Capítulo 1

Estadística Descriptiva

La *Estadística* actual es el resultado de la unión de dos disciplinas que evolucionaron independientemente hasta el siglo XIX, estas son el *Cálculo de Probabilidades* que tuvo sus inicios como ciencia en el estudio de los juegos de azar en el siglo XVII, mientras que la *Estadística Descriptiva* tiene sus origenes en los censos del mundo antiguo. De la unión de estas disciplinas se obtuvieron resultados, que permiten el *Contraste de hipótesis* y la *Inferencia estadística* en ambiente de incertidumbre.

Por tanto, la Estadística es la ciencia de los datos; implica la colección, clasificación, síntesis, organización, análisis e interpretación de los datos. Esta ciencia suele aplicarse a dos tipos de problemas:

1. *Resumir*, *describir* y *explorar* datos.
2. Utilizar datos de muestra para *inferir* la naturaleza del conjunto de datos del que se escogió la muestra.

La rama de la Estadística que se dedica a la organización, síntesis y descripción de conjuntos de datos es la **Estadística Descriptiva**, mientras que la rama de la Estadística que se ocupa de utilizar datos de muestra para inferir algo acerca de una población se denomina **Estadística Inferencial**.

1.1 Conceptos elementales

A continuación introducimos los conceptos generales básicos en Estadística:

Población

Universo, Población estadística, Colectivo o simplemente *Población* es el conjunto de elementos que son objeto de estudio. Las poblaciones podrán ser finitas o infinitas según 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 que se refiere el estudio estadístico.

Muestra

Una *muestra* es un subconjunto de individuos de la población. La muestra, debidamente elegida, se somete a observación científica, en representación del conjunto, con el propósito de obtener resultados válidos para toda la población.

El número de elementos que componen la muestra se denomina *tamaño muestral* (N) y si coincide con el tamaño de la población, la muestra se denomina *censo*. Las dificultades para realizar un censo (población infinita, dificultad de acceso a todos los individuos, coste económico, capacidad de trabajo y tiempo necesario, etc.) hacen que sea preferible el muestreo. En este caso, las técnicas de Inferencia Estadística nos 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.

Caracteres

Los *caracteres* son las cualidades de los individuos de la población que son objeto de estudio. Los caracteres pueden ser cualitativos (nacionalidad o color del pelo) o cuantitativos (nº de hijos o m^2 de vivienda).

Los caracteres cualitativos reciben el nombre de *atributos* y se designan utilizando las primeras letras del alfabeto en mayúsculas (A,B,C,...). Los caracteres cuantitativos se denominan *variables estadísticas* y se designan utilizando las últimas letras del alfabeto en mayúsculas (...,X,Y,Z). A su vez, las variables pueden ser *discretas* (nº de acciones vendidas un día en la Bolsa de Valores, nº de estudiantes matriculados en una Universidad) o *continuas* (vida media de los tubos de televisión producidos por una fábrica, longitud de 1000 tornillos producidos por una empresa, temperaturas medidas tomadas en un observatorio cada media hora) según la naturaleza de los valores numéricos.

Modalidades

Los diferentes valores que puede tomar un carácter se denominan *modalidades*. Éstas deben estar bien definidas de tal manera que cada individuo pertenezca a una única modalidad. Se denotan haciendo uso de la letra minúscula correspondiente al nombre de la variable y afectada por un subíndice de orden. Por ejemplo, $x_1, x_2, ..., x_k$ denota las distintas modalidades del carácter X.

Ejemplo: *Consideremos la* población *formada por todos los automóviles de un cierto modelo producidos por una fábrica. Un conjunto de 100 automóviles extraídos de dicha población constituye una* muestra *de tamaño 100.*

Realizamos una encuesta *que consiste en medir la compresión del motor en cada uno de los 100 automóviles bajo determinadas condiciones. El resultado es una muestra de 100 valores del* carácter *"compresión del motor" que resulta ser una* variable continua *cuyas* modalidades *corresponden a todas las posibles relaciones volumétricas.*

1.2 Distribuciones de un carácter

Uno de los conceptos sobre el que descansarán muchas definiciones posteriores y que simplifica la presentación de los datos es el de *frecuencia* que no es sino el número de veces que aparece una determinada modalidad de un carácter. La utilización de las frecuencias en las tablas estadísticas nos permite organizar y resumir el conjunto de datos de modo que sea más comprensible y significativo.

1.2.1 Frecuencias

En adelante consideraremos una población o muestra de tamaño N en la que observaremos el carácter cuantitativo o variable estadística X que presenta las modalidades $x_1, x_2, ..., x_k$. Las siguientes definiciones son también válidas para caracteres cualitativos.

Frecuencia Absoluta (n_i). Llamamos frecuencia absoluta de un valor x_i de la variable X al número de individuos observados que presentan esta modalidad.

Frecuencia Relativa (f_i). Llamamos frecuencia relativa de un valor x_i de la variable X al cociente entre la frecuencia absoluta y el total de individuos. La frecuencia relativa representa la proporción de individuos que presentan una determinada modalidad.

$$f_i = \frac{n_i}{N} \qquad\qquad i = 1, 2, ..., k$$

Frecuencias Acumuladas Absolutas (N_i) o Relativas (F_i). Llamamos frecuencia acumulada de un valor x_i de la variable X a la suma de las frecuencias de los valores que sean inferiores o iguales a el. Si los valores x_i están ordenados de forma creciente entonces

$$N_i = \sum_{j=1}^{i} n_j \qquad\qquad F_i = \sum_{j=1}^{i} f_j = \frac{N_i}{N} \qquad\qquad i = 1, 2, ..., k$$

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

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

Propiedades de las frecuencias. De las definiciones anteriores destacamos las siguientes propiedades:

$$1)\quad 0 \le n_i \le N \qquad 2)\quad \sum_{i=1}^{k} n_i = N \qquad 3)\quad n_i = N_i - N_{i-1}$$

$$4)\quad 0 \le f_i \le 1 \qquad 5)\quad \sum_{i=1}^{k} f_i = 1 \qquad 6)\quad f_i = F_i - F_{i-1}$$

1.2.2 Distribuciones y tablas de frecuencias

Una vez recogidos los datos de la muestra se realiza una primera clasificación, denominada *distribución de frecuencias*, en la que aparecen las distintas modalidades observadas del carácter junto a su frecuencia. Esta presentación en la que se agrupan los datos y se disponen de manera

ordenada se denomina *tabla estadística o de frecuencias* y contiene las primeras informaciones obtenidas de los datos de la muestra.

Las distribuciones de frecuencias de una sola variable las clasificaremos en tres tipos, eligiéndose uno u otro tipo en consideración al número de observaciones y al número de valores distintos que toma la variable.

Distribuciones de tipo I.

Son aquellas distribuciones que constan de un reducido número de observaciones y, en consecuencia, de un reducido número de valores distintos que toma la variable.

En este caso, como hay pocas observaciones, la forma de presentarlas no requiere ningún artificio especial. Para construir la tabla estadística basta simplemente con anotar ordenadamente las observaciones en fila o en columna, generalmente de menor a mayor.

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

Ejemplo 1.1 *Para realizar un estudio sobre la venta de ordenadores al día, en una determinada empresa de informática, se observa, durante 5 semanas, el número de ordenadores vendidos y se obtienen los siguientes resultados: 10, 12, 20, 6 y 10. Representar su distribución de frecuencias.*

Distribuciones de tipo II.

Son aquellas distribuciones donde el número de observaciones es grande, pero el número de valores distintos que toma la variable es pequeño.

Para construir la tabla estadística correspondiente basta con poner en una primera columna los pocos valores distintos de la variable, y en una segunda, correspondiéndose con la primera, las frecuencias que estemos interesados en mostrar. Los valores ordenados de menor a mayor se disponen como figura en la siguiente tabla:

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

Ejemplo 1.2 *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 automóviles. Representar la distribución de frecuencias si los resultados fueron:*

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

Distribuciones de tipo III.

Son aquellas distribuciones en las que el número de observaciones es grande y el número de valores distintos que toma la variable es también grande y, por tanto, no es posible escribirlos todos ellos en una columna, como se hizo anteriormente.

Para tabular estos datos conviene *agruparlos* en unos cuantos *intervalos* y determinar el número de individuos que pertenecen a cada uno de ellos. Tomar el intervalo como unidad de estudio, en lugar de cada valor de la variable, supone una simplificación pero resulta una perdida de información. Por lo tanto, es importante elegir un número adecuado de intervalos que equilibre estos dos aspectos.

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 intervalo* y se denota por a_i que puede ser variable o constante para todos los intervalos. Los unión de todos los intervalos ha de recubrir a todos los valores de la variable (exhaustivo) pero sin solaparse (excluyente).

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 tal y como figura en la siguiente tabla:

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

Normalmente, se suele dividir el intervalo entre 5 y 20 clases de igual anchura. El número de clases es arbitrario, pero se obtiene una mejor descripción gráfica si se utilizan pocas clases cuando el conjunto de datos es grande. Podemos utilizar la siguiente regla empírica:

Número de observaciones	Número de clases
Menos de 25	5 ó 6
Entre 25 y 50	7 - 14
Más de 50	15 - 20

Ejemplo 1.3 *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 utilizando intervalos de amplitud 10.

1.3 Representaciones gráficas

Con el fin de comunicar rápidamente una imagen visual de los datos, se representan las frecuencias (absolutas, relativas o acumuladas) mediante distintos tipos de gráficas. A continuación se relacionan los tipos de representación más utilizados que conviene conocer para elegir el más adecuado a cada caso.

Caracteres cualitativos

Los tipos de representación que se muestran a continuación pueden utilizarse indistintamente. La creatividad y la originalidad puede dar lugar a otros tipos de gráficas, siempre y cuando cumplan con el objetivo de garantizar una imagen sencilla y real de los datos.

Diagrama de rectángulos. Para cada modalidad, se representa un rectángulo cuya altura coincide con la frecuencia absoluta (o relativa).

Ejemplo: *Representar gráficamente las velocidades (m/seg) orbitales de los planetas del sistema solar: Mercurio (29'7), Venus (21'8), Tierra (18'5), Marte (15'0), Júpiter (8'1), Saturno (6'0), Urano (4'2), Neptuno (3'4), Plutón (3'0).*

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 $\alpha_i = 360 \cdot f_i$.

Ejemplo: *Representar gráficamente las áreas (millones de millas cuadradas) de los océanos: Pacífico (63'8), Atlántico (31'5), Índico (28'4), Antártico (7'6), Ártico (4'8).*

Pictograma y cartogramas.

Ejemplo: *Representar gráficamente los precios medios de la vivienda (millones de pesetas) de tres provincias españolas: Madrid (28'5), Málaga (15'3), Guadalajara (8'9).*

Caracteres cuantitativos

Este tipo de representaciones gráficas se realiza sobre los ejes de coordenadas y para que sea más significativa, puede ser interesante el cambio de escala en los ejes o el cambio de inicio de la escala, si bien esto último debe indicarse (p.e. mediante una línea en zigzag en el eje correspondiente) para no inducir a engaño.

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.

Ejemplos 1.1 *de la venta de ordenadores* y 1.2 *de la encuesta de tráfico.*

Histograma. Se utiliza para representar los datos agrupados en intervalos. Para cada clase, se dibuja un rectángulo apoyado en 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.

Ejemplo: 1.3 *sobre las calificaciones en Matemáticas.*

Polígono de frecuencias. Se trata de unir los extremos de las barras en el diagrama de barras o los puntos medios superiores de los rectángulos en el histograma.

Ejemplos 1.1, 1.2 y 1.3.

Diagrama de frecuencias acumuladas. Igual que el polígono de frecuencia pero utilizando las frecuencias acumuladas.

Ejemplos 1.1, 1.2 y 1.3.

1.4 Medidas de tendencia central

Hasta ahora hemos mostrado distintas formas de presentar los datos de manera clara y ordenada. A veces conviene reducir toda esta información en uno o varios valores cuantitativos que sean más o menos representativos y que nos permitan comparar distintas muestras.

1.4.1 Media

Consideremos una variable X (tipo II) que toma los valores $x_1, x_2, ..., x_k$ con las frecuencias $n_1, n_2, ..., n_k$ respectivamente haciendo un total de N datos.

La *media aritmética* o simplemente *media* es la suma de los valores de todos los datos dividido entre el número total de datos. Se denota por

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

Por tanto, esta médida sólo es válida para caracteres cuantitativos. 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 de las observaciones respecto de la media es cero.

Ejemplo: *Calcular la media aritmética en el ejemplo* 1.2.

Si la variable X es de tipo I, entonces la frecuencia para cada uno de sus valores es 1.

Ejemplo: *Calcular la media aritmética en el ejemplo* 1.1.

Si la variable X es de tipo III, entonces consideraremos las marcas de clase como valores de la variable cuya frecuencia quedará determinada por el número de datos contenidos en el intervalo correspondiente.

Ejemplo: *Calcular la media aritmética en el ejemplo* 1.3.

Comportamiento de la media frente a transformaciones lineales. 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: *Los salarios de los 6 obreros de una empresa son 80.000, 110.000, 120.000, 140.000, 160.000 y 170.000 ptas. Calcular la media aritmética de los mismos.*

Otros promedios. Aunque la media aritmética es la más utilizada, existen otras medidas de tendencia central denominadas medias que pueden resultar interesantes para determinados casos.

Como ejemplo, se utiliza la *media ponderada* cuando asociamos ciertos factores $(w_1, w_2, ...)$ peso a cada valor $(x_1, x_2, ...)$ de la variable con el fin de dar más relevancia a unos que a otros.

$$MP = \frac{\sum w_i x_i}{\sum w_i}$$

Otro tipo de medias lo constituye un grupo denominado φ-medias que se obtienen aplicando la fórmula

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

para alguna función φ que sea continua y monótona. Como ejemplos:

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

Entre ellas se establece la siguiente relación: $H\leq G\leq \bar{x}\leq MQ$

1.4.2 Mediana

Como se muestra en el siguiente ejemplo, la media aritmética es un promedio muy sensible a los valores extremos de la variable. En tal caso será conveniente emplear la mediana como medida de tendencia central.

Ejemplo: *Diez medidas del diámetro de un cilindro fueron anotadas por un científico como 3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98, 40.6 cm. Hallar la media aritmética de tales medidas y discutir si dicho promedio es significativo.*

La *mediana* o valor mediano es aquel que divide a la población en dos partes de igual tamaño, la mitad son mayores que él y la otra mitad inferior a él. Si N es impar, existirá dicho valor y coincidirá con un término de la población, mientras que si es par, se tomarán los dos valores centrales y se calculará la media.

Ejemplo: *Calcular la mediana de los siguientes datos: 3, 6, 4, 4, 8, 8, 8, 5 y 10.*
Ejemplo: *Calcular la mediana de los siguientes datos: 15, 5, 7, 18, 11, 12, 5 y 9.*
Ejemplo: *Calcular la mediana en el ejemplo* 1.1 y 1.2.

En el caso de que los datos vengan agrupados por intervalos, se calculará primero el intervalo que contenga la mediana (intervalo mediano), para posteriormente interpolar en él mediante la fórmula:

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

Ejemplo: *Calcular de dos formas distintas (datos agrupados y sin agrupar) la mediana de las calificaciones finales en Matemáticas en el ejemplo* 1.3.

1.4.3 Moda

La *moda* de un conjunto de datos es el valor de la variable que presenta mayor frecuencia. La moda puede no ser única o incluso no existir. 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: *Calcular la moda de los siguientes datos: 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12 y 18.*
Ejemplo: *Calcular la moda de los siguientes datos: 3, 5, 8, 10, 12, 15 y 16.*
Ejemplo: *Calcular la moda de los siguientes datos: 2, 3, 4, 4, 4, 5, 5, 7, 7, 7 y 9.*

Ejemplo: *Calcular la moda en el ejemplo* 1.1 y 1.2.

En el caso de datos agrupados en intervalos, se toma como *intervalo modal* el que resulta con mayor altura ($h_i = n_i/a_i$) en el histograma, y se calcula

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

Ejemplo: *Calcular de dos formas distintas (datos agrupados y sin agrupar) la moda de las calificaciones finales en Matemáticas del ejemplo* 1.3.

1.4.4 Cuantiles

Los *cuantiles* son parámetros que dividen a la población en partes. En general, un *cuantil de orden* 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 tres tipos de cuantiles que dividen a la población en 4, 10 o 100 partes.

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

Deciles: Son 9 y dividen a la población en 10 partes iguales. Se llama decil de orden d al valor D_d que divide a la población en dos partes de tal forma que $k = d/10$ 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 al valor P_c que divide a la población en dos partes de tal forma que $k = c/100$ sea menor que él y el resto mayor.

Ejemplo: *Calcular algunos cuantiles para los ejemplos* 1.1 y 1.2.

En el caso de datos agrupados en intervalos, para calcular el cuantil de orden k se elige el intervalo que contiene al valor $N \cdot k$ que buscamos y se calcula

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

Ejemplo: *Calcular algunos cuantiles para el ejemplo* 1.3.

1.5 Medidas de dispersión

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.

Para las definiciones que siguen consideramos la variable X que toma los valores $x_1, x_2, ..., x_k$ con las frecuencias $n_1, n_2, ..., n_k$ respectivamente haciendo un total de N datos.

1.5.1 Rango

La medida de dispersión más simple es el *rango* o *recorrido* que corresponde a la diferencia entre el mayor valor observado de la variable y el menor.

Ejemplo: Calcular el rango en los ejemplos 1.1, 1.2 y 1.3.

En algunas ocasiones, con el objetivo de evitar la influencia de los valores extremos de la variables, se utilizan otros rangos que corresponden a los distintos cuantiles:

Rango intercuartílico: Diferencia entre el cuartil de orden 3 y el de orden 1.

Rango interdecílico: Diferencia entre el decil de orden 9 y el de orden 1.

Rango intercentílico: Diferencia entre el percentil de orden 99 y el de orden 1.

Ejemplo: Calcular los rangos intercuantílicos en los ejemplos 1.1, 1.2 y 1.3.

1.5.2 Desviación media

Otra medida de la dispersión de los datos de la muestra se puede obtener midiendo las distancias desde cada uno de los valores hasta un punto elegido previamente. Por tanto, 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 tomaremos una medida de tendencia central (media o mediana) como valor del parámetro. La media aritmética de estas desviaciones respecto del promedio nos asegurarán una medida de la dispersión de la muestra.

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

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

Ejemplo: Calcular la desviación media en los ejemplos 1.1, 1.2 y 1.3.

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 perderemos de vista la idea de medir desviaciones respecto de un promedio.

1.5.3 Varianza 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.

Se define la *varianza* de una conjunto de datos como

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

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

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

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 σ^2 es la varianza de la variable X, entonces $a^2\sigma^2$ es la varianza de la variable $aX + b$.

De la definición de varianza se puede deducir una fórmula más simple para su cálculo que consiste en la media de los cuadrados menos el cuadrado de la media.

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

Una relación importante que mantiene la media y la desviación típica es la siguiente regla empírica para distribuciones cuya forma se aproxima a la distribución normal

Porcentaje de la población	Intervalo al que pertenecen
68%	$\bar{x} \pm \sigma$
95%	$\bar{x} \pm 2\sigma$
99%	$\bar{x} \pm 3\sigma$

Variable tipificada. Por último, haciendo uso de la media y de la desviación típica de la variable X, podemos considerar una nueva variable que viene dada por:

$$Z = \frac{X - \bar{x}}{\sigma} \qquad \text{es decir} \qquad z_i = \frac{x_i - \bar{x}}{\sigma}$$

Esta nueva variable obtenida es adimensional (independiente de las unidades usadas), se denomina *variable tipificada* 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 distribuciones.

Ejemplo: *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 ?*

1.5.4 Coeficiente de variación.

Hasta ahora, las medidas de dispersión que hemos visto vienen expresadas de forma absoluta en las unidades de la variable; por tanto, no son útiles si queremos establecer una comparación entre las dispersiones de dos muestras que vengan expresadas en unidades distintas.

El *coeficiente de variación de Pearson* es el cociente entre la desviación típica y la media:

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

Ejemplo: Calcular este coeficiente de variación en los ejemplos 1.1, 1.2 y 1.3.

Este coeficiente pierde representatividad cuando la media se acerca a cero. 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.

A veces, este coeficiente aparece multiplicado por 100, para mayor comodidad en el manejo de las cifras, trabajando así con porcentajes.

Ejemplo: *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 σ_A=280 horas y σ_B=310. Comparar las dispersiones de las dos poblaciones en términos absolutos y relativos.*

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 suele ser la media o la mediana.

1.5.5 Momentos

Además de determinar medidas de dispersión, los momentos resultan muy utiles para calcular determinados parámetros y forman parte de la definición de algunos coeficientes. De hecho, generalizan algunas definiciones (media y varianza) ya vistas.

Definimos el *momento de orden r repecto al punto c* de la forma:

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

En particular, resultan de especial interés, dos casos:

Momentos ordinarios: Si $c = 0$ entonces $m_r = \sum_{i=1}^{k} x_i^r f_i$

Momentos centrales: Si $c = \bar{x}$ entonces $\mu_r = \sum_{i=1}^{k}(x_i - \bar{x})^r f_i$.

Destacamos las siguientes propiedades:

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) $\mu_2 = \sigma^2 = m_2 - \bar{x}^2$

y relaciones entre los momentos centrales y ordinarios:

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

1.6 Medidas de simetría

Diremos que una distribución de frecuencias es simétrica cuando los valores de la variable que equidistan de un valor central tienen las mismas frecuencias. En tal caso se verifica que $\bar{x} = M_e = M_o$.

Una distribución de frecuencias es *asimétrica* si no es simétrica. La asimetría puede presentarse a la derecha o a la izquierda.

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 $M_o \leq M_e \leq \bar{x}$.

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 $\bar{x} \leq M_e \leq M_o$.

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

Coeficiente de asimetría de Pearson

De acuerdo a las relaciones entre media, mediana y moda, establecidas para las distintas asimetrías, definimos el coeficiente de sesgo de Pearson

$$A_P = \frac{\bar{x} - M_o}{\sigma} \approx \frac{3(\bar{x} - M_e)}{\sigma}$$

donde

$$\begin{cases} A_P > 0 & \text{Asimetría a la derecha o positiva} \\ A_P = 0 & \text{Simetría} \\ A_P < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

Ejemplo: *Utilizar el coeficiente de Pearson para determinar el sesgo en los ejemplos* 1.1, 1.2 y 1.3.

Coeficiente de asimetría de Fisher

Otro coeficiente adimensional que mide el sesgo, haciendo uso del momento central de orden 3, es

$$A_F = \frac{\mu_3}{\sigma^3}$$

donde

$$\begin{cases} A_F > 0 & \text{Asimetría a la derecha o positiva} \\ A_F = 0 & \text{Simetría} \\ A_F < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

Ejemplo: *Utilizar el coeficiente de Fisher para determinar el sesgo en los ejemplos* 1.1, 1.2 y 1.3.

1.7 Medidas de Apuntamiento

El *apuntamiento* o la *curtosis* mide si la forma de la distribución es más o menos afilada o aplastada que la de la distribución normal (campana de Gauss) con igual media y varianza.

Se calcula como

$$g_2 = \frac{\mu_4}{\sigma^4}$$

que se interpreta así:

$$\begin{cases} g_2 > 3 & \text{Más apuntamiento que la normal: leptocúrtica} \\ g_2 = 3 & \text{Igual apuntamiento que la normal: mesocúrtica} \\ g_2 < 3 & \text{Menos apuntamiento que la normal: platicúrtica} \end{cases}$$

A veces se define $g_2 = \frac{\mu_4}{\sigma^4} - 3$ y se compara con 0.

Ejemplo: *Utilizar este coeficiente para determinar la curtosis en los ejemplos* 1.1, 1.2 y 1.3.

1.8 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 confiabilidad de un sistema en particular, se prueban 100 componentes de computadora 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 confiabilidad del sistema de cómputo ?

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 que utilizan alcohol como combustible.

   (d) Número de caracteres impresos por línea de salida de computadora en 20 impresoras de línea.

   (e) Número de miembros de una familia.

   (f) Estado civil de una persona.

   (g) Tiempo de vuelo de un misil.

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. En cada caso, determinar el parámetro más apropiado e interpretar los resultados.

   (a) Resistencia a la tensión (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 subterraneo.

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
Defecto de materiales	4
Defctos 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 empaque.

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) Investigadores de Massachusetts Institute of Technology (MIT) estudiaron las propiedades espectroscópicas de asteroides de la franja principal con un diámetro menor a los diez kilómetros. Aquí se presentan el número de exposiciones de imagen espectral independientes para 40 observaciones de asteroides:

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. 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.

6. El grupo A comprende 10 puntuaciones y la media y la mediana son respectivamente 15.5 y 13. El grupo B comprende 20 puntuaciones, y la media y la mediana son respectivamente 11.4 y 10. ¿ Cuáles son la media y la mediana de las 30 puntuaciones obtenidas combinando los grupos A y B ?

7. La variable Y es igual a la variable X multiplicada por la constante 4 y la variable Z es igual a la X más la constante 4. Compara las dispersiones de las variables Y y Z.

8. Consideremos una cierta población de obreros de la cual conocemos los salarios. Se va a tratar un convenio colectivo y se presentan dos alternativas:

   (a) Aumento lineal de "a" ptas.

   (b) Aumento del "b" tanto por ciento del sueldo de cada uno.

   Estudiar que modalidad es más social, es decir, la que iguala más los salarios.

9. 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?

10. Se tiene dos distribuciones campaniformes y simétricas, de ellas se sabe
    Distrib. de X: Me=10 σ_x^2=4 n=2 $\sum f_i x_i^4$=12416
    Distrib. de Y: Mo=8 σ_y^2=4 n=82 $\sum f_i x_i^4$= 8000
    ¿ Cuál tiene mayor dispersión ?

11. Consideremos los siguientes valores obtenidos en una muestra.

    $$2 \quad , \quad 4 \quad , \quad 6 \quad , \quad 8$$

    Se pide:

    (a) Calcular la media y la varianza.

    (b) Hallar los valores tipificados de la variable.

    (c) Comprobar que los nuevos valores de la variable tienen media 0 y desviación típica 1. ¿ Crees que este resultado puede ser una propiedad de la variable tipificada ?

    (d) Observemos qué ocurre cuando perdemos, ganamos o modificamos algún dato de la muestra.

    Caso1 Descubrimos que el valor 8 observado es erróneo y lo eliminamos.

    Caso2 Contamos con un nuevo valor, el 5, para nuestra muestra.

    Caso3 Descubrimos que el valor 8 observado es erróneo y lo cambiamos por el verdadero valor que es el 9.

    En cada caso, ¿ cómo cambia el valor de la media y la varianza sin tener que aplicar nuevamente las fórmulas a todos los datos ? Está claro que esta experiencia que vamos a realizar tiene más sentido cuando tenemos una gran cantidad de datos.

12. Calcular los parámetros en funcion de los datos:

    (a) Si $N = 2$, $\bar{x} = 2'625$ y $\sigma = 1'125$, ¿ cuáles son los datos de la muestra ?

    (b) Si $CV = 0'5$, $\bar{x} = 2$ y $m_3 = 14$, ¿ cuánto vale μ_3 ?