

# 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.

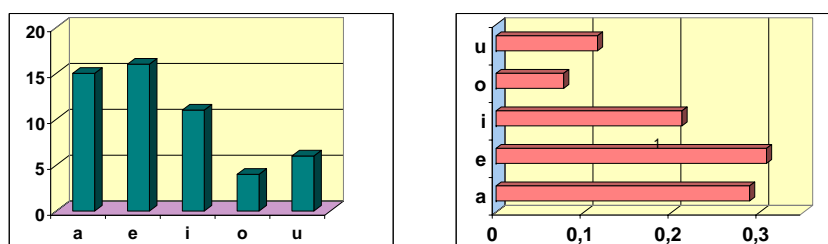


Figura 1.6: Diagrama de rectángulos

**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.

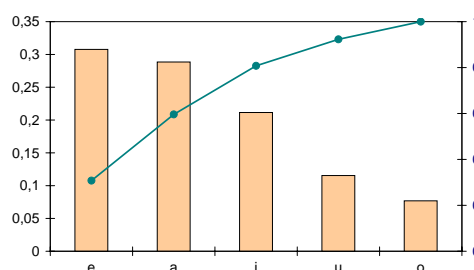


Figura 1.7: Diagrama de Pareto

**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.

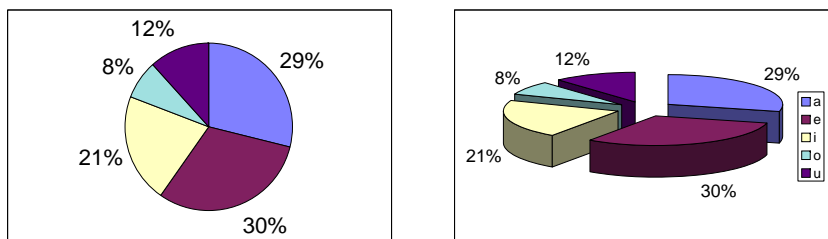


Figura 1.8: Diagrama de sectores

**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.

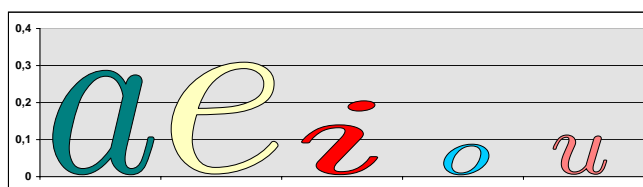


Figura 1.9: Pictograma

### 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).

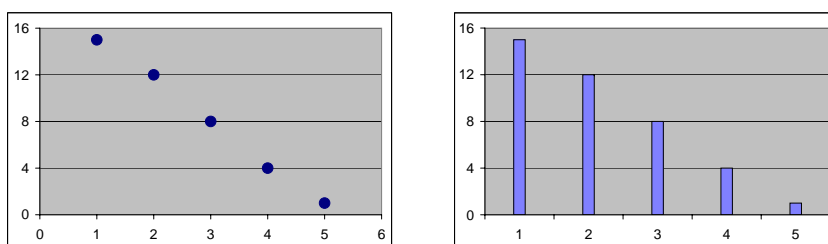


Figura 1.10: Diagrama de puntos – diagrama de barras

**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).

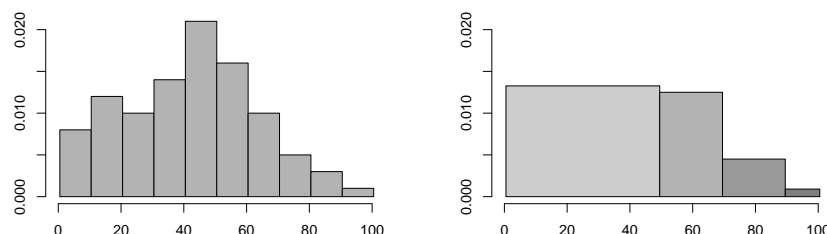


Figura 1.11: Histogramas

**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).

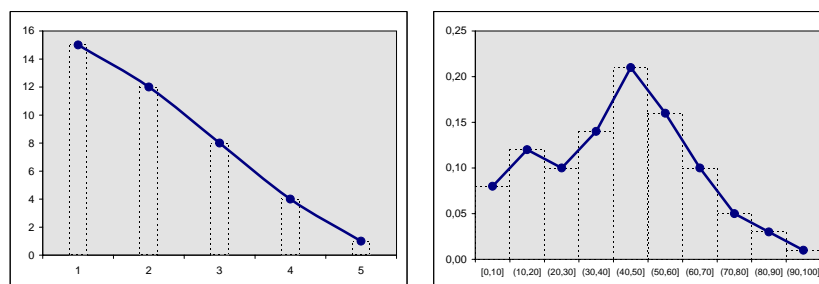


Figura 1.12: Polígonos de frecuencias

**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.

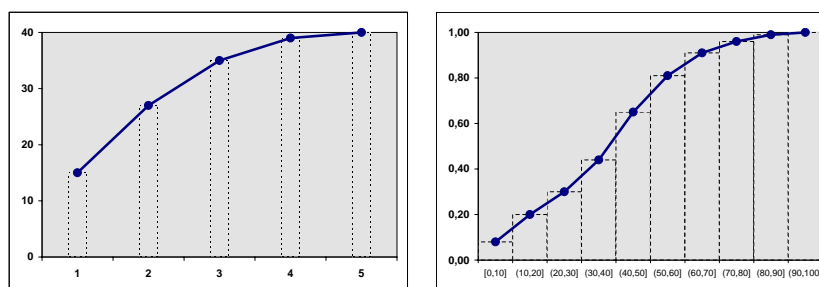


Figura 1.13: Diagrama de frecuencias (absolutas/relativas) acumuladas

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  $1\frac{1}{2}$ . 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.

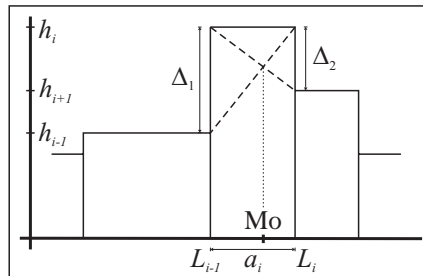


Figura 1.14: Cálculo de la moda en el histograma

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.

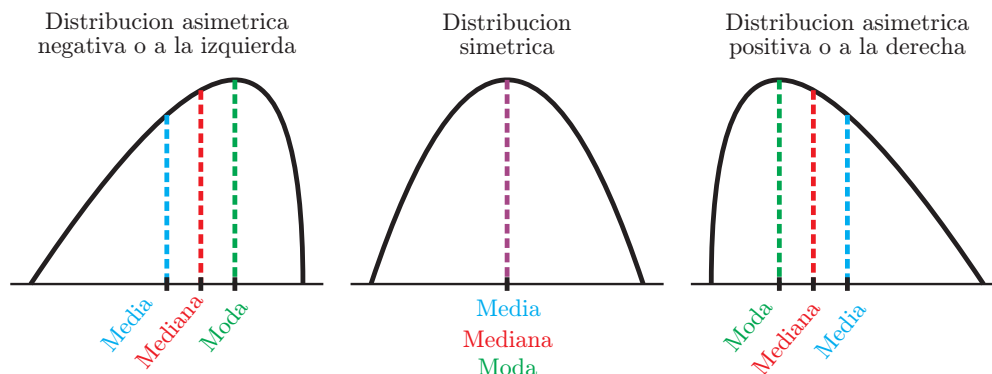


Figura 1.17: Formas de la distribución de frecuencias

**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). □

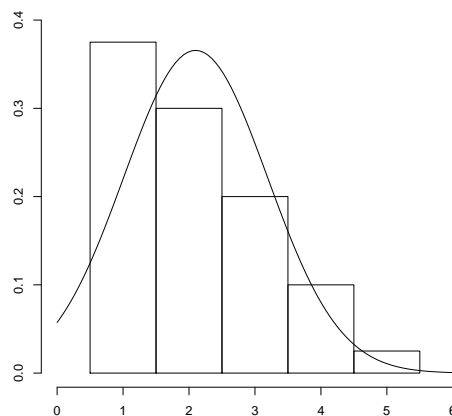


Figura 1.18: Formas de la distribución de frecuencias del ejemplo 1.5

**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.

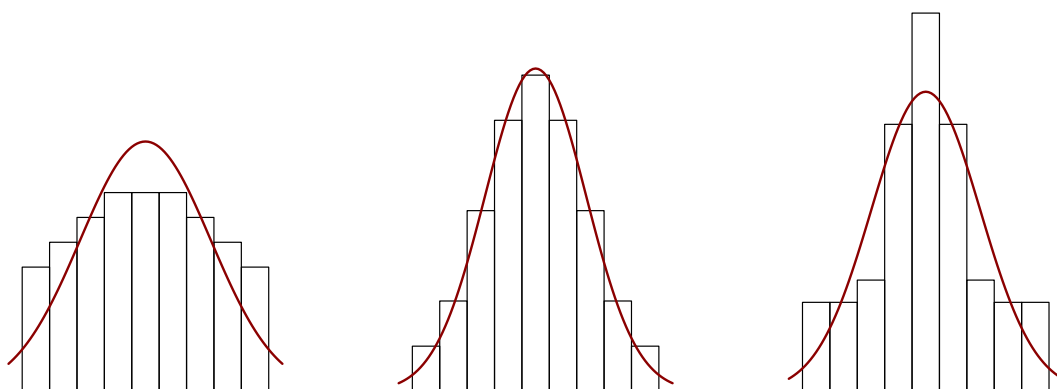


Figura 1.19: Formas de la distribución de frecuencias

<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 inoxidables 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.

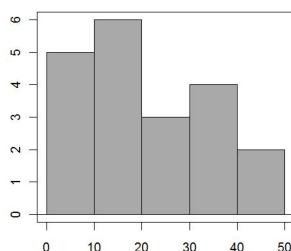
- b) Hallar los valores tipificados de la variable  $X$  y comprobar que la media de estos nuevos valores es 0 y la varianza es 1.
- c) Demostrar que el resultado del apartado anterior constituye una propiedad de cualquier variable tipificada.

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	$\sum y_i^4 f_i = 5648$

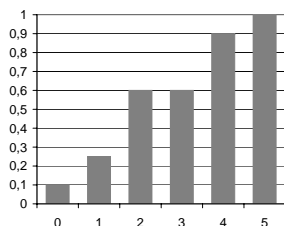
Determinar los dos valores de  $X$ , y comparar la dispersión y la curtosis de ambas variables.

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.

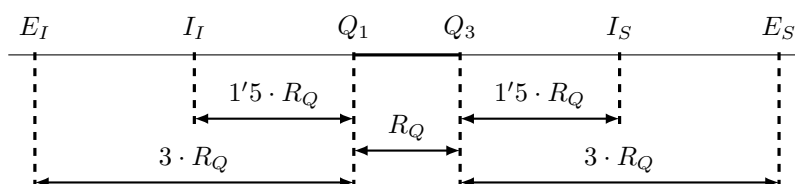


Figura 1.20: Intervalo para detectar datos anómalos

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)
```