie de tres lanzamientos. Obtener y representar su distribución de probabilidad.

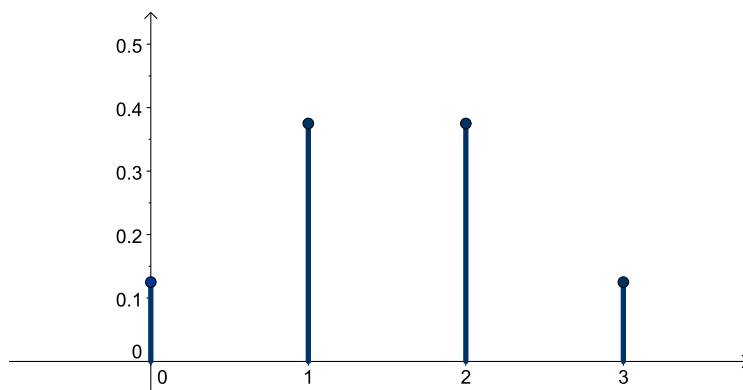
La variable  $X$  toma los valores 0, 1, 2 y 3, que constituyen el soporte. Para calcular la probabilidad de cada uno de ellos recurrimos a los sucesos correspondientes:

$$\begin{aligned} p(0) = P(X = 0) &= P(\{TTT\}) = 1/8 \\ p(1) = P(X = 1) &= P(\{TTH, THT, HTT\}) = 3/8 \\ p(2) = P(X = 2) &= P(\{THH, HTH, HHT\}) = 3/8 \\ p(3) = P(X = 3) &= P(\{HHH\}) = 1/8 \end{aligned}$$

Por tanto la distribución de probabilidad de la variable  $X$  es:

$x$	$p(x)$
0	1/8
1	3/8
2	3/8
3	1/8

y su representación gráfica mediante diagrama de barras es:



Obsérvese la analogía de la distribución de probabilidad de esta variable aleatoria discreta  $X$ , con las distribuciones de frecuencia estudiadas en el tema de estadística descriptiva.  $\square$

A continuación vamos a definir los conceptos de función de distribución, media, varianza y momentos de una variable aleatoria discreta a partir de su distribución de probabilidad. Por analogía, usando el concepto frecuentista de la probabilidad, podríamos definir el resto de las medidas de centralización, dispersión, simetría y apuntamiento tal y como se hizo en el tema de estadística descriptiva.



Para las definiciones que siguen a continuación, consideraremos una variable aleatoria discreta  $X$  que toma los valores en el conjunto  $S_x = \{x_1, x_2, \dots\}$  con probabilidades  $p(x_1), p(x_2), \dots$ . Si el número de valores que toma la variable es infinito numerable es necesario asegurarse de que las series correspondientes, que aparecen en las fórmulas, son absolutamente convergentes.

### 5.3.2. Función de distribución

Dada una variable aleatoria discreta  $X$ , para todo número real  $x$  se define su *función de distribución* asociada  $F(x)$ , de la siguiente manera:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

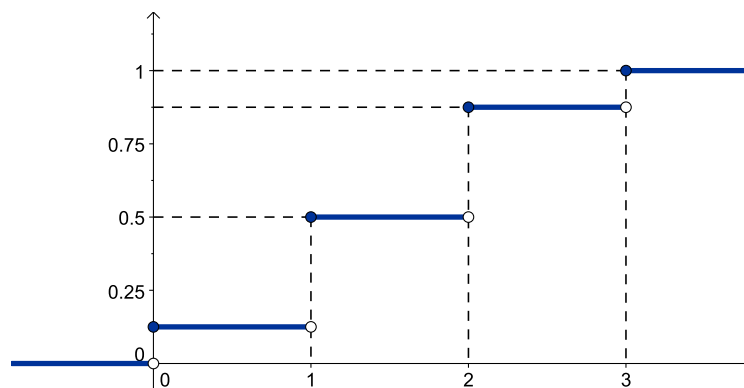
Gráficamente, esta función “acumulativa” adopta una forma de escalera, donde los saltos se producen en los puntos del soporte, siendo  $F(x)$  continua por la derecha, en cada uno de ellos. Además, la altura del salto en cada punto corresponde con la probabilidad de que la variable tome ese valor.

**Ejemplo 5.3** *Obtener y representar la función de distribución de la variable aleatoria definida en el ejemplo 5.2 de la página 176.*

La función de distribución de la variable aleatoria que determina el número de caras que aparecen en cada serie de tres lanzamientos de una moneda perfecta es:

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1/8 & \text{si } 0 \leq x < 1 \\ 4/8 & \text{si } 1 \leq x < 2 \\ 7/8 & \text{si } 2 \leq x < 3 \\ 1 & \text{si } x \geq 3 \end{cases}$$

y su representación gráfica es



que tiene forma de escalera donde los saltos se producen en los puntos del soporte, y la altura del salto corresponde con la probabilidad en el punto.  $\square$

### 5.3.3. Función generatriz de probabilidad

Cuando el soporte de una variable aleatoria discreta  $X$  es el conjunto  $\mathbb{N} = \{0, 1, 2, \dots\}$ , podemos definir la *función generatriz de probabilidad* de la variable aleatoria  $X$  como la serie de potencias

$$G(s) = \sum_{n=0}^{\infty} s^n \cdot p_n \quad \text{con} \quad s \in (-1, 1)$$

donde  $p_n = p(n) = P(X = n)$ . Además, se suelen utilizar la abreviatura “f.g.p.” para referirse a la función generatriz de probabilidad.

La función generatriz es infinitamente derivable y nos permite obtener una de las propiedades más importantes conocida como *teorema de inversión* que establece la distribución de probabilidad de la variable aleatoria  $X$  en términos de su función generatriz:

$$p_n = \frac{G^{(n)}(0)}{n!} \quad \text{para todo } n = 0, 1, 2, \dots$$

**Ejemplo 5.4** Sea  $X$  la variable aleatoria discreta que determina el número de caras ( $H$ ) antes de obtener la primera cruz ( $T$ ) en lanzamientos consecutivos de una misma moneda equilibrada. Determine la función generatriz de probabilidad, compruebe que se verifica el teorema de inversión y responda a la siguiente pregunta: ¿cómo cambiaría esta función si la moneda estuviese trucada con probabilidad  $p$  de salir cara?

El soporte de la variable  $X$  es el conjunto de los números naturales  $S_x = \{0, 1, 2, \dots\}$  y su distribución de probabilidad se determina así:

$$p_n = P(HH \dots HT) = P(H) \cdot P(H) \cdot \dots \cdot P(H) \cdot P(T) = \left(\frac{1}{2}\right)^n \frac{1}{2} = \left(\frac{1}{2}\right)^{n+1}$$

siendo  $p_n = p(n) = P(X = n)$ . Por lo tanto, la función generatriz de probabilidad de  $X$  es:

$$G(s) = \sum_{n=0}^{\infty} s^n \cdot p_n = \sum_{n=0}^{\infty} s^n \frac{1}{2} \left(\frac{1}{2}\right)^n = \sum_{n=0}^{\infty} \frac{1}{2} \left(\frac{s}{2}\right)^n \stackrel{*}{=} \frac{1}{2-s}$$

Obsérvese (\*) que para obtener la expresión explícita de la función en términos de funciones elementales, hemos utilizado que la serie de potencias correspondiente era una serie geométrica convergente para  $s \in (-1, 1)$ .

Calculando las derivadas sucesivas de la función generatriz de probabilidad  $G(s)$ , se puede comprobar que

$$G^{(n)}(s) = \frac{n!}{(2-s)^{n+1}} \quad \longrightarrow \quad G^{(n)}(0) = \frac{n!}{2^{n+1}}$$

y, por lo tanto,

$$p_n = \frac{G^{(n)}(0)}{n!} = \frac{n!/2^{n+1}}{n!} = \left(\frac{1}{2}\right)^{n+1}$$

que pone de manifiesto el teorema de inversión.

Si la moneda estuviese trucada y fuese  $p$  la probabilidad de salir cara, entonces

$$p_n = P(HH \dots HT) = P(H) \cdot P(H) \cdot \dots \cdot P(H) \cdot P(T) = p^n(1-p)$$

y, por lo tanto, la función generatriz de probabilidad sería:

$$G(s) = \sum_{n=0}^{\infty} s^n \cdot p_n = \sum_{n=0}^{\infty} s^n p^n (1-p) = \frac{1-p}{1-ps}$$

para todo  $s \in (-1, 1)$ . □

## 5.4. Variable aleatoria continua

Muchas variables aleatorias que se observan en la vida real no son discretas porque pueden tomar cualquier valor en un intervalo de números, o en uniones de ellos. Por ejemplo, el tiempo de espera en una cola, la durabilidad de un componente electrónico, la velocidad de un automóvil o la resistencia a la tensión de una nueva aleación. A las variables de este tipo las definiremos como *variables aleatorias continuas*.

Matemáticamente, una variable aleatoria  $X$  se dice que es *continua* si su función de distribución  $F(x)$  correspondiente es continua. Se suelen utilizar la abreviatura “v.a.c.” para referirse a las variables aleatorias continuas.

Asociada a cada variable aleatoria continua, existe una función, llamada *función de densidad* que determina la distribución de probabilidad de la variable aleatoria. Veamos, en primer lugar, esta función de densidad y, después, estudiaremos la función de distribución y la relación entre ambas funciones.

### 5.4.1. Función de densidad

Dada una variable aleatoria continua  $X$ , decimos que una función real  $f(x)$ , integrable y no negativa, es la *función de densidad de probabilidad* (o simplemente *función de densidad*) de la variable aleatoria  $X$  si el área encerrada entre la curva y el eje OX es igual a la unidad y, además, la probabilidad de que  $X$  se encuentre entre dos valores  $x_1$  y  $x_2$  con  $x_1 \leq x_2$  es igual al área comprendida entre estos dos valores, es decir,

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad \text{y} \quad P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

Y además, podemos calcular la probabilidad de que la variable tome valores en cualquier otro intervalo. Por ejemplo,

$$P(X \leq x_1) = \int_{-\infty}^{x_1} f(x) dx \quad \text{y} \quad P(X > x_2) = \int_{x_2}^{\infty} f(x) dx$$

El soporte de una variable aleatoria continua es el conjunto de números reales donde la función de densidad  $f(x)$  sea estrictamente positiva. Si este soporte es un intervalo, por ejemplo  $S_x = (a, b)$ , entonces las integrales impropias se reducen a integrales definidas. De esta manera

$$\int_a^b f(x) dx = 1$$

y si  $c$  es un número comprendido entre  $a$  y  $b$  ( $a < c < b$ ) entonces

$$P(X \leq c) = \int_a^c f(x) dx \quad \text{o bien} \quad P(X \geq c) = \int_c^b f(x) dx$$

Obsérvese que la probabilidad de que una variable aleatoria continua tome un valor particular es cero, aunque sea posible. Es decir, la probabilidad medirá intervalos de ocurrencia de la variable, no instancias puntuales. Por lo tanto, no será relevante que una desigualdad sea o no estricta. Por ejemplo  $P(X \leq x) = P(X < x)$  y  $P(X \geq x) = P(X > x)$  o bien

$$P(x_1 < X < x_2) = P(x_1 \leq X < x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X \leq x_2)$$

### 5.4.2. Función de distribución

Dada una variable aleatoria continua  $X$ , a la función acumulativa

$$F(x) = P(X \leq x)$$

se la denomina *función de distribución de  $X$* , y su representación gráfica corresponde a una función continua, creciente, definida en el intervalo  $(-\infty, \infty)$  y con asíntotas horizontales para los valores de  $y = 0$  e  $y = 1$ .

La función de distribución se define en términos de la función de densidad, de la siguiente manera:

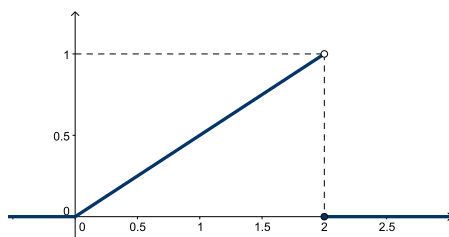
$$F(x) = \int_{-\infty}^x f(t) dt$$

y por tanto, en los valores de  $x$  donde exista la derivada de  $F(x)$ , se verifica la igualdad

$$f(x) = F'(x)$$

que relaciona las funciones de distribución y densidad.

**Ejemplo 5.5** Consideremos la variable aleatoria  $X$  que determina la duración en unidades de tiempo (u.t.) de un componente electrónico y cuya función de densidad viene representada en el siguiente gráfico:



Determinar y representar su función de distribución, y calcular las probabilidades de que el componente dure más de 1 u.t., exactamente 1 u.t. y más de una unidad de tiempo sabiendo que dura menos de 1'5 u.t.

A la vista de la representación gráfica, deducimos que la función de densidad es

$$f(x) = \begin{cases} x/2 & \text{si } 0 < x < 2 \\ 0 & \text{en el resto} \end{cases}$$

y, a partir de ella, podemos calcular la función de distribución  $F(x) = P(X \leq x)$  de la siguiente manera

$$\text{Si } x < 0 \quad \text{entonces } F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt = 0$$

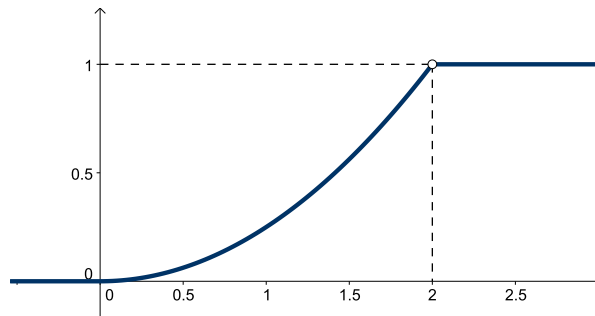
$$\text{Si } 0 \leq x < 2 \quad \text{entonces } F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt = \int_0^x \frac{t}{2} dt = \frac{x^2}{4}$$

$$\text{Si } x \geq 2 \quad \text{entonces } F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt = \int_0^2 \frac{t}{2} dt = 1$$

y, por lo tanto,

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ x^2/4 & \text{si } 0 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

cuya gráfica es



que representa una función continua pues corresponde a una variable aleatoria continua.

Ahora utilizamos estas dos funciones (densidad y distribución) para calcular las probabilidades. En primer lugar, calculamos la probabilidad de que el componente dure más de 1 u.t.

$$P(X > 1) = \int_1^{\infty} f(x) dx = \int_1^2 \frac{x}{2} dx = \frac{3}{4}$$

o bien

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - \frac{1}{4} = \frac{3}{4}$$

En segundo lugar, la probabilidad de que el componente dure exactamente 1 u.t. es cero pues la variable aleatoria es continua. Y, por último, la probabilidad de que el componente dure más de 1 u.t. sabiendo que dura menos de 1'5 u.t., es una probabilidad condicionada que se calcula así:

$$P(X > 1 | X < 1'5) = \frac{P(1 < x < 1'5)}{P(X < 1'5)} = \frac{F(1'5) - F(1)}{F(1'5)} = \frac{9/16 - 1/4}{9/16} = \frac{5}{9}$$

Obsérvese que las probabilidades se han calculado utilizando la función de distribución.  $\square$

## 5.5. Esperanza matemática y otras medidas

En esta sección vamos a introducir el concepto de esperanza matemática que permite definir los momentos de una variable aleatoria. La analogía con las variables estadísticas nos permitirá deducir las principales medidas de centralización, dispersión, simetría y apuntamiento.

Para las definiciones que siguen consideraremos la variable aleatoria  $X$  con soporte  $S_x$ .

### 5.5.1. Esperanza matemática

Para definir la *esperanza matemática* (o simplemente, *esperanza*) distinguiremos entre variables aleatorias discretas y continuas.

- Si  $X$  es una variable aleatoria discreta entonces su esperanza matemática es:

$$E[X] = \sum_{x_i \in S_x} x_i \cdot p(x_i)$$

- Si  $X$  es una variable aleatoria continua entonces su esperanza matemática es:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

La esperanza matemática está definida a partir de una serie (en el caso discreto) o de una integral impropia (en el caso continuo), de manera que la esperanza matemática no existe, o no se puede definir, si la serie o integral correspondiente no es convergente. Por lo tanto, algunas de las definiciones que se presentan a continuación, donde interviene la esperanza matemática, están condicionadas a la existencia de esta esperanza matemática.

### 5.5.2. Momentos

Se llama *momento de orden  $k$  respecto del parámetro  $c$* , y se denota por  $M_k(c)$ , a la esperanza matemática de la variable  $(X - c)^k$ , es decir

$$M_k(c) = E[(X - c)^k]$$

Y en función de que la variable aleatoria sea discreta o continua, se define respectivamente así:

$$M_k(c) = \sum_{x_i \in S_x} (x_i - c)^k \cdot p(x_i) \quad , \quad M_k(c) = \int_{-\infty}^{\infty} (x - c)^k \cdot f(x) dx$$

Como casos particulares, y por su importancia, se definen los dos siguientes tipos de momentos:

- Si  $c = 0$  tenemos los *momentos ordinarios* que representamos por  $m_k$  y definidos como  $m_k = E[X^k]$  para cada uno de los tipos de variables, discretas y continuas, respectivamente:

$$m_k = \sum_{x_i \in S_x} x_i^k \cdot p(x_i) \quad , \quad m_k = \int_{-\infty}^{\infty} x^k \cdot f(x) dx$$

- Si  $c = \mu_x$  tenemos los *momentos centrales* que representamos por  $\mu_k$  y definidos como  $\mu_k = E[(X - \mu_x)^k]$  para cada uno de los tipos de variables, discretas y continuas, respectivamente:

$$\mu_k = \sum_{x_i \in S_x} (x_i - \mu_x)^k \cdot p(x_i) \quad , \quad \mu_k = \int_{-\infty}^{\infty} (x - \mu_x)^k \cdot f(x) dx$$

Los momentos son de gran importancia porque forman parte de la definición de muchas medidas, por ejemplo, la media, la varianza o los coeficientes de asimetría o aplastamiento. Veamos, ahora, una función asociada a cada variable aleatoria que la caracteriza porque permite calcular sus momentos ordinarios.

### 5.5.3. Función generatriz de momentos

La función generatriz de probabilidad sólo se define para variables aleatorias discretas que toman valores en  $\mathbb{N}$ . Por lo tanto, se hace necesario definir una función más general, asociada a cualquier tipo de variable aleatoria, continua o discreta, y que caracterice la distribución de probabilidad de esa variable.

Sea  $X$  una variable aleatoria. Se define la *función generatriz de momentos* asociada a la variable  $X$  como la función

$$M(t) = E(e^{tX})$$

siempre que la esperanza exista en un entorno del cero  $(-t_0, t_0)$ . Se suelen utilizar la abreviatura “f.g.m.” para referirse a la función generatriz de momentos, y según sea discreta o continua la variable aleatoria la expresión, respectivamente será:

$$M(t) = \sum_{x_i \in S_x} e^{tx_i} \cdot p(x_i) \quad , \quad M(t) = \int_{-\infty}^{\infty} e^{tx} \cdot f(x) dx$$

Si  $X$  es una variable aleatoria con función generatriz de momentos  $M(t)$  que es finita para  $|t| < t_0$  con  $t_0 > 0$ , entonces  $X$  posee momentos ordinarios de todos los órdenes y además

$$E(X^n) = M^{(n)}(0)$$

Esta propiedad justifica el nombre de esta función generatriz pues determina los momentos ordinarios a partir de las derivadas sucesivas de la función en el cero.

**Ejemplo 5.6** Sea  $X$  la variable aleatoria discreta, definida en el ejemplo 5.4 de la página 178, que determina el número de caras antes de obtener la primera cruz en el lanzamiento de una moneda equilibrada. Determine la función generatriz de momentos y, a partir de ella, calcule los momentos ordinarios de primer y segundo orden.

En el ejemplo 5.4 determinamos que la distribución de probabilidad de la variable  $X$  era:

$$p(n) = \left(\frac{1}{2}\right)^{n+1} \quad \text{para todo } n \in \mathbb{N}$$

y, por lo tanto, su función generatriz de momentos es:

$$M(t) = \sum_{n=0}^{\infty} e^{tn} p(n) = \sum_{n=0}^{\infty} e^{tn} \left(\frac{1}{2}\right)^{n+1} = \sum_{n=0}^{\infty} \frac{1}{2} \left(\frac{e^t}{2}\right)^n = \frac{1}{2 - e^t}$$

Ahora, calculando el valor en 0 de las derivadas sucesivas de esta función, obtenemos los momentos ordinarios. Por ejemplo, la primera derivada

$$M'(t) = \frac{e^t}{(2 - e^t)^2} \quad \longrightarrow \quad M'(0) = 1$$

determina que  $E[X] = 1$  que, como veremos, corresponde a la media de la variable.  $\square$

#### 5.5.4. Medidas de posición

A continuación, definimos las principales medidas de posición.

##### Media

La esperanza matemática de la variable aleatoria  $X$  recibe el nombre de *media* de la variable y se denota por  $\bar{x}$ , o bien,  $\mu_x$ . La estructura de su fórmula y la interpretación de su valor es similar a la media aritmética definida en estadística descriptiva pero sustituyendo las frecuencias relativas (de los datos que se han observado) por la probabilidad de los valores de la variable (resultados posibles).

El comportamiento de la esperanza respecto de las transformaciones lineales es el siguiente:

$$\text{Si } Y = a + bX \quad \text{entonces} \quad E[Y] = a + bE[X]$$

##### Moda

La *moda* de una variable aleatoria  $X$  es el valor del soporte que tiene mayor probabilidad (variable discreta) o densidad (variable continua).

##### Cuantiles

El cuantil de orden  $k$  de una variable aleatoria  $X$  es el punto  $c_k$  del soporte que verifica las dos siguientes condiciones:

$$P(X \leq c_k) \geq k \quad \text{y} \quad P(X \geq c_k) \geq 1 - k \quad \text{con} \quad 0 < k < 1$$

que pueden resumirse en la siguiente condición

$$k \leq F(c_k) \leq k + P(X = c_k)$$

y que, en el caso de una variable continua, equivale a  $F(c_k) = k$ .



En general, el cuantil de orden  $k$  no es único y, además, si  $c_k$  y  $c'_k$  son dos cuantiles de orden  $k$  de una misma variable aleatoria, con  $c_k < c'_k$ , entonces cualquier valor del intervalo  $(c_k, c'_k)$  es también, un cuantil de orden  $k$ .

A partir de esta definición, y por analogía a la definición de las medidas de estadística descriptivas, podemos considerar los distintos cuantiles (cuartiles, deciles y percentiles).

Como caso particular, definimos la *mediana* de una variable aleatoria  $X$  como el punto “Me” que verifica las dos siguientes condiciones:  $P(X \leq \text{Me}) \geq 1/2$  y  $P(X \geq \text{Me}) \geq 1/2$ . Obsérvese que si  $X$  es una variable aleatoria continua entonces la mediana verifica que  $F(\text{Me}) = 1/2$ .

**Ejemplo 5.7** Calcule la media, la mediana y la moda de la variable aleatoria discreta definida en el ejemplo 5.2 de la página 176.

**Ejemplo 5.8** Calcule la media y la mediana de la variable aleatoria continua definida en el ejemplo 5.5 de la página 180, y compruebe que no existe la moda de esta distribución.

### 5.5.5. Medidas de dispersión

A continuación, definimos las principales medidas de dispersión y veremos que la estructura de sus fórmulas y la interpretación de sus valores son similares a las de sus homónimos en estadística descriptiva pero sustituyendo las frecuencias relativas (de los datos que se han observado) por la probabilidad de los valores de la variable (resultados posibles).

#### Rangos

El *rango* de una variable aleatoria es la diferencia entre los valores extremos del soporte, si son finitos, e infinito, en otro caso.

A partir de los cuantiles, también podemos definir los rangos intercuartílico, interdecílico e intercentílico, análogamente a como se define en estadística descriptiva.

#### Varianza y desviación típica

La varianza de una variable aleatoria  $X$ , que denotaremos por  $\sigma_x^2$ , o bien, por  $V[X]$ , se define como el momento central de orden 2, es decir

$$\sigma_x^2 = V[X] = E[(X - E[X])^2]$$

Según sea discreta o continua la variable aleatoria la expresión, respectivamente será:

$$\sigma_x^2 = \sum_{x_i \in S_x} (x_i - \bar{x})^2 \cdot p(x_i) \quad , \quad \sigma_x^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 \cdot f(x) dx$$

Si desarrollamos el cuadrado y aplicamos las propiedades de la esperanza (serie o integral) obtenemos la siguiente fórmula:

$$\sigma_x^2 = V[X] = E[X^2] - (E[X])^2$$

que permite calcular la varianza de una manera más sencilla.

También se define la *desviación típica* de una variable aleatoria discreta  $X$  como la raíz cuadrada de la varianza. Según sea discreta o continua la variable aleatoria la expresión, respectivamente será:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\sum_{x_i \in S_x} (x_i - \bar{x})^2 \cdot p(x_i)} \quad , \quad \sigma_x = \sqrt{\sigma_x^2} = \sqrt{\int_{-\infty}^{\infty} (x - \bar{x})^2 \cdot f(x) dx}$$

El comportamiento de la varianza respecto de las transformaciones lineales es el siguiente:

$$\text{Si } Y = a + bX \quad \text{entonces} \quad V[Y] = b^2 V[X]$$

### Coefficiente de variación

A partir de los conceptos de media ( $\bar{x}$ ) y desviación típica ( $\sigma_x$ ) de una variable aleatoria  $X$ , se define el coeficiente de variación de la siguiente manera:

$$CV(X) = \frac{\sigma_x}{|\bar{x}|}$$

siempre que la media de la variable sea distinta de cero.

Este coeficiente nos permite comparar la dispersión de dos variables aleatorias.

**Ejemplo 5.9** Calcule el rango intercuartílico y el coeficiente de variación de la variable aleatoria discreta definida en el ejemplo 5.2 de la página 176.

**Ejemplo 5.10** Calcule el rango intercuartílico y el coeficiente de variación de la variable aleatoria continua definida en el ejemplo 5.5 de la página 180.

#### 5.5.6. Medidas de forma

La simetría y el apuntamiento de una variable aleatoria se estudia de manera similar al de una variable estadística y para medir ambas características se utilizan los mismos coeficientes adimensionales, con sus mismas interpretaciones en función de su valor.

### Coefficiente de asimetría

Para medir la simetría de una variable aleatoria  $X$  se define el coeficiente de asimetría de Fisher, que se denota por  $g_1$ , de la siguiente manera:

$$g_1 = \frac{\mu_3}{\sigma^3}$$

### Coefficiente de aplastamiento

Para medir la curtosis de una variable aleatoria  $X$  se define el coeficiente de aplastamiento de Fisher, que se denota por  $g_2$ , de la siguiente manera:

$$g_2 = \frac{\mu_4}{\sigma^4} - 3$$

**Ejemplo 5.11** *Estudiar la simetría y la curtosis de la variable aleatoria discreta definida en el ejemplo 5.2 de la página 176.*

**Ejemplo 5.12** *Estudiar la simetría y la curtosis de la variable aleatoria continua definida en el ejemplo 5.5 de la página 180.*

## 5.6. Variable aleatoria bidimensional

Vamos a generalizar el concepto de variable aleatoria y de función de distribución para considerar el estudio conjunto de dos variables aleatorias. Los resultados obtenidos reflejan un paralelismo con los contenidos del tema de regresión y correlación.

Sea  $(\Omega, \mathcal{A}, P)$  un espacio probabilizable y sean  $X$  e  $Y$  dos variables aleatorias definidas sobre ese espacio. Una *variable aleatoria bidimensional* es una aplicación  $(X, Y) : \Omega \rightarrow \mathbb{R}^2$  que verifica la siguiente propiedad:

para todo  $(x, y) \in \mathbb{R}^2$  el conjunto  $\{\omega \in \Omega / X(\omega) \leq x, Y(\omega) \leq y\} \in \mathcal{A}$ .

**Ejemplo 5.13** *Consideremos el experimento consistente en lanzar una moneda al aire tres veces. Sea  $X$  la variable aleatoria que determina el número de caras ( $H$ ) obtenidas, y sea  $Y$  la variable aleatoria que toma los valores 0, si la primera vez salió cara ( $H$ ), y 1, si la primera vez salió cruz ( $T$ ). Determine la variable aleatoria bidimensional  $(X, Y)$ .*

El espacio muestral del experimento consistente en lanzar tres veces una moneda se puede representar así:

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

y, para cada uno de estos sucesos, representamos el valor de la variable  $(X, Y)$  en la siguiente tabla:

$\omega \in \Omega$	$HHH$	$HHT$	$HTH$	$THH$	$HTT$	$THT$	$TTH$	$TTT$
$(X, Y)(\omega)$	(3, 0)	(2, 0)	(2, 0)	(2, 1)	(1, 0)	(1, 1)	(1, 1)	(0, 1)

que corresponde a una variable aleatoria discreta cuyo soporte es  $S_{xy} = \{0, 1, 2, 3\} \times \{0, 1\}$ .  $\square$

### 5.6.1. Función de distribución

Sea  $(\Omega, \mathcal{A}, P)$  un espacio probabilizable y sean  $(X, Y)$  una variable aleatoria bidimensional definida en ese espacio. Llamaremos *función de distribución conjunta* de la variable  $(X, Y)$  a la función  $F : \mathbb{R}^2 \rightarrow \mathbb{R}$  definida por

$$F(x, y) = P(X \leq x, Y \leq y)$$

Las propiedades de esta función de distribución conjunta son similares a las de la función de distribución de una variable aleatoria unidimensional:

1.  $F(-\infty, -\infty) = \lim_{(x,y) \rightarrow (-\infty, -\infty)} F(x, y) = 0$ , y además,
  - a)  $F(-\infty, y) = \lim_{x \rightarrow -\infty} F(x, y) = 0$  para todo  $y \in \mathbb{R}$ , y
  - b)  $F(x, -\infty) = \lim_{y \rightarrow -\infty} F(x, y) = 0$  para todo  $x \in \mathbb{R}$ .
2.  $F(\infty, \infty) = \lim_{(x,y) \rightarrow (\infty, \infty)} F(x, y) = 1$ .
3.  $F(x, y)$  es monótona no decreciente respecto a cada una de sus variables, es decir
  - a) Si  $x_1 < x_2$  entonces  $F(x_1, y) \leq F(x_2, y)$  para todo  $y \in \mathbb{R}$
  - b) Si  $y_1 < y_2$  entonces  $F(x, y_1) \leq F(x, y_2)$  para todo  $x \in \mathbb{R}$
4.  $F(x, y)$  es continua a la derecha respecto a cada una de sus variables, es decir,
  - a)  $\lim_{h \rightarrow 0^+} F(x + h, y) = F(x, y)$  para todo  $y \in \mathbb{R}$
  - b)  $\lim_{k \rightarrow 0^+} F(x, y + k) = F(x, y)$  para todo  $x \in \mathbb{R}$

La función de distribución permite calcular la probabilidad de cualquier rectángulo de  $\mathbb{R}^2$  de la forma  $(x_1, x_2] \times (y_1, y_2]$ , de la siguiente manera:

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$$

Además, si  $(X, Y)$  es una variable aleatoria bidimensional con función de distribución conjunta  $F(x, y)$ , siendo  $F_1(x)$  y  $F_2(y)$  las funciones de distribución de las variables aleatorias  $X$  e  $Y$ , respectivamente, entonces decimos que estas variables son independientes si, y sólo si,

$$F(x, y) = F_1(x) \cdot F_2(y) \quad \text{para todo } (x, y) \in \mathbb{R}^2$$

### 5.6.2. Tipos de variables aleatorias bidimensionales

Existen varios tipos de variables aleatorias bidimensionales en función de la naturaleza (discreta, continua o mixta) de las variables que la componen vamos a estudiar dos casos: las variables aleatorias bidimensionales discretas y las continuas.

#### Variables aleatorias bidimensionales discretas

Una variable aleatoria bidimensional  $(X, Y)$  se dice que es discreta si  $X$  e  $Y$  son variables aleatorias discretas.

Supongamos que  $X$  toma los valores  $\{x_1, x_2, \dots, x_k\}$ , e  $Y$  toma los valores  $\{y_1, y_2, \dots, y_p\}$ . Entonces la distribución de probabilidad de la variable  $(X, Y)$  viene determinada por la tabla de doble entrada

$X \backslash Y$	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_p$	
$x_1$	$p_{11}$	$p_{12}$	$\dots$	$p_{1j}$	$\dots$	$p_{1p}$	$p_{1\cdot}$
$x_2$	$p_{21}$	$p_{22}$	$\dots$	$p_{2j}$	$\dots$	$p_{2p}$	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_i$	$p_{i1}$	$p_{i2}$	$\dots$	$p_{ij}$	$\dots$	$p_{ip}$	$p_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_k$	$p_{k1}$	$p_{k2}$	$\dots$	$p_{kj}$	$\dots$	$p_{kp}$	$p_{k\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	$\dots$	$p_{\cdot j}$	$\dots$	$p_{\cdot p}$	1

siendo

$$p_{ij} = p(x_i, y_j) = P(X = x_i, Y = y_j) \quad \text{con} \quad \sum_{i=1}^k \sum_{j=1}^p p_{ij} = 1$$

$$p_{i\cdot} = \sum_{j=1}^p p_{ij} \quad \text{y} \quad p_{\cdot j} = \sum_{i=1}^k p_{ij}$$

La función de distribución de la variable aleatoria bidimensional  $(X, Y)$  se define así:

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} P(X = x_i, Y = y_j)$$

El momento de orden  $(r, s)$  respecto al punto  $(a, b)$ , de la variable aleatoria bidimensional  $(X, Y)$ , se definen así:

$$M_{rs}(a, b) = E[(X - a)^r (Y - b)^s] = \sum_{i=1}^k \sum_{j=1}^p (x_i - a)^r \cdot (y_j - b)^s \cdot p_{ij}$$

denotando por  $m_{rs}$  los momentos ordinarios (cuando  $a = 0$  y  $b = 0$ ) y por  $\mu_{rs}$  los momentos centrales (cuando  $a = \bar{x}$  y  $b = \bar{y}$ ). Estos momentos definen, entre otras medidas, las medias y varianzas de las variables  $X$  e  $Y$ , así como su covarianza:

$$\bar{x} = m_{10} \quad , \quad \bar{y} = m_{01} \quad , \quad \sigma_x^2 = \mu_{20} \quad , \quad \sigma_y^2 = \mu_{02} \quad , \quad \sigma_{xy} = \mu_{11}$$

Los conceptos de distribución marginal, distribución condicionada e independencia de variables son similares a los de las variables estadísticas cambiando frecuencia por probabilidad.

**Ejemplo 5.14** Estudiar la variable aleatoria bidimensional  $(X, Y)$  cuya distribución de probabilidad se muestra en la siguiente tabla:

$Y \backslash X$	0	1	2	3	
-1	0'06	0'02	0'04	0'08	0'2
0	0'15	0'05	0'10	0'20	0'5
1	0'09	0'03	0'06	0'12	0'3
	0'3	0'1	0'2	0'4	1

En primer lugar observamos que la variable bidimensional  $(X, Y)$  es discreta porque son también discretas sus dos componentes, y el soporte es el producto cartesiano de los soportes de cada una de sus componentes, es decir, el conjunto  $S_{xy} = \{0, 1, 2, 3\} \times \{-1, 0, 1\}$ .

La probabilidad de cualquier región de  $\mathbb{R}^2$  se calcula sumando las probabilidades  $p(x, y)$  correspondientes a todos los puntos  $(x, y) \in S_{xy}$  que pertenecen a la región. Por ejemplo,

$$\begin{aligned} P((X-1)^2 + Y^2 \leq 1) &= p(0, 0) + p(1, 0) + p(2, 0) + p(1, 1) + p(1, -1) = \\ &= 0'15 + 0'05 + 0'10 + 0'03 + 0'02 = \\ &= 0'35 \end{aligned}$$

Las distribuciones marginales, que denotaremos por  $p_x(x_i)$  y por  $p_y(y_j)$ , aparecen representadas en el margen de la tabla y son las siguientes:

$x_i$	$p_x(x_i)$	$y_j$	$p_y(y_j)$
0	0'3	-1	0'2
1	0'1	0	0'5
2	0'2	1	0'3
3	0'4		

y, a partir de ellas, podemos comprobar que las variables son independientes pues

$$p(x_i, y_j) = p_x(x_i) \cdot p_y(y_j) \quad \text{para todo } (x_i, y_j) \in S_{xy}$$

Además, podemos calcular cualquier medida de cada una de las variables, por ejemplo, la media y la varianza de la variable  $X$ :

$$\begin{aligned} E[X] &= \sum_{x=0}^3 x \cdot p(x) = 1'7 \\ E[X^2] &= \sum_{x=0}^3 x^2 \cdot p(x) = 4'5 \\ V[X] &= E[X^2] - (E[X])^2 = 4'5 - (1'7)^2 = 1'61 \end{aligned}$$

También podemos calcular las distribuciones condicionadas. Por ejemplo, la distribución de probabilidad de la variable  $Y$  condicionada al valor 2 de la variable  $X$ , que denotamos por  $p_2(y_j)$ , es:

$$p_2(y_j) = p(y_j | X = 2) = \frac{p(x, y)}{p_x(2)} \quad \longrightarrow \quad \begin{array}{c|c} y_j & p_2(y_j) \\ \hline -1 & 0'2 \\ 0 & 0'5 \\ 1 & 0'3 \end{array}$$

que coincide con la distribución marginal de la variable  $Y$ , pues las variables son independientes.  $\square$

### Variables aleatorias bidimensionales continuas

Una variable aleatoria bidimensional  $(X, Y)$  se dice que es continua si  $X$  e  $Y$  son variables aleatorias continuas.

La distribución de probabilidad de la variable  $(X, Y)$  viene determinada por una función de densidad  $f(x, y)$ , integrable y no negativa, que verifica:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du \quad \text{para todo } (x, y) \in \mathbb{R}^2$$

y, además,

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

para todo punto  $(x, y) \in \mathbb{R}^2$  donde exista esta derivada de segundo orden.

La función de densidad permite calcular la probabilidad de cualquier rectángulo de  $\mathbb{R}^2$  de la forma  $(x_1, x_2] \times (y_1, y_2]$ , de la siguiente manera:

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y) dy dx$$

y, en general, se puede calcular la probabilidad de cualquier región  $D \subset \mathbb{R}^2$  integrando (integrales dobles) la función de densidad sobre la región:

$$P(D) = \iint_D f(x, y) dx dy$$

El momento de orden  $(r, s)$  respecto al punto  $(a, b)$ , de la variable aleatoria bidimensional  $(X, Y)$ , se definen así:

$$M_{rs}(a, b) = E[(X - a)^r (Y - b)^s] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - a)^r \cdot (y - b)^s \cdot f(x, y) dy dx$$

denotando por  $m_{rs}$  los momentos ordinarios (cuando  $a = 0$  y  $b = 0$ ) y por  $\mu_{rs}$  los momentos centrales (cuando  $a = \mu_x$  y  $b = \mu_y$ ). Estos momentos definen, entre otras medidas, las medias y varianzas de las variables  $X$  e  $Y$ , así como su covarianza:

$$\mu_x = m_{10} \quad , \quad \mu_y = m_{01} \quad , \quad \sigma_x^2 = \mu_{20} \quad , \quad \sigma_y^2 = \mu_{02} \quad , \quad \sigma_{xy} = \mu_{11}$$

Las distribuciones marginales de la variable aleatoria bidimensional  $(X, Y)$  son

$$F_1(x) = F(x, \infty) = \int_{-\infty}^x \int_{-\infty}^{\infty} f(u, v) dv du \quad , \quad F_2(y) = F(\infty, y) = \int_{-\infty}^{\infty} \int_{-\infty}^y f(u, v) dv du$$

siendo

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{y} \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

las funciones de densidad de las distribuciones de las variables  $X$  e  $Y$ , respectivamente.

Las distribuciones condicionadas de la variable aleatoria bidimensional  $(X, Y)$  son

$$F_y(x) = F(x|y) = P(X \leq x | Y = y) = \frac{\int_{-\infty}^x f(u, y) du}{f_2(y)}$$

$$F_x(y) = F(y|x) = P(Y \leq y | X = x) = \frac{\int_{-\infty}^y f(x, v) dv}{f_1(x)}$$

siendo

$$f_y(x) = f(x|y) = \frac{f(x,y)}{f_2(y)} \quad , \quad f_x(y) = f(y|x) = \frac{f(x,y)}{f_1(x)}$$

donde  $f_y(x)$  es la función de densidad de la variable  $X$  condicionada al valor  $y$  de la variable  $Y$ , y  $f_x(y)$  es la función de densidad de la variable  $Y$  condicionada al valor  $x$  de la variable  $X$ . Obsérvese que para poder definir estas funciones condicionadas es necesario que sea positivo el correspondiente valor de la función de densidad marginal que aparece en el denominador.

Y, por último, diremos que las variables aleatorias  $X$  e  $Y$  son independientes si, y sólo si,

$$F(x,y) = F_1(x) \cdot F_2(y) \quad \text{o bien} \quad f(x,y) = f_1(x) \cdot f_2(y) \quad \text{para todo} \quad (x,y) \in \mathbb{R}^2$$

**Ejemplo 5.15** Estudiar la variable aleatoria bidimensional  $(X,Y)$  con función de densidad

$$f(x,y) = \begin{cases} cxy & \text{si } 0 \leq y \leq 1-x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

En primer lugar determinamos el soporte  $S_{xy}$  que resulta ser la región de  $\mathbb{R}^2$  con forma de triángulo de vértices  $(0,0)$ ,  $(1,0)$  y  $(0,1)$ . Después, podemos determinar el valor de la constante  $c$  aplicando que la integral de la función de densidad sobre el soporte es 1.

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = \int_0^1 \int_0^{1-x} cxy dy dx = \int_0^1 \frac{cx(x-1)^2}{2} dx = \frac{c}{24}$$

y determinamos el valor  $c = 24$  resolviendo la ecuación correspondiente.

Ahora, podemos calcular las distribuciones marginales:

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x,y) dy = \int_0^{1-x} 24xy dy = 12x(x-1)^2 \quad \text{si } 0 \leq x \leq 1 \\ f_2(y) &= \int_{-\infty}^{\infty} f(x,y) dx = \int_0^{1-y} 24xy dx = 12y(y-1)^2 \quad \text{si } 0 \leq y \leq 1 \end{aligned}$$

y, por lo tanto,

$$f_1(x) = \begin{cases} 12x(x-1)^2 & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases} \quad \text{y} \quad f_2(y) = \begin{cases} 12y(y-1)^2 & \text{si } 0 \leq y \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

y comprobar que las variables no son independientes pues  $f(x,y) \neq f_1(x)f_2(y)$ .

Con las distribuciones marginales podemos calcular la media y la varianza de cada una de las variables:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf_1(x) dx = \int_0^1 12x^2(x-1)^2 dx = \frac{2}{5} \\ E[X^2] &= \int_{-\infty}^{\infty} x^2 f_1(x) dx = \int_0^1 12x^3(x-1)^2 dx = \frac{1}{5} \\ V[X] &= E[X^2] - (E[X])^2 = \frac{1}{5} - \left(\frac{2}{5}\right)^2 = \frac{1}{25} \end{aligned}$$

y, por simetría, deducimos que  $E[Y] = 2/5$  y que  $V[Y] = 1/25$ .



Y también podemos calcular las distribuciones condicionadas. Por ejemplo, la distribución de probabilidad de la variable  $Y$  condicionada al valor  $1/2$  de la variable  $X$  es

$$f_{1/2}(y) = f(y | X = 1/2) = \frac{f(1/2, y)}{f_1(1/2)} = \frac{24 \cdot (1/2) \cdot y}{3/2} = 8y \quad \text{para todo } 0 \leq y \leq \frac{1}{2}$$

y, por lo tanto,

$$f_{1/2}(y) = \begin{cases} 8y & \text{si } 0 \leq y \leq \frac{1}{2} \\ 0 & \text{en el resto} \end{cases}$$

Con todas estas funciones de densidad calculadas podríamos obtener la probabilidad de cualquier conjunto y las medidas de cualquiera de las variables aplicando las fórmulas correspondientes.  $\square$

### Regresión y correlación

Los conceptos de regresión y correlación de variables aleatorias son similares a los de las variables estadísticas cambiando frecuencia por probabilidad. El objetivo es el mismo: “encontrar y medir una relación entre las variables  $X$  e  $Y$ , que nos permita predecir una de ellas en función de la otra”. Para ello, determinaremos la línea de regresión, que en el caso lineal, consiste en encontrar los valores de  $a$  y  $b$  en el modelo  $Y^* = a + bX$  que minimice  $E[(Y - a - bX)^2]$ . Y el resultado es

$$b = \frac{\sigma_{xy}}{\sigma_x^2}, \quad a = \mu_y - b\mu_x \quad \text{y} \quad r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

siendo  $r$  un número real en el intervalo  $[-1, 1]$ , que se denomina coeficiente de correlación lineal y que determina la bondad del ajuste.

**Ejemplo 5.16** *Obtener la recta de regresión de  $Y/X$  para las variables  $X$  e  $Y$  estudiadas en el ejemplo 5.13 de la página 187 y determinar la bondad del ajuste.*

La distribución de probabilidad de la variable  $(X, Y)$  se representa en la siguiente tabla:

$Y \backslash X$	0	1	2	3	
0	0	1/8	2/8	1/8	1/2
1	1/8	2/8	1/8	0	1/2
	1/8	3/8	3/8	1/8	1

y, a partir de ella, determinamos la curva general de regresión:

$x$	$y$	$p(x, y)$
0	1	1/8
1	2/3	3/8
2	1/3	3/8
3	0	1/8

que nos permite calcular, de forma más sencilla, la recta de regresión, pues los puntos de esta curva están alineados y la única recta que pasa por ellos es la recta buscada:

$$y = 1 - \frac{1}{3}x$$

Si queremos estudiar la bondad del ajuste tenemos que calcular el coeficiente de correlación lineal, utilizando los datos de la distribución de probabilidad de la variable  $(X, Y)$ , presentados en la primera tabla:

$$\bar{x} = 1'5 \quad , \quad \bar{y} = 0'5 \quad , \quad \sigma_x^2 = 0'75 \quad , \quad \sigma_y^2 = 0'25 \quad , \quad \sigma_{xy} = -0'25$$

y el coeficiente de correlación lineal de Pearson es

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{-1/4}{\sqrt{3/4} \cdot \sqrt{1/4}} = -\frac{1}{\sqrt{3}} = -0'577$$

que se interpreta de forma similar a su homónimo en estadística descriptiva.  $\square$

**Ejemplo 5.17** *Obtener la recta de regresión de  $Y/X$  para las variables  $X$  e  $Y$  estudiadas en el ejemplo 5.15 de la página 192 y determinar la bondad del ajuste.*

En el ejemplo 5.15 ya habíamos calculado la media y la varianza de cada una de las variables que resultaban ser:

$$\bar{x} = \bar{y} = \frac{2}{5} \quad \text{y} \quad \sigma_x^2 = \sigma_y^2 = \frac{1}{25}$$

Si calculamos la covarianza

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f(x, y) dx dy = \int_0^1 \int_0^{1-x} 24x^2 y^2 dy dx = \dots = \frac{2}{15}$$

$$\sigma_{xy} = E[XY] - E[X] \cdot E[Y] = \frac{2}{15} - \frac{2}{5} \cdot \frac{2}{5} = -\frac{2}{75}$$

ya tenemos todas las medidas para determinar al recta de regresión de  $Y/X$

$$y = a + bx \quad \longrightarrow \quad \left\{ \begin{array}{l} b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-2/75}{1/25} = -\frac{2}{3} \\ a = \bar{y} - b\bar{x} = \frac{2}{5} + \frac{2}{3} \cdot \frac{2}{5} = \frac{2}{3} \end{array} \right\} \quad \longrightarrow \quad y = \frac{2}{3} - \frac{2}{3}x$$

y la bondad del ajuste queda determinada por el coeficiente de correlación lineal de Pearson:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{-2/75}{1/5 \cdot 1/5} = -\frac{2}{3}$$

que se interpreta de forma similar a su homónimo en estadística descriptiva.  $\square$

### 5.7. Relación de problemas

1. Sea  $X$  el número de años que deben transcurrir antes de que un tipo particular de máquina necesite reemplazo. Supóngase que la distribución de probabilidad de  $X$  es  $P(1) = 0'3$ ,  $P(2) = 0'4$ ,  $P(3) = 0'2$  y  $P(4) = 0'1$ . Calcule y represente la función de distribución.
2. Dado el experimento consistente en lanzar un par de dados, consideramos las siguientes variables aleatorias:

$X$  = “máximo de la puntuación obtenida entre los dos dados”.

$Y$  = “diferencia (en valor absoluto) de los puntos obtenidos en los dados”.

Para cada una de las variables que hemos definido, se pide:

- a) Calcular y representar la distribución de probabilidad.
  - b) Calcular y representar la función de distribución.
  - c) Calcular la esperanza, la varianza y desviación típica.
  - d) Calcular:  $P(X \leq \bar{x})$ ,  $P(X > \bar{x})$ ,  $P(2 < X \leq 4)$ .
  - e) Calcular:  $P(Y \leq 2)$ ,  $P(Y > 2)$ ,  $P(Y = 2)$ ,  $P(Y > 7)$ .
  - f) Determinar la mediana, la moda y los cuartiles.
3. Se lanza cuatro veces una moneda trucada que tiene  $2/3$  de probabilidad de salir cara ( $H$ ) y  $1/3$  de probabilidad de salir cruz ( $T$ ). Consideramos las siguientes variables aleatorias:

$X$  = “mayor número de caras consecutivas obtenidas en los cuatro lanzamientos”.

$Y$  = “número total de caras obtenidas en los cuatros lanzamientos”.

- a) Para cada una de las variables que hemos definido
    - 1) Calcule y represente la distribución de probabilidad.
    - 2) Calcule y represente la función de distribución.
    - 3) Calcule la esperanza, la varianza y la desviación típica.
  - b) Utilice alguna de las variables  $X$  o  $Y$  que hemos definido, para calcular las siguientes probabilidades:
    - 1) Probabilidad de que salgan a lo sumo dos caras consecutivas.
    - 2) Probabilidad de que salgan, al menos, dos caras (no necesariamente consecutivas).
4. Consideremos la variable aleatoria  $X$  con función de distribución

$$F(x) = \begin{cases} 0 & \text{si } x < -1 \\ 0'3 & \text{si } -1 \leq x < 0 \\ 0'7 & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

- a) Dibuje la función de distribución.
- b) Calcule la distribución de probabilidad.
- c) Calcule las probabilidades  $P(X > 0)$ ,  $P(X \leq 2)$  y  $P(X = 1 | X \geq 0)$ .

5. *Distribución degenerada.* Sea  $X$  una variable aleatoria que sólo toma el valor  $x_0$ . Determine su distribución de probabilidad, su función de distribución, su media, su varianza y su función generatriz de momentos.
6. *Distribución de Bernoulli.* Consideramos la variable aleatoria  $X$  que sólo toma los valores 0 y 1, y que la probabilidad asociada al punto  $x = 1$  es un valor  $p \in [0, 1]$ .
- Calcule la media, la varianza y las funciones generatrices de probabilidad y de momentos. Particularice los resultados para  $p = 1/2$ .
  - Determine experimentos y variables que tengan esta distribución de probabilidad.
  - Compruebe que si  $p = 0$ , o bien, si  $p = 1$ , entonces la distribución de probabilidad de nuestra variable es degenerada.
7. *Distribución uniforme discreta.* Consideremos la variable aleatoria discreta  $U$  que toma los valores  $\{1, 2, \dots, n\}$ , todos ellos con la misma probabilidad. Calcule la distribución de probabilidad, su media y su varianza.

Nota: para calcular la varianza se necesita saber que  $\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$

8. *Distribución geométrica I de parámetro  $p$ .* Sea  $X$  la variable aleatoria que determina el número de fallos antes del primer éxito, siendo  $p$  la probabilidad de éxito y  $q = 1 - p$  la probabilidad de fracaso. Pensemos, por ejemplo, en lanzamientos consecutivos de una moneda, siendo éxito, por ejemplo, el suceso F. Se pide:
- Determinar la distribución de probabilidad de la variable aleatoria  $X$ .
  - Demostrar que  $\bar{x} = q/p$  y que  $\sigma_x^2 = q/p^2$ .
  - Determinar la función generatriz de probabilidad y comprobar que se verifica el teorema de inversión.
  - Determinar la función generatriz de momentos y utilizarla para comprobar los resultados obtenidos para la media y la varianza de la variable.
  - Particularizar los resultados para  $p = 1/2$ .
  - ¿Qué sucede si  $p = 0$ , o bien, si  $p = 1$ ?
9. *Distribución geométrica II de parámetro  $p$ .* Sea  $X$  una variable aleatoria discreta con función generatriz de probabilidad

$$G(s) = \frac{ps}{1 - sq}$$

para algún  $p \in [0, 1]$  con  $q = 1 - p$ . Se pide:

- Determinar la distribución de probabilidad de la variable aleatoria  $X$ .
- Demostrar que  $\bar{x} = 1/p$  y que  $\sigma_x^2 = q/p^2$ .
- Comprobar que la variable  $X$  representa el número de pruebas necesarias para obtener el primer éxito, siendo  $p$  la probabilidad de éxito y  $q = 1 - p$  la probabilidad de fracaso.
- Determinar la función generatriz de momentos y utilizarla para comprobar los resultados obtenidos para la media y la varianza de la variable.
- Particularizar los resultados para  $p = 1/2$ .

f) ¿Qué sucede si  $p = 0$ , o bien, si  $p = 1$ ?

10. *Distribución de Poisson.* Sea  $X$  una v.a.d que toma los valores  $0, 1, 2, \dots$  con probabilidad

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

para algún valor real  $\lambda > 0$ .

- Demuestre que  $E[x] = V[X] = \lambda$ .
- Determine la función generatriz de probabilidad y compruebe que se verifica el teorema de inversión.
- Determine la función generatriz de momentos y utilícela para comprobar los resultados obtenidos para la media y la varianza de la variable.
- Particularice los resultados para  $\lambda = 2$ .

11. Sea  $X$  una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} x/8 & \text{si } 0 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

Se pide:

- Representar la función de densidad de  $X$ .
  - Calcular y representar la función de distribución de  $X$ .
  - Calcular la esperanza, la varianza y desviación típica de  $X$ .
  - Calcular la mediana, la moda y el rango intercuartílico.
  - Calcular las probabilidades  $P(1 \leq X \leq 3)$ ,  $P(X \leq 1)$ ,  $P(X \geq 3)$ ,  $P(X > 0)$  y  $P(X \geq 5)$ , y las probabilidades condicionadas  $P(X > 1 | X < 3)$  y  $P(X \geq Q_1 | X \leq Me)$ .
12. *Distribución Uniforme Continua.* Obtenga  $k$  para que  $f(x) = k$ , sea una función de densidad en el intervalo  $[0, 1]$ . Halle su función de distribución, su media y su varianza. Obtenga los mismos resultados para el caso en el que la función esté definida en el intervalo  $[a, b]$ .
13. La demanda diaria de gasolina sin plomo (en litros) en cierta estación de servicio es una variable aleatoria  $X$ . Supóngase que  $X$  tiene la densidad

$$f(x) = \begin{cases} k & \text{si } 4000 < x < 9000 \\ 0 & \text{en el resto} \end{cases}$$

Se pide:

- Calcular el valor de  $k$ .
- Representar la función de densidad de  $X$ .
- Calcular y representar la función de distribución de  $X$ .
- Calcular la esperanza, la varianza y desviación típica de  $X$ .
- Calcular la probabilidad de vender más de 5000 litros.

14. Una variable aleatoria  $X$  tiene por función de densidad:

$$f(x) = \begin{cases} k \cdot x^2 & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

Determine el valor de  $k$  y encuentre el número  $c$  tal que  $F(c) = 72'9\%$ .

15. Una variable aleatoria  $X$  tiene por función de densidad:

$$f(x) = \begin{cases} c \cdot e^{-x} & \text{si } x > 0 \\ 0 & \text{en el resto} \end{cases}$$

Calcule el valor de  $c$  y determine la función de distribución, la media, la mediana, la varianza y la función generatriz de momentos.

16. Una variable aleatoria  $X$  tiene por función de distribución:

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ k \cdot x^n & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1 \end{cases}$$

Determine la función de densidad, la media, la mediana y la varianza, para cualquier valor entero de  $n$  que sea mayor que 1.

17. La función de densidad de una variable aleatoria  $X$  es  $f(x) = kx$  si  $x \in (0, 1)$  y  $f(x) = 0$  en el resto. Halle:

- La función de distribución.
- $P(X < 2/3)$ .
- $P(1/3 < X < 1/2)$ .
- El valor  $a$  tal que  $P(X < a) = 0'25$  e interpretar el resultado.
- Su media y varianza.

18. Sea  $X$  el espesor (en milímetros) de las arandelas que produce una máquina. Supóngase que  $X$  tiene una densidad  $f(x) = kx$  si  $x \in (1'9, 2'1)$  y  $f(x) = 0$  en el resto. Halle:

- La función de distribución.
- La probabilidad de que una arandela tenga espesor  $1'95$ .
- $P(1'95 < X < 2'05)$ .
- El valor  $a$  tal que  $P(X < a) = 0'25$  e interpretar el resultado.
- Su media y varianza.

19. Una máquina fabrica ejes, cuya medida del radio ( $X$ ) se distribuyen según la función de densidad  $f(x) = k \cdot (x - 9'9) \cdot (x - 10'1)$  si  $x \in (9'9, 10'1)$  y cero en caso contrario ( $x$  en milímetros).

- Determine el valor de  $k$ , y calcule la media y la varianza.
- Si se desechan todos los ejes cuyos radios se desvían en más de  $0'03$  mm. de la media, calcule la proporción de ejes fabricados que se rechazarán.

- c) Determine la nueva función de densidad  $f(h)$ , si los ejes se midiesen en centímetros, es decir, si  $h = x/10$ .

20. Sea  $X$  una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} 1/4 & \text{si } 0 \leq x < 1 \\ 1/2 & \text{si } 1 \leq x < 2 \\ a & \text{si } 2 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

- a) Determine el valor de  $a$ .  
 b) Determine y representa la función de distribución.  
 c) Calcule la media, la mediana y la moda.  
 d) Calcule la varianza  
 e) Estudie la simetría y la curtosis.

21. Sea  $X$  una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} 0 & \text{si } x < -1 \\ a + x & \text{si } -1 \leq x < 0 \\ a - x & \text{si } 0 \leq x < 1 \\ 0 & \text{si } x \geq 1 \end{cases}$$

- a) Determine el valor de  $a$ .  
 b) Determine y representa la función de distribución.  
 c) Calcule la media, la mediana y la moda.  
 d) Calcule la varianza  
 e) Estudie la simetría y la curtosis.

22. Sea  $X$  una variable aleatoria con función de distribución

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x^2 & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

Se pide:

- a) Dibujar la función de distribución.  
 b) Calcular y dibujar la función de densidad.  
 c) Calcular las probabilidades  $P(X < 0'25)$  y  $P(X < 0'25 | X < 0'5)$ .

23. El tiempo de reparación (en horas) de un tipo de máquina, tiene la función de distribución:

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x/2 & \text{si } 0 \leq x < 1 \\ 1/2 & \text{si } 1 \leq x < 2 \\ x/4 & \text{si } 2 \leq x < 4 \\ 1 & \text{si } x \geq 4 \end{cases}$$

Se pide:

- a) Dibujar la función de distribución.
- b) Calcular, dibujar e interpretar la función de densidad.
- c) Calcular la probabilidad de que si el tiempo de reparación es superior a una hora, lo sea de 3'5 horas ( $P(X \geq 3'5 | X \geq 1)$ ).

24. *Variable aleatoria mixta (v.a.m).* La mayoría de los problemas se modelizan utilizando distribuciones discretas o continuas. Sin embargo, en ocasiones es necesario considerar una mezcla de las dos distribuciones. Una variable aleatoria  $X$  se dice que es mixta si su distribución de probabilidad está determinada por la probabilidad en un conjunto de puntos  $D = \{x_1, x_2, \dots\}$ , a lo sumo numerable, y por una función no negativa  $g(x)$  (a modo de función de densidad) que determina la probabilidad de los intervalos de números reales que no contengan puntos de  $D$ , de manera que

$$\sum_{x_i \in D} P(X = x_i) + \int_{-\infty}^{\infty} g(x) dx = 1$$

Obsérvese que  $g(x)$  no es una función de densidad pues  $\int_{-\infty}^{\infty} g(x) dx < 1$ . La función de distribución se define de la manera habitual, y la media y la varianza se definen así:

$$E(X) = \mu_x = \sum_{x_i \in D} x_i \cdot P(X = x_i) + \int_{-\infty}^{\infty} x g(x) dx \quad , \quad V(X) = E((X - \mu_x)^2)$$

Sea  $X$  una variable aleatoria mixta cuya distribución de probabilidad está definida por

$$P(0) = 0'1 \quad , \quad P(1) = 0'2 \quad , \quad g(x) = \begin{cases} 0'05x + 0'2 & \text{si } x \in [2, 4] \\ 0 & \text{en el resto} \end{cases}$$

Se pide:

- a) Determinar las probabilidades:  $P(X < 0)$ ,  $P(X \geq 0'5)$  y  $P(3 \leq X \leq 7)$ .
- b) Calcular y representar la función de distribución.
- c) Calcular la media, y la varianza.
- d) Calcular la mediana y el rango intercuartílico.
- e) Estudiar la simetría y la curtosis de la distribución de la variable.

25. Consideremos las tres funciones de distribución:

$$F_1(x) = \begin{cases} 0 & \text{si } x < 1 \\ (x-1)/2 & \text{si } 1 \leq x < 2 \\ 1/2 & \text{si } 2 \leq x < 3 \\ (x-2)/2 & \text{si } 3 \leq x < 4 \\ 1 & \text{si } x \geq 4 \end{cases} \quad , \quad F_2(x) = \begin{cases} 0 & \text{si } x < 1 \\ (x-1)/3 & \text{si } 1 \leq x < 2 \\ 1/3 & \text{si } 2 \leq x < 3 \\ (x-1)/3 & \text{si } 3 \leq x < 4 \\ 1 & \text{si } x \geq 4 \end{cases}$$

$$\text{y} \quad F_3(x) = \begin{cases} 0 & \text{si } x < 1 \\ 1/6 & \text{si } 1 \leq x < 2 \\ 1/2 & \text{si } 2 \leq x < 3 \\ 5/6 & \text{si } 3 \leq x < 4 \\ 1 & \text{si } x \geq 4 \end{cases}$$

Para cada una de estas distribuciones, se pide:



- a) Representar la función y determinar el tipo de variable aleatoria correspondiente (discreta, continua o mixta).
- b) Calcular las siguientes probabilidades:  
 $P(X = 1)$ ,  $P(X < 3)$ ,  $P(X \leq 3)$ ,  $P(2 < X \leq 3)$ ,  $P(X \geq 4)$ ,  $P(X < 3 | X \geq 2)$ .
- c) Calcular la media, la mediana, la varianza y el rango intercuartílico.
- d) Determinar la simetría de las distribuciones.
26. Propiedades de las funciones generatrices: Sea  $X$  una variable aleatoria discreta tal que  $S_x \subset \mathbb{N}$ , y sean  $G(s)$  y  $M(t)$  sus funciones generatrices de probabilidad y momentos, respectivamente. Demuestre que se verifican las siguientes propiedades
- a)  $G(1) = 1$
- b)  $M(0) = 1$
- c)  $M(t) = G(e^t)$

27. *Distribución exponencial de parámetro  $\lambda$* . Sea  $X$  una v.a.c con función de densidad

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

para algún valor  $\lambda > 0$ .

- a) Represente la función de densidad y verifique sus propiedades.
- b) Calcule y represente la función de distribución.
- c) Demuestre que  $E(x) = 1/\lambda$  y que  $V(X) = 1/\lambda^2$ .
- d) Determine la función generatriz de momentos y utilícela para comprobar los resultados obtenidos en el apartado anterior.
- e) Determine la asimetría de la distribución.
- f) Particularice los resultados para  $\lambda = 2$ .
28. Consideremos la variable aleatoria bidimensional  $(X, Y)$  con distribución de probabilidad

$X \backslash Y$	1	2	3	4
1	$k$	0	0'1	0
2	0'3	0	0'1	0'2
3	0	0'2	0	0

- a) Determine el valor de  $k$ .
- b) Calcule las probabilidades:  $P(1 < X \leq 3, Y = 2)$  y  $P(X \geq 2 | Y < 2)$ .
- c) Calcule  $F(2, 2)$ .
- d) ¿Son  $X$  e  $Y$  variables independientes?
- e) ¿Qué variable está más dispersa, la  $X$  o la  $Y$ ?
- f) Compruebe que  $Y/X=3$  es una variable aleatoria degenerada.

29. Consideremos la variable aleatoria bidimensional  $(X, Y)$  con distribución de probabilidad

$Y \backslash X$	0	1	4	9
0	0	0	0	$k$
1	0	0	$1/4$	0
2	0	$1/4$	0	0
3	$1/4$	0	0	0

- Determine el valor de  $k$ .
  - ¿Qué variable está más dispersa, la  $X$  o la  $Y$ ?
  - ¿Son  $X$  e  $Y$  variables independientes?
  - Ajuste el modelo de regresión  $Y = a + b\sqrt{X}$ .
30. Consideremos la variable aleatoria bidimensional  $(X, Y)$  con distribución de probabilidad

$$p(x, y) = \begin{cases} c(x + y) & \text{si } (x, y) \in \{0, 1, 2, 3\} \times \{0, 1, 2\} \\ 0 & \text{en el resto} \end{cases}$$

- Determine el valor de  $c$ .
  - Ajuste el modelo lineal de regresión  $Y = a + bX$ .
  - Calcule y represente la función de distribución de la variable  $X$ .
31. *Distribución uniforme bidimensional discreta.* Consideramos la variable  $(U, V)$  que toma todos los valores en el conjunto  $\{1, 2, 3, 4, 5\} \times \{1, 2, 3, 4\}$  con la misma probabilidad.
- Determine la distribución de probabilidad conjunta.
  - Compruebe que las distribuciones marginales son también de tipo uniforme.
  - Calcule las rectas de regresión  $Y/X$  y  $X/Y$  y estudie su correlación lineal.
  - Estudie la independencia de las distribuciones marginales.
32. Calcule las rectas de regresión  $Y/X$  y determine la bondad de los ajustes para los pares de variables  $X$  e  $Y$  de los ejercicios 2 y 3 de esta relación de problemas.
33. Consideremos la variable aleatoria continua  $(X, Y)$  con función de densidad

$$f(x, y) = \begin{cases} cx^2y & \text{si } x^2 \leq y \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

- Dibuje la región de  $\mathbb{R}^2$  que representa el soporte de la variable.
- Determine el valor de  $c$ .
- Calcule las probabilidades  $P(X \geq 0)$ ,  $P(Y \leq 1/4)$ ,  $P(Y^2 < X)$  y  $P(X^2 \leq Y < X)$ .
- Calcule la probabilidad correspondiente al cuadrado de lado 1 que tiene su centro en el origen de coordenadas.
- Calcule la probabilidad correspondiente al círculo de radio 1, centrado en el origen.
- Determine las distribuciones marginales de las variables  $X$  e  $Y$ .
- ¿Son  $X$  e  $Y$  variables independientes?
- Calcule las rectas de regresión  $X/Y$  e  $Y/X$  y determine la bondad de los ajustes.

34. Consideremos la variable aleatoria continua  $(X, Y)$  con función de densidad

$$f(x, y) = \begin{cases} cy^2 & \text{si } 0 \leq x \leq 2 \text{ y } 0 \leq y \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

- Dibuje la región de  $\mathbb{R}^2$  que representa el soporte de la variable.
- Determine el valor de  $c$ .
- Determine la función de distribución.
- Calcule las probabilidades  $P(X \geq 0)$ ,  $P\left(Y \leq \frac{1}{2}\right)$  y  $P\left(\frac{1}{2} < X < \frac{3}{2}, \frac{1}{4} < Y < \frac{3}{4}\right)$ .
- Calcule las probabilidades condicionadas  $P(Y \leq X | X > 1)$  y  $P(Y \leq X | Y \leq 1/2)$ .
- Calcule la probabilidad correspondiente al rectángulo que tiene su centro en el origen de coordenadas y cuya base y altura miden respectivamente dos unidades y una unidad.
- Compruebe que la variable aleatoria  $X$  se distribuye de manera uniforme.

35. Consideremos la variable aleatoria continua  $(X, Y)$  con función de densidad

$$f(x, y) = \begin{cases} c(x^2 + y) & \text{si } 0 \leq y \leq 1 - x^2 \\ 0 & \text{en el resto} \end{cases}$$

- Dibuje la región de  $\mathbb{R}^2$  que representa el soporte de la variable.
- Determine el valor de  $c$ .
- Calcule las probabilidades  $P(X \geq 0)$  y  $P(Y \geq |3X/2|)$ .
- Calcule la probabilidad correspondiente al cuadrado de lado 1 que tiene su centro en el origen de coordenadas.
- Determine las distribuciones marginales de las variables  $X$  e  $Y$ .
- ¿Son  $X$  e  $Y$  variables independientes?
- Calcule la recta de regresión  $Y/X$  y determine la bondad del ajuste.

36. Consideremos la variable aleatoria continua  $(X, Y)$  con función de distribución

$$F(x, y) = \begin{cases} 0 & \text{si } x < 0 \text{ o } y < 0 \\ kxy(x + y) & \text{si } 0 \leq x < 2 \text{ y } 0 \leq y < 2 \\ 2kx(x + 2) & \text{si } 0 \leq x < 2 \text{ y } y \geq 2 \\ 2ky(2 + y) & \text{si } x \geq 2 \text{ y } 0 \leq y < 2 \\ 1 & \text{si } x \geq 2 \text{ y } y \geq 2 \end{cases}$$

- Determine el valor de  $k$ .
- Determine la función de densidad y el soporte de la variable  $(X, Y)$ .
- Calcule las probabilidades  $P(X \leq 1, Y \leq 1)$  y  $P(0 \leq X \leq 1, 0 \leq Y \leq 1)$ .
- Calcule la probabilidad condicionada  $P(0 \leq X \leq 1 | 0 \leq Y \leq 1)$ .
- Calcule la probabilidad correspondiente al cuadrado de lado 2 que tiene su centro en el origen de coordenadas.
- Calcule la probabilidad correspondiente al círculo de radio 1, centrado en el origen.

- g) Determine las distribuciones marginales (función de densidad y de distribución) y calcule la media y la varianza de cada una de las variables.
- h) ¿Son  $X$  e  $Y$  variables independientes?
- i) Determine la distribución de  $Y$  condicionada al valor  $x = 1$  de la variable  $X$ .
- j) Determine la distribución de  $X$  condicionada al valor  $y = 1$  de la variable  $Y$ .
- k) Calcule la recta de regresión  $Y/X$  y determine la bondad del ajuste.
37. *Distribución uniforme bidimensional continua.* Sea  $(X, Y)$  la variable aleatoria continua con función de densidad constante en todo el soporte  $S_{xy} = [0, 1] \times [0, 1]$ . Se pide:
- a) Determinar el valor constante de la función de densidad.
- b) Determinar la función de distribución.
- c) Determinar las distribuciones marginales de  $X$  e  $Y$  y calcular sus medias y varianzas.
- d) ¿Son  $X$  e  $Y$  variables independientes?
- e) Calcular las probabilidades  $P(X \leq 0'5)$ ,  $P(Y \leq 0'5)$  y  $P(X \leq 0'5, Y \leq 0'5)$ .
- f) Calcular las probabilidades  $P(Y < X^2)$ ,  $P(X \leq Y)$  y  $P(X \leq Y \leq X^2)$ .
- g) Calcular la probabilidad condicionada  $P(X \leq 0'5 | Y \geq 0'5)$ .
- h) Calcular la probabilidad correspondiente al cuadrado de lado 1 que tiene su centro en el origen de coordenadas.
- i) Calcular la probabilidad correspondiente al círculo de radio 1, centrado en el origen.
- j) Repetir los cuatro primeros apartados de este ejercicio suponiendo que el soporte original de la distribución uniforme hubiese sido el conjunto  $S_{xy} = [a, b] \times [c, d]$
38. Consideremos la variable aleatoria bidimensional  $(X, Y)$  cuya función de densidad es
- $$f(x, y) = \begin{cases} kxe^{-xy} & \text{si } 0 \leq x \leq 1 \text{ e } y > 0 \\ 0 & \text{en el resto} \end{cases}$$
- a) Determine el valor de la constante  $k$ .
- b) Determinar la función de distribución.
- c) Determinar las distribuciones marginales de  $X$  e  $Y$  y calcular sus medias y varianzas.
- d) ¿Son  $X$  e  $Y$  variables independientes?
- e) Calcular las probabilidades  $P(X \leq 0'5)$ ,  $P(Y \leq 1)$  y  $P(X \leq 0'5, Y \leq 1)$ .

# Apuntes de ESTADÍSTICA

## Distribuciones de probabilidad



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento a los profesores Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.*



## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Capítulo 6

# Distribuciones de probabilidad

En el capítulo anterior hemos visto el concepto de variable aleatoria distinguiendo los tipos discreto y continuo, en variables unidimensionales y bidimensionales. En este capítulo vamos a presentar las distribuciones de probabilidad de algunas variables aleatorias particulares que son de especial importancia por representar los modelos teóricos de muchos fenómenos aleatorios.

### 6.1. Distribuciones uniformes

Las distribuciones uniformes se caracterizan por repartir la probabilidad, de manera uniforme, en todo el soporte. Por lo tanto sus distribuciones de probabilidad se representan mediante funciones constantes. Es decir, si la variable es discreta, la distribución uniforme asigna la misma probabilidad a todos los puntos del soporte; y si la variable es continua, la función de densidad es constante.

#### 6.1.1. Distribución uniforme discreta

Una variable aleatoria discreta  $X$  que toma los valores  $x_1, x_2, x_3, \dots, x_n$  con probabilidades

$$P[X = x_k] = \frac{1}{n} \quad \text{con} \quad k = 1, 2, \dots, n$$

recibe el nombre de *variable uniforme discreta*, su distribución de probabilidad *distribución uniforme discreta* y se denota por  $X \rightsquigarrow U(x_1, x_2, \dots, x_n)$ .

Por ejemplo, los resultados que se obtienen al lanzar un dado o elegir al azar entre varias posibilidades, se modelizan con una distribución uniforme. En ellos, se trata de representar el caso en el que no tenemos información sobre la importancia de un resultado u otro, de ahí que se les asigne la misma probabilidad a todos los valores de la variable.

En el caso particular de que la variable tome como valores los primeros números naturales:

$$P[X = k] = \frac{1}{n} \quad \text{con} \quad k = 1, 2, \dots, n$$

entonces su media, varianza y desviación típica son:

$$\mu_x = \frac{n+1}{2} \quad , \quad \sigma_x^2 = \frac{n^2-1}{12} \quad , \quad \sigma_x = \sqrt{\frac{n^2-1}{12}}$$

Un caso muy particular de distribución uniforme lo constituye la distribución de probabilidad degenerada que sólo toma un único valor con probabilidad 1. En este caso, la media es el propio valor, y la varianza y la desviación típica son 0.

### 6.1.2. Distribución uniforme continua

Se dice que la variable aleatoria continua  $X$  sigue una *distribución uniforme* en el intervalo  $[a, b]$  y se denota por  $X \rightsquigarrow U[a, b]$  cuando su función de densidad es

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{si } x \notin [a, b] \end{cases}$$

y su media, varianza y desviación típica son

$$\mu_x = \frac{a+b}{2} \quad ; \quad \sigma_x^2 = \frac{(b-a)^2}{12} \quad ; \quad \sigma_x = \frac{b-a}{\sqrt{12}}$$

Uno de los ejemplos más comunes de esta distribución es la elección de un número al azar entre 0 y 1 que constituye una variable con distribución  $U[0, 1]$ . En muchos lenguajes de programación y programas de cálculo matemático se implementan funciones que permiten generar números aleatorios.

### 6.1.3. Distribución uniforme bidimensional

Las distribuciones uniformes bidimensionales puede ser también discreta o continua y su definición es análoga a la distribución unidimensional correspondiente.

#### Distribución uniforme discreta bidimensional

Una variable aleatoria discreta  $(X, Y)$  con soporte  $S_{xy} = \{x_1, x_2, \dots, x_k\} \times \{y_1, y_2, \dots, y_p\}$  se distribuye de manera uniforme si su distribución de probabilidad es

$$p(x_i, y_j) = P[X = x_i, Y = y_j] = \frac{1}{k \cdot p} \quad \text{para todo } (x_i, y_j) \in S_{xy}$$

#### Distribución uniforme continua bidimensional

Una variable aleatoria continua  $(X, Y)$  con función de densidad

$$f(x, y) = \begin{cases} \frac{1}{(b-a)(d-c)} & \text{si } x \in [a, b] \times [c, d] \\ 0 & \text{si } x \notin [a, b] \times [c, d] \end{cases}$$



se dice que se distribuye uniformemente en su soporte  $s_{xy} = [a, b] \times [c, d]$  y se puede comprobar que las distribuciones marginales son también uniformes, y que las variables  $X$  e  $Y$  son independientes, es decir, que  $f(x, y) = f_1(x) \cdot f_2(y)$ , siendo

$$f_1(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{si } x \notin [a, b] \end{cases} \quad \text{y} \quad f_2(y) = \begin{cases} \frac{1}{d-c} & \text{si } y \in [c, d] \\ 0 & \text{si } y \notin [c, d] \end{cases}$$

## 6.2. Distribución Binomial

Muchos experimentos están asociados a fenómenos aleatorios con sólo dos posibles resultados. En esta sección veremos que la distribución de Bernoulli modeliza estos experimentos, mientras que repetición de cualquiera de ellos se modeliza con la distribución binomial. Por ejemplo, lanzar una moneda al aire es un experimento aleatorio que sólo tiene dos posibles resultados y utilizaremos la distribución de Bernoulli para modelizarlo. Sin embargo, si lanzamos al aire una moneda 10 veces, entonces la distribución binomial modeliza el número de veces que sale cara o cruz.

Además, también veremos tres distribuciones más que están relacionadas con la distribución binomial: la distribución multinomial que la generaliza y las distribuciones Hipergeométrica y binomial negativa.

### 6.2.1. Distribución de Bernoulli

Un experimento que sólo admite 2 resultados posibles excluyentes:

- Suceso  $E$  (representa el éxito) con probabilidad  $P(E) = p$ .
- Suceso  $F$  (representa el fracaso) con probabilidad  $P(F) = 1 - p = q$ .

recibe el nombre de *prueba de Bernoulli*.

Consideremos la variable aleatoria discreta  $X$  asociada al experimento que asocia el valor 1 al suceso  $E$  con probabilidad  $p$  y el valor 0 al suceso  $F$  con probabilidad  $q$ . Esta variable recibe el nombre de *variable de Bernoulli* y se denota por  $X \sim \text{Ber}(p)$ .

La distribución de probabilidad es:

$$p(1) = P(X = 1) = p \quad \text{y} \quad p(0) = P(X = 0) = 1 - p = q \quad \text{con} \quad p + q = 1$$

y su media, varianza y desviación típica son:

$$\mu_x = p \quad , \quad \sigma_x^2 = p \cdot q \quad , \quad \sigma_x = \sqrt{p \cdot q}$$

Por ejemplo, estudiar los resultados de lanzar una moneda perfecta o trucada, el sexo de un colectivo, la validez de una pieza fabricada, etc., son experimentos que se modelizan con la distribución de Bernoulli. En todos ellos, sólo hay dos resultados posibles e incompatibles, y no necesariamente de igual probabilidad.

### 6.2.2. Distribución Binomial

Supongamos que se realizan  $n$  pruebas de Bernoulli sucesivas e independientes. Entonces, la variable aleatoria discreta

$X =$  “número de veces que ocurre el suceso  $E$  (éxito) en las  $n$  pruebas”

se denomina *variable binomial* de parámetros  $n$  y  $p$  y se denota por  $X \sim B(n, p)$  donde  $p$  es la probabilidad de éxito en cada prueba de Bernoulli. La variable binomial  $X$  se puede considerar como la suma de  $n$  variables independientes de Bernoulli, es decir

$$X = X_1 + X_2 + \dots + X_n \quad \text{con} \quad X_i \sim \text{Ber}(p) \quad \text{para todo } i=1, 2, \dots, n$$

La variable aleatoria definida toma los valores  $\{0, 1, 2, \dots, n\}$  con probabilidad

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \quad \text{con} \quad \begin{cases} n = 1, 2, 3, \dots \\ k = 0, 1, 2, \dots, n \\ 0 < p < 1 \\ q = 1 - p \end{cases}$$

y su media, varianza y desviación típica son:

$$\mu_x = n \cdot p \quad , \quad \sigma_x^2 = n \cdot p \cdot q \quad , \quad \sigma_x = \sqrt{n \cdot p \cdot q}$$

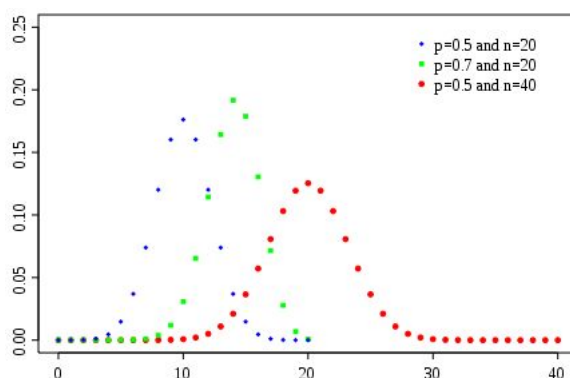


Figura 6.1: Distribuciones binomiales

**Ejemplo 6.1** De una caja de 25 fósforos de los cuales 5 tienen la cabeza blanca, se eligen 4 fósforos al azar con reposición. ¿Qué probabilidad hay de que, exactamente, uno de ellos tenga la cabeza blanca?

El número de fósforos con la cabeza blanca, entre los cuatro elegidos, sigue una distribución binomial de parámetros  $n = 4$  y  $p = 5/25$ . Por lo tanto, la probabilidad de que, exactamente, uno de ellos tenga la cabeza blanca es:

$$P(X = 1) = \binom{4}{1} \left(\frac{5}{25}\right)^1 \left(\frac{20}{25}\right)^3 = 0,4096$$

□

**Observaciones:**

- Si  $n = 1$  entonces  $B(1, p) \equiv Ber(p)$
- La distribución de probabilidad es simétrica si  $p = q$ . Si  $p < q$  presenta asimetría a la derecha; si  $p > q$ , asimetría a la izquierda (ver figura 6.1).
- Aproximaciones: Si  $n$  es “grande” ( $n > 30$ ) la distribución binomial se aproxima por una distribución de Poisson (si  $p$  ó  $q$  son “pequeños”) o por una distribución normal (en otro caso) con los siguientes parámetros:
  - a) Si  $n > 30$  y  $np < 5$  entonces  $B(n, p) \approx P(np)$
  - b) Si  $n > 30$  y  $nq < 5$  entonces  $B(n, q) \approx P(nq)$
  - c) Si  $n > 30$ ,  $np \geq 5$  y  $nq \geq 5$  entonces  $B(n, p) \approx N(np, \sqrt{npq})$

En las secciones 6.3.1 y 6.4.1 referidas a la distribución de Poisson y a la distribución Normal respectivamente, se detallan estas aproximaciones. Además, veremos que en el último caso, cuando utilicemos la distribución normal para aproximar a la binomial, será necesario hacer una corrección de continuidad.

- Valores tabulados: Los valores de  $P(X = k)$  se encuentran tabulados para algunos valores de  $p$  entre 0 y 0,5. Para buscarlos se considera:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k} = b(n, k, p)$$

Si el valor de  $p$  es mayor que 0,5 entonces hay que tener en cuenta la siguiente propiedad

$$b(n, k, p) = \binom{n}{k} \cdot p^k \cdot q^{n-k} = \binom{n}{n-k} \cdot q^{n-k} \cdot p^k = b(n, n-k, q)$$

es decir, para encontrar en la tabla  $P(X = k)$  con  $p > 0,5$  se busca en la tabla correspondiente a  $q = 1 - p$  la probabilidad  $P(X = n - k)$ .

Interpolación: Si el valor de  $p$  es menor que 0,5 pero no está tabulado se interpola entre los valores inferior y superior más próximos a  $p$ .

**6.2.3. Distribución Multinomial**

La distribución Multinomial o Polinomial es una generalización de la distribución binomial cuando en cada prueba se consideran  $k$  sucesos excluyentes  $A_1, A_2, \dots, A_k$  con probabilidades  $p_1, p_2, \dots, p_k$  respectivamente, siendo  $p_1 + p_2 + \dots + p_k = 1$ .

Supongamos que se realizan sucesivamente  $n$  pruebas independientes de este tipo y consideramos las siguientes variables aleatorias discretas:

$X_i$  = “número de veces que ocurre el suceso  $A_i$  en las  $n$  pruebas” con  $i = 1, 2, \dots, k$ .

A la variable  $k$ -dimensional  $X = (X_1, X_2, \dots, X_k)$  se le denomina *variable polinomial o multinomial*. Su función de probabilidad es:

$$P[X_1 = n_1; X_2 = n_2; \dots; X_k = n_k] = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \quad \text{con} \quad \sum_{i=1}^k n_i = n$$

**Ejemplo 6.2** Una agencia de publicidad ha determinado que, en una encuesta televisada, la probabilidad de que una persona vote por tres candidatos  $A$ ,  $B$  y  $C$  es, respectivamente,  $0'1$ ,  $0'4$  y  $0'5$ . Suponiendo que se realiza la encuesta a diez personas, se pide: (1) Probabilidad de que el candidato  $B$  no obtenga ningún voto, y el  $A$  y el  $C$  el mismo número de votos, (2) Probabilidad de que el  $A$  obtenga los diez votos, (3) Probabilidad de que  $A$  obtenga al menos 5 votos, y (4) Probabilidad de que  $B$  obtenga más votos que  $C$ .

Solución: ...

□

#### 6.2.4. Distribución Hipergeométrica

Consideremos una población con  $N$  elementos de dos clases distintas de los cuales  $D$  elementos son de la clase  $E$  y  $N - D$  elementos son de la clase complementaria  $F$ .

Al tomar un elemento de esta población, la probabilidad de que proceda de una u otra clase es

$$\begin{aligned} P(E) &= \frac{D}{N} = p \Rightarrow D = p \cdot N \\ P(F) &= \frac{N - D}{N} = q = 1 - p \Rightarrow N - D = q \cdot N \end{aligned}$$

Consideremos el experimento consistente en tomar, sin reemplazamiento,  $n$  elementos consecutivamente de esta población. A la variable

$X =$  “número de elementos de la clase  $E$  en una muestra de tamaño  $n$ ”

se la denomina *variable hipergeométrica*. Esta variable toma los valores  $0, 1, 2, \dots, n$  con probabilidad

$$P[X = k] = \frac{\binom{D}{k} \cdot \binom{N-D}{n-k}}{\binom{N}{n}} = \frac{\binom{p \cdot N}{k} \cdot \binom{q \cdot N}{n-k}}{\binom{N}{n}} \quad \text{con} \quad \begin{cases} N = 1, 2, 3, \dots \\ n = 1, 2, \dots, N \\ k = 0, 1, 2, \dots, n \end{cases}$$

Esta distribución de probabilidad se denomina *distribución hipergeométrica* de parámetros  $N$ ,  $D$  y  $n$  y se denota con la expresión  $X \rightsquigarrow HGeo(N, D, n)$ . Su media, varianza y desviación típica son

$$\mu_x = n \cdot p \quad ; \quad \sigma_x^2 = n \cdot p \cdot q \cdot \frac{N - n}{N - 1} \quad ; \quad \sigma_x = \sqrt{n \cdot p \cdot q \cdot \frac{N - n}{N - 1}}$$

**Ejemplo 6.3** Considérese un fabricante de automóviles que compra los motores a una compañía donde se fabrican bajo estrictas condiciones. El fabricante recibe un lote de 40 motores. Su plan para aceptar el lote consiste en seleccionar ocho, de manera aleatoria, y someterlos a prueba. Si encuentra que ninguno de los motores presenta serios defectos, el fabricante acepta el lote; de otra forma lo rechaza. Si el lote contiene dos motores con serios defectos, ¿cuál es la probabilidad de que sea aceptado?

La distribución del número de motores sin defectos serios en el lote de 8 de los 40 motores es hipergeométrica de parámetros  $N = 40$ ,  $D = 2$  y  $n = 8$ . Por lo tanto, la probabilidad de que no encuentre ningún motor con defectos ( $k = 0$ ) es  $P(0) = P(X = 0) = 0'6359$ . □

La diferencia entre las distribuciones hipergeométrica y binomial es que, en la distribución binomial, las probabilidades permanecen constantes a lo largo de todas las pruebas (extracciones con reemplazamiento), mientras que en la distribución hipergeométrica, las probabilidades varían de una a otra prueba (extracciones sin reemplazamiento). Sin embargo, si  $N$  es “grande” respecto a  $n$ , las probabilidades varían muy poco de una prueba a la siguiente, por lo que en estos casos ( $n/N < 0.1$ ) se puede decir que la variable hipergeométrica sigue aproximadamente una distribución binomial

$$P[X = k] = \frac{\binom{p \cdot N}{k} \cdot \binom{q \cdot N}{n-k}}{\binom{N}{n}} \xrightarrow{N \rightarrow \infty} \binom{n}{k} p^k q^{n-k}$$

**Ejemplo 6.4** *Un fabricante asegura que sólo el 1 % de su producción total se encuentra defectuosa. Supónganse que se ordenan 1000 artículos y se seleccionan 25 al azar para inspeccionarlos. Si el fabricante se encuentra en lo correcto, ¿cuál es la probabilidad de observar dos o más artículos defectuosos en la muestra?*

El experimento del ejemplo se modeliza con una distribución hipergeométrica de parámetros  $N = 1000$ ,  $D = p \cdot N = 10$  y  $n = 25$  que se aproxima por una distribución binomial de parámetros  $n = 25$  y  $p = 0.01$ . Por lo tanto, la probabilidad de observar dos o más artículos defectuosos es 0.0258.  $\square$

### 6.2.5. Distribución Binomial negativa

Consideremos un experimento que consiste en realizar sucesivas pruebas de Bernoulli. La variable

$$X = \text{“número de fracasos antes de obtener el } n\text{-ésimo éxito”}$$

se denomina *binomial negativa*. La distribución de probabilidad asociada es

$$P[X = k] = \binom{n+k-1}{k} \cdot p^n \cdot q^k \quad \text{con} \quad \begin{cases} k = 0, 1, 2, 3, \dots \\ n = 1, 2, \dots \\ 0 < p < 1 \end{cases}$$

y se denomina *distribución binomial negativa* de parámetro  $n$  y  $p$ , se denota por  $X \sim Bn(n, p)$ , su media, varianza y desviación típica son

$$\mu_x = \frac{n \cdot q}{p} \quad ; \quad \sigma_x^2 = \frac{n \cdot q}{p^2} \quad ; \quad \sigma_x = \frac{\sqrt{n \cdot q}}{p}$$

y sus funciones generatrices de probabilidad y de momentos son

$$G(s) = \left( \frac{p}{1 - sq} \right)^n \quad \text{y} \quad M(t) = \left( \frac{p}{1 - qe^t} \right)^n$$

**Ejemplo 6.5** *Para obtener el permiso de conducir se realiza un test con veinte preguntas. Se sabe que una determinada persona tiene una probabilidad de 0.8 de contestar bien a cada pregunta. Para aprobar el test es necesario contestar bien a diez preguntas. ¿Cuál es la probabilidad de que apruebe al contestar la décimo segunda pregunta?*

El experimento del ejemplo se modeliza con una distribución binomial negativa de parámetros  $n = 20$  y  $p = 0.8$ , y la probabilidad que nos piden es 0.24.  $\square$

La distribución binomial negativa se relaciona con la distribución binomial de la siguiente manera:

$$\text{Si } X \rightsquigarrow Bn(n, p) \text{ entonces } \left\{ \begin{array}{l} P(X \leq k) = P(Y \geq n) \\ P(X = k) = P(Y = n - 1) \cdot p \end{array} \right\} \text{ siendo } Y \rightsquigarrow B(n + k - 1, p)$$

que permite calcular las probabilidades de la distribución binomial negativa a partir de las probabilidades de la distribución binomial.

**Ejemplo 6.6** *Calcular la probabilidad de obtener cinco cruces antes de la tercera cara.*

Si la variable  $X$  representa el número de cruces antes de la tercer cara, entonces

$$X \rightsquigarrow Bn(3, 1/2) \quad \text{y} \quad P(X = 5) = \binom{7}{5} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^5 \approx 0'082$$

Pero si queremos utilizar la distribución binomial, entonces

$$P(X = 5) = P(Y = 2) \cdot \frac{1}{2} = \binom{7}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^5 \frac{1}{2} \approx 0'082$$

sabiendo que  $Y \rightsquigarrow B(7, 1/2)$ . □

Por último, debemos indicar que para poder utilizar la distribución binomial negativa en aquellos ejemplos de extracciones de una urna, estas extracciones han de ser con reemplazamiento.

## 6.3. Distribuciones asociadas a fenómenos aleatorios de espera

Cuando la demanda de un servicio excede la capacidad del servidor de atender a las demandas, se produce una cola. Pensemos, por ejemplo en la cola de clientes que se forma en las cajas de un supermercado. A continuación presentamos tres distribuciones que está íntimamente relacionadas con los fenómenos de espera que se estudian en la teoría de colas: la distribución de Poisson, la exponencial y la geométrica.

La distribución de Poisson surge cuando estudiamos el número de demandas (clientes) que acceden a un sistema de colas por unidad de tiempo y la distribución exponencial representa el tiempo que transcurre entre la llegada de dos demandas consecutivas. Ambas distribuciones están asociadas a sistemas de colas en tiempo continuo. Sin embargo, la distribución geométrica está asociada a sistemas de colas en tiempo discreto donde los eventos sólo pueden ocurrir en los extremos de intervalos de longitud fija.

### 6.3.1. Distribución de Poisson

Una variable aleatoria discreta  $X$  se dice que sigue una *distribución de probabilidad de Poisson* de parámetro  $\lambda$  si toma todos los valores enteros  $0, 1, 2, \dots$  con probabilidades

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad \text{con} \quad \left\{ \begin{array}{l} k = 0, 1, 2, \dots \\ \lambda > 0 \end{array} \right.$$

y se denota por  $X \sim P(\lambda)$ . Su media, varianza y desviación típica son:

$$\mu_x = \lambda \quad , \quad \sigma_x^2 = \lambda \quad , \quad \sigma_x = \sqrt{\lambda}$$

y sus funciones generatrices de probabilidad y de momentos son

$$G(s) = e^{\lambda(s-1)} \quad \text{y} \quad M(t) = e^{\lambda(e^t-1)}$$

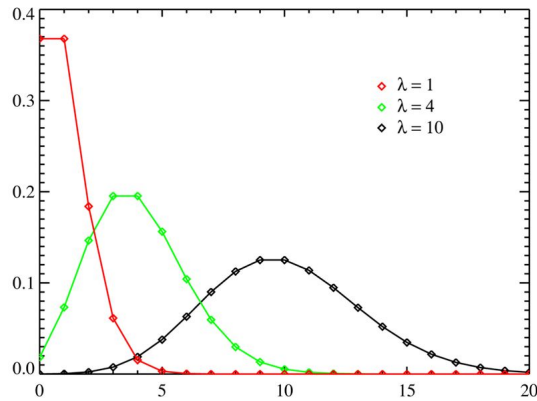


Figura 6.2: Distribuciones de Poisson

La distribución de Poisson representa el “número de ocurrencias de un fenómeno aleatorio durante un periodo de tiempo fijo”, cuando se verifican estas tres propiedades: (1) el número de ocurrencias sólo depende de la amplitud del intervalo de tiempo y no del instante desde donde se mide (proceso estacionario), (2) el número de ocurrencias en un intervalo es independiente del número de ocurrencias en cualquier otro intervalo de tiempo anterior o posterior (propiedad markoviana), y (3) podemos dividir el intervalo de tiempo en subintervalos donde la probabilidad de una ocurrencia en cada uno de ellos es proporcional (con constante  $\lambda$ ) a su longitud.

En este caso, cuando se verifican las tres condiciones, el parámetro  $\lambda$  de la distribución de Poisson es el número esperado de ocurrencias por unidades tiempo, y el número medio de ocurrencias en un intervalo de amplitud  $\Delta t$  es  $t\Delta t$ .

Muchos de los ejemplos que modeliza esta distribución están asociados a fenómenos de espera (teoría de colas) como, por ejemplo, el número de llamadas telefónicas a la hora que recibe una central telefónica, el número de piezas defectuosas en una gran muestra tomada de un lote en el que la proporción de piezas defectuosas es pequeña, el número de clientes que llegan a una ventanilla de pagos de un banco por periodos de diez minutos, el número de emisiones de partículas radioactivas durante un periodo dado, el número de accidentes durante un periodo de tiempo, etc.

La distribución de Poisson se presenta en casos de probabilidad pequeña. Si un suceso  $E$  tiene una probabilidad  $p$  (pequeña) de ocurrir al realizar una prueba elemental, la variable

$$X = \text{“número de veces que ocurre el suceso } E \text{ durante un gran número de pruebas”}$$

sigue una distribución de Poisson de parámetro  $\lambda = n \cdot p$ . Por ello, esta distribución se utiliza como aproximación de la distribución binomial cuando  $n$  es grande y  $p$  o  $q$  son pequeños. En

general, cuando  $n > 30$  y  $np < 5$  la distribución binomial de parámetros  $n$  y  $p$  se aproxima por una distribución de Poisson de parámetro  $\lambda = np$ , o bien, si  $n > 30$  y  $nq < 5$  la distribución binomial de parámetros  $n$  y  $q$  se aproxima por una distribución de Poisson de parámetro  $\lambda = nq$ .

**Ejemplo 6.7** *Aproximación de una distribución binomial por una distribución de Poisson.*

Solución: ...

□

### 6.3.2. Distribución Geométrica o de Pascal

Consideremos un experimento que consiste en realizar sucesivas pruebas de Bernoulli, todas ellas independientes y con probabilidad  $p$  de éxito. En este caso, la variable

$X$  = “número de pruebas necesaria para obtener el primer éxito”

se denomina *variable geométrica*. La distribución de probabilidad asociada es

$$P[X = k] = p \cdot q^{k-1} \quad \text{con} \quad \begin{cases} k = 1, 2, 3, \dots \\ 0 < p < 1 \quad ; \quad q = 1 - p \end{cases}$$

y se denomina *distribución geométrica o de Pascal* de parámetro  $p$  y se denota por  $X \sim \text{Geo}(p)$ . Su media, varianza y desviación típica son

$$\mu_x = \frac{1}{p} \quad ; \quad \sigma_x^2 = \frac{q}{p^2} \quad ; \quad \sigma_x = \frac{\sqrt{q}}{p}$$

y sus funciones generatrices de probabilidad y de momentos son

$$G(s) = \frac{ps}{1 - sq} \quad \text{si} \quad s < \frac{1}{q} \quad \text{y} \quad M(t) = \frac{pe^t}{1 - qe^t} \quad \text{si} \quad e^t < \frac{1}{q}$$

El número de lanzamientos de una moneda, que son necesarios para obtener la primera cara, o el número de extracciones (con reemplazamiento) de una urna, que son necesarias para encontrar la bola blanca entre varias bolas negras, son ejemplos que se modelizan con la distribución geométrica.

**Ejemplo 6.8** *Para obtener el permiso de conducir se realiza un test con veinte preguntas. Se sabe que una determinada persona tiene una probabilidad de 0'8 de contestar bien a cada pregunta. Calcule la probabilidad de que la primera pregunta que contesta bien sea la tercera que hace.*

El experimento del ejemplo se modeliza con una distribución geométrica de parámetro  $p = 0'8$  y la probabilidad que nos piden es 0'032. □

También se denomina geométrica a la variable

$X$  = “número de fracasos antes de obtener el primer éxito”



y, en este caso, su distribución de probabilidad asociada es

$$P[X = k] = p \cdot q^k \quad \text{con} \quad \begin{cases} k = 0, 1, 2, 3, \dots \\ 0 < p < 1 \quad ; \quad q = 1 - p \end{cases}$$

su media, varianza y desviación típica son

$$\mu_x = \frac{q}{p} \quad ; \quad \sigma_x^2 = \frac{q}{p^2} \quad ; \quad \sigma_x = \frac{\sqrt{q}}{p}$$

y sus funciones generatrices de probabilidad y de momentos son

$$G(s) = \frac{p}{1 - sq} \quad \text{si} \quad s < \frac{1}{q} \quad \text{y} \quad M(t) = \frac{p}{1 - qe^t} \quad \text{si} \quad e^t < \frac{1}{q}$$

### 6.3.3. Distribución Exponencial

Se dice que la variable aleatoria continua  $X$  sigue una *distribución exponencial* de parámetro  $\lambda > 0$  y se denota por  $X \rightsquigarrow \text{Exp}(\lambda)$  si su función de densidad es de la forma:

$$X \rightsquigarrow \text{Exp}(\lambda) \quad \text{si} \quad f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{en el resto} \end{cases}$$

y su media, varianza, desviación típica y función generatriz de momentos son

$$\mu_x = \frac{1}{\lambda} \quad ; \quad \sigma_x^2 = \frac{1}{\lambda^2} \quad ; \quad \sigma_x = \frac{1}{\lambda} \quad ; \quad M(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}$$

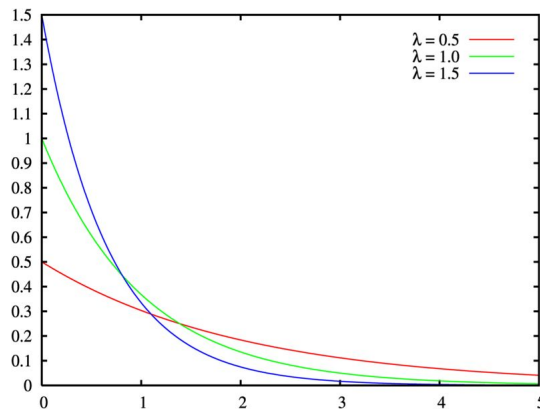


Figura 6.3: Distribuciones exponenciales

La variable aleatoria exponencial representa el tiempo de espera entre dos sucesos, cuando el momento en que ocurre el primero no influye en la distribución de tiempos de espera; es decir,

Si  $X \rightsquigarrow \text{Exp}(\lambda)$  entonces  $P(X > a + x | X \geq a) = P(X > x)$  para todo  $a > 0$  y  $x > 0$ .

Esta propiedad se denomina *falta o pérdida de memoria* pues la probabilidad del tiempo de espera no depende del momento en el que empiece a considerarse.

Existe una relación entre las variables Geométrica y Poisson con la distribución Exponencial relacionada con los fenómenos de espera (teoría de colas). Por un lado, si la variable  $X \rightsquigarrow P(\lambda)$  representa el número de ocurrencias por unidad de tiempo, entonces  $Y \rightsquigarrow \text{Exp}(\lambda)$  representa el tiempo transcurrido entre ocurrencias consecutivas. Por otro lado, la distribución geométrica se puede asociar a fenómenos aleatorios de espera en los que el tiempo sólo puede darse en intervalos de longitud fija pues si  $X \rightsquigarrow \text{Exp}(\lambda)$  entonces la distribución que asocia a cada  $n \in \mathbb{N}$  la probabilidad  $P(X \in (n, n+1])$  es una geométrica de parámetro  $p = 1 - e^{-\lambda}$ .

## 6.4. Distribuciones normales

En esta sección vamos a presentar la distribución de probabilidad más importante: La distribución normal. Hay dos razones fundamentales que acreditan la importancia de esta distribución:

1. Por un lado, modeliza la distribución de probabilidad de muchas variables aleatorias que se presentan en los estudios científicos (ingeniería, medicina, economía, ...).
2. Por otro lado, aproxima a la distribución de la media de muestras aleatorias de una misma distribución (teorema central del límite) que es un resultado básico para la inferencia estadística.

En esta sección vamos a presentar las distribuciones normales, unidimensional y bidimensional, y el teorema central del límite.

### 6.4.1. Distribución Normal o de Laplace-Gauss

Se dice que la variable aleatoria continua  $X$  sigue una *distribución normal* de media  $\mu$  y desviación típica  $\sigma$  y se denota por  $X \rightsquigarrow N(\mu, \sigma)$  cuando su función de densidad es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{con} \quad \begin{cases} -\infty < \mu < \infty \\ \sigma > 0 \end{cases}$$

y su media, varianza y desviación típica son

$$\mu_x = \mu \quad ; \quad \sigma_x^2 = \sigma^2 \quad ; \quad \sigma_x = \sigma$$

#### Características de la distribución:

- Representación gráfica: La función de densidad  $f(x)$  presenta un máximo en  $x = \mu$ , dos puntos de inflexión en  $x = \mu - \sigma$  y  $x = \mu + \sigma$  y tiene al eje OX como asíntota. Además, es simétrica respecto de la recta  $x = \mu$  y por tanto, la media, la mediana y la moda coinciden en este punto (ver figura 6.4).
- Aditividad: La suma de dos variables aleatorias normales independientes es otra variable aleatoria normal, es decir

$$\text{Si } X_1 \rightsquigarrow N(\mu_1, \sigma_1) \text{ y } X_2 \rightsquigarrow N(\mu_2, \sigma_2) \text{ entonces } X_1 \pm X_2 \rightsquigarrow N\left(\mu_1 \pm \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

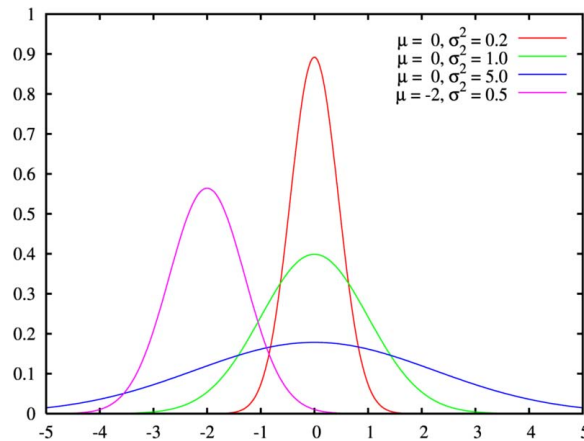


Figura 6.4: Distribuciones normales

Más general, si tomamos muestras de tamaño  $n$  de una población  $N(\mu, \sigma)$  entonces  $\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

### Variable normal tipificada

Si la variable  $X$  es  $N(\mu, \sigma)$  entonces la nueva variable

$$Z = \frac{X - \mu}{\sigma}$$

sigue también una distribución normal de media  $\mu_z = 0$  y desviación típica  $\sigma_z = 1$ , es decir,  $Z \rightsquigarrow N(0, 1)$ . Esta variable  $Z$  se denomina *variable normal tipificada* y su función de densidad es

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad \text{con} \quad -\infty < z < \infty$$

La distribución de la variable  $Z$  se encuentra tabulada aunque sólo aparecen valores de  $Z$  no negativos, o áreas  $\alpha = P(Z \geq z_\alpha) \leq 0,5$ . En otro caso se utiliza la simetría

$$Z_\alpha = -Z_{1-\alpha} \quad \text{y por tanto} \quad \begin{cases} P(Z \geq -Z_\alpha) = P(Z \geq Z_{1-\alpha}) = 1 - \alpha \\ P(Z \leq Z_\alpha) = 1 - P(Z \geq Z_\alpha) = 1 - \alpha \end{cases}$$

La gran utilidad de la variable normal tipificada  $Z$  es que nos permite calcular áreas (probabilidades) de cualquier variable con distribución normal, es decir, si  $X \rightsquigarrow N(\mu, \sigma)$  entonces

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right)$$

### Aproximación

La distribución normal de media  $np$  y desviación típica  $\sqrt{npq}$  se utiliza como aproximación de la distribución binomial de parámetros  $n$  y  $p$  cuando  $n$  es grande y  $np \geq 5$  y  $nq \geq 5$ .

Para utilizar correctamente la aproximación de una variable aleatoria discreta  $X$  con distribución binomial por una variable aleatoria continua  $Y$  con distribución normal es necesario hacer una corrección de continuidad de tal manera que:

$$\begin{aligned} P(X = a) &= P(a - 0'5 \leq Y \leq a + 0'5) \\ P(a < X < b) &= P(a + 0'5 \leq Y \leq b - 0'5) \\ P(a \leq X \leq b) &= P(a - 0'5 \leq Y \leq b + 0'5) \\ P(a < X \leq b) &= P(a + 0'5 \leq Y \leq b + 0'5) \\ P(a \leq X < b) &= P(a - 0'5 \leq Y \leq b - 0'5) \end{aligned}$$

### 6.4.2. Distribución normal bidimensional

Se dice que la variable aleatoria  $(X, Y)$  sigue una *distribución normal bidimensional* de medias  $\mu_x$  y  $\mu_y$ , desviaciones típicas  $\sigma_x$  y  $\sigma_y$ , y covarianza  $\rho\sigma_x\sigma_y$  (correlación  $\rho$ ), si su función de densidad es

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$$

con  $-\infty < \mu_x < \infty$ ,  $-\infty < \mu_y < \infty$ ,  $\sigma_x > 0$ ,  $\sigma_y > 0$  y  $-1 < \rho < 1$ .

Las distribuciones marginales de las variables  $X$  e  $Y$  son distribuciones normales de media  $\mu_x$  y  $\mu_y$ , y de desviación  $\sigma_x$  y  $\sigma_y$ , respectivamente.

Si  $X$  e  $Y$  no están correlacionadas ( $\rho = 0$ ) entonces la distribución conjunta se puede factorizar como producto de las distribuciones marginales y, por lo tanto, las variables son independientes. Y viceversa, es decir, si las variables aleatorias  $X$  e  $Y$  son independientes y sus distribuciones son normales, entonces la distribución conjunta es una distribución normal bidimensional. Esta relación entre la independencia y la correlación, que se verifica para las distribuciones normales, no es cierta en general, es decir, que dos variables aleatorias cualesquiera pueden estar no correlacionadas ( $\rho = 0$ ) sin que sean independientes.

Por último, y como consecuencia de los resultados anteriores, podemos deducir que si  $Z_1$  y  $Z_2$  son variables aleatorias independientes con distribución normal tipificada, entonces la función de densidad conjunta es

$$f(z_1, z_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2 + z_2^2)}$$

que corresponde a una distribución normal bidimensional.

### 6.4.3. Teorema central del límite

El teorema central del límite no es un resultado concreto. Es el nombre genérico por el que se conocen una serie de resultados que establecen la convergencia de la distribución de probabilidad de una suma creciente de variables aleatorias hacia la distribución normal. Existen diferentes versiones del teorema, en función de las condiciones utilizadas para asegurar la convergencia. Una de las más simples establece que es suficiente que las variables que se suman sean independientes, idénticamente distribuidas, con valor esperado y varianza finitas.

Sea  $\{X_n\}$  una sucesión de variables aleatorias independientes e idénticamente distribuidas, todas ellas con media  $\mu$  y desviación típica  $\sigma$ , ambas finitas. Sea  $S_n = X_1 + \dots + X_n$  la sucesión

de sumas parciales (con media  $n\mu$  y varianza  $n\sigma^2$ ). Entonces la distribución de probabilidad de su variable tipificada converge a la distribución normal de media 0 y desviación 1, es decir

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \longrightarrow N(0, 1)$$

También podemos expresar este resultado en términos de la media aritmética de las variables, de la siguiente manera. Sea  $X_1, X_2, \dots, X_n$  un conjunto de variables aleatorias independientes e idénticamente distribuidas, todas ellas con media  $\mu$  y desviación típica  $\sigma$ . Si  $n$  es suficientemente grande ( $n > 30$ ), entonces la distribución de probabilidad de la media aritmética de las variables ( $\bar{X}$ ) es aproximadamente una distribución normal de media  $\mu_{\bar{x}} = \mu$  y desviación típica  $\sigma_{\bar{x}}^2 = \sigma^2/n$ , es decir,

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \longrightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Obsérvese que  $\{X_1, X_2, \dots, X_n\}$  puede representar una muestra aleatoria de la distribución de probabilidad de una determinada variable de una población y el resultado nos garantiza que la media muestral se distribuye según una distribución normal con la misma media que la variable poblacional estudiada. Y este resultado es independiente de la distribución poblacional de partida. Además, la desviación de la media muestral, que se conoce como error típico o estándar, disminuye conforme aumenta el tamaño de la muestra.

## 6.5. Distribuciones derivadas de la normal

En esta sección vamos a presentar tres distribuciones de probabilidad de tipo continuo que serán esenciales en el desarrollo de la inferencia estadística: la distribución  $\chi^2$  de Pearson, la distribución  $t$  de Student y la distribución  $F$  de Fisher-Snedecor. Como veremos, estas tres distribuciones surgen a partir de la distribución normal.

### 6.5.1. Distribución $\chi^2$ de Pearson

Si  $X_1, X_2, \dots, X_n$  son  $n$  variables aleatorias  $N(0, 1)$  independientes entre sí, entonces la variable positiva

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$$

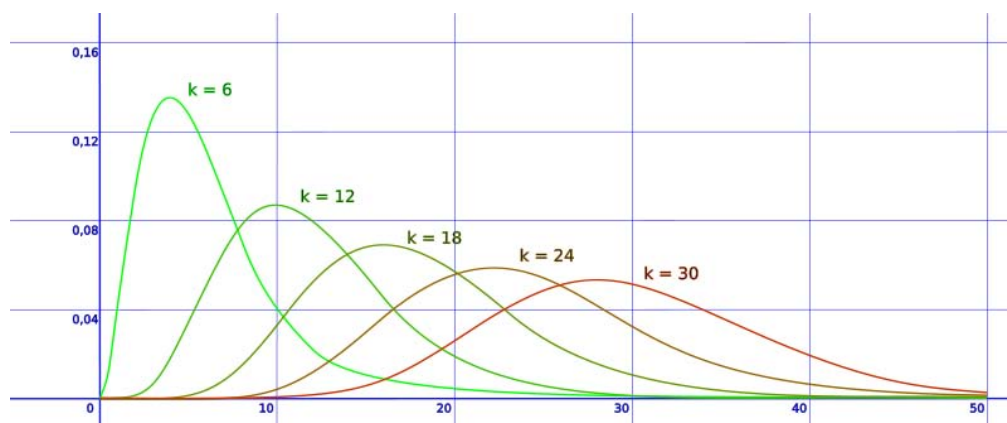
recibe el nombre de  $\chi^2$  de Pearson con  $n$  grados de libertad, se denota por  $\chi_n^2$  y su función de densidad es

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} \quad \text{con } x > 0$$

siendo  $\Gamma$  la función gamma definida así:  $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$  para todo  $x > 0$ . Se puede comprobar que  $\Gamma(1) = 1$ ,  $\Gamma(1/2) = \sqrt{\pi}$  y que para todo  $k > 0$  se verifica que  $\Gamma(k+1) = k \cdot \Gamma(k)$ .

La media, la varianza y la desviación típica de la distribución  $\chi_n^2$  es

$$\mu_x = n \quad ; \quad \sigma_x^2 = 2n \quad ; \quad \sigma_x = \sqrt{2n}$$

Figura 6.5: Distribuciones  $\chi_k^2$ **Características de la distribución:**

- La variable sólo toma valores positivos por tratarse de la suma de los cuadrados de  $n$  variables (ver figura 6.5).
- Aditividad: La suma de dos variables aleatorias independientes  $\chi^2$  con  $n_1$  y  $n_2$  grados de libertad respectivamente es una nueva variable aleatoria  $\chi^2$  con  $n_1 + n_2$  grados de libertad, es decir,

$$\chi_{n_1}^2 + \chi_{n_2}^2 = \chi_{n_1+n_2}^2$$

- Aproximación: Las distribuciones  $\chi^2$  de Pearson son asimétricas a la derecha y se aproximan asintóticamente a la distribución normal (ver figura 6.5). Para  $n > 30$  la variable

$$\sqrt{2 \cdot \chi_n^2} \xrightarrow{n \rightarrow \infty} N(\sqrt{2n-1}, 1)$$

- Si tomamos muestras de tamaño  $n$  con media  $\bar{x}$  y cuasivarianza  $s^2$  de una población  $N(\mu, \sigma)$ , la variable

$$\chi_{n-1}^2 = (n-1) \cdot \frac{s^2}{\sigma^2}$$

es una  $\chi^2$  con  $n-1$  grados de libertad.

- Valores tabulados: Para el uso de las tablas consideramos un punto  $\chi_{\alpha;n}^2$  (punto crítico) que representa el valor de la abscisa que tiene a la derecha una área igual a  $\alpha$  (nivel de significación) en una  $\chi^2$  de Pearson con  $n$  grados de libertad. Es decir,

$$P(\chi_n^2 \geq \chi_{\alpha;n}^2) = \alpha$$

Para áreas a la izquierda se tiene:

$$P(\chi_n^2 \leq \chi_{\alpha;n}^2) = 1 - P(\chi_n^2 \geq \chi_{\alpha;n}^2) = 1 - \alpha$$

### 6.5.2. Distribución $t$ de Student

Si  $X_1, X_2, \dots, X_n$  y  $X$  son  $n + 1$  variables que se distribuyen según una  $N(0, \sigma)$  entonces la variable

$$t_n = \frac{X}{\sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2}} = \frac{Z}{\sqrt{\chi_n^2/n}}$$

se denomina  $t$  de Student con  $n$  grados de libertad, y su función de densidad es

$$f(x) = \frac{1}{\sqrt{n} \cdot \beta\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{con} \quad \begin{cases} n = 1, 2, \dots \\ -\infty < x < \infty \end{cases}$$

siendo  $\beta$  la función beta que se define a partir de la función gamma de la siguiente manera:  $\beta(x, y) = \Gamma(x) \cdot \Gamma(y) / \Gamma(x + y)$ .

La media de la distribución  $t$  de Student es 0 si  $n > 1$  y su varianza es  $\frac{n}{n-2}$  si  $n > 2$ .

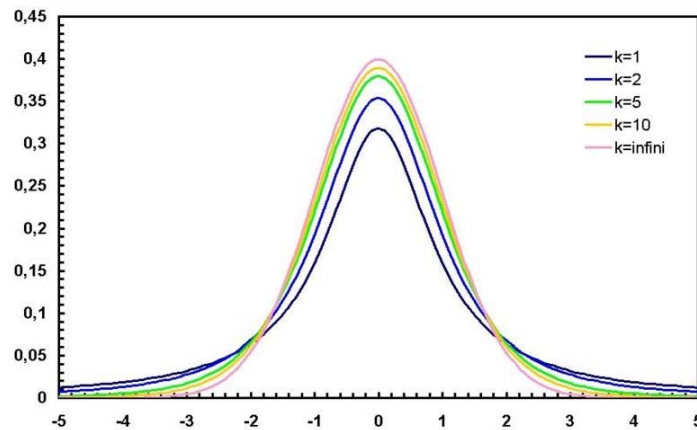


Figura 6.6: Distribuciones  $t_k$  de Student

#### Características de la distribución:

- La variable toma todos los valores de la recta real y es simétrica respecto al eje OY (ver figura 6.6).
- La distribución  $t$  de Student se aproxima asintóticamente ( $n \rightarrow \infty$ ) a la distribución normal tipificada (ver figura 6.6).
- Si tomamos muestras de tamaño  $n$  con media  $\bar{x}$  y cuasivarianza  $s^2$  de una población  $N(\mu, \sigma)$ , la variable

$$t_{n-1} = \frac{\bar{x} - \mu}{s} \cdot \sqrt{n}$$

es una  $t$  de Student con  $n - 1$  grados de libertad.

- Valores tabulados: Para el uso de las tablas consideramos un punto  $t_{\alpha;n}$  (punto crítico) que representa el valor de la abscisa que tiene a la derecha una área igual a  $\alpha$  (nivel de significación) en una  $t$  de Student con  $n$  grados de libertad. Es decir,

$$P(t_n \geq t_{\alpha;n}) = \alpha$$

En la tabla sólo se encuentran valores  $t \geq 0$  (o áreas  $\alpha \leq 0,5$ ) por lo que es necesario utilizar las relaciones:

$$t_{\alpha;n} = -t_{1-\alpha;n} \quad \text{y} \quad P(t_n \leq t_{\alpha;n}) = 1 - \alpha$$

### 6.5.3. Distribución $F$ de Fisher-Snedecor

Sean  $X_1$  y  $X_2$  dos variables  $\chi^2$  de Pearson con  $n_1$  y  $n_2$  grados de libertad respectivamente, independientes entre sí. Entonces a la variable

$$F_{n_1, n_2} = \frac{X_1/n_1}{X_2/n_2} = \frac{\chi_{n_1}^2/n_1}{\chi_{n_2}^2/n_2}$$

se le denomina  $F$  de Fisher-Snedecor con  $n_1$  y  $n_2$  grados de libertad, y su función de densidad es

$$f(x) = \frac{\Gamma((n_1 + n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} n_1^{n_1/2} n_2^{n_2/2} \frac{x^{(n_1/2)-1}}{(n_1 x + n_2)^{(n_1+n_2)/2}} \quad \text{con } x > 0$$

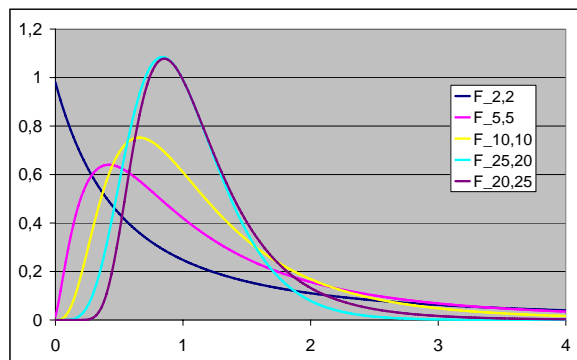


Figura 6.7: Distribuciones  $F_{n_1, n_2}$  de Snedecor

#### Características de la distribución:

- La variable sólo toma valores positivos y su distribución es asimétrica a la derecha (ver figura 6.7).
- Valores tabulados: Para el uso de las tablas consideramos un punto  $F_{\alpha;n_1;n_2}$  (punto crítico) que representa el valor de la abscisa que tiene a la derecha una área igual a  $\alpha$  (nivel de significación) en una  $F$  de Fisher-Snedecor con  $n_1$  y  $n_2$  grados de libertad. Es decir,

$$P(F_{n_1; n_2} \geq F_{\alpha; n_1; n_2}) = \alpha$$



Sólo disponemos de tablas para los siguientes valores de  $\alpha$ : 0'1, 0'05, 0'025, 0'01 y 0'005. Para otros valores de  $\alpha$  entre 0'005 y 0'1 será necesario interpolar. Sin embargo, cuando necesitemos valores de  $\alpha$  próximos a uno, utilizaremos la relación:

$$F_{\alpha;n_1;n_2} = \frac{1}{F_{1-\alpha;n_2;n_1}}$$

## 6.6. Simulación y Método de Montecarlo

En esta sección presentamos el *Método de Montecarlo* que agrupa una serie de procedimientos basados en la simulación de distribuciones de probabilidad. Este método se aplica a una gran variedad de problemas tanto aleatorios como deterministas, que resultan complicados de abordar de manera analítica o donde la experimentación directa con la realidad puede presentar inconvenientes (coste elevado, tiempo, pruebas destructivas o imposibles, etc.). En estos casos, se realizan experimentos en un ordenador, utilizando muestras aleatorias, para modelizar el problema y obtener soluciones aproximadas.

El nombre de Método de Montecarlo hace referencia al casino que se ubica en el principado de Mónaco, al tomar una ruleta como un generador simple de números aleatorios. Aunque su origen es anterior, su desarrollo se produce a mediados del siglo XX coincidiendo con el desarrollo de los ordenadores. Una de las primera aplicaciones fue la resolución de integrales que no se pueden resolver por métodos analíticos, usando números aleatorios. Posteriormente se utilizó para cualquier esquema que emplease números aleatorios, usando variables aleatorias con distribuciones de probabilidad conocidas.

Veamos un sencillo ejemplo que pone de manifiesto el método y sus posibles aplicaciones.

**Ejemplo 6.9** *Consideremos el círculo centrado en el origen y de radio unidad. Sea  $S$  el sector circular correspondiente al área del círculo dibujada en el primer cuadrante. Determine un valor aproximado del área del sector circular.*

En primer lugar, consideramos el cuadrado de vértices  $(0,0)$ ,  $(1,0)$ ,  $(1,1)$  y  $(0,1)$  donde se inscribe el sector circular  $S$ . Ahora vamos a simular una distribución uniforme sobre el cuadrado. Para ellos generamos dos números aleatorios en el intervalo  $[0,1]$  que nos determinan un punto del cuadrado. Este punto podrá pertenecer o no al sector circular. Repetimos el experimento  $N$  veces, generando  $N$  puntos en el cuadrado, y resulta que  $n$  de ellos ( $n < N$ ) también pertenecían al sector circular. Si aplicamos la regla de Laplace podemos determinar que la relación entre el área del sector circular y el área del cuadrado es, aproximadamente,  $n/N$ . Como el área del cuadrado es 1, entonces  $n/N$  es una aproximación del área del sector circular y, por lo tanto,  $4n/N$  aproxima a  $\pi$ .  $\square$

Obsérvese que el procedimiento empleado en el ejemplo es fácilmente generalizable para el cálculo aproximado de la integral definida de cualquier función acotada.

Los resultados obtenidos con este procedimiento son aproximados, sin embargo el error absoluto de la estimación decrece en la relación  $1/\sqrt{N}$ , siendo  $N$  el tamaño de la muestra simulada, en virtud del teorema central del límite.

El Método de Montecarlo se basa en la simulación de distribuciones. En el ejemplo, hemos simulado una distribución uniforme bidimensional generando números aleatorios en su soporte

$[0, 1] \times [0, 1]$  y, para ello, generábamos pares de números aleatorios en el intervalo  $[0, 1]$  (el producto de las distribuciones uniformes de dos variables aleatorias independientes es una distribución uniforme bidimensional cuyo soporte es el producto cartesiano de los soportes de las variables independientes).

En la mayoría de los lenguajes de programación y de los programas específicos de cálculo matemático, están implementadas funciones (rand, random, aleat, ...) para la generación de números aleatorios en determinados intervalos. Existen multitud de algoritmos generadores de estos números aleatorios, y de métodos generales y específicos para la simulación de cualquier distribución de probabilidad.

A modo de ejemplo, presentamos un sencillo procedimiento, conocido como *método de inversión*, para la simulación de algunas distribuciones, en concreto, aquellas para cuya función de distribución, sea sencillo calcular la inversa. Sea  $X$  una variable aleatoria continua con función de distribución  $F_x$  estrictamente creciente, y sea  $U$  una variable aleatoria con distribución uniforme en el intervalo  $(0,1)$ . Entonces, la variable aleatoria  $F_x^{-1}(U)$  tiene a  $F_x$  como función de distribución.

**Ejemplo 6.10** *Utilice el método de inversión para determinar un procedimiento que permita generar muestras aleatorias de tamaño  $n$  de una distribución exponencial de parámetro 2.*

Si  $X \sim \text{Exp}(2)$  entonces  $F(x) = 1 - e^{-2x}$  con  $x > 0$ . Si igualamos la expresión de la función a la variable  $U$  (uniforme) y despejamos la variable  $x$  obtenemos la expresión de la inversa de la función  $F$ :

$$F(x) = 1 - e^{-2x} = u \quad \longrightarrow \quad x = F^{-1}(u) = -\frac{1}{2} \log(1 - u)$$

Por lo tanto, el procedimiento consiste en generar  $n$  números aleatorios  $u_i \in (0,1)$  con  $i = 1, \dots, n$  (valores de  $n$  variables aleatorias independientes con distribución  $U[0, 1]$ ), de manera que los valores  $x_i = -\log(1 - u_i)/2$  constituyen la muestra aleatoria buscada.  $\square$

## 6.7. Relación de problemas

1. Distribución uniforme discreta. Consideramos la variable  $X \rightsquigarrow U(1, 2, \dots, n)$ , se pide:
  - a) Probar que es una distribución de probabilidad (la suma de probabilidades es 1) y representarla.
  - b) Calcular y representar la función de distribución.
  - c) Deducir la esperanza, varianza y desviación típica.
  - d) Calcular la mediana y la moda.
2. Distribución de Bernoulli. Se pide:
  - a) Probar que es una distribución de probabilidad (la suma de probabilidades es 1) y representarla.
  - b) Calcular y representar la función de distribución.
  - c) Deducir la esperanza, varianza y desviación típica.
  - d) Calcular la mediana y la moda.
3. El 20% de los hogares de una ciudad están asegurados contra incendios. Una compañía de seguros está realizando una campaña de publicidad informando a los hogares de sus ofertas. Si cada tarde contacta al azar con 5 hogares, se pide:
  - a) ¿Que distribución de probabilidad modeliza el número de hogares, de esos 5, que aún no están asegurados?
  - b) Determinar el número de hogares que se espera que no estén asegurados.
  - c) Probabilidad de que sólo estén asegurados dos hogares.
  - d) Probabilidad de que estén asegurados al menos tres hogares.
  - e) Probabilidad de que ninguno esté asegurado.
  - f) Probabilidad de que alguno esté asegurado.
4. La probabilidad de ganar a un determinado juego es 0'1. Si jugamos diez partidas
  - a) ¿Qué distribución de probabilidad representa el número de partidas ganadas?
  - b) ¿Cuántas partidas esperamos ganar?
  - c) ¿Qué probabilidad hay de perder todas las partidas?
  - d) ¿Qué probabilidad hay de ganar (exactamente) una partida?
  - e) ¿Qué probabilidad hay de ganar alguna una partida?
  - f) ¿Qué probabilidad hay de ganar (exactamente) dos partidas?
  - g) ¿Qué probabilidad hay de ganar, al menos, dos partidas?
  - h) ¿Qué probabilidad hay de ganar más de la mitad de las partidas?
5. Buscar en las tablas las siguientes probabilidades correspondientes a variables aleatorias discretas que siguen distribuciones de Poisson con distintos parámetros.

a) Si $X \rightsquigarrow P(2'6)$ calcular $P(X = 4)$	b) Si $X \rightsquigarrow P(1'1)$ calcular $P(X = 13)$
c) Si $X \rightsquigarrow P(9)$ calcular $P(X = 16)$	d) Si $X \rightsquigarrow P(2'3)$ calcular $P(X = 5)$
e) Si $X \rightsquigarrow P(2'6)$ calcular $P(X \leq 2)$	f) Si $X \rightsquigarrow P(1'1)$ calcular $P(X \geq 13)$
g) Si $X \rightsquigarrow P(7)$ calcular $P(X > 16)$	h) Si $X \rightsquigarrow P(1'5)$ calcular $P(X < 12)$

6. En una gasolinera la llegada de vehículos sigue una distribución de Poisson de parámetro 1'6. Calcule las probabilidades de los siguientes sucesos:

- a) Que lleguen dos vehículos.
- b) Que llegue algún vehículo.
- c) Que lleguen más de tres vehículos.
- d) Que el número de vehículos que lleguen esté comprendido entre 2 y 5 (ambos inclusive).

7. La probabilidad de ganar a un determinado juego es 0'1. Si jugamos 40 veces

- a) ¿Cuántas partidas esperamos ganar?
- b) ¿Qué probabilidad hay de ganar exactamente 16 partidas?
- c) ¿Qué probabilidad hay de ganar, al menos, 16 partidas?
- d) ¿Qué probabilidad hay de perder todas las partidas?
- e) ¿Qué probabilidad hay de ganar alguna partida?

Compare los resultados de este ejercicio con los obtenidos en el ejercicio 4 y extraiga conclusiones.

8. Si  $X \sim B(150; 0'02)$ , calcule las siguientes probabilidades

- a)  $P(X = 2)$                       b)  $P(X < 3)$
- c)  $P(X \geq 4)$                       d)  $P(X \geq 3 | X \leq 4)$

9. Distribución uniforme continua. Supongamos que  $X \sim U[a, b]$ .

- a) Represente la función

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{si } x \notin [a, b] \end{cases}$$

y pruebe que es una función de densidad.

- b) Calcule y represente la función de distribución.
- c) Deduzca las fórmulas de la esperanza, la varianza y la desviación típica.
- d) Calcule la mediana y la moda.
- e) Aplique los resultados obtenidos a la variable  $A \sim U[0, 1]$ , que representa la elección al azar de un número aleatorio entre 0 y 1.

10. Un profesor propone un cuestionario de cien preguntas tipo test a un curso con 200 alumnos. Suponiendo que las puntuaciones  $X$  obtenidas por los alumnos siguen una distribución normal de media 60 puntos y varianza 100. Calcule las siguientes probabilidades:

- a)  $P(X \geq 70)$                       b)  $P(X \leq 80)$                       c)  $P(X \leq 30)$
- d)  $P(X \geq 46)$                       e)  $P(39 \leq X \leq 80)$                       f)  $P(80 \leq X \leq 82'5)$
- g)  $P(30 \leq X \leq 40)$                       h)  $P(|X - 60| \leq 20)$                       i)  $P(|X - 60| \geq 20)$

11. Consideremos el mismo enunciado del ejercicio anterior. Se pide:

- a) Número de alumnos que obtuvieron 70 o más puntos.  
 b) Hallar el rango intercuartílico, interdecílico e intercentílico. Interpretar resultados.  
 c) Nota mínima correspondiente al 30'5 % de los alumnos con mejor nota.  
 d) Nota mínima correspondiente al 83'65 % de los alumnos con mejor nota.  
 e) Nota máxima correspondiente al 2'17 % de los alumnos con peor nota.  
 f) Si eliminamos al 25 % de los alumnos con peores notas y al 10 % de los alumnos con mejores notas, ¿entre qué notas están el resto de los alumnos?
12. Si  $X \sim N(\mu, \sigma)$ , pruebe e interprete las siguientes igualdades:
- a)  $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0'6826$   
 b)  $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0'9544$   
 c)  $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0'9973$
13. Si  $X \sim N(\mu, \sigma)$ , calcule  $\alpha$  para que la probabilidad  $P(\mu - \alpha\sigma \leq X \leq \mu + \alpha\sigma)$  sea igual a 0'9, 0'95 ó 0'99.
14. Una marca de automóviles decide otorgar un premio a los distribuidores que vendan más de 250 automóviles en un año. El número de automóviles vendidos al año por los distribuidores A y B está normalmente distribuido de la forma siguiente:

Distribuidor	Media	Desviación
A	190	28
B	165	45

Se pide:

- a) A priori, sin hacer cálculos, ¿qué distribuidor parece tener más posibilidad de obtener un premio?  
 b) Determine a qué distribuidor beneficia más la decisión de la empresa, calculando el porcentaje de años que obtendrá premio cada uno de los dos distribuidores.  
 c) ¿Qué cantidad mínima de automóviles debería determinar la marca, si quiere que ambos distribuidores tengan la misma probabilidad de llevarse el premio?  
 d) Si se asocian los dos distribuidores A y B, ¿qué porcentaje de los años obtendrán premio por vender más de 500 automóviles?
15. Si  $X \sim B(1500; 0'02)$ , calcule las siguientes probabilidades
- a)  $P(X = 20)$       b)  $P(X < 30)$   
 c)  $P(X \geq 40)$       d)  $P(X \geq 30 | X \leq 40)$

Compare los resultados de este ejercicio con los obtenidos en el ejercicio 8 y extraiga conclusiones.

16. Variable aleatoria con distribución  $\chi^2$  de Pearson.

- a) Calcule los puntos críticos:

$$\chi_{0'90;5}^2, \quad \chi_{0'01;26}^2, \quad \chi_{0'025;8}^2, \quad \chi_{0'08;10}^2, \quad \chi_{0'015;41}^2$$

b) Calcule las probabilidades:

$$\begin{aligned} P(\chi_8^2 \geq 3'49) & \quad , \quad P(\chi_8^2 \leq 15'507) & \quad , \quad P(\chi_{10}^2 \geq 4) \\ P(\chi_{20}^2 \leq 29) & \quad , \quad P(7'255 \leq \chi_{17}^2 \leq 30'191) & \quad , \quad P(\chi_{61}^2 \leq 50) \end{aligned}$$

17. Variable aleatoria con distribución  $t$  de Student.

a) Calcule los puntos críticos:

$$t_{0'20;20} \quad , \quad t_{0'99;10} \quad , \quad t_{0'25;10} \quad , \quad t_{0'05;90} \quad , \quad t_{0'15;35}$$

b) Calcule las probabilidades:

$$\begin{aligned} P(t_{10} \geq 1'372) & \quad , \quad P(t_8 \leq 2'896) & \quad , \quad P(t_{20} \geq -1'325) \\ P(t_8 \leq 1'2) & \quad , \quad P(-0'5 \leq t_6 \leq 0'6) & \quad , \quad P(|t_{24}| \geq 2) \end{aligned}$$

18. Variable aleatoria con distribución  $F$  de Fisher-Snedecor.

a) Calcule los puntos críticos:

$$F_{0'10;10;12} \quad , \quad F_{0'05;5;24} \quad , \quad F_{0'01;50;30} \quad , \quad F_{0'90;28;30} \quad , \quad F_{0'02;7;20} \quad , \quad F_{0'92;24;20}$$

b) Calcule las probabilidades:

$$\begin{aligned} P(F_{6;12} \geq 2'331) & \quad , \quad P(F_{2;8} \leq 4'459) & \quad , \quad P(F_{25;50} \geq -1'2) \\ P(F_{10;20} \geq 3) & \quad , \quad P(F_{5;4} \leq 5) & \quad , \quad P(2 \leq F_{10;20} \leq 2'25) \end{aligned}$$

19. Calcule los cuartiles de las siguientes distribuciones:

$$\begin{aligned} \text{a) } B(10, 1/2) & \quad \text{b) } B(40, 1/10) & \quad \text{c) } B(40, 1/20) & \quad \text{d) } B(100, 0'85) \\ \text{e) } P(1) & \quad \text{f) } P(2'5) & \quad \text{g) } N(0, 1) & \quad \text{h) } N(1'25, 0'05) \end{aligned}$$

20. Calcule el rango interdecílico de las siguientes distribuciones:

$$\text{a) } \chi_7^2 \quad \text{b) } \chi_{85}^2 \quad \text{c) } t_{10} \quad \text{d) } F_{30,6}$$

### 6.8. Relación de problemas II – Temas 4, 5 y 6

1. El juego A se gana si al lanzar 100 veces dos dados se obtiene al menos tres veces un “doble seis”. Otro juego B se gana si al lanzar 100 veces un dado se obtiene al menos 15 veces el “seis”.
  - a) Determine qué juego es más favorable.
  - b) Si el 40 % de los jugadores optan por el juego A, mientras que el resto juegan al B, ¿cuál es el porcentaje de ganadores?
  - c) En las mismas condiciones del apartado anterior, ¿en qué proporción se encuentran los que han jugado al juego A, entre los ganadores?
2. La distribución de las puntuaciones de los 200 candidatos a una sección de aprendizaje en un test es una normal de media 32'3 y desviación 8'5. Se decide que el 15 % de los candidatos serán orientados a otra sección por tener un nivel demasiado alto y el 25 % a otra por tener el nivel demasiado bajo.
  - a) ¿Entre qué límites habrá que tener la nota para ser admitido en esta sección?
  - b) De los candidatos admitidos a esta sección, ¿cuántos superan la puntuación 35?
3. En el control de calidad de una fábrica, se ha determinado que el porcentaje de cigarrillos defectuosos es del 1 %. Si una máquina los envasa en paquetes de 20 unidades, se pide:
  - a) Calcular la probabilidad de que un paquete tenga a lo sumo 1 cigarrillo defectuoso.
  - b) Si los paquetes se envasan en cartones de 10 paquetes, calcular la probabilidad de que existan al menos 2 paquetes con más de un cigarrillo defectuoso.
  - c) Si los cartones se envasan en cajas de 100 cartones, calcular la probabilidad de que exista alguna caja con al menos 2 paquetes con más de un cigarrillo defectuoso.
4. Al analizar el efecto de un repelente para insectos, se encontró que los frutos no tratados, eran atacados en un 10 %, mientras que solo lo eran en un 1 % si habían recibido el tratamiento. Los frutos se envasan en cajas de 200 unidades.
  - a) Encuentre la probabilidad de que en una caja que contiene frutos tratados, se encuentren más de 20 atacados.
  - b) Halle la probabilidad de que en una caja cuyos frutos no fueron tratados, se encuentren más de 20 atacados.
  - c) A un almacén llega un 30 % de cajas con frutos tratados. ¿Cuál es la probabilidad de que una caja con 22 frutos atacados, no haya sido tratada?
  - d) ¿Cuál es la probabilidad de que una caja con más de 20 frutos atacados, haya recibido el tratamiento?
  - e) Halle la probabilidad de que de 5 frutos extraídos al azar de una caja, encontremos exactamente 2 atacados.
  - f) Encuentre la probabilidad de obtener 2 atacados al extraer 5 frutos de una caja con exactamente 22 atacados.

5. Una variable aleatoria  $X$ , se distribuye según una normal de media 5 y varianza 4. Halle las probabilidades de los siguientes sucesos:

$$a) P(X < 1) \quad b) P(2 < X < 7) \quad c) P(X > 5'6)$$

y determine el valor de  $a$  para que se verifique  $P(X > a) = 0'05$

6. ¿Cuál es la probabilidad de que de 18 000 lanzamientos de un dado, el número de ases esté comprendido entre 2 900 y 3 100? Comparar los resultados entendiendo si el número de ases está comprendido entre 2900 y 3100 estrictamente o no estrictamente.
7. La probabilidad de que una máquina falle determinado día, es de 0'0375 si se trata de un día soleado y de 0'05 si es lluvioso. El servicio técnico debe atender las averías de las 150 máquinas instaladas. Si el 20 % de los días resultan ser lluviosos, determine las siguientes probabilidades:
- a) Probabilidad de que una máquina concreta se averíe en un día.
  - b) Probabilidad de recibir un día más de 7 avisos.
  - c) Probabilidad de no recibir ninguna llamada de avería.
8. Una fábrica produce un 5 % de piezas defectuosas. Un control de calidad previo al envasado, es capaz de detectar el 80 % de las piezas defectuosas, que son retiradas, pero también retira equivocadamente el 1 % de las piezas correctas.
- a) Calcular la proporción de piezas defectuosas envasadas.
  - b) Si se colocan en paquetes de 40 piezas. ¿Qué probabilidad existe de obtener 2 o más piezas defectuosas por paquete?
9. La proporción de tabletas de aspirinas que resultan defectuosas (están partidas, tienen diferente peso, ...) es del 3 %.
- a) Si las aspirinas se envasan en tubos de 10 tabletas. ¿Cuál es la probabilidad de que un tubo contenga a lo sumo una tableta defectuosa?
  - b) Si los tubos se colocan en cajas de 300 unidades (tubos). ¿Cuál es la probabilidad de que una caja contenga exactamente 45 tubos con más de una tableta defectuosa?
  - c) Calcule la probabilidad del apartado anterior si los tubos contienen 50 tabletas.
10. Una central telefónica distingue entre dos tipos de usuarios (particulares y empresas). La probabilidad de que la línea esté ocupada entre la 9 y las 14 horas para particulares es del 2 % mientras que para las empresas este porcentaje de ocupación es del 15 %
- a) Se desea contactar con 150 particulares, hallar la probabilidad de que 10 o más tengan la línea ocupada.
  - b) Se desea contactar con 150 empresas, hallar la probabilidad de que 15 o más tengan la línea ocupada.
  - c) Si las empresas constituyen el 25 % de los usuarios. ¿Cuál será la probabilidad de encontrar ocupada la línea si marcamos un número al azar?
  - d) Si llamamos al azar a un teléfono y resulta ocupado, ¿Cuál será la probabilidad de que pertenezca a una empresa? ¿y a un particular? Observa como ha cambiado esta información las probabilidades asignadas a priori tanto a los particulares como a las empresas.



11. El tiempo que tarda una máquina en perforar un material de tipo 1 se distribuye según una normal de media 2 y desviación 0'5, un material de tipo 2 según una normal de media 3 y desviación 0'1 y un material de tipo 3 según una normal de media 4 y desviación 2. Una empresa recibe una partida de placas de los tres tipos de material donde el 20 % de las placas son de tipo 1 y el 70 % del resto son de tipo 2.
- a) Calcular la probabilidad de que la máquina tarde más de tres segundos en perforar una placa elegida al azar.
  - b) Si se ha tardado más de 3 segundos en perforar una placa elegida al azar, ¿con qué material es más probable que esté fabricada?
  - c) En el control de calidad se rechaza una placa si se ha tardado más de 3 segundos en perforarla. Definimos la variable aleatoria  $X$  que cuenta el número de placas rechazadas. Si analizamos un lote de 100, calcular la probabilidad de rechazar más de 40 sabiendo que el número de placas rechazadas es mayor de 20 y menor o igual que 60, es decir, calcular  $P[X > 40 | 20 < X \leq 60]$
12. Una fábrica produce listones de madera para cerillas, para lo que se necesita que se corten a 3 cm. La cortadora (cortadora/envasadora) que tienen desde hace 5 años, proporciona listones cuya longitud se distribuye de forma normal de media 3 y varianza 0'25.
- La dirección adquirió hace un mes un nuevo modelo de cortadora que proporciona listones cuya longitud se distribuye de forma normal de media 3 y varianza 0'23.
- Un listón es considerado defectuoso si su longitud es menor que 2'365 cm. y tendremos que revisar toda la producción diaria si al examinar una caja de cerillas de 100 unidades (todas procedentes de la misma cortadora) encontramos al menos 13 defectuosas.
- Si la nueva cortadora produce un 20 % más de listones que la antigua:
- a) Por término medio, ¿cuántos días al año habrá que revisar la producción diaria?
  - b) Si la producción de ayer tuvo que ser revisada, ¿cuánto vale la probabilidad de que la caja analizada contenga cerillas de la cortadora antigua? y ¿de qué máquina es más probable que provenga la caja analizada?
13. Se pretenden estudiar las especificaciones de fábrica de un sistema automático de vigilancia para exteriores que aseguran que es capaz de detectar al 90 % de los intrusos que se acerquen en días soleados, pero el aparato resulta muy sensible a la humedad y sólo es capaz de detectar al 50 % de los intrusos si el día es lluvioso. Se pretenden verificar las especificaciones de fábrica del sistema.
- a) Si acercamos 36 individuos al local en un día soleado, ¿qué probabilidad hay de que el sistema no detecte a 10 o más individuos?
  - b) Calcular la misma probabilidad si el día fuese lluvioso.
- Instalamos el sistema en un local situado en la Costa del Sol donde la proporción de días soleados es 9 veces mayor que la de días lluviosos.
- c) Calcular la probabilidad de que el sistema no sea capaz de detectar a 10 o más intrusos en un día cualquiera.
  - d) Si el sistema no ha detectado a 10 o más intrusos, ¿es más probable que el día haya sido soleado o lluvioso?

14. El tiempo  $t$  (en minutos) que se retrasa un avión de Iberia que cubre la línea Málaga-Madrid es una variable aleatoria continua con densidad de probabilidad

$$f(t) = \begin{cases} k \cdot (25 - t^2) & \text{si } -5 < t < 5 \\ 0 & \text{en cualquier otro instante.} \end{cases}$$

[Nota: Un valor negativo de  $t$  significa que el avión adelantó su llegada.]

- Calcular el valor de  $k$ .
- ¿Qué retraso se espera que tenga el avión?
- ¿Qué probabilidad hay de que llegue con más de tres minutos de retraso?
- Si en la ventanilla de información nos confirman que el avión trae retraso, ¿qué probabilidad hay de que llegue más de 3 minutos tarde?
- Si en los dos apartados anteriores se pregunta lo mismo, ¿por qué se obtienen resultados distintos?

Supongamos que la compañía Aviaco también realiza vuelos en la misma línea Málaga-Madrid pero transporta la mitad de viajeros que Iberia (consideraremos que no hay más compañías que cubran esa línea) y su tiempo de retraso sigue una distribución  $t$  de Student con 3 grados de libertad.

- Sin saber a qué compañía (Iberia o Aviaco) pertenece el próximo avión que llega a Málaga procedente de Madrid, ¿qué probabilidad hay de que llegue con más de 3 minutos de adelanto sobre el horario previsto?
  - Si el avión llegó con más de tres minutos de adelanto, ¿a qué compañía es más probable que perteneciera?
15. El tiempo  $t$  (en segundos) que tarda una máquina en perforar un material de tipo I es una variable aleatoria continua que se distribuye según la siguiente función de densidad:

$$f(t) = \begin{cases} k \cdot (t^2 - 4t) & \text{si } 0 \leq t \leq 4 \\ 0 & \text{en cualquier otro instante.} \end{cases}$$

Se pide:

- Calcular el valor de  $k$ .
- ¿Qué probabilidad hay de que tarde menos de 3 segundos en perforar una placa de material tipo I?
- ¿Cómo varía la misma probabilidad del apartado anterior, si sabemos de antemano que tardará más de un segundo?
- Calcular el tiempo medio que tarda en perforar placas de tipo I.

Supongamos que el tiempo que tarda esta misma máquina en perforar un material de tipo II es una variable aleatoria con distribución normal de media 2 y desviación 0'5. Además, de todas las placas que perfora la máquina en un mismo día, el 20 % son de material tipo I y el 80 % son de material tipo II.

- Elegimos al azar una placa. Calcular la probabilidad de que la máquina tarde más de tres segundos en perforarla.
- Si la máquina tardó más de tres segundos en perforar esta placa, ¿qué probabilidad hay de ser una material tipo I?

## 6.9. Anexo I: Justificación de algunos resultados

En esta sección vamos a presentar la justificación de algunos de los resultados que hemos visto en este tema. Incluiremos aquellas demostraciones que utilizan resultados básicos de matemáticas o aquellas que se apoyan en los conocimientos aprendidos en otras asignaturas de matemáticas de la titulación.

### 6.9.1. Distribución Binomial

La suma de las probabilidades es 1 ..... (por el binomio de Newton)

### 6.9.2. Propiedades de la función Gamma

La función gamma Euler se define de la siguiente manera:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

Esta función es continua, está definida (integral convergente) para todo  $x > 0$  y, entre sus propiedades, destacan las siguientes:

1.  $\Gamma(1) = 1$
2.  $\Gamma(x+1) = x \cdot \Gamma(x)$
3.  $\Gamma(1/2) = \sqrt{\pi}$

Además, a partir de las dos primeras propiedades se puede deducir que si  $n \in \mathbb{N}$  entonces  $\Gamma(n+1) = n!$  por lo que, de alguna manera, esta función generaliza a la función factorial.

Veamos la demostración de estas tres propiedades. En primer lugar vamos a demostrar que  $\Gamma(1) = 1$  de la siguiente manera:

$$\Gamma(1) = \int_0^{\infty} e^{-t} dt = \left[ -e^{-t} \right]_0^{\infty} = 0 - (-1) = 1$$

Para demostrar la segunda propiedad  $\Gamma(x+1) = x \cdot \Gamma(x)$  utilizamos el método de integración por partes tomando  $u = t^x$  y  $dv = e^{-t} dt$ , y por lo tanto  $du = x t^{x-1}$  y  $v = -e^{-t}$ , obtenido

$$\Gamma(x+1) = \int_0^{\infty} t^x e^{-t} dt = \left[ -t^x e^{-t} \right]_0^{\infty} + \int_0^{\infty} x t^{x-1} e^{-t} dt = 0 + x \int_0^{\infty} t^{x-1} e^{-t} dt = x \Gamma(x)$$

Y para demostrar la tercera propiedad  $\Gamma(1/2) = \sqrt{\pi}$  será necesario demostrar el siguiente resultado previo:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$$



# Apuntes de ESTADÍSTICA

## Inferencia estadística



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento al profesor Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.*

## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Capítulo 7

# Inferencia estadística

Cuando queremos obtener información sobre una población y disponemos de los datos de todos los individuos (censo), entonces podemos utilizar la estadística descriptiva que tiene como objeto el estudio de un conjunto de elementos con alguna característica común a todos ellos.

Sin embargo, cuando no podemos tener acceso a los datos de todos los individuos, utilizaremos la inferencia estadística que tiene por objeto extraer conclusiones de la totalidad de la población, a partir de los datos de una muestra de ella.

Los dos problemas fundamentales que estudia la inferencia estadística son el “problema de la estimación” y el “problema del contraste de hipótesis”. Cuando se conoce la distribución que sigue la variable aleatoria objeto de estudio y sólo tenemos que estimar los parámetros que la determinan, estamos ante un problema de inferencia estadística paramétrica; por el contrario, cuando no se conoce la distribución que sigue la variable aleatoria objeto de estudio, estamos ante un problema de inferencia estadística no paramétrica.

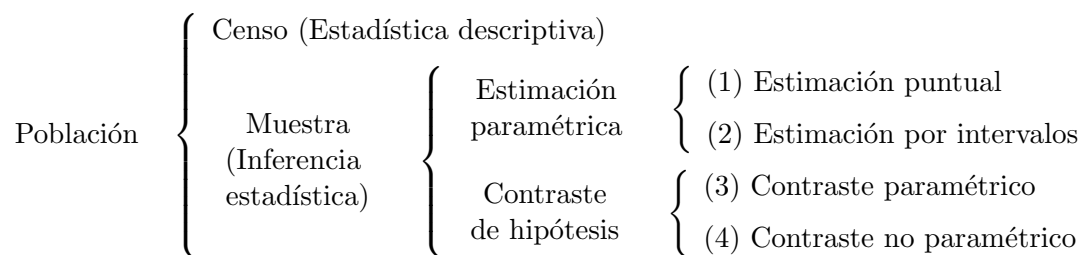
En todos los problemas que estudia la inferencia estadística juega un papel fundamental la “teoría de muestras” que estudia las técnicas y procedimientos que debemos emplear para que las muestras sean representativas de la población que pretendemos estudiar, de forma que los errores en la determinación de los parámetros de la población objeto de estudio sean mínimos.

### 7.1. Inferencia estadística

La Inferencia Estadística es la parte de la estadística matemática que se encarga del estudio de los métodos para la obtención del modelo de probabilidad (forma funcional y parámetros que determinan la función de distribución) que sigue una variable aleatoria de una determinada población, a través de una muestra (parte de la población) obtenida de la misma.

En la inferencia estadística se distinguen básicamente dos tipos de objetivos:

1. **Inferencia paramétrica:** Deducir características (parámetros) de la población a partir de los datos de una muestra.
2. **Contraste de Hipótesis:** Analizar la concordancia o no de los resultados muestrales con determinadas hipótesis sobre la población.



En este tema estudiaremos algunos problemas tanto de inferencia paramétrica (1, 2 y 3) como de inferencia no paramétrica (4). En inferencia estadística paramétrica nos vamos a limitar a problemas donde la variable aleatoria objeto de estudio sigue una distribución binomial, Poisson o normal, y nuestro objetivo será tratar de estimar los parámetros que la determinan, es decir, el parámetro  $p$  de la binomial, el parámetro  $\lambda$  de la Poisson, y los parámetros  $\mu$  y  $\sigma$  de la normal. En los problemas de estimación no paramétrica nos limitaremos al estudio de la bondad de un ajuste, la homogeneidad de varias muestras y la independencia de caracteres, como aplicaciones de la  $\chi^2$ .

### 7.1.1. Teoría de muestras

En la práctica, suele ocurrir con frecuencia que no es posible estudiar todos los elementos de la población, por distintas razones:

- Si el número de elementos de la población es muy elevado, el estudio llevaría tanto tiempo que sería impracticable o económicamente inviable.
- El estudio puede implicar la destrucción del elemento objeto de estudio. Por ejemplo, estudiar la vida media de una partida de bombillas, o la tensión de rotura de cables.
- Los elementos pueden existir conceptualmente, pero no en la realidad. Por ejemplo, la proporción de piezas defectuosas que producirá una máquina.

En estas ocasiones, lo que se hace es seleccionar una muestra de la población, de manera que, de la observación del comportamiento individual de cada uno de los elementos, se puedan obtener unas leyes generales de comportamiento de tipo promedio o de tipo predominante para todos los elementos de la población.

La teoría de muestras estudia los procedimientos para tomar muestras de manera apropiada, es decir, las muestras tienen que ser representativas de la población. Y para conseguirlo, se deben cumplir dos principios básicos:

1. Independencia en la selección de los individuos que forman la muestra
2. Que todos los individuos tengan la misma probabilidad de ser incluidos en la muestra

Para conseguir estos objetivos se emplean distintas técnicas de muestreo. De los distintos métodos que existen para la obtención de muestras, destacamos tres de los más utilizados:



- *Muestreo aleatorio simple*. Se eligen al azar los elementos para garantizar que todos los individuos de la población tienen la misma oportunidad de ser incluidos en dicha muestra. Puede ser de dos tipos: con o sin reposición.
- *Muestreo estratificado*. Los elementos de la población se dividen en clases o estratos. La muestra se toma asignando un número o cuota de miembros a cada estrato (proporcional a su tamaño relativo o su variabilidad) y escogiendo los elementos por muestreo aleatorio simple dentro del estrato.
- *Muestreo sistemático*. Los elementos de la población están ordenados en listas. Se divide la población en tantas partes como el tamaño muestral y se elige al azar un número de orden. La muestra se obtiene tomando el elemento que ocupa ese número de orden en cada parte de la población.

En adelante, en los problemas de inferencia estadística consideraremos que las muestras son suficientemente representativas para inferir o estimar las características poblacionales.

Si consideramos una muestra de tamaño  $n$  representativa de la población, puesto que los  $n$  elementos que integran la muestra son elegidos aleatoriamente, es evidente que sus medidas o características son, a su vez, variables aleatorias, ya que dependen de los valores aleatorios de los valores muestrales tomados al azar.

Por tanto, una *muestra* es un vector aleatorio  $(X_1, X_2, \dots, X_n) \in E^n$ , que tendrá asociada una probabilidad de ser elegido.

Llamaremos *estadístico* a una función  $F : E^n \rightarrow \mathbb{R}$ , es decir, una “fórmula” de las variables que transforma los valores tomados de la muestra en un número real. Además, a la distribución de  $F$  se le llama distribución del estadístico en el muestreo. Por ejemplo, la función

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

es un estadístico que permitirá obtener la media de los valores muestrales, cuando dispongamos de los datos de la muestra.

## 7.2. Estimación paramétrica

Cuando se realiza una afirmación acerca de los parámetros de la población en estudio, basándose en la información contenida en la muestra se dice que realizamos una estimación puntual pero si señalamos un intervalo de valores dentro del cual se tiene confianza de que esté el valor del parámetro decimos que estamos realizando una estimación por intervalos.

### 7.2.1. Estimación puntual

El proceso de estimación puntual utiliza un estadístico, que llamaremos *estimador puntual*, para obtener algún parámetro de la población. Como estadístico que es, el estimador puntual es una variable aleatoria que tiene una distribución en el muestreo que depende, en general, del parámetro en cuestión.

Se utilizan dos criterios esenciales para medir la bondad del estimador:

- a) Que sea *centrado* o *insesgado*, es decir, que su media coincida con el parámetro a estimar.
- b) Que sea de *mínima varianza* o que tenga la menor varianza entre todos los estimadores del parámetro.

Si verifica las dos condiciones diremos que el estimador es *eficiente*. A continuación, relacionamos los estadísticos eficientes más usuales, así como su distribución de probabilidad que nos permitirá obtener los intervalos de confianza. Para todos ellos, consideraremos que la muestra de tamaño  $n$  es  $\{x_1, x_2, \dots, x_n\}$ .

### La proporción muestral en una distribución binomial

La proporción muestral del suceso  $E$

$$\hat{p} = \frac{\text{frecuencia absoluta del suceso } E}{n}$$

estima la proporción  $p$  de la población que presenta una determinada característica  $E$  (éxito) frente a los que no la presentan  $F$  (fracaso). Las propiedades más importantes son:

1. El estimador es insesgado, es decir, la distribución en el muestreo de  $\hat{p}$  tiene de media  $p$ .
2. El estimador es de varianza mínima igual a  $\frac{p \cdot q}{n}$  con  $q = 1 - p$ .
3. Para valores grandes del tamaño de la muestra (en la práctica  $n > 30$ ), la proporción muestral  $\hat{p}$  se distribuye según una distribución normal:

$$\text{Si } n > 30 \quad \text{entonces} \quad \hat{p} \rightsquigarrow N\left(p, \sqrt{\frac{pq}{n}}\right) \iff \frac{\hat{p} - p}{\sqrt{pq/n}} \rightsquigarrow N(0, 1)$$

### La media muestral en una distribución de Poisson

La media muestral

$$\hat{\lambda} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

es un estimador puntual del parámetro  $\lambda$  de una población cuya característica estudiada sigue una distribución de Poisson de parámetro  $\lambda$  (= media de la población). Las propiedades más importantes son:

1. El estimador es insesgado, es decir, la distribución en el muestreo de  $\hat{\lambda}$  tiene de media  $\lambda$ .
2. El estimador es de varianza mínima.
3. Si el tamaño de la muestra es suficientemente grande, el estimador  $\hat{\lambda}$  se distribuye según una distribución normal:

$$\hat{\lambda} \rightsquigarrow N\left(\lambda, \sqrt{\frac{\lambda}{n}}\right)$$

### La Cuasivarianza muestral en una distribución normal

La cuasivarianza o varianza muestral

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 f_i = \frac{n}{n-1} \cdot \left( \sum_{i=1}^n x_i^2 f_i - \bar{x}^2 \right)$$

es un estimador de la varianza  $\sigma^2$  de una población cuya característica en estudio sigue una distribución normal  $N(\mu, \sigma)$ . Las propiedades más importantes son:

1. El estimador es insesgado, es decir,  $E(s^2) = \sigma^2$ .
2. El estimador es de varianza mínima.
3. La variable  $\frac{(n-1)s^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$

### La media muestral en una distribución normal

La media muestral

$$\hat{\mu} = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

es un estimador de la media  $\mu$  de una población cuya característica en estudio sigue una distribución normal  $N(\mu, \sigma)$ . Las propiedades más importantes son:

1. El estimador es insesgado, es decir,  $E(\bar{x}) = \mu$ .
2. El estimador es de varianza mínima.
3. Para valores grandes del tamaño de la muestra (en la práctica  $n > 30$ ), la media muestral  $\bar{x}$  se distribuye según una distribución normal que depende del tamaño  $N_p$  de la población:

$$\text{Si } n > 30 \quad \text{entonces} \quad \bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N_p - n}{N_p - 1}}\right)$$

4. Si la población es infinita o el muestreo es con reposición, la segunda raíz vale 1, es decir,

$$\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

lo que permite considerar las siguientes tipificaciones del estimador de la media:

- Si  $\sigma$  es conocido entonces  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$  pues  $\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- Si  $\sigma$  es desconocido y  $n > 30$  entonces

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \rightsquigarrow N(0, 1) \quad \text{pues} \quad \bar{x} \rightsquigarrow N\left(\mu, \frac{s}{\sqrt{n}}\right)$$

- Si  $\sigma$  es desconocido y  $n \leq 30$  entonces  $z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \rightsquigarrow t_{n-1}$

### 7.2.2. Estimación por intervalos

En la práctica, no sólo interesa dar una estimación puntual de un parámetro  $\theta$  sino un intervalo de valores dentro del cual se tiene confianza de que esté el estimador  $\hat{\theta}$  del parámetro. Por tanto, lo que buscamos es un estimador denominado “estimador por intervalo” compuesto de una pareja de estadísticos  $L_i$  (límite inferior) y  $L_s$  (límite superior) tales que

$$P(L_i \leq \hat{\theta} \leq L_s) = 1 - \alpha \quad \text{con} \quad 0 < \alpha < 1$$

donde  $1 - \alpha$  se llama **nivel de confianza** y  $\alpha$  se denomina **nivel de significación**. Es decir, llamamos **intervalo de confianza** para el parámetro  $\theta$  con nivel de confianza  $1 - \alpha$ , a una expresión del tipo  $L_i \leq \hat{\theta} \leq L_s$  donde los límites  $L_i$  y  $L_s$  dependen de la muestra y se calculan de manera tal que si construimos muchos intervalos, cada vez con distintos valores muestrales, el  $100(1 - \alpha)\%$  de ellos contendrán el verdadero valor del parámetro.

Sin embargo, cuando tenemos el intervalo de confianza de una muestra concreta, o este intervalo pertenece al  $100(1 - \alpha)\%$  de los que contienen al parámetro y, por lo tanto, el parámetro está en el intervalo con probabilidad 1; o bien, este intervalo pertenece al  $100\alpha\%$  de los que no contienen al parámetro y, por lo tanto, el parámetro está en el intervalo con probabilidad 0. Pero como difícilmente se llegará a saber con exactitud si el intervalo concreto es de uno u otro tipo, entonces el nivel de confianza  $100(1 - \alpha)\%$  nos determinará una medida de la bondad del intervalo.

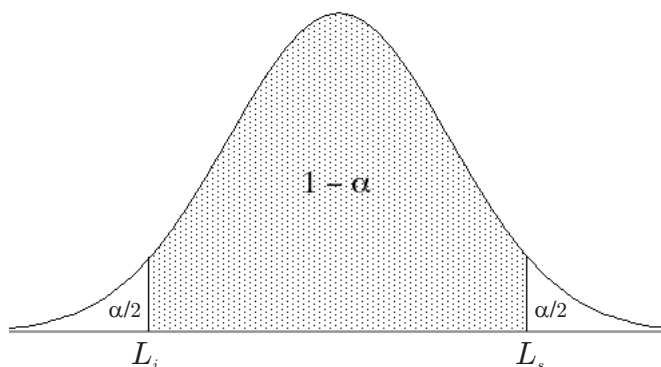


Figura 7.1: Intervalo de confianza

La amplitud del intervalo está íntimamente relacionada con los niveles de confianza y significación. Si la amplitud del intervalo es pequeña entonces la afirmación de que el parámetro pertenece al intervalo tiene gran significación ( $\alpha$  es grande) pero ofrece poca confianza ( $1 - \alpha$  es pequeña). Pero si la amplitud del intervalo es grande entonces la afirmación de que el parámetro pertenece al intervalo tiene menor significación ( $\alpha$  es pequeño) aunque ofrece mucha confianza ( $1 - \alpha$  es grande). Por ejemplo, la afirmación “la altura media de una población está entre 1’69 y 1’71 metros” con  $\alpha = 0’25$  es más significativa que la afirmación “la altura media de una población está entre 1’60 y 1’80 metros” con  $\alpha = 0’01$ , aunque esta última afirmación ofrece más confianza  $1 - \alpha = 0’99$  que la primera  $1 - \alpha = 0’75$ .

Las tablas del anexo presentan los principales intervalos de confianza para los parámetros  $\mu$  y  $\sigma$  de la distribución normal  $N(\mu, \sigma)$ , el parámetro  $p$  de la distribución binomial  $B(n, p)$ , y el parámetro  $\lambda$  de la distribución de Poisson  $P(\lambda)$ . Si no se especifica o se deduce lo contrario,

supondremos que la distribuciones consideradas son de tipo normal, y que el nivel de confianza es del 95 %.

**Ejemplo 7.1** *Obtener dos intervalos de confianza, uno al 99 % y otro al 95 %, para el consumo medio de combustible de un determinado tipo de coche, sabiendo que los consumos observados en 5 ensayos fueron 5'2, 4'3, 5'1, 4'7 y 4'9.*

En primer lugar, suponemos (puesto que ni se dice, ni se deduce lo contrario) que el consumo medio de gasolina de ese determinado tipo de vehículo sigue una distribución normal  $N(\mu, \sigma)$  con  $\mu$  y  $\sigma$  desconocidos.

En este caso, como el tamaño de la muestra es pequeño ( $n \leq 30$ ) y  $\sigma$  es desconocido entonces

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \rightsquigarrow t_{n-1}$$

lo que nos permite determinar los extremos del intervalo de confianza para  $\mu$  que resulta ser

$$I = \left[ \bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right]$$

A partir de la muestra, se obtiene que  $n = 5$ ,  $\bar{x} = 4'84$  y  $s = 0'358$  y si el nivel de significación es  $\alpha = 0'01$  (Nivel de confianza del 99 %) entonces  $t_{\alpha/2, n-1} = 4'604$ . Con estos datos ya podemos obtener el intervalo de confianza al 99 %, que resulta ser

$$\left[ 4'84 \pm 4'604 \cdot \frac{0'358}{\sqrt{5}} \right] = [4'103, 5'577]$$

Si el nivel de significación es  $\alpha = 0'05$  (Nivel de confianza del 95 %) entonces  $t_{\alpha/2, n-1} = 2'776$  y con este dato ya podemos obtener el intervalo de confianza al 95 %, que resulta ser

$$\left[ 4'84 \pm 2'776 \cdot \frac{0'358}{\sqrt{5}} \right] = [4'396, 5'284]$$

Obsérvese que al disminuir el nivel de confianza, también disminuye la amplitud del intervalo pues se pierde confianza de que el parámetro esté en el intervalo, aunque se gana significación pues la región precisa más el rango de posibles valores del parámetro. Como se puede observar en la fórmula, otra forma de reducir la amplitud del intervalo es aumentar el tamaño de la muestra.

□

### 7.3. Contraste de Hipótesis

Otro objetivo fundamental de la Teoría de Muestras, es confirmar o rechazar hipótesis sobre un parámetro poblacional, mediante el empleo de muestras. Es decir, contrastar una hipótesis estadísticamente es juzgar si cierta propiedad supuesta para cierta población es compatible con lo observado en una muestra de ella.

Supongamos que el parámetro de la población, que es objeto de estudio, es  $\theta$ . El procedimiento que se sigue para contratar un valor de  $\theta$  es el siguiente. En primer lugar, se establece a priori, antes de tomar la muestra, la hipótesis que queremos contrastar, es decir, la suposición

que queremos ver si se cumple o no. Esta hipótesis es una igualdad referida al parámetro  $\theta$ , se denomina *hipótesis nula*, se denota por  $H_0$  y será rechazada o no a la vista de los datos de la muestra.

En segundo lugar, se establece, también previamente, la llamada *hipótesis alternativa* que se denota por  $H_a$  y que será admitida cuando  $H_0$  sea rechazada. La hipótesis alternativa puede ser de dos tipos: de tipo desigualdad “mayor que” ( $>$ ) o “menor que” ( $<$ ), y de tipo negación ( $\neq$ ). Como veremos, cada uno de estos tipos de hipótesis dan lugar a un tipo de contraste (unilateral y bilateral, respectivamente)

En tercer lugar, se define un estadístico  $\hat{\theta}$  relacionado con la hipótesis que queremos contrastar. Por ello,  $\hat{\theta}$  se denomina *estadístico de contraste*. La distribución de probabilidad de este estadístico es la que nos permitirá establecer el criterio de aceptación o rechazo de la hipótesis.

A continuación, suponiendo que  $H_0$  es verdadera, se calculan dos regiones complementarias: la *región de aceptación* y la *región crítica* ( $R$ ) o *región de rechazo* de la hipótesis nula. Para establecer estas regiones se fija un valor de probabilidad  $\alpha$  (suficientemente pequeño) que denominaremos *nivel de significación* y que representa la probabilidad de que el estadístico de contraste tome un valor en la región crítica.

Por último, a partir de los valores de la muestra, calculamos el valor  $\hat{\theta}_0$  que toma el estadístico para esos valores y tomamos la decisión final con el siguiente criterio:

- Si  $\hat{\theta}_0 \in R$  entonces rechazamos  $H_0$  y aceptamos  $H_a$ .
- Si  $\hat{\theta}_0 \notin R$  entonces no podemos rechazar  $H_0$ .

Obsérvese que, en el segundo supuesto, no rechazamos la hipótesis nula. Sin embargo, eso no quiere decir que podamos afirmar que sea  $H_0$  sea cierta, aunque tampoco podemos descartarlo y, por lo tanto, admitimos que  $H_0$  es cierta, por una cuestión de simplicidad.

La decisión de rechazar o no la hipótesis nula está basada en los datos de la muestra y, por lo tanto, podemos cometer dos tipos de errores:

1. Error de tipo I: Rechazar  $H_0$  cuando es cierta. La probabilidad de cometer este error es lo que hemos denominado nivel de significación ( $\alpha$ ).

$$\alpha = P(\text{rechazar } H_0 \mid H_0 \text{ es cierta}) = P(\text{aceptar } H_a \mid H_0 \text{ es cierta})$$

2. Error de tipo II: No rechazar  $H_0$  cuando es falsa. La probabilidad de cometer este error se denota con la letra  $\beta$ .

$$\beta = P(\text{no rechazar } H_0 \mid H_0 \text{ es falsa})$$

Estos errores están íntimamente relacionados pues cuando  $\alpha$  decrece entonces  $\beta$  crece, y no es posible encontrar contrastes que permitan simultáneamente hacer ambos errores tan pequeños como queramos. Por lo tanto, será necesario destacar una de las hipótesis de manera que no será rechazada salvo que su falsedad se haga muy evidente. En los contrastes, la hipótesis considerada es  $H_0$  que sólo será rechazada cuando la evidencia de su falsedad supere el  $100(1 - \alpha)\%$ , que denominamos, *nivel de confianza*.

Al tomar un valor de  $\alpha$  pequeño tendremos que  $\beta$  se aproxima a uno. Lo ideal a la hora de definir un contraste es encontrar un compromiso satisfactorio entre  $\alpha$  y  $\beta$ , aunque siempre, a favor de  $H_0$ . Denominamos *potencia del contraste* a la cantidad  $1 - \beta$ , es decir:

$$1 - \beta = P(\text{rechazar } H_0 \mid H_0 \text{ es falsa})$$

En la siguiente tabla se recogen las distintas situaciones que se pueden dar en función de la decisión que tomemos y con las probabilidades correspondientes:

	no rechazar $H_0$	rechazar $H_0$
$H_0$ es cierta	Acierto $1 - \alpha$	Error tipo I $\alpha$
$H_0$ es falsa	Error tipo II $\beta$	Acierto $1 - \beta$

En muchos casos resulta indiferente qué hipótesis se considera la nula y cual la alternativa. Sin embargo, cuando la decisión que tomemos tenga graves consecuencias, entonces tomaremos como hipótesis nula la más desfavorable, es decir, aquella cuyas consecuencias por rechazarla cuando es cierta son más graves que las de aceptarla cuando sea falsa.

Por ejemplo, pensemos que tenemos que decidir si un acusado es inocente o culpable, si un paciente mejora o empeora ante un tratamiento, o si un vehículo de pasajeros tendrá o no un accidente. En estos ejemplos debemos tomar como hipótesis nula que el acusado es inocente, que el enfermo empeora o que el vehículo tendrá un accidente, pues en todas ellas, es más grave rechazarla cuando es cierta que admitirla siendo falsa.

En estos casos, hay que elegir la hipótesis nula a menos que la evidencia a favor de la hipótesis alternativa sea muy significativa. Es decir, sólo se aceptará la hipótesis alternativa para  $\alpha$  próximo a cero, aunque para ellos sea necesario que  $\beta$  sea próximo a uno, ya que las consecuencias del error tipo I (condenar a un inocente, creer equivocadamente que el enfermo mejora ante el tratamiento, o pensar erróneamente que el vehículo no tendrá un accidente), son más graves que las del error de tipo II (liberar a un culpable, creer equivocadamente que el enfermo empeora, o pensar erróneamente que el vehículo tendrá un accidente).

Y ahora, veamos un ejemplo que pone de manifiesto tanto los conceptos y reflexiones que hemos planteado, como el procedimiento que se sigue en el contraste de hipótesis de un problema estadístico.

**Ejemplo 7.2** *Consideremos un proceso de fabricación que en condiciones correctas produce componentes cuya resistencia eléctrica se distribuye normalmente con media 20 Ohm y desviación típica 0'5 Ohm. A veces, y de forma imprevisible, el proceso se desajusta, produciendo un aumento o disminución de la resistencia media de los componentes, pero sin variar la desviación típica. Para contrastar si el proceso funciona correctamente se toma una muestra de cinco unidades midiendo su resistencia, resultando 18'4, 19'2, 20'3, 19'5 y 20'1. ¿Podríamos concluir con estos datos que el proceso está desajustado?*

El problema nos dice que la distribución de probabilidad de la resistencia eléctrica de un componente es de tipo normal y el parámetro objeto de estudio es la media  $\mu$ . Por lo tanto, en primer lugar, elegimos las siguientes hipótesis nula y alternativa que nos permitirán responder

a la pregunta

$$\begin{cases} H_0 & : \mu = 20 \\ H_a & : \mu \neq 20 \end{cases}$$

En segundo lugar, elegimos el estadístico de contraste  $\bar{x}$  asociado a nuestro parámetro  $\mu$  y cuya distribución de probabilidad es:

$$\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{o bien} \quad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$$

En tercer lugar, fijamos el nivel de significación  $\alpha = 0'05$  (valor por defecto) y suponiendo que  $H_0$  es verdadera ( $\mu = 20$ ), se calculan la región de aceptación (con probabilidad  $1 - \alpha$ ) y la región crítica o de rechazo (con probabilidad  $\alpha$ ) a partir de la distribución de probabilidad del estadístico, de la siguiente manera:

$$1 - \alpha = P(\text{"región de aceptación"}) = P(-z_{\alpha/2} \leq \frac{\bar{x} - 20}{0'5/\sqrt{n}} \leq z_{\alpha/2})$$

y, por lo tanto, la región de aceptación para nuestro estadístico de contraste  $z$  es el intervalo  $[-1'96, 1'96]$ . De manera que la región crítica o de rechazo es su complementario, es decir,  $(-\infty, -1'96) \cup (1'96, \infty)$  que se obtendría así:

$$\alpha = P(\text{"región crítica"}) = P\left(\left|\frac{\bar{x} - 20}{0'5/\sqrt{n}}\right| > z_{\alpha/2}\right)$$

Por último, a partir de los valores de la muestra, calculamos el valor del estadístico y tomamos la decisión final. Como  $n = 5$  y  $\bar{x} = 19'5$  entonces el estadístico de contraste  $z = 2'236$  pertenece a la región crítica y, por lo tanto, rechazamos la hipótesis nula y aceptamos la hipótesis alternativa ( $\mu \neq 20$ ), es decir, tendremos que suponer que el proceso se ha desajustado.  $\square$

Obsérvese que, en este ejemplo, la región de rechazo estaba constituida por la unión de dos intervalos. Por esta razón, este tipo de contrastes se denominan bilaterales y se producen cuando la hipótesis alternativa es la negación de la hipótesis nula, es decir, cuando la hipótesis nula es de tipo “=” y la alternativa es de tipo “ $\neq$ ”. Veamos ahora un ejemplo de contraste unilateral

**Ejemplo 7.3** Con los datos del ejemplo 7.2, ¿podemos concluir que el proceso se ha desajustado por exceso?

En este caso, las hipótesis nula y alternativa que nos permitirán responder a la pregunta son

$$\begin{cases} H_0 & : \mu = 20 \\ H_a & : \mu > 20 \end{cases}$$

Elegimos el estadístico de contraste  $\bar{x}$  asociado a nuestro parámetro  $\mu$  y cuya distribución de probabilidad es:

$$\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{o bien} \quad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$$

Si fijamos el nivel de significación  $\alpha = 0'05$  y suponemos que  $H_0$  es verdadera ( $\mu = 20$ ) entonces se puede calcular la región crítica o de rechazo (con probabilidad  $\alpha$ ) a partir de la distribución de probabilidad del estadístico, de la siguiente manera:

$$\alpha = P(\text{"región crítica"}) = P\left(\frac{\bar{x} - 20}{0'5/\sqrt{n}} > z_{\alpha}\right)$$



y, por lo tanto, la región de rechazo es  $(1'645, \infty)$ . Como  $n = 5$  y  $\bar{x} = 19'5$  entonces el estadístico de contraste  $z = -2'236$  no pertenece a la región crítica y, por lo tanto, no podemos rechazar la hipótesis nula. Eso no significa que debamos aceptarla, aunque en estos casos, es bastante común rechazar la hipótesis alternativa, de manera que afirmaríamos que la media no ha aumentado, es decir, que  $\mu \leq 20$ .  $\square$

Las tablas del anexo presentan los principales contrastes de hipótesis para los parámetros  $\mu$  y  $\sigma$  de la distribución normal, y el parámetro  $p$  de la distribución binomial. Para cada uno de ellos, se presentan las regiones críticas o de rechazo, de los distintos contrastes unilaterales y bilaterales. Si no se especifica o se deduce lo contrario, supondremos que la distribuciones consideradas son de tipo normal, y que el nivel de confianza es del 95 %.

## 7.4. Inferencia no paramétrica

Por lo general, para estudiar un carácter en una población, se examina solamente una muestra tomada de la población. Cualquiera que sea la población teórica que se considere, siempre existirán desviaciones entre la distribución teórica y la distribución empírica u observada. El problema consiste, por tanto, en saber en qué medida estas desviaciones son debidas a:

1. *El azar.* Estas diferencias tienden a desaparecer si el número de observaciones (tamaño de la muestra) es suficientemente grande.
2. *Tomar una distribución teórica inadecuada.*

En este último caso, la distribución  $\chi^2$  de Pearson se puede aplicar para ver si un conjunto de datos observados coincide o no con un conjunto de datos esperados.

A continuación se enumeran las principales aplicaciones de la  $\chi^2$ . En cada una de ella se trata de contrastar si una cierta hipótesis  $H_0$  es coherente con los datos obtenidos en la muestra.

1. *Bondad de ajuste:* Se trata de determinar si la hipótesis sobre el tipo de distribución teórica (binomial, poisson, normal, etc.) que rige un experimento es consistente con los datos que aparecen en la muestra.
2. *Contraste de homogeneidad de varias muestras:* Se trata de contrastar si varias muestras con un mismo carácter han sido o no tomadas de una misma población.
3. *Contraste de dependencia o independencia de caracteres:* Se trata de comparar si dos o más distribuciones empíricas son comparables a una misma distribución teórica. Y esto se utiliza para ver si dos caracteres son o no independientes.

En todos los casos se realiza el test de la  $\chi^2$  que consiste en lo siguiente: Supongamos que al tomar una muestra los posibles sucesos  $x_1, x_2, \dots, x_k$  se presentan con frecuencias  $o_1, o_2, \dots, o_k$ , llamadas frecuencias observadas, y que según las leyes de la probabilidad, se esperaba que apareciesen con frecuencias  $e_1, e_2, \dots, e_k$ , llamadas frecuencias esperadas o teóricas. Una medida de la discrepancia entre las frecuencias esperadas y las observadas viene proporcionada por el estadístico  $\hat{\chi}^2$  definido por

$$\hat{\chi}^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k}$$

de manera que si  $\hat{\chi}^2 = 0$  entonces las frecuencias observadas y teóricas coinciden completamente, mientras que si  $\hat{\chi}^2 > 0$ , estas frecuencias no coinciden exactamente. A mayor valor de  $\hat{\chi}^2$  mayor discrepancia entre las frecuencias esperadas y las observadas.

Para contrastar si las frecuencias observadas difieren significativamente de las esperadas utilizaremos que la distribución del estadístico  $\hat{\chi}^2$  se aproxima muy bien si  $k \geq 5$  y  $e_i \geq 5$  por la distribución  $\chi_v^2$ . El número de grados de libertad viene dado por  $v = k - 1 - m$ , siendo  $m$  el número de parámetros de la población que ha sido necesario estimar, a partir de estadísticos de la muestra, para poder calcular las frecuencias teóricas.

### 7.4.1. Bondad de ajuste. Tabla de contingencia

Consideremos en una población el carácter  $X$  que admite las modalidades  $x_1, \dots, x_k$  excluyentes (o una variable continua y dividimos el recorrido en  $k$  clases). Se toma una muestra de tamaño  $n$  de la población, siendo  $o_i$  el número de elementos que presentan la modalidad  $x_i$  (frecuencia observada de  $x_i$ ). Si denotamos por  $p_i$  la probabilidad que teóricamente asignamos a la modalidad  $x_i$ , entonces las frecuencias esperadas para cada  $x_i$  serán  $e_i = n \cdot p_i$  con  $i = 1, \dots, k$ .

Con estos datos podemos construir la siguiente tabla

$X$	$x_1$	$x_2$	$\dots$	$x_i$	$\dots$	$x_k$
Frec. Observada	$o_1$	$o_2$	$\dots$	$o_i$	$\dots$	$o_k$
Frec. Esperada	$e_1$	$e_2$	$\dots$	$e_i$	$\dots$	$e_k$

que recibe el nombre de *tabla de contingencia*  $1 \times K$  y cuyos elementos verifican:

$$\sum_{i=1}^k o_i = n \quad , \quad \sum_{i=1}^k p_i = 1 \quad \text{y} \quad \sum_{i=1}^k e_i = n$$

Ahora consideramos la hipótesis  $H_0$  que consiste en suponer que la distribución teórica escogida representa bien a la distribución empírica y que, por tanto, las desviaciones entre las frecuencias observadas y las teóricas son debidas al azar. Veamos en qué condiciones podemos aceptar o rechazar la hipótesis  $H_0$ . Para ello, definimos el estadístico

$$\hat{\chi}^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k} = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

que sigue aproximadamente una distribución  $\chi^2$  de Pearson con  $k - 1$  grados de libertad si no existe diferencia significativa entre las frecuencias observadas y las teóricas. Así pues, a un nivel de significación  $\alpha$ , tenemos que:

$$\begin{cases} \text{Si } \hat{\chi}^2 < \chi_{\alpha; k-1}^2 & \text{se acepta la hipótesis a nivel } \alpha. \\ \text{Si } \hat{\chi}^2 \geq \chi_{\alpha; k-1}^2 & \text{se rechaza la hipótesis a nivel } \alpha. \end{cases}$$

Para calcular el estadístico  $\hat{\chi}^2$  podemos utilizar la siguiente igualdad:

$$\sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{o_i^2}{e_i} - n$$

En el test de la  $\chi^2$  hay que hacer las siguientes consideraciones:

1. Si la distribución que queremos ajustar es continua, determinaremos, siempre que sea posible,  $k$  clases excluyentes, con  $k \geq 5$ , que determinarán las modalidades de la variable.
2. Si hay alguna modalidad que tenga alguna frecuencia esperada menor que cinco se agrupan dos o más modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor o igual que cinco.
3. Si para obtener las frecuencias esperadas, necesitamos calcular  $m$  parámetros de la distribución teórica entonces los grados de libertad de la distribución  $\chi^2$  son  $k - m - 1$ .
4. Si el estadístico  $\hat{\chi}^2$  es demasiado próximo a cero, debe mirarse con suspicacia el experimento, pues es raro que las frecuencias observadas coincidan demasiado bien con las frecuencias esperadas. Para estudiar estas situaciones podemos examinar si el valor de  $\hat{\chi}^2$  es menor que  $\chi_{0'95;v}^2$  ó  $\chi_{0'99;v}^2$ , en cuyo caso decidimos que el acuerdo es demasiado bueno al nivel de significación 0'05 ó 0'01 respectivamente.

**Ejemplo 7.4** La siguiente tabla contiene las notas (sobre 100) que han obtenido los alumnos en Estadística en los últimos 5 años clasificadas en rangos de 10 puntos:

Rango Nota	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frecuencia	2	30	80	145	250	245	140	75	28	5

¿Se puede afirmar al 95% que las distribución de las notas es de tipo normal?

El problema nos plantea si la distribución de los datos corresponde a una distribución normal  $N(\mu, \sigma)$ . En primer lugar, utilizaremos la muestra para determinar los parámetros de la distribución normal, mediante estimación puntual.

$$\hat{\mu} = \bar{x} = 49'84 \quad \text{y} \quad \hat{\sigma} = s = 16'088$$

En segundo lugar, como el propio enunciado ya establece las modalidades, construimos la tabla de contingencia  $1 \times 10$

$x_i$	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
$o_i$	2	30	80	145	250	245	140	75	28	5
$e_i$	6'64	25'18	76'94	161'64	233'57	232'18	158'77	74'67	24'14	6'28

a partir de los valores de probabilidad  $p_i$  obtenidos de la distribución normal  $N(49'84, 16'088)$ , para cada uno de los intervalos de notas (modalidades), los cuales permiten calcular los valores esperados  $e_i = 1000p_i$ . Por ejemplo, el valor esperado 25'18 para la modalidad 10-20 se ha obtenido multiplicando 1000 por 0'02518, siendo  $0'02518 = P(10 \leq X \leq 20)$  con  $X \sim N(49'84, 16'088)$ .

Ahora, se calcula el estadístico de contraste

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{o_i^2}{e_i} - n = 10'956$$

y se compara con el valor de la  $\chi^2$  con 7 grados de libertad ( $v = k - m - 1 = 10$  modalidades - 2 parámetros estimados - 1) para el valor por defecto  $\alpha = 0'05$ :

$$\hat{\chi}^2 = 10'956 < 14'067 = \chi_{7,0'05}^2$$

y, por lo tanto, se acepta que la distribución de las notas es de tipo normal.  $\square$

### 7.4.2. Contraste de homogeneidad de varias muestras

Una muestra es homogénea cuando todas las observaciones se rigen por la misma distribución de probabilidades. En otro caso se dice que la muestra es heterogénea.

Las causas más importantes por las cuales una muestra no es homogénea son:

- La población es heterogénea respecto a la variable estudiada. Por ejemplo el nivel de renta en una población difiere según se trate de una zona urbana o rural.
- La población es homogénea respecto a la variable del estudio, pero en el proceso de muestreo se producen errores o cambios en el sistema de medida, a consecuencia de lo cual ciertos datos de la muestra son heterogéneos.

El objetivo es determinar si varias muestras de un mismo carácter  $X$  han sido o no tomadas de una misma población. Para ello usaremos el test de la  $\chi^2$  de Pearson de la siguiente manera:

Supongamos que se tienen  $k$  muestras con  $n_1, n_2, \dots, n_k$  elementos cada una. Las cuales tienen, respectivamente,  $o_1, o_2, \dots, o_k$  elementos con una determinada característica  $A$ .

Hacemos la hipótesis  $H_0$ , que consiste en suponer que todas las muestras proceden de la misma población. Bajo esta hipótesis, la proporción  $p$  de elementos con la característica  $A$  es

$$p = \frac{o_1 + o_2 + \dots + o_k}{n_1 + n_2 + \dots + n_k}$$

y el número de elementos esperados en la muestra que poseen la característica  $A$  es:

$$e_i = n_i \cdot p \quad \text{para todo } i = 1, 2, \dots, k.$$

El problema ahora es determinar si la diferencia entre las frecuencias observadas y las esperadas se debe al azar o si se debe a que las muestras no se pueden considerar como procedentes de una misma población. Para ello, definimos el estadístico:

$$\chi_{k-1}^2 = \frac{1}{p(1-p)} \cdot \sum_{i=1}^k \frac{(o_i - e_i)^2}{n_i}$$

que, si  $H_0$  es cierta, sigue aproximadamente una  $\chi^2$  con  $k - 1$  grados de libertad. El número de grados de libertad es  $k - 1$  ya que tenemos  $2k$  variables (frecuencias esperadas) y hay que restar  $k + 1$  parámetros que hemos obtenido de la muestra (el parámetro  $p$  y los  $k$  parámetros  $n_i - n_i \cdot p$ ).

Luego, al nivel de significación  $\alpha$  podemos establecer:

$$\begin{cases} \text{Si } \chi_{k-1}^2 < \chi_{\alpha; k-1}^2 & \text{Se acepta } H_0 \\ \text{Si } \chi_{k-1}^2 \geq \chi_{\alpha; k-1}^2 & \text{Se rechaza } H_0 \end{cases}$$

De manera análoga podemos contrastar si la frecuencia de un elemento de la población se mantiene constante a lo largo de las extracciones o, lo que es lo mismo, las muestras provienen de una población determinada. Así, en una población binomial se puede contrastar la hipótesis

de que la proporción de elementos con una característica  $A$  es constante e igual a  $p$ . Entonces el estadístico:

$$\hat{\chi}^2 = \frac{1}{p(1-p)} \cdot \sum_{i=1}^k \frac{(o_i - n_i p)^2}{n_i}$$

sigue aproximadamente una distribución  $\chi^2$  con  $k$  grados de libertad si la hipótesis es verdadera. Luego si  $\chi_k^2 < \chi_{\alpha; k}^2$  se acepta la hipótesis y en otro caso se rechaza a un nivel de significación  $\alpha$ .

En una población de Poisson se puede contrastar la hipótesis de que el número medio de elementos con la característica  $A$  en cada muestra es constante, es decir:

$$\lambda = \bar{o} = \frac{\sum_{i=1}^k o_i}{k} = \text{constante}$$

entonces, el estadístico

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(o_i - \bar{o})^2}{\bar{o}} = \frac{1}{\bar{o}} \cdot \sum_{i=1}^k o_i^2 - \sum_{i=1}^k o_i$$

sigue aproximadamente una distribución  $\chi^2$  con  $k - 1$  grados de libertad si la hipótesis es verdadera.

#### 7.4.3. Contraste de dependencia o independencia de caracteres. Tablas de contingencia $K \times M$

Hasta ahora hemos utilizado el test de la  $\chi^2$  para saber si una serie de datos se ajustaban o no a una distribución teórica. Podemos igualmente comparar dos o más distribuciones empíricas entre sí si cada una de ellas es comparable a una misma distribución teórica.

Supongamos que queremos comparar dos caracteres  $X$  e  $Y$  en una misma población, que admiten las modalidades siguientes:  $X = \{x_1, x_2, \dots, x_k\}$  e  $Y = \{y_1, y_2, \dots, y_m\}$ . Para ello, tomamos una muestra de tamaño  $n$ , siendo  $o_{ij}$  el número de elementos que presentan la modalidad  $x_i$  de  $X$  e  $y_j$  de  $Y$  (frecuencia observada).

Si consideramos la hipótesis  $H_0$  que consiste en suponer que no existen diferencias significativas entre las dos distribuciones empíricas de  $X$  e  $Y$ , entonces con cada frecuencia observada  $o_{ij}$  tenemos una frecuencia teórica o esperada  $e_{ij}$  que podemos calcular mediante la expresión

$$e_{ij} = p_{ij} \cdot n = \frac{o_{x_i} \cdot o_{y_j}}{n} \quad \text{para todo } i = 1, 2, \dots, k \text{ y } j = 1, 2, \dots, m$$

siendo  $p_{ij}$  las probabilidades de que un elemento tomado de la muestra presente las modalidades  $x_i$  de  $X$  e  $y_j$  de  $Y$ , es decir

$$p_{ij} = \frac{o_{x_i}}{n} \cdot \frac{o_{y_j}}{n}$$

Con estos datos podemos construir la siguiente tabla

$X \setminus Y$	$y_1 \dots$	$y_j \dots$	$y_m$	Frecuencia $o_{x_i}$
$x_1$	$o_{11} \dots$	$o_{1j} \dots$	$o_{1m}$	$o_{x_1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i$	$o_{i1} \dots$	$o_{ij} \dots$	$o_{im}$	$o_{x_j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_k$	$o_{k1} \dots$	$o_{kj} \dots$	$o_{km}$	$o_{x_k}$
Frecuencias $o_{y_j}$	$o_{y_1} \dots$	$o_{y_j} \dots$	$o_{y_m}$	$n$

que recibe el nombre de *tabla de contingencia*  $K \times M$  y cuyos elementos verifican:

$$\sum_{i=1}^k \sum_{j=1}^m o_{ij} = n \quad , \quad \sum_{i=1}^k o_{x_i} = \sum_{j=1}^m o_{y_j} = n \quad , \quad \sum_{i=1}^k \sum_{j=1}^m p_{ij} = 1 \quad y \quad \sum_{i=1}^k \sum_{j=1}^m e_{ij} = n$$

Análogamente a los casos anteriores, definimos el estadístico:

$$\hat{\chi}^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^k \sum_{j=1}^m \frac{o_{ij}^2}{e_{ij}} - n$$

que sigue aproximadamente una distribución  $\chi^2_{(k-1)(m-1)}$  si es cierta  $H_0$ , con  $e_{ij} > 5$ , para todo  $1 \leq i \leq k, 1 \leq j \leq m$ ; en otro caso es preciso agrupar filas o columnas contiguas. Así pues, a nivel de significación  $\alpha$  podemos contratar la hipótesis  $H_0$ :

$$\begin{cases} \text{Si } \chi^2_{(k-1) \cdot (m-1)} < \chi^2_{\alpha; (k-1) \cdot (m-1)} & \text{se acepta } H_0 \\ \text{Si } \chi^2_{(k-1) \cdot (m-1)} \geq \chi^2_{\alpha; (k-1) \cdot (m-1)} & \text{se rechaza } H_0 \end{cases}$$

Este contraste no paramétrico se utiliza muy frecuentemente para ver si existe o no relación entre los caracteres  $X$  e  $Y$ , es decir, si son o no independientes. Entonces recibe el nombre de *contraste de independencia de caracteres*:

$$\begin{cases} \text{Si } \chi^2_{(k-1) \cdot (m-1)} < \chi^2_{\alpha; (k-1) \cdot (m-1)} & XyY \text{ son independientes al nivel } \alpha \\ \text{Si } \chi^2_{(k-1) \cdot (m-1)} \geq \chi^2_{\alpha; (k-1) \cdot (m-1)} & XeY \text{ no son independientes al nivel } \alpha \end{cases}$$

### Coefficiente de contingencia

Una media del grado de relación o dependencia entre dos caracteres  $X$  e  $Y$  en una tabla de contingencia viene dada por el *coeficiente de contingencia*  $C$  que se define por

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

A mayor valor de  $C$  más alto es el grado de dependencia entre las dos variables  $X$  e  $Y$ .

## 7.5. Relación de problemas

1. Un nuevo modelo de automóvil realiza 10 pruebas de consumo con 9 litros de gasolina, obteniéndose: 137'4, 136, 132, 141, 129, 130'8, 140, 129'7, 133 y 136 kilómetros recorridos en cada prueba.
  - a) Utilizar los resultados de las pruebas anteriores para estimar la media y la varianza del consumo de gasolina (suponer que los datos están normalmente distribuidos).
  - b) Estimar un intervalo de confianza para la media de kilómetros recorridos con 9 litros de combustible.
2. Una muestra de 10 medidas de las constantes recuperadoras de muelles para amortiguadores da una media de 15 Nw/mm con desviación típica de 0'2. Encontrar intervalos de confianza al 5 % de la media y de la varianza.
3. El contenido medio en grasa para dos tipos de queso A y B es  $\bar{x}_A=33'2\%$  y  $s_A=3'4\%$  con  $n = 27$  y  $\bar{x}_B=35'4\%$  y  $s_B=3\%$  con  $n = 42$ . Se pide:
  - a) Construir un intervalo de confianza al 95 %, para el porcentaje de grasa de los tipos A y B.
  - b) Construir un intervalo de confianza para la diferencia en el contenido en grasa de ambos tipos.
  - c) ¿Se observan diferencias significativas?
4. Lanzamos una moneda 200 veces.
  - a) Halle un intervalo donde se encontrará el número de caras obtenidas con una probabilidad del 99 %, supuesta la moneda equilibrada.
  - b) Si se obtienen 110 veces caras, ¿deberá suponerse al 99 % que está trucada?
  - c) Si se obtiene una proporción de caras del 45 %, ¿cual debe ser el número mínimo de tiradas para rechazar la hipótesis de estar equilibrada?
5. En un sondeo a 500 votantes del barrio A y 300 del barrio B, un candidato resultó preferido por el 43 % de los de A y el 42 % de los de B. Al nivel  $\alpha = 5\%$ .
  - a) Obtener intervalos de confianza para los resultados esperados en A y en B.
  - b) ¿Puede admitirse que el candidato obtendrá mejores resultados en A que en B?
6. Una compañía aseguradora comprueba que la probabilidad, para determinado grupo de riesgo, de tener un accidente mortal en un periodo del año es de 0'003. Cada accidente provoca un pago fijo de 100 000 euros. Si la compañía tiene 10 000 asegurados.
  - a) Estimar la prima anual que, a un nivel del 1 %, asegure que no se provocarán pérdidas en la empresa.
  - b) Responder a la pregunta anterior si tuviese 100 000 asegurados.
7. Para comprobar si un fármaco es útil en el tratamiento de una enfermedad, de la que datos anteriores nos dan un plazo de recuperación de 34 días con una desviación de 7, tomamos una muestra de 50 pacientes, suministrándosele a 25 de ellos (grupo A) un placebo y a los otros 25 (grupo B) el tratamiento.

El grupo A tuvo un periodo medio de recuperación de 25 días con desviación de 5, mientras el grupo B obtuvo una media de 24 días y desviación 5. Contrastar:

- a) Que el tratamiento es eficaz sobre los métodos anteriores.
  - b) Que su eficacia es psicológica, pues no difiere de los individuos no tratados, que creen que si lo son (grupo A).
8. Un equipo médico sostiene que su tratamiento ha conseguido sanar al 90 % de los pacientes de una enfermedad en 3 días. Realizado un experimento con 400 pacientes sanaron 342 en dicho plazo.
- a) Utilizar el resultado del experimento para estimar la proporción de pacientes que reaccionan favorablemente al tratamiento.
  - b) Contrastar si la hipótesis que sostiene el equipo médico es correcta.
9. Para contrastar el nivel de Matemáticas de los alumnos de dos centros de enseñanza se selecciona un grupo de alumnos de cada centro y se les somete a una prueba de nivel. Las calificaciones obtenidas por los grupos de 40 y 30 alumnos de los centros de enseñanza A y B dan una media de  $5'4$  y  $5'7$  respectivamente; mientras que las desviaciones típicas resultan ser respectivamente de  $1'3$  y  $0'9$ . Contrastar al nivel del 90 % que no existen diferencias en el nivel de conocimientos de Matemáticas entre ambos centros de enseñanza.
10. Para analizar el efecto de un tratamiento contra la procesionaria del pino, se divide el terreno en 100 parcelas de las que aleatoriamente se tratan 40, obteniéndose en las tratadas 20 árboles atacados de 230 observados, mientras en las no tratadas se observaron 7 atacados de 300 observados. ¿Se puede deducir que el tratamiento es eficaz al 10 %? ¿Y al 5 %?
11. Para estimar el número de castaños de un bosque con  $500 \text{ km}^2$ , se seleccionan aleatoriamente 10 parcelas de  $1 \text{ km}^2$  cada una, contando exhaustivamente los castaños existentes. Obteniéndose 25, 27, 32, 28, 23, 20, 28, 19, 17 y 20 en cada una de las parcelas.
- a) Hallar un intervalo de confianza al 5 % para el número medio de castaños por  $\text{km}^2$  y para el bosque completo.
  - b) Hallar si puede aceptarse que existen más de 15.000 castaños en el bosque (con nivel de significación  $\alpha = 0,01$ ).
  - c) Si en vez de 10 parcelas de  $1 \text{ km}^2$ , se hubiesen considerado 100 parcelas de  $0'1 \text{ km}^2$ , obteniéndose una media de  $1/10$  de la anterior y una desviación típica de  $1/10$  de la anterior. ¿Se obtiene un intervalo de confianza del número de castaños del bosque más preciso?
12. En una muestra de 50 neumáticos de cierta clase se obtuvo una vida media de 32 000 km. y una desviación estándar ( $s$ ) de 4 000 km.
- a) ¿Puede afirmar el fabricante que la vida media de esos neumáticos es mayor que 30 000 km.? Establezca y pruebe la hipótesis correspondiente en un nivel del 5 %, suponiendo normalidad.
  - b) ¿Hasta qué número de kilómetros podríamos afirmar que llega la vida media de los neumáticos con el mismo nivel de confianza?



13. Si mediciones simultáneas de una tensión eléctrica por medio de dos tipos diferentes de voltímetros proporcionan las diferencias (en volts) 0'8, 0'2, -0'3, 0'1, 0'0, 0'5, 0'7 y 0'2, ¿puede afirmarse al 4 % que no existen diferencias significativas, en la calibración de los dos tipos de instrumentos?
14. Supóngase que, en un equipo eléctrico alimentado con baterías, es más económico reemplazar todas estas a intervalos fijos que reemplazar cada batería por separado cuando se agota, ello ocurre cuando la desviación estándar de la vida de las mismas sea mayor que cierto límite, esto es, mayor que 5 horas. Plantee y aplique una prueba apropiada, utilizando una muestra de 28 valores de vida con desviación estándar  $s = 3'5$  horas y suponiendo normalidad. Tome  $\alpha = 6\%$ .
15. Suponga que las marcas I y II de focos eléctricos tienen el mismo precio y son de la misma calidad, excepto, tal vez, por su vida útil. Un cliente compró 100 focos de cada marca y comprobó que los focos de la marca I tenían una vida media de 1 120 horas con una desviación estándar de 75 horas; y para los focos de la marca II los valores correspondientes fueron 1 064 y 82 horas, respectivamente. ¿Es significativa la diferencia en media, de la vida útil de los focos de ambas marcas?
16. *Determinación del tamaño muestral.* En los ejercicios anteriores, hemos supuesto conocido el tamaño de la muestra. El problema de determinar el tamaño muestral es crucial ya que un tamaño de la muestra excesivamente elevado puede resultar costoso en tiempo y dinero, sin embargo, si la muestra es demasiado pequeña podemos no encontrar el grado deseado de fiabilidad (la amplitud del intervalo es inversamente proporcional a la raíz cuadrada del tamaño de la muestra). Se trata de despejar la variable  $n$  en los estadísticos de los extremos del intervalo de confianza correspondiente
  - a) Consideremos el intervalo de confianza para la media de una distribución normal de varianza conocida. Determinar el tamaño de la muestra que debemos considerar de forma que la diferencia entre la media poblacional y la media muestral sea, en valor absoluto, menor que un cierto error ( $\epsilon$ ) a un determinado nivel de confianza  $(1-\alpha)$ .
  - b) Una muestra aleatoria de 196 datos extraídos de una población normal de varianza igual a 100, presenta una media muestral de 160.
    - 1) Determinar al 95 % un intervalo de confianza para la media poblacional y señalar la diferencia máxima entre la media muestral y la desconocida media poblacional.
    - 2) Si se quiere tener una confianza del 95 % de que la estimación de la media se encuentra a una distancia de 1'2 de la verdadera media poblacional, ¿debemos tomar más observaciones adicionales? ¿Cuántas?
17. *Nivel crítico.* En los ejercicios anteriores, hemos realizado los cálculos para un determinado nivel de confianza. A partir de los datos de una muestra (tamaño, media, varianza, proporción, etc.) podemos estar interesados en determinar el nivel de confianza crítico a partir del cual se acepta o rechaza una determinada hipótesis sin más que aumentar o disminuir este nivel de confianza. Se trata pues de despejar el valor de  $\alpha$  en el contraste de hipótesis correspondiente.

El cálculo del nivel crítico resulta útil para poder manipular el resultado de un determinado contraste, puesto que una vez calculado, se puede fácilmente establecer el nivel de confianza para que los resultados de la muestra respalden una determinada hipótesis.

Ejemplo: Un determinado partido político realiza un sondeo preelectoral y obtiene un 55 % de éxitos en intención de voto. Con estos resultados, ¿cuál es el mínimo nivel de confianza que le permite asegurar que podrá obtener mayoría absoluta, si la consulta se realizó sobre 240 votantes?

18. El 70 % de las bellotas que produce un árbol son comidas por los animales y el resto germina en un 60 %

- a) Hallar un intervalo de confianza con  $\alpha = 0'02$  para el número de bellotas germinadas, procedentes de un árbol determinado que produjo 20 000 bellotas.
- b) Un ingeniero agrónomo afirma que en determinado tipo de suelo la proporción de las que germinan es del 75 %. Para ello se dejan caer 100 bellotas impidiendo el acceso de animales, de las cuales germinan 67.
  - 1) Contrastar al nivel  $\alpha = 0'01$ , que en ese tipo de suelo se produce un aumento del porcentaje de germinación.
  - 2) Contrastar que la hipótesis del ingeniero es cierta.

19. Analizada la operación de montaje de una máquina de un equipo, se observa que puede ser realizada en dos secuencias diferentes A y B. Para evitar la posible influencia del entrenamiento de los operarios, se seleccionaron aleatoriamente 18, que desconocían el proceso de montaje, asignándoles aleatoriamente al aprendizaje del montaje de una u otra secuencia, tras un mes de aprendizaje, se realizaron mediciones obteniéndose los siguientes tiempos de montaje:

Procedimiento A:	32	37	35	28	41	44	35	31	34
Procedimiento B:	35	31	29	25	34	40	27	32	31

- a) Obtener intervalos al nivel de confianza del 99 % para la media del tiempo de montaje por uno y otro método.
  - b) Contrastar al nivel  $\alpha = 0'10$  la igualdad de varianzas de ambos métodos.
  - c) De acuerdo al resultado obtenido en el apartado anterior, contrastar al nivel  $\alpha = 0'1$  la igualdad de las medias de ambos métodos.
20. Se desea contrastar si la temperatura del agua del mar en Alicante es mayor que en Málaga y para ello se realizaron mediciones cada dos meses durante un año, resultando:

Alicante	14°	16°	18°	21°	22°	14°
Málaga	12°	16°	19°	21°	21°	13°

Realizar el contraste al nivel  $\alpha = 0'05$  de significación.

21. Un estudio del precio de los pisos en una ciudad resultó que en el año 1992 se distribuían normalmente con media 100 000 ptas/m<sup>2</sup> y desviación típica de 8 000 ptas/m<sup>2</sup>.
- a) Estimar el precio mínimo por metro cuadrado que no alcanzan el 25 % de los pisos.
  - b) Si elegimos una muestra al azar de 10 pisos, hallar la probabilidad de que alguno cueste más de 125 000 ptas/m<sup>2</sup>.

- c) En un estudio posterior (1997) se estudian 30 pisos al azar, obteniéndose una media de 105 000 ptas/m<sup>2</sup>, con  $s = 10\,000$  ptas/m<sup>2</sup>. Estudiar si es admisible, al nivel de significación  $\alpha = 0'1$  que la varianza se ha mantenido.
- d) Analizar si puede admitirse que la media ha aumentado, con las mismas hipótesis del apartado anterior.
22. Se quiere estudiar si la velocidad media de lectura es mayor en ambiente urbano que en rural; para ello, se toma una muestra de 500 personas de tipo urbano, resultando en palabras por minuto:

$$\sum_{i=1}^{500} p_i = 75000 \quad , \quad \sum_{i=1}^{500} p_i^2 = 14'23 \cdot 10^6$$

siendo  $p_i$  el número de palabras por minuto del individuo  $i$ -ésimo, mientras que para una muestra de 300 personas de ambiente rural, dieron unos resultados de:

$$\sum_{i=1}^{300} p_i = 43500 \quad , \quad \sum_{i=1}^{300} p_i^2 = 6'83 \cdot 10^6$$

Dar un intervalo de confianza para la diferencia de las velocidades medias en ambos ambientes.

23. Un método de depuración de aguas residuales mediante tratamiento con cloro deja un contenido medio de impurezas de  $1'48 \text{ mg/m}^3$  con  $\sigma = 0'13$ .

Un método alternativo con metano produce, mediante muestreo aleatorio simple, los siguientes resultados en dos sectores de una ciudad:

Sector A: Media=1'45       $n_A=10$        $s_A=0'3$

Sector B: Media=1'43       $n_B=20$        $s_B=0'35$

Contrastar si existen diferencias en media y varianza para la muestra total al nivel 0'1 entre el método con metano y el método con cloro.

Sugerencia: Calcular primero la media y cuasivarianza para el total de los 30 datos muestrales

24. La elasticidad del plástico puede variar dependiendo del proceso por el cual se prepara. Para comparar la elasticidad del plástico producido por dos procesos diferentes se tomaron seis muestras extraídas de cada uno de los procesos, obteniéndose los siguientes resultados:

Proceso A:	6'1	9'2	8'7	7'5	9'0	7'3
Proceso B:	9'2	8'1	6'9	7'9	6'5	9'0

- a) Calcular dos intervalos de confianza al 95 %, uno para la elasticidad media y otro para la varianza de los datos obtenidos en el proceso A. Interpretar los resultados.
- b) Cuestión teórica: Deseamos ser más precisos en nuestras afirmaciones sobre la media y varianza de la elasticidad de los plásticos fabricados de acuerdo al proceso A, es decir, queremos ofrecer intervalos de confianza con menor amplitud. ¿Qué dos soluciones se pueden plantear? Razonar la respuesta.
- c) ¿Presentan los datos suficiente evidencia para poder asegurar que existe diferencia entre las elasticidades medias de los dos procesos? Usar  $\alpha = 0'05$ . Si la respuesta es afirmativa, contrastar qué proceso obtiene un plástico de mayor resistencia.

- d) Obtener un intervalo de confianza al 95 % para la diferencia de las medias de elasticidad de los procesos. A la vista del intervalo calculado, ¿qué respuesta se puede dar a la pregunta del apartado anterior? Comentarla.
25. Los científicos consideran que el benceno es un agente químico que puede causar el cáncer. Diversos estudios han comprobado que la gente que trabaja con benceno durante más de 5 años, tiene 20 veces más probabilidad de contraer leucemia. Como resultado se impuso una Ley para limitar el nivel medio de benceno en el ambiente de trabajo a un máximo de 1 ppm.

Un estudio en una planta productora consiste en tomar 10 muestras del aire en periodos de tiempo regulares (días sucesivos) obteniéndose:

0'95 , 0'97 , 0'90 , 0'88 , 1'00 , 1'05 , 1'18 , 1'13 , 1'15 , 1'09

Estimar si debe aceptarse al 95 % la hipótesis de estar violando el límite medio permitido. ¿Se podría afirmar lo mismo al 80 %? ¿Por qué?

26. El *tiempo de respuesta de un ordenador* se define como el tiempo que un usuario debe esperar mientras el ordenador accede a la información en el disco. Suponga que un centro de datos desea comparar los tiempos de respuesta de dos unidades de disco de ordenador. Se seleccionaron muestras aleatorias independientes de 13 tiempos de respuesta para el Disco 1 y 16 tiempos de respuesta para el Disco 2. A continuación, se presentan los datos registrados en milisegundos:

Disco 1:	59	92	54	102	73	60	73	75	74	84	47	33	61			
Disco 2:	71	38	47	53	63	48	41	68	40	60	44	39	34	75	86	73

Se pide:

- Calcular dos intervalos de confianza al 95 %, uno para el tiempo medio de respuesta y otro para la varianza del tiempo de respuesta del Disco 1. Interpretar los resultados.
  - Cuestión teórica: Deseamos ser más precisos en nuestras afirmaciones sobre la media y varianza del tiempo de respuesta, es decir, queremos ofrecer intervalos de confianza de menor amplitud. ¿Qué dos soluciones se pueden plantear? Razonar la respuesta.
  - Contrastar si podemos considerar que los tiempos medios de respuesta de ambos discos son iguales. Si no lo son, establecer la hipótesis y contrastar cuál de ellos es más rápido.
27. Pensamos que el porcentaje de piezas defectuosas fabricadas por una determinada máquina es del 45 %. Para contrastar nuestra hipótesis se han seleccionado 25 piezas detectándose entre ellas 16 defectuosas.
- ¿Estábamos en lo cierto al 95 %? ¿Y al 99 %?
  - Dar una explicación si las respuestas a los apartados anteriores son distintas.
  - Si alguna respuesta es negativa, proponer una afirmación sobre si el porcentaje real es mayor o menor que lo que pensábamos y contrastar dicha afirmación.

28. En 200 tiradas de una moneda, han salido 115 caras y 85 cruces. Contrastar la hipótesis de que la moneda es buena, con nivel de significación (a) 0'05 y (b) 0'01. Utilice, en primer lugar un contraste paramétrico, y compare los resultados con los que se obtendrían utilizando un contraste no paramétrico.
29. En 120 lanzamientos de un dado las distintas caras del dado han aparecido con frecuencias: 25, 17, 15, 23, 24 y 16. Contrastar al nivel 0'05 que el dado no está trucado.
30. En 360 tiradas de un par de dados, han salido 80 siete y 30 onces. Al nivel de significación del 0'05 contrastar que los dados no están sesgados.
31. Para contrastar una hipótesis no paramétrica se ha realizado tres veces un mismo experimento. Los valores de  $\chi^2$  son 2'37, 2'86 y 3'54 cada uno con un grado de libertad. Verificar que aunque  $H_0$  no se puede rechazar al nivel 0'05 usando un único experimento de los anteriores, sí se puede rechazar cuando se combinan los tres.
32. Se lanzan cinco monedas 1000 veces. Se considera  $o_i$  el número de veces que han salido  $i$  caras en el experimento, resultando la sucesión

$$o_0 = 38 \quad , \quad o_1 = 144 \quad , \quad o_2 = 342 \quad , \quad o_3 = 287 \quad , \quad o_4 = 164 \quad \text{y} \quad o_5 = 25$$

Ajustar una distribución binomial y contrastar la bondad del ajuste.

33. El número de individuos que poseen los cuatro grupos sanguíneos debe estar en las proporciones  $q^2 : p^2 + 2pq : r^2 + 2qr : 2pr$ , siendo  $p + q + r = 1$ . Dadas las frecuencias observadas 180, 360, 132 y 98, verificar la compatibilidad de los resultados con  $p = 0'4$ ,  $q = 0'5$  y  $r = 0'1$ .
34. Las leyes de la herencia de Mendel predicen la aparición de tipos de guisantes en la relación 9 : 3 : 3 : 1 para las clases lisa y amarilla, lisa y verde, arrugada y amarilla, arrugada y verde. En un experimento se obtuvieron, respectivamente, 315, 108, 101 y 32. A un nivel de 0'05, ¿coinciden los datos con la teoría?
35. En un laboratorio se observó el número de partículas  $\alpha$  que llegan a una determinada zona procedentes de una sustancia radiactiva en un corto espacio de tiempo, siempre igual, anotándose los resultados en la siguiente tabla:

Número de partículas	0	1	2	3	4	5
Número de periodos de tiempo	120	200	140	20	10	2

- a) Ajuste una distribución de Poisson.
- b) Calcule la probabilidad con que llegan.
- c) Verifique si el ajuste es correcto mediante una  $\chi^2$ , con un nivel  $\alpha = 0'05$ .
36. En un examen de estadística, se obtuvieron las siguientes calificaciones:

60, 70, 90, 85, 90, 50, 75, 90, 80, 70, 60, 75, 75, 75, 80, 60, 65, 60  
90, 70, 60, 70, 65, 50, 85, 80, 90, 85, 80, 75, 50, 55, 60, 65, 70, 75

Comprobar si las calificaciones obtenidas se distribuyen según una normal a un nivel 0'05.

37. En un hospital se ensayó la eficacia de cinco medicamentos en un grupo de pacientes, con el objeto de determinar si al final del tratamiento un paciente determinado mejoraba o no. Las observaciones que se encontraron están anotadas en la siguiente tabla:

Tratamientos	A	B	C	D	E	Total
Número de Pacientes	51	54	48	49	48	250
Pacientes mejorados	12	8	10	15	5	50

¿Existe diferencia entre los diferentes medicamentos a un nivel de 0'05?

38. En un experimento con 164 personas resfriadas, se administró un medicamento a la mitad de ellas y a la otra mitad se les dio una píldora de azúcar. Con los datos de la siguiente tabla, verificar la hipótesis de que este medicamento no es mejor que la píldora de azúcar para curar los resfriados.

	Beneficiosa	Perjudicial	Sin efecto
Fármaco	50	10	22
Azúcar	42	12	28

39. Una fábrica de automóviles quiere averiguar si el sexo de sus posibles clientes tiene relación con la preferencia de modelo. Se toma una muestra de dos mil posibles clientes y se clasifican así:

Sexo / Modelo	A	B	C
Mujer	340	400	260
Varón	350	270	380

¿Se puede decir que el sexo influye en el modelo elegido a un nivel  $\alpha = 0'01$ ?

40. Una zapatería se abastece de cuatro fabricantes. Cada zapato es inspeccionado antes de ponerlo en venta. Hay tres defectos diferentes que causarían la devolución al fabricante. En una muestra se encontraron los siguientes defectos:

Fabricante / Defecto	I	II	III
A	17	10	13
B	10	10	10
C	18	15	17
D	15	5	10

¿Se puede decir que los defectos son independientes del fabricante a un nivel  $\alpha = 0'01$ ?

41. En dos ciudades  $A$  y  $B$ , se observó el color del pelo y de los ojos de sus habitantes, encontrándose las siguientes tablas:

Ojos / Pelo	Rubio	No rubio
Azul	47	23
No Azul	31	93

Ojos / Pelo	Rubio	No rubio
Azul	54	30
No azul	42	80

Se pide:

- Hallar los coeficientes de contingencia de las dos ciudades.
- ¿En cuál de las dos ciudades podemos afirmar que hay mayor dependencia entre el color del pelo y de los ojos?

# Apuntes de ESTADÍSTICA

## ANEXO

### Tablas de los Intervalos de confianza



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento al profesor Carlos Cerezo Casermeiro, por sus correcciones y sugerencias en la elaboración de estos apuntes.*


## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.



## Anexo A

# Tablas de intervalos de confianza

Intervalos de confianza para la media  $\mu$  de una distribución normal  $N(\mu, \sigma)$

Varianza Conocida	Varianza desconocida	
	Muestras grandes $n > 30$	Muestras pequeñas $n \leq 30$
$I = \left[ \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$	$I = \left[ \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$	$I = \left[ \bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right]$

Intervalo de confianza para la varianza  $\sigma^2$  de una distribución normal  $N(\mu, \sigma)$

$$I = \left[ \frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2}, \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2} \right]$$

Intervalo de confianza para el parámetro  $p$  de una distribución binominal  $B(n, p)$

$$I = \left[ \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Intervalo de confianza para el parámetro  $\lambda$  de una distribución de Poisson  $P(\lambda)$

$$I = \left[ \hat{\lambda} \pm z_{\alpha/2} \sqrt{\frac{\hat{\lambda}}{n}} \right]$$

**Intervalo de confianza para la diferencia de medias  $(\mu_1 - \mu_2)$  de dos distribuciones normales  $N(\mu_1, \sigma_1)$  y  $N(\mu_2, \sigma_2)$**

Varianzas	Muestras	Varianzas	Intervalo
Conocidas			$I = \left[ (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$
Desconocidas	grandes $n_1 + n_2 > 30$ $n_1 \simeq n_2$		$I = \left[ (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$
	Pequeñas $n_1 + n_2 \leq 30$	Iguales	$I = \left[ (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1 + n_2 - 2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$
		Distintas	$I = \left[ (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, f} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$

donde

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \text{y} \quad f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 + 1} + \frac{(s_2^2/n_2)^2}{n_2 + 1}} - 2$$

son respectivamente la media ponderada de las varianzas muestrales y la aproximación de Welch.

**Intervalo de confianza para la razón de varianzas  $\sigma_1^2/\sigma_2^2$  de dos poblaciones normales  $N(\mu_1, \sigma_1)$  y  $N(\mu_2, \sigma_2)$**

$$I = \left[ \frac{s_1^2/s_2^2}{F_{\alpha/2; n_1-1, n_2-1}}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2; n_1-1, n_2-1}} \right]$$

**Intervalo de confianza para la diferencia de parámetros  $(p_1 - p_2)$  de dos distribuciones binomiales  $B(n_1, p_1)$  y  $B(n_2, p_2)$**

$$I = \left[ \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right]$$

# Apuntes de ESTADÍSTICA

## ANEXO

### Tablas de los Contrastes de hipótesis



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento al profesor Carlos Cerezo Casermeiro, por sus correcciones y sugerencias en la elaboración de estos apuntes.*

## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Anexo B

### Tablas de contrastes de hipótesis (regiones de rechazo)

Contraste de hipótesis para la media ( $\mu = \mu_0$ ) de una población normal  $N(\mu, \sigma)$

Varianza	Muestras	$H_0 : \mu \geq \mu_0$ $H_a : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_a : \mu > \mu_0$
conocida		$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$	$\frac{ \bar{x} - \mu_0 }{\sigma/\sqrt{n}} > z_{\alpha/2}$	$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$
desconocida	grandes $n > 30$	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -z_\alpha$	$\frac{ \bar{x} - \mu_0 }{s/\sqrt{n}} > z_{\alpha/2}$	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > z_\alpha$
desconocida	pequeñas $n \leq 30$	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{\alpha, n-1}$	$\frac{ \bar{x} - \mu_0 }{s/\sqrt{n}} > t_{\alpha/2, n-1}$	$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha, n-1}$

Contraste de hipótesis para la varianza ( $\sigma^2 = \sigma_0^2$ ) de una población normal  $N(\mu, \sigma)$

$H_0 : \sigma^2 \geq \sigma_0^2$ $H_a : \sigma^2 < \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_a : \sigma^2 \neq \sigma_0^2$	$H_0 : \sigma^2 \leq \sigma_0^2$ $H_a : \sigma^2 > \sigma_0^2$
$\frac{(n-1)s^2}{\sigma_0^2} < \chi_{1-\alpha, n-1}^2$	$\frac{(n-1)s^2}{\sigma_0^2} \notin [\chi_{1-\frac{\alpha}{2}, n-1}^2; \chi_{\frac{\alpha}{2}, n-1}^2]$	$\frac{(n-1)s^2}{\sigma_0^2} > \chi_{\alpha, n-1}^2$

**Contraste de hipótesis de la igualdad de medias ( $\mu_1 = \mu_2$ ) de dos poblaciones normales  $N(\mu_1, \sigma_1)$  y  $N(\mu_2, \sigma_2)$  de varianzas  $\sigma_1^2$  y  $\sigma_2^2$  conocidas**

$H_0 : \mu_1 \geq \mu_2$ $H_a : \mu_1 < \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_a : \mu_1 > \mu_2$
$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -z_\alpha$	$\frac{ \bar{x}_1 - \bar{x}_2 }{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_\alpha$

**Contraste de hipótesis de la igualdad de medias ( $\mu_1 = \mu_2$ ) de dos poblaciones normales  $N(\mu_1, \sigma_1)$  y  $N(\mu_2, \sigma_2)$  de varianzas  $\sigma_1^2$  y  $\sigma_2^2$  desconocidas para muestras grandes ( $n_1 + n_2 > 30$ ,  $n_1 \simeq n_2$ )**

$H_0 : \mu_1 \geq \mu_2$ $H_a : \mu_1 < \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_a : \mu_1 > \mu_2$
$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} < -z_\alpha$	$\frac{ \bar{x}_1 - \bar{x}_2 }{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > z_{\alpha/2}$	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > z_\alpha$

**Contraste de hipótesis de la igualdad de medias ( $\mu_1 = \mu_2$ ) de dos poblaciones normales  $N(\mu_1, \sigma_1)$  y  $N(\mu_2, \sigma_2)$  de varianzas  $\sigma_1^2$  y  $\sigma_2^2$  desconocidas pero iguales ( $\sigma_1^2 = \sigma_2^2$ ) para muestras pequeñas ( $n_1 + n_2 \leq 30$ )**

$H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \leq \mu_2$
$H_a : \mu_1 < \mu_2$	$H_a : \mu_1 \neq \mu_2$	$H_a : \mu_1 > \mu_2$
$\frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{\alpha, n_1+n_2-2}$	$\frac{ \bar{x}_1 - \bar{x}_2 }{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\frac{\alpha}{2}, n_1+n_2-2}$	$\frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha, n_1+n_2-2}$

donde

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

es la media ponderada de las cuasivarianzas muestrales.

**Contraste de hipótesis de la igualdad de medias ( $\mu_1 = \mu_2$ ) de dos poblaciones normales  $N(\mu_1, \sigma_1)$  y  $N(\mu_2, \sigma_2)$  de varianzas  $\sigma_1^2$  y  $\sigma_2^2$  desconocidas y distintas ( $\sigma_1^2 \neq \sigma_2^2$ ) para muestras pequeñas ( $n_1 + n_2 \leq 30$ )**

$H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \leq \mu_2$
$H_a : \mu_1 < \mu_2$	$H_a : \mu_1 \neq \mu_2$	$H_a : \mu_1 > \mu_2$
$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} < -t_{\alpha, f}$	$\frac{ \bar{x}_1 - \bar{x}_2 }{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{\alpha/2, f}$	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{\alpha, f}$

donde

$$f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 + 1} + \frac{(s_2^2/n_2)^2}{n_2 + 1}} - 2$$

es la aproximación de Welch.

**Contraste de hipótesis de la igualdad de varianzas ( $\sigma_1^2 = \sigma_2^2$ ) de dos poblaciones normales  $N(\mu_1, \sigma_1)$  y  $N(\mu_2, \sigma_2)$**

$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_a : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_a : \sigma_1^2 \neq \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_a : \sigma_1^2 > \sigma_2^2$
$\frac{s_1^2}{s_2^2} < F_{1-\alpha; n_1-1, n_2-1}$	$\frac{s_1^2}{s_2^2} \notin [F_{1-\frac{\alpha}{2}; n_1-1, n_2-1}, F_{\frac{\alpha}{2}; n_1-1, n_2-1}]$	$\frac{s_1^2}{s_2^2} > F_{\alpha; n_1-1, n_2-1}$

**Contraste de hipótesis para el parámetro  $p$  de una distribución binomial  $B(n, p)$**

$H_0 : p \geq p_0$ $H_a : p < p_0$	$H_0 : p = p_0$ $H_a : p \neq p_0$	$H_0 : p \leq p_0$ $H_a : p > p_0$
$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < -z_\alpha$	$\frac{ \hat{p} - p_0 }{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > z_{\alpha/2}$	$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > z_\alpha$

**Contraste de hipótesis para la igualdad de los parámetros ( $p_1 = p_2$ ) de dos distribuciones binomiales  $B_1(n_1, p_1)$  y  $B_2(n_2, p_2)$  para muestras grandes**

$H_0 : p_1 \geq p_2$ $H_a : p_1 < p_2$	$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$	$H_0 : p_1 \leq p_2$ $H_a : p_1 > p_2$
$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < -z_\alpha$	$\frac{ \hat{p}_1 - \hat{p}_2 }{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > z_{\alpha/2}$	$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > z_\alpha$



# Apuntes de ESTADÍSTICA

## ANEXO

### Tablas de las Distribuciones de probabilidad



*Sixto Sánchez Merino*  
Dpto. de Matemática Aplicada  
Universidad de Málaga



*Mi agradecimiento al profesor Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.*

## *Apuntes de Estadística*

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

*Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.*

**Usted es libre de:**

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

**Bajo las condiciones siguientes:**

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

## Anexo C

# Tablas de las distribuciones de probabilidad

En este anexo se incluyen las tablas de las distribuciones de probabilidad más usuales.



# Distribución Binomial $B(n, p)$

$$b(n, k, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$n$	$k$	$p$												
		0,01	0,05	0,10	0,15	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,49	0,50
2	0	0,9801	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4444	0,4225	0,3600	0,3025	0,2601	0,2500
2	1	0,0198	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4444	0,4550	0,4800	0,4950	0,4998	0,5000
2	2	0,0001	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1111	0,1225	0,1600	0,2025	0,2401	0,2500
3	0	0,9703	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2963	0,2746	0,2160	0,1664	0,1327	0,1250
3	1	0,0294	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4444	0,4436	0,4320	0,4084	0,3823	0,3750
3	2	0,0003	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2222	0,2389	0,2880	0,3341	0,3674	0,3750
3	3	0,0000	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0370	0,0429	0,0640	0,0911	0,1176	0,1250
4	0	0,9606	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1975	0,1785	0,1296	0,0915	0,0677	0,0625
4	1	0,0388	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3951	0,3845	0,3456	0,2995	0,2600	0,2500
4	2	0,0006	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,2963	0,3105	0,3456	0,3675	0,3747	0,3750
4	3	0,0000	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,0988	0,1115	0,1536	0,2005	0,2400	0,2500
4	4	0,0000	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0123	0,0150	0,0256	0,0410	0,0576	0,0625
5	0	0,9510	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1317	0,1160	0,0778	0,0503	0,0345	0,0313
5	1	0,0480	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3292	0,3124	0,2592	0,2059	0,1657	0,1563
5	2	0,0010	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3292	0,3364	0,3456	0,3369	0,3185	0,3125
5	3	0,0000	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1646	0,1811	0,2304	0,2757	0,3060	0,3125
5	4	0,0000	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0412	0,0488	0,0768	0,1128	0,1470	0,1563
5	5	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0041	0,0053	0,0102	0,0185	0,0282	0,0313
6	0	0,9415	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0878	0,0754	0,0467	0,0277	0,0176	0,0156
6	1	0,0571	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2634	0,2437	0,1866	0,1359	0,1014	0,0938
6	2	0,0014	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3292	0,3280	0,3110	0,2780	0,2436	0,2344
6	3	0,0000	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2195	0,2355	0,2765	0,3032	0,3121	0,3125
6	4	0,0000	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0823	0,0951	0,1382	0,1861	0,2249	0,2344
6	5	0,0000	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0165	0,0205	0,0369	0,0609	0,0864	0,0938
6	6	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0014	0,0018	0,0041	0,0083	0,0138	0,0156
7	0	0,9321	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0585	0,0490	0,0280	0,0152	0,0090	0,0078
7	1	0,0659	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,2048	0,1848	0,1306	0,0872	0,0604	0,0547
7	2	0,0020	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,3073	0,2985	0,2613	0,2140	0,1740	0,1641
7	3	0,0000	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2561	0,2679	0,2903	0,2918	0,2786	0,2734
7	4	0,0000	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1280	0,1442	0,1935	0,2388	0,2676	0,2734
7	5	0,0000	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0384	0,0466	0,0774	0,1172	0,1543	0,1641
7	6	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0064	0,0084	0,0172	0,0320	0,0494	0,0547
7	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0006	0,0016	0,0037	0,0068	0,0078
8	0	0,9227	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0390	0,0319	0,0168	0,0084	0,0046	0,0039
8	1	0,0746	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1561	0,1373	0,0896	0,0548	0,0352	0,0313
8	2	0,0026	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2731	0,2587	0,2090	0,1569	0,1183	0,1094
8	3	0,0001	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2731	0,2786	0,2787	0,2568	0,2273	0,2188
8	4	0,0000	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1707	0,1875	0,2322	0,2627	0,2730	0,2734
8	5	0,0000	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0683	0,0808	0,1239	0,1719	0,2098	0,2188
8	6	0,0000	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0171	0,0217	0,0413	0,0703	0,1008	0,1094
8	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0024	0,0033	0,0079	0,0164	0,0277	0,0313
8	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0002	0,0007	0,0017	0,0033	0,0039
9	0	0,9135	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0260	0,0207	0,0101	0,0046	0,0023	0,0020
9	1	0,0830	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1171	0,1004	0,0605	0,0339	0,0202	0,0176
9	2	0,0034	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2341	0,2162	0,1612	0,1110	0,0776	0,0703
9	3	0,0001	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2731	0,2716	0,2508	0,2119	0,1739	0,1641
9	4	0,0000	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2048	0,2194	0,2508	0,2600	0,2506	0,2461
9	5	0,0000	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1024	0,1181	0,1672	0,2128	0,2408	0,2461
9	6	0,0000	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0341	0,0424	0,0743	0,1160	0,1542	0,1641
9	7	0,0000	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0073	0,0098	0,0212	0,0407	0,0635	0,0703
9	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0009	0,0013	0,0035	0,0083	0,0153	0,0176
9	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003	0,0008	0,0016	0,0020
10	0	0,9044	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0173	0,0135	0,0060	0,0025	0,0012	0,0010
10	1	0,0914	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0867	0,0725	0,0403	0,0207	0,0114	0,0098
10	2	0,0042	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1951	0,1757	0,1209	0,0763	0,0494	0,0439
10	3	0,0001	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2601	0,2522	0,2150	0,1665	0,1267	0,1172
10	4	0,0000	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2276	0,2377	0,2508	0,2384	0,2130	0,2051
10	5	0,0000	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1366	0,1536	0,2007	0,2340	0,2456	0,2461
10	6	0,0000	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0569	0,0689	0,1115	0,1596	0,1966	0,2051
10	7	0,0000	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0163	0,0212	0,0425	0,0746	0,1080	0,1172
10	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0030	0,0043	0,0106	0,0229	0,0389	0,0439
10	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0005	0,0016	0,0042	0,0083	0,0098
10	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0010



# Distribución de Poisson $P(\lambda)$

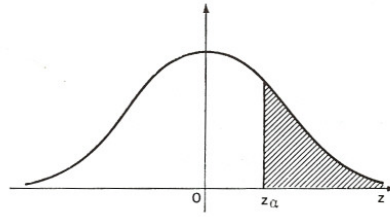
$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$\lambda$	0	1	2	3	4	5	6	7	8	9	10	11	12
0,1	0,9048	0,0905	0,0045	0,0002	0,0000								
0,2	0,8187	0,1637	0,0164	0,0011	0,0001	0,0000							
0,3	0,7408	0,2222	0,0333	0,0033	0,0003	0,0000							
0,4	0,6703	0,2681	0,0536	0,0072	0,0007	0,0001	0,0000						
0,5	0,6065	0,3033	0,0758	0,0126	0,0016	0,0002	0,0000						
0,6	0,5488	0,3293	0,0988	0,0198	0,0030	0,0004	0,0000						
0,7	0,4966	0,3476	0,1217	0,0284	0,0050	0,0007	0,0001	0,0000					
0,8	0,4493	0,3595	0,1438	0,0383	0,0077	0,0012	0,0002	0,0000					
0,9	0,4066	0,3659	0,1647	0,0494	0,0111	0,0020	0,0003	0,0000					
1,0	0,3679	0,3679	0,1839	0,0613	0,0153	0,0031	0,0005	0,0001	0,0000				
1,1	0,3329	0,3662	0,2014	0,0738	0,0203	0,0045	0,0008	0,0001	0,0000				
1,2	0,3012	0,3614	0,2169	0,0867	0,0260	0,0062	0,0012	0,0002	0,0000				
1,3	0,2725	0,3543	0,2303	0,0998	0,0324	0,0084	0,0018	0,0003	0,0001	0,0000			
1,4	0,2466	0,3452	0,2417	0,1128	0,0395	0,0111	0,0026	0,0005	0,0001	0,0000			
1,5	0,2231	0,3347	0,2510	0,1255	0,0471	0,0141	0,0035	0,0008	0,0001	0,0000			
1,6	0,2019	0,3230	0,2584	0,1378	0,0551	0,0176	0,0047	0,0011	0,0002	0,0000			
1,7	0,1827	0,3106	0,2640	0,1496	0,0636	0,0216	0,0061	0,0015	0,0003	0,0001	0,0000		
1,8	0,1653	0,2975	0,2678	0,1607	0,0723	0,0260	0,0078	0,0020	0,0005	0,0001	0,0000		
1,9	0,1496	0,2842	0,2700	0,1710	0,0812	0,0309	0,0098	0,0027	0,0006	0,0001	0,0000		
2,0	0,1353	0,2707	0,2707	0,1804	0,0902	0,0361	0,0120	0,0034	0,0009	0,0002	0,0000		
2,2	0,1108	0,2438	0,2681	0,1966	0,1082	0,0476	0,0174	0,0055	0,0015	0,0004	0,0001	0,0000	
2,4	0,0907	0,2177	0,2613	0,2090	0,1254	0,0602	0,0241	0,0083	0,0025	0,0007	0,0002	0,0000	
2,6	0,0743	0,1931	0,2510	0,2176	0,1414	0,0735	0,0319	0,0118	0,0038	0,0011	0,0003	0,0001	0,0000
2,8	0,0608	0,1703	0,2384	0,2225	0,1557	0,0872	0,0407	0,0163	0,0057	0,0018	0,0005	0,0001	0,0000
3,0	0,0498	0,1494	0,2240	0,2240	0,1680	0,1008	0,0504	0,0216	0,0081	0,0027	0,0008	0,0002	0,0001
3,2	0,0408	0,1304	0,2087	0,2226	0,1781	0,1140	0,0608	0,0278	0,0111	0,0040	0,0013	0,0004	0,0001
3,4	0,0334	0,1135	0,1929	0,2186	0,1858	0,1264	0,0716	0,0348	0,0148	0,0056	0,0019	0,0006	0,0002
3,6	0,0273	0,0984	0,1771	0,2125	0,1912	0,1377	0,0826	0,0425	0,0191	0,0076	0,0028	0,0009	0,0003
3,8	0,0224	0,0850	0,1615	0,2046	0,1944	0,1477	0,0936	0,0508	0,0241	0,0102	0,0039	0,0013	0,0004
4,0	0,0183	0,0733	0,1465	0,1954	0,1954	0,1563	0,1042	0,0595	0,0298	0,0132	0,0053	0,0019	0,0006
5,0	0,0067	0,0337	0,0842	0,1404	0,1755	0,1755	0,1462	0,1044	0,0653	0,0363	0,0181	0,0082	0,0034
6,0	0,0025	0,0149	0,0446	0,0892	0,1339	0,1606	0,1606	0,1377	0,1033	0,0688	0,0413	0,0225	0,0113
7,0	0,0009	0,0064	0,0223	0,0521	0,0912	0,1277	0,1490	0,1490	0,1304	0,1014	0,0710	0,0452	0,0263
8,0	0,0003	0,0027	0,0107	0,0286	0,0573	0,0916	0,1221	0,1396	0,1396	0,1241	0,0993	0,0722	0,0481
9,0	0,0001	0,0011	0,0050	0,0150	0,0337	0,0607	0,0911	0,1171	0,1318	0,1318	0,1186	0,0970	0,0728
10,0	0,0000	0,0005	0,0023	0,0076	0,0189	0,0378	0,0631	0,0901	0,1126	0,1251	0,1251	0,1137	0,0948
	13	14	15	16	17	18	19	20	21	22	23	24	25
5,0	0,0013	0,0005	0,0002	0,0000									
6,0	0,0052	0,0022	0,0009	0,0003	0,0001	0,0000							
7,0	0,0142	0,0071	0,0033	0,0014	0,0006	0,0002	0,0001	0,0000					
8,0	0,0296	0,0169	0,0090	0,0045	0,0021	0,0009	0,0004	0,0002	0,0001	0,0000			
9,0	0,0504	0,0324	0,0194	0,0109	0,0058	0,0029	0,0014	0,0006	0,0003	0,0001	0,0000		
10,0	0,0729	0,0521	0,0347	0,0217	0,0128	0,0071	0,0037	0,0019	0,0009	0,0004	0,0002	0,0001	0,0000





# Distribución Normal $N(0, 1)$

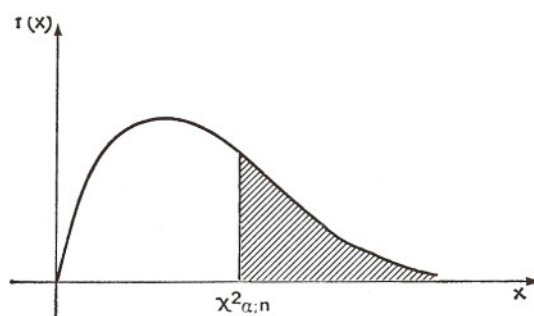


$z_\alpha$	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014

	0,00	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
3	1,35E-03	9,68E-04	6,87E-04	4,83E-04	3,37E-04	2,33E-04	1,59E-04	1,08E-04	7,24E-05	4,81E-05
4	3,17E-05	2,07E-05	1,34E-05	8,55E-06	5,42E-06	3,40E-06	2,11E-06	1,30E-06	7,94E-07	4,80E-07
5	2,87E-07	1,70E-07	9,98E-08	5,80E-08	3,34E-08	1,90E-08	1,07E-08	6,01E-09	3,33E-09	1,82E-09
6	9,90E-10	5,32E-10	2,83E-10	1,49E-10	7,80E-11	4,04E-11	2,07E-11	1,05E-11	5,26E-12	2,62E-12



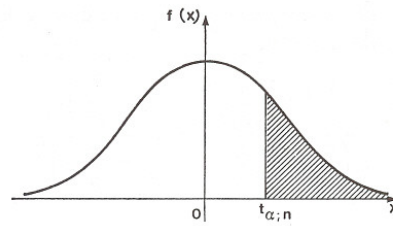
# Distribución $\chi^2$



$n \backslash \alpha$	0,995	0,99	0,98	0,975	0,95	0,90	0,10	0,05	0,025	0,02	0,01
1	3,927E-05	1,571E-04	6,285E-04	9,821E-04	0,0039	0,0158	2,706	3,841	5,024	5,412	6,635
2	0,0100	0,0201	0,0404	0,0506	0,103	0,211	4,605	5,991	7,378	7,824	9,210
3	0,072	0,115	0,185	0,216	0,352	0,584	6,251	7,815	9,348	9,837	11,345
4	0,207	0,297	0,429	0,484	0,711	1,064	7,779	9,488	11,143	11,668	13,277
5	0,412	0,554	0,752	0,831	1,145	1,610	9,236	11,070	12,833	13,388	15,086
6	0,676	0,872	1,134	1,237	1,635	2,204	10,645	12,592	14,449	15,033	16,812
7	0,989	1,239	1,564	1,690	2,167	2,833	12,017	14,067	16,013	16,622	18,475
8	1,344	1,646	2,032	2,180	2,733	3,490	13,362	15,507	17,535	18,168	20,090
9	1,735	2,088	2,532	2,700	3,325	4,168	14,684	16,919	19,023	19,679	21,666
10	2,156	2,558	3,059	3,247	3,940	4,865	15,987	18,307	20,483	21,161	23,209
11	2,603	3,053	3,609	3,816	4,575	5,578	17,275	19,675	21,920	22,618	24,725
12	3,074	3,571	4,178	4,404	5,226	6,304	18,549	21,026	23,337	24,054	26,217
13	3,565	4,107	4,765	5,009	5,892	7,042	19,812	22,362	24,736	25,472	27,688
14	4,075	4,660	5,368	5,629	6,571	7,790	21,064	23,685	26,119	26,873	29,141
15	4,601	5,229	5,985	6,262	7,261	8,547	22,307	24,996	27,488	28,259	30,578
16	5,142	5,812	6,614	6,908	7,962	9,312	23,542	26,296	28,845	29,633	32,000
17	5,697	6,408	7,255	7,564	8,672	10,085	24,769	27,587	30,191	30,995	33,409
18	6,265	7,015	7,906	8,231	9,390	10,865	25,989	28,869	31,526	32,346	34,805
19	6,844	7,633	8,567	8,907	10,117	11,651	27,204	30,144	32,852	33,687	36,191
20	7,434	8,260	9,237	9,591	10,851	12,443	28,412	31,410	34,170	35,020	37,566
21	8,034	8,897	9,915	10,283	11,591	13,240	29,615	32,671	35,479	36,343	38,932
22	8,643	9,542	10,600	10,982	12,338	14,041	30,813	33,924	36,781	37,659	40,289
23	9,260	10,196	11,293	11,689	13,091	14,848	32,007	35,172	38,076	38,968	41,638
24	9,886	10,856	11,992	12,401	13,848	15,659	33,196	36,415	39,364	40,270	42,980
25	10,520	11,524	12,697	13,120	14,611	16,473	34,382	37,652	40,646	41,566	44,314
26	11,160	12,198	13,409	13,844	15,379	17,292	35,563	38,885	41,923	42,856	45,642
27	11,808	12,879	14,125	14,573	16,151	18,114	36,741	40,113	43,195	44,140	46,963
28	12,461	13,565	14,847	15,308	16,928	18,939	37,916	41,337	44,461	45,419	48,278
29	13,121	14,256	15,574	16,047	17,708	19,768	39,087	42,557	45,722	46,693	49,588
30	13,787	14,953	16,306	16,791	18,493	20,599	40,256	43,773	46,979	47,962	50,892



# Distribución $t$ de Student



$n \backslash \alpha$	0,40	0,30	0,20	0,10	0,050	0,025	0,010	0,005	0,001	0,0005
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,21	12,92
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,526	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,416
100	0,254	0,526	0,845	1,290	1,660	1,984	2,364	2,626	3,174	3,390
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,340
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,107	3,310
1E+05	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291



Distribución  $F$  de Fisher-Snedecor para  $\alpha = 0'1$

$n_2 \setminus n_1$	1	2	3	4	5	6	7	8	9	10	12	15	24	30	40	60	120	1E+05
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	62,00	62,26	62,53	62,79	63,06	63,33
2	8,526	9,000	9,162	9,243	9,293	9,326	9,349	9,367	9,381	9,392	9,408	9,425	9,450	9,458	9,466	9,475	9,483	9,491
3	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,216	5,200	5,176	5,168	5,160	5,151	5,143	5,134
4	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,896	3,870	3,831	3,817	3,804	3,790	3,775	3,761
5	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,268	3,238	3,191	3,174	3,157	3,140	3,123	3,105
6	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,905	2,871	2,818	2,800	2,781	2,762	2,742	2,722
7	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,668	2,632	2,575	2,555	2,535	2,514	2,493	2,471
8	3,458	3,113	2,924	2,806	2,726	2,668	2,624	2,589	2,561	2,538	2,502	2,464	2,404	2,383	2,361	2,339	2,316	2,293
9	3,360	3,006	2,813	2,693	2,611	2,551	2,505	2,469	2,440	2,416	2,379	2,340	2,277	2,255	2,232	2,208	2,184	2,159
10	3,285	2,924	2,728	2,605	2,522	2,461	2,414	2,377	2,347	2,323	2,284	2,244	2,178	2,155	2,132	2,107	2,082	2,055
11	3,225	2,860	2,660	2,536	2,451	2,389	2,342	2,304	2,274	2,248	2,209	2,167	2,100	2,076	2,052	2,026	2,000	1,972
12	3,177	2,807	2,606	2,480	2,394	2,331	2,283	2,245	2,214	2,188	2,147	2,105	2,036	2,011	1,986	1,960	1,932	1,904
13	3,136	2,763	2,560	2,434	2,347	2,283	2,234	2,195	2,164	2,138	2,097	2,053	1,983	1,958	1,931	1,904	1,876	1,846
14	3,102	2,726	2,522	2,395	2,307	2,243	2,193	2,154	2,122	2,095	2,054	2,010	1,938	1,912	1,885	1,857	1,828	1,797
15	3,073	2,695	2,490	2,361	2,273	2,208	2,158	2,119	2,086	2,059	2,017	1,972	1,899	1,873	1,845	1,817	1,787	1,755
16	3,048	2,668	2,462	2,333	2,244	2,178	2,128	2,088	2,055	2,028	1,985	1,940	1,866	1,839	1,811	1,782	1,751	1,718
17	3,026	2,645	2,437	2,308	2,218	2,152	2,102	2,061	2,028	2,001	1,958	1,912	1,836	1,809	1,781	1,751	1,719	1,686
18	3,007	2,624	2,416	2,286	2,196	2,130	2,079	2,038	2,005	1,977	1,933	1,887	1,810	1,783	1,754	1,723	1,691	1,657
19	2,990	2,606	2,397	2,266	2,176	2,109	2,058	2,017	1,984	1,956	1,912	1,865	1,787	1,759	1,730	1,699	1,666	1,631
20	2,975	2,589	2,380	2,249	2,158	2,091	2,040	1,999	1,965	1,937	1,892	1,845	1,767	1,738	1,708	1,677	1,643	1,607
21	2,961	2,575	2,365	2,233	2,142	2,075	2,023	1,982	1,948	1,920	1,875	1,827	1,748	1,719	1,689	1,657	1,623	1,586
22	2,949	2,561	2,351	2,219	2,128	2,060	2,008	1,967	1,933	1,904	1,859	1,811	1,731	1,702	1,671	1,639	1,604	1,567
23	2,937	2,549	2,339	2,207	2,115	2,047	1,995	1,953	1,919	1,890	1,845	1,796	1,716	1,686	1,655	1,622	1,587	1,549
24	2,927	2,538	2,327	2,195	2,103	2,035	1,983	1,941	1,906	1,877	1,832	1,783	1,702	1,672	1,641	1,607	1,571	1,533
25	2,918	2,528	2,317	2,184	2,092	2,024	1,971	1,929	1,895	1,866	1,820	1,771	1,689	1,659	1,627	1,593	1,557	1,518
26	2,909	2,519	2,307	2,174	2,082	2,014	1,961	1,919	1,884	1,855	1,809	1,760	1,677	1,647	1,615	1,581	1,544	1,504
27	2,901	2,511	2,299	2,165	2,073	2,005	1,952	1,909	1,874	1,845	1,799	1,749	1,666	1,636	1,603	1,569	1,531	1,491
28	2,894	2,503	2,291	2,157	2,064	1,996	1,943	1,900	1,865	1,836	1,790	1,740	1,656	1,625	1,592	1,558	1,520	1,478
29	2,887	2,495	2,283	2,149	2,057	1,988	1,935	1,892	1,857	1,827	1,781	1,731	1,647	1,616	1,583	1,547	1,509	1,467
30	2,881	2,489	2,276	2,142	2,049	1,980	1,927	1,884	1,849	1,819	1,773	1,722	1,638	1,606	1,573	1,538	1,499	1,456
40	2,835	2,440	2,226	2,091	1,997	1,927	1,873	1,829	1,793	1,763	1,715	1,662	1,574	1,541	1,506	1,467	1,425	1,377
60	2,791	2,393	2,177	2,041	1,946	1,875	1,819	1,775	1,738	1,707	1,657	1,603	1,511	1,476	1,437	1,395	1,348	1,292
120	2,748	2,347	2,130	1,992	1,896	1,824	1,767	1,722	1,684	1,652	1,601	1,545	1,447	1,409	1,368	1,320	1,265	1,193
1E+05	2,706	2,303	2,084	1,945	1,847	1,774	1,717	1,670	1,632	1,599	1,546	1,487	1,383	1,342	1,295	1,240	1,169	1,008





Distribución  $F$  de Fisher-Snedecor para  $\alpha = 0'05$

$n_2 \setminus n_1$	1	2	3	4	5	6	7	8	9	10	12	15	24	30	40	60	120	1E+05
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786	8,745	8,703	8,639	8,617	8,594	8,572	8,549	8,526
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,912	5,858	5,774	5,746	5,717	5,688	5,658	5,628
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,678	4,619	4,527	4,496	4,464	4,431	4,398	4,365
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,000	3,938	3,841	3,808	3,774	3,740	3,705	3,669
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,575	3,511	3,410	3,376	3,340	3,304	3,267	3,230
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347	3,284	3,218	3,115	3,079	3,043	3,005	2,967	2,928
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,073	3,006	2,900	2,864	2,826	2,787	2,748	2,707
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,913	2,845	2,737	2,700	2,661	2,621	2,580	2,538
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,788	2,719	2,609	2,570	2,531	2,490	2,448	2,405
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,687	2,617	2,505	2,466	2,426	2,384	2,341	2,296
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,604	2,533	2,420	2,380	2,339	2,297	2,252	2,206
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,534	2,463	2,349	2,308	2,266	2,223	2,178	2,131
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,475	2,403	2,288	2,247	2,204	2,160	2,114	2,066
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,425	2,352	2,235	2,194	2,151	2,106	2,059	2,010
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,381	2,308	2,190	2,148	2,104	2,058	2,011	1,960
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,342	2,269	2,150	2,107	2,063	2,017	1,968	1,917
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378	2,308	2,234	2,114	2,071	2,026	1,980	1,930	1,878
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,278	2,203	2,082	2,039	1,994	1,946	1,896	1,843
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420	2,366	2,321	2,250	2,176	2,054	2,010	1,965	1,916	1,866	1,812
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,342	2,297	2,226	2,151	2,028	1,984	1,938	1,889	1,838	1,783
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,320	2,275	2,204	2,128	2,005	1,961	1,914	1,865	1,813	1,757
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,300	2,255	2,183	2,108	1,984	1,939	1,892	1,842	1,790	1,733
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282	2,236	2,165	2,089	1,964	1,919	1,872	1,822	1,768	1,711
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,265	2,220	2,148	2,072	1,946	1,901	1,853	1,803	1,749	1,691
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305	2,250	2,204	2,132	2,056	1,930	1,884	1,836	1,785	1,731	1,672
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291	2,236	2,190	2,118	2,041	1,915	1,869	1,820	1,769	1,714	1,654
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,223	2,177	2,104	2,027	1,901	1,854	1,806	1,754	1,698	1,638
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,092	2,015	1,887	1,841	1,792	1,740	1,683	1,622
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077	2,003	1,924	1,793	1,744	1,693	1,637	1,577	1,509
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,917	1,836	1,700	1,649	1,594	1,534	1,467	1,389
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910	1,834	1,750	1,608	1,554	1,495	1,429	1,352	1,254
1E+05	3,842	2,996	2,605	2,372	2,214	2,099	2,010	1,939	1,880	1,831	1,752	1,666	1,517	1,459	1,394	1,318	1,222	1,010



Distribución  $F$  de Fisher-Snedecor para  $\alpha = 0'025$

$n_2 \setminus n_1$	1	2	3	4	5	6	7	8	9	10	12	15	24	30	40	60	120	1E+05
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	976,7	984,9	997,2	1001,4	1005,6	1009,8	1014,0	1018,3
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,46	39,46	39,47	39,48	39,49	39,50
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,12	14,08	14,04	13,99	13,95	13,90
4	12,22	10,65	9,979	9,605	9,364	9,197	9,074	8,980	8,905	8,844	8,751	8,657	8,511	8,461	8,411	8,360	8,309	8,257
5	10,01	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681	6,619	6,525	6,428	6,278	6,227	6,175	6,123	6,069	6,015
6	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523	5,461	5,366	5,269	5,117	5,065	5,012	4,959	4,904	4,849
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823	4,761	4,666	4,568	4,415	4,362	4,309	4,254	4,199	4,142
8	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,357	4,295	4,200	4,101	3,947	3,894	3,840	3,784	3,728	3,670
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026	3,964	3,868	3,769	3,614	3,560	3,505	3,449	3,392	3,333
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779	3,717	3,621	3,522	3,365	3,311	3,255	3,198	3,140	3,080
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588	3,526	3,430	3,330	3,173	3,118	3,061	3,004	2,944	2,883
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436	3,374	3,277	3,177	3,019	2,963	2,906	2,848	2,787	2,725
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,312	3,250	3,153	3,053	2,893	2,837	2,780	2,720	2,659	2,596
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,209	3,147	3,050	2,949	2,789	2,732	2,674	2,614	2,552	2,487
15	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123	3,060	2,963	2,862	2,701	2,644	2,585	2,524	2,461	2,395
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	3,049	2,986	2,889	2,788	2,625	2,568	2,509	2,447	2,383	2,316
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,985	2,922	2,825	2,723	2,560	2,502	2,442	2,380	2,315	2,248
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,929	2,866	2,769	2,667	2,503	2,445	2,384	2,321	2,256	2,187
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,880	2,817	2,720	2,617	2,452	2,394	2,333	2,270	2,203	2,133
20	5,871	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837	2,774	2,676	2,573	2,408	2,349	2,287	2,223	2,156	2,085
21	5,827	4,420	3,819	3,475	3,250	3,090	2,969	2,874	2,798	2,735	2,637	2,534	2,368	2,308	2,246	2,182	2,114	2,042
22	5,786	4,383	3,783	3,440	3,215	3,055	2,934	2,839	2,763	2,700	2,602	2,498	2,331	2,272	2,210	2,145	2,076	2,003
23	5,750	4,349	3,750	3,408	3,183	3,023	2,902	2,808	2,731	2,668	2,570	2,466	2,299	2,239	2,176	2,111	2,041	1,968
24	5,717	4,319	3,721	3,379	3,155	2,995	2,874	2,779	2,703	2,640	2,541	2,437	2,269	2,209	2,146	2,080	2,010	1,935
25	5,686	4,291	3,694	3,353	3,129	2,969	2,848	2,753	2,677	2,613	2,515	2,411	2,242	2,182	2,118	2,052	1,981	1,906
26	5,659	4,265	3,670	3,329	3,105	2,945	2,824	2,729	2,653	2,590	2,491	2,387	2,217	2,157	2,093	2,026	1,954	1,878
27	5,633	4,242	3,647	3,307	3,083	2,923	2,802	2,707	2,631	2,568	2,469	2,364	2,195	2,133	2,069	2,002	1,930	1,853
28	5,610	4,221	3,626	3,286	3,063	2,903	2,782	2,687	2,611	2,547	2,448	2,344	2,174	2,112	2,048	1,980	1,907	1,829
29	5,588	4,201	3,607	3,267	3,044	2,884	2,763	2,669	2,592	2,529	2,430	2,325	2,154	2,092	2,028	1,959	1,886	1,807
30	5,568	4,182	3,589	3,250	3,026	2,867	2,746	2,651	2,575	2,511	2,412	2,307	2,136	2,074	2,009	1,940	1,866	1,787
40	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,452	2,388	2,288	2,182	2,007	1,943	1,875	1,803	1,724	1,637
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334	2,270	2,169	2,061	1,882	1,815	1,744	1,667	1,581	1,482
120	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,222	2,157	2,055	1,945	1,760	1,690	1,614	1,530	1,433	1,311
1E+05	5,024	3,689	3,116	2,786	2,567	2,408	2,288	2,192	2,114	2,048	1,945	1,833	1,640	1,566	1,484	1,388	1,269	1,012



Distribución  $F$  de Fisher-Snedecor para  $\alpha = 0'01$

$n_2 \setminus n_1$	1	2	3	4	5	6	7	8	9	10	12	15	24	30	40	60	120	1E+05
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,888	9,722	9,466	9,379	9,291	9,202	9,112	9,021
6	13,75	10,92	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,718	7,559	7,313	7,229	7,143	7,057	6,969	6,880
7	12,25	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,469	6,314	6,074	5,992	5,908	5,824	5,737	5,650
8	11,26	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,667	5,515	5,279	5,198	5,116	5,032	4,946	4,859
9	10,56	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	5,111	4,962	4,729	4,649	4,567	4,483	4,398	4,311
10	10,04	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,706	4,558	4,327	4,247	4,165	4,082	3,996	3,909
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,397	4,251	4,021	3,941	3,860	3,776	3,690	3,603
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	4,155	4,010	3,780	3,701	3,619	3,535	3,449	3,361
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191	4,100	3,960	3,815	3,587	3,507	3,425	3,341	3,255	3,166
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030	3,939	3,800	3,656	3,427	3,348	3,266	3,181	3,094	3,004
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,666	3,522	3,294	3,214	3,132	3,047	2,959	2,869
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780	3,691	3,553	3,409	3,181	3,101	3,018	2,933	2,845	2,753
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791	3,682	3,593	3,455	3,312	3,084	3,003	2,920	2,835	2,746	2,653
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597	3,508	3,371	3,227	2,999	2,919	2,835	2,749	2,660	2,566
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523	3,434	3,297	3,153	2,925	2,844	2,761	2,674	2,584	2,489
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	3,231	3,088	2,859	2,778	2,695	2,608	2,517	2,421
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640	3,506	3,398	3,310	3,173	3,030	2,801	2,720	2,636	2,548	2,457	2,360
22	7,945	5,719	4,817	4,313	3,988	3,758	3,587	3,453	3,346	3,258	3,121	2,978	2,749	2,667	2,583	2,495	2,403	2,306
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539	3,406	3,299	3,211	3,074	2,931	2,702	2,620	2,535	2,447	2,354	2,256
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,256	3,168	3,032	2,889	2,659	2,577	2,492	2,403	2,310	2,211
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217	3,129	2,993	2,850	2,620	2,538	2,453	2,364	2,270	2,170
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421	3,288	3,182	3,094	2,958	2,815	2,585	2,503	2,417	2,327	2,233	2,132
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388	3,256	3,149	3,062	2,926	2,783	2,552	2,470	2,384	2,294	2,198	2,097
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358	3,226	3,120	3,032	2,896	2,753	2,522	2,440	2,354	2,263	2,167	2,064
29	7,598	5,420	4,538	4,045	3,725	3,499	3,330	3,198	3,092	3,005	2,868	2,726	2,495	2,412	2,325	2,234	2,138	2,034
30	7,562	5,390	4,510	4,018	3,699	3,473	3,304	3,173	3,067	2,979	2,843	2,700	2,469	2,386	2,299	2,208	2,111	2,006
40	7,314	5,179	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801	2,665	2,522	2,288	2,203	2,114	2,019	1,917	1,805
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632	2,496	2,352	2,115	2,028	1,936	1,836	1,726	1,601
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559	2,472	2,336	2,192	1,950	1,860	1,763	1,656	1,533	1,381
1E+05	6,635	4,605	3,782	3,319	3,017	2,802	2,640	2,511	2,408	2,321	2,185	2,039	1,791	1,697	1,592	1,473	1,325	1,015



Distribución  $F$  de Fisher-Snedecor para  $\alpha = 0'005$

$n_2 \setminus n_1$	1	2	3	4	5	6	7	8	9	10	12	15	24	30	40	60	120	1E+05
1	16211	19999	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24940	25044	25148	25253	25359	25464
2	198,5	199,0	199,2	199,2	199,3	199,3	199,4	199,4	199,4	199,4	199,4	199,4	199,5	199,5	199,5	199,5	199,5	199,5
3	55,55	49,80	47,47	46,19	45,39	44,84	44,43	44,13	43,88	43,69	43,39	43,08	42,62	42,47	42,31	42,15	41,99	41,83
4	31,33	26,28	24,26	23,15	22,46	21,97	21,62	21,35	21,14	20,97	20,70	20,44	20,03	19,89	19,75	19,61	19,47	19,32
5	22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,77	13,62	13,38	13,15	12,78	12,66	12,53	12,40	12,27	12,14
6	18,63	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,39	10,25	10,03	9,814	9,474	9,358	9,241	9,122	9,001	8,879
7	16,24	12,40	10,88	10,05	9,522	9,155	8,885	8,678	8,514	8,380	8,176	7,968	7,645	7,534	7,422	7,309	7,193	7,076
8	14,69	11,04	9,596	8,805	8,302	7,952	7,694	7,496	7,339	7,211	7,015	6,814	6,503	6,396	6,288	6,177	6,065	5,951
9	13,61	10,11	8,717	7,956	7,471	7,134	6,885	6,693	6,541	6,417	6,227	6,032	5,729	5,625	5,519	5,410	5,300	5,188
10	12,83	9,427	8,081	7,343	6,872	6,545	6,302	6,116	5,968	5,847	5,661	5,471	5,173	5,071	4,966	4,859	4,750	4,639
11	12,23	8,912	7,600	6,881	6,422	6,102	5,865	5,682	5,537	5,418	5,236	5,049	4,756	4,654	4,551	4,445	4,337	4,226
12	11,75	8,510	7,226	6,521	6,071	5,757	5,525	5,345	5,202	5,085	4,906	4,721	4,431	4,331	4,228	4,123	4,015	3,904
13	11,37	8,186	6,926	6,233	5,791	5,482	5,253	5,076	4,935	4,820	4,643	4,460	4,173	4,073	3,970	3,866	3,758	3,647
14	11,06	7,922	6,680	5,998	5,562	5,257	5,031	4,857	4,717	4,603	4,428	4,247	3,961	3,862	3,760	3,655	3,547	3,436
15	10,80	7,701	6,476	5,803	5,372	5,071	4,847	4,674	4,536	4,424	4,250	4,070	3,786	3,687	3,585	3,480	3,372	3,260
16	10,58	7,514	6,303	5,638	5,212	4,913	4,692	4,521	4,384	4,272	4,099	3,920	3,638	3,539	3,437	3,332	3,224	3,112
17	10,38	7,354	6,156	5,497	5,075	4,779	4,559	4,389	4,254	4,142	3,971	3,793	3,511	3,412	3,311	3,206	3,097	2,984
18	10,22	7,215	6,028	5,375	4,956	4,663	4,445	4,276	4,141	4,030	3,860	3,683	3,402	3,303	3,201	3,096	2,987	2,873
19	10,07	7,093	5,916	5,268	4,853	4,561	4,345	4,177	4,043	3,933	3,763	3,587	3,306	3,208	3,106	3,000	2,891	2,776
20	9,944	6,986	5,818	5,174	4,762	4,472	4,257	4,090	3,956	3,847	3,678	3,502	3,222	3,123	3,022	2,916	2,806	2,691
21	9,830	6,891	5,730	5,091	4,681	4,393	4,179	4,013	3,880	3,771	3,602	3,427	3,147	3,049	2,947	2,841	2,730	2,614
22	9,727	6,806	5,652	5,017	4,609	4,322	4,109	3,944	3,812	3,703	3,535	3,360	3,081	2,982	2,880	2,774	2,663	2,546
23	9,635	6,730	5,582	4,950	4,544	4,259	4,047	3,882	3,750	3,642	3,475	3,300	3,021	2,922	2,820	2,713	2,602	2,484
24	9,551	6,661	5,519	4,890	4,486	4,202	3,991	3,826	3,695	3,587	3,420	3,246	2,967	2,868	2,765	2,658	2,546	2,428
25	9,475	6,598	5,462	4,835	4,433	4,150	3,939	3,776	3,645	3,537	3,370	3,196	2,918	2,819	2,716	2,609	2,496	2,377
26	9,406	6,541	5,409	4,785	4,384	4,103	3,893	3,730	3,599	3,492	3,325	3,151	2,873	2,774	2,671	2,563	2,450	2,330
27	9,342	6,489	5,361	4,740	4,340	4,059	3,850	3,687	3,557	3,450	3,284	3,110	2,832	2,733	2,630	2,522	2,408	2,287
28	9,284	6,440	5,317	4,698	4,300	4,020	3,811	3,649	3,519	3,412	3,246	3,073	2,794	2,695	2,592	2,483	2,369	2,247
29	9,230	6,396	5,276	4,659	4,262	3,983	3,775	3,613	3,483	3,377	3,211	3,038	2,759	2,660	2,557	2,448	2,333	2,210
30	9,180	6,355	5,239	4,623	4,228	3,949	3,742	3,580	3,450	3,344	3,179	3,006	2,727	2,628	2,524	2,415	2,300	2,176
40	8,828	6,066	4,976	4,374	3,986	3,713	3,509	3,350	3,222	3,117	2,953	2,781	2,502	2,401	2,296	2,184	2,064	1,932
60	8,495	5,795	4,729	4,140	3,760	3,492	3,291	3,134	3,008	2,904	2,742	2,570	2,290	2,187	2,079	1,962	1,834	1,689
120	8,179	5,539	4,497	3,921	3,548	3,285	3,087	2,933	2,808	2,705	2,544	2,373	2,089	1,984	1,871	1,747	1,606	1,431
1E+05	7,880	5,299	4,280	3,715	3,350	3,091	2,897	2,745	2,621	2,519	2,359	2,187	1,898	1,789	1,669	1,533	1,364	1,016

